

The
Handbook
of
Brain Theory
and
Neural Networks

Michael A. Arbib
editor

second edition

The Handbook of
Brain Theory
and Neural Networks

This Page Intentionally Left Blank

The Handbook of Brain Theory and Neural Networks

Second Edition

EDITED BY

Michael A. Arbib

EDITORIAL ADVISORY BOARD

Shun-ichi Amari • John Barnden • Andrew Barto • Ronald Calabrese
Avis Cohen • Joaquín Fuster • Stephen Grossberg • John Hertz
Marc Jeannerod • Mitsuo Kawato • Christof Koch • Wolfgang Maass
James McClelland • Kenneth Miller • Terrence Sejnowski
Noel Sharkey • DeLiang Wang

EDITORIAL ASSISTANT

Prudence H. Arbib

A Bradford Book

THE MIT PRESS

Cambridge, Massachusetts

London, England

© 2003 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Times Roman by Impressions Book and Journal Services, Inc., Madison, Wisconsin, and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

The handbook of brain theory and neural networks / Michael A. Arbib,
editor—2nd ed.

p. cm.

“A Bradford book.”

Includes bibliographical references and index.

ISBN 0-262-01197-2

1. Neural networks (Neurobiology)—Handbooks, manuals, etc.
2. Neural networks (Computer science)—Handbooks, manuals, etc.

I. Arbib, Michael A.

QP363.3.H36 2002

612.8'2—dc21

2002038664

CIP

Contents

Preface to the Second Edition ix
Preface to the First Edition xi
How to Use This Book xv

Part I: Background: The Elements of Brain Theory and Neural Networks 1

How to Use Part I 3

- I.1. Introducing the Neuron 3
 - The Diversity of Receptors* 4
 - Basic Properties of Neurons* 4
 - Receptors and Effectors* 7
 - Neural Models* 7
 - More Detailed Properties of Neurons* 9
- I.2. Levels and Styles of Analysis 10
 - A Historical Fragment* 10
 - Brains, Machines, and Minds* 11
 - Levels of Analysis* 12
 - Schema Theory* 13
- I.3. Dynamics and Adaptation in Neural Networks 15
 - Dynamic Systems* 15
 - Continuous-Time Systems* 15
 - Discrete-Time Systems* 16
 - Stability, Limit Cycles, and Chaos* 16
 - Hopfield Nets* 17
 - Adaptation in Dynamic Systems* 18
 - Adaptive Control* 18
 - Pattern Recognition* 18
 - Associative Memory* 19
 - Learning Rules* 19
 - Hebbian Plasticity and Network Self-Organization* 19
 - Perceptrons* 20
 - Network Complexity* 20
 - Gradient Descent and Credit Assignment* 21
 - Backpropagation* 21
 - A Cautionary Note* 22
 - Envoi* 23

Part II: Road Maps: A Guided Tour of Brain Theory and Neural Networks 25

How to Use Part II 27

- II.1. The Meta-Map 27
- II.2. Grounding Models of Neurons and Networks 29
 - Grounding Models of Neurons* 29
 - Grounding Models of Networks* 31

- II.3. Brain, Behavior, and Cognition 31
 - Neuroethology and Evolution* 31
 - Mammalian Brain Regions* 34
 - Cognitive Neuroscience* 37
- II.4. Psychology, Linguistics, and Artificial Intelligence 40
 - Psychology* 40
 - Linguistics and Speech Processing* 42
 - Artificial Intelligence* 44
- II.5. Biological Neurons and Networks 47
 - Biological Neurons and Synapses* 47
 - Neural Plasticity* 49
 - Neural Coding* 52
 - Biological Networks* 54
- II.6. Dynamics and Learning in Artificial Networks 55
 - Dynamic Systems* 55
 - Learning in Artificial Networks* 58
 - Computability and Complexity* 64
- II.7. Sensory Systems 65
 - Vision* 65
 - Other Sensory Systems* 70
- II.8. Motor Systems 71
 - Robotics and Control Theory* 71
 - Motor Pattern Generators* 73
 - Mammalian Motor Control* 74
- II.9. Applications, Implementations, and Analysis 77
 - Applications* 77
 - Implementation and Analysis* 78

Part III: Articles 81

The articles in Part III are arranged alphabetically by title. To retrieve articles by author, turn to the contributors list, which begins on page 1241.

- Action Monitoring and Forward Control of Movements 83
- Activity-Dependent Regulation of Neuronal Conductances 85
- Adaptive Resonance Theory 87
- Adaptive Spike Coding 90
- Amplification, Attenuation, and Integration 94
- Analog Neural Nets: Computational Power 97
- Analog VLSI Implementations of Neural Networks 101
- Analogy-Based Reasoning and Metaphor 106
- Arm and Hand Movement Control 110
- Artificial Intelligence and Neural Networks 113

Associative Networks	117	Dendritic Learning	320
Auditory Cortex	122	Dendritic Processing	324
Auditory Periphery and Cochlear Nucleus	127	Dendritic Spines	332
Auditory Scene Analysis	132	Development of Retinotectal Maps	335
Axonal Modeling	135	Developmental Disorders	339
Axonal Path Finding	140	Diffusion Models of Neuron Activity	343
Backpropagation: General Principles	144	Digital VLSI for Neural Networks	349
Basal Ganglia	147	Directional Selectivity	353
Bayesian Methods and Neural Networks	151	Dissociations Between Visual Processing Modes	358
Bayesian Networks	157	Dopamine, Roles of	361
Biologically Inspired Robotics	160	Dynamic Link Architecture	365
Biophysical Mechanisms in Neuronal Modeling	164	Dynamic Remapping	368
Biophysical Mosaic of the Neuron	170	Dynamics and Bifurcation in Neural Nets	372
Brain Signal Analysis	175	Dynamics of Association and Recall	377
Brain-Computer Interfaces	178	Echolocation: Cochleotopic and Computational Maps	381
Canonical Neural Models	181	EEG and MEG Analysis	387
Cerebellum and Conditioning	187	Electrolocation	391
Cerebellum and Motor Control	190	Embodied Cognition	395
Cerebellum: Neural Plasticity	196	Emotional Circuits	398
Chains of Oscillators in Motor and Sensory Systems	201	Energy Functionals for Neural Networks	402
Chaos in Biological Systems	205	Ensemble Learning	405
Chaos in Neural Systems	208	Equilibrium Point Hypothesis	409
Cognitive Development	212	Event-Related Potentials	412
Cognitive Maps	216	Evolution and Learning in Neural Networks	415
Cognitive Modeling: Psychology and Connectionism	219	Evolution of Artificial Neural Networks	418
Collective Behavior of Coupled Oscillators	223	Evolution of Genetic Networks	421
Collicular Visuomotor Transformations for Gaze Control	226	Evolution of the Ancestral Vertebrate Brain	426
Color Perception	230	Eye-Hand Coordination in Reaching Movements	431
Command Neurons and Command Systems	233	Face Recognition: Neurophysiology and Neural Technology	434
Competitive Learning	238	Face Recognition: Psychology and Connectionism	438
Competitive Queuing for Planning and Serial Performance	241	Fast Visual Processing	441
Compositionality in Neural Systems	244	Feature Analysis	444
Computing with Attractors	248	Filtering, Adaptive	449
Concept Learning	252	Forecasting	453
Conditioning	256	Gabor Wavelets and Statistical Pattern Recognition	457
Connectionist and Symbolic Representations	260	Gait Transitions	463
Consciousness, Neural Models of	263	Gaussian Processes	466
Constituency and Recursion in Language	267	Generalization and Regularization in Nonlinear Learning Systems	470
Contour and Surface Perception	271	GENESIS Simulation System	475
Convolutional Networks for Images, Speech, and Time Series	276	Geometrical Principles in Motor Control	476
Cooperative Phenomena	279	Global Visual Pattern Extraction	482
Cortical Hebbian Modules	285	Graphical Models: Parameter Learning	486
Cortical Memory	290	Graphical Models: Probabilistic Inference	490
Cortical Population Dynamics and Psychophysics	294	Graphical Models: Structure Learning	496
Covariance Structural Equation Modeling	300	Grasping Movements: Visuomotor Transformations	501
Crustacean Stomatogastric System	304	Habituation	504
Data Clustering and Learning	308	Half-Center Oscillators Underlying Rhythmic Movements	507
Databases for Neuroscience	312		
Decision Support Systems and Expert Systems	316		

Hebbian Learning and Neuronal Regulation	511	Motor Primitives	701
Hebbian Synaptic Plasticity	515	Motor Theories of Perception	705
Helmholtz Machines and Sleep-Wake Learning	522	Multiagent Systems	707
Hemispheric Interactions and Specialization	525	Muscle Models	711
Hidden Markov Models	528	Neocognitron: A Model for Visual Pattern Recognition	715
Hippocampal Rhythm Generation	533	Neocortex: Basic Neuron Types	719
Hippocampus: Spatial Models	539	Neocortex: Chemical and Electrical Synapses	725
Hybrid Connectionist/Symbolic Systems	543	Neural Automata and Analog Computational Complexity	729
Identification and Control	547	Neuroanatomy in a Computational Perspective	733
Imaging the Grammatical Brain	551	Neuroethology, Computational	737
Imaging the Motor Brain	556	Neuroinformatics	741
Imaging the Visual Brain	562	Neurolinguistics	745
Imitation	566	Neurological and Psychiatric Disorders	751
Independent Component Analysis	569	Neuromanifolds and Information Geometry	754
Information Theory and Visual Plasticity	575	Neuromodulation in Invertebrate Nervous Systems	757
Integrate-and-Fire Neurons and Networks	577	Neuromodulation in Mammalian Nervous Systems	761
Invertebrate Models of Learning: <i>Aplysia</i> and <i>Hermissenda</i>	581	Neuromorphic VLSI Circuits and Systems	765
Ion Channels: Keys to Neuronal Specialization	585	NEURON Simulation Environment	769
Kalman Filtering: Neural Implications	590	Neuropsychological Impairments	773
Laminar Cortical Architecture in Visual Perception	594	Neurosimulation: Tools and Resources	776
Language Acquisition	600	NMDA Receptors: Synaptic, Cellular, and Network Models	781
Language Evolution and Change	604	NSL Neural Simulation Language	784
Language Evolution: The Mirror System Hypothesis	606	Object Recognition	788
Language Processing	612	Object Recognition, Neurophysiology	792
Layered Computation in Neural Networks	616	Object Structure, Visual Processing	797
Learning and Generalization: Theoretical Bounds	619	Ocular Dominance and Orientation Columns	801
Learning and Statistical Inference	624	Olfactory Bulb	806
Learning Network Topology	628	Olfactory Cortex	810
Learning Vector Quantization	631	Optimal Sensory Encoding	815
Lesioned Networks as Models of Neuropsychological Deficits	635	Optimality Theory in Linguistics	819
Limb Geometry, Neural Control	638	Optimization, Neural	822
Localized Versus Distributed Representations	643	Optimization Principles in Motor Control	827
Locomotion, Invertebrate	646	Orientation Selectivity	831
Locomotion, Vertebrate	649	Oscillatory and Bursting Properties of Neurons	835
Locust Flight: Components and Mechanisms in the Motor	654	PAC Learning and Neural Networks	840
Markov Random Field Models in Image Processing	657	Pain Networks	843
Memory-Based Reasoning	661	Past Tense Learning	848
Minimum Description Length Analysis	662	Pattern Formation, Biological	851
Model Validation	666	Pattern Formation, Neural	859
Modular and Hierarchical Learning Systems	669	Pattern Recognition	864
Motion Perception: Elementary Mechanisms	672	Perception of Three-Dimensional Structure	868
Motion Perception: Navigation	676	Perceptrons, Adalines, and Backpropagation	871
Motivation	680	Perspective on Neuron Model Complexity	877
Motoneuron Recruitment	683	Phase-Plane Analysis of Neural Nets	881
Motor Control, Biological and Theoretical	686	Philosophical Issues in Brain Theory and Connectionism	886
Motor Cortex: Coding and Decoding of Directional Operations	690	Photonic Implementations of Neurobiologically Inspired Networks	889
Motor Pattern Generation	696		

Population Codes	893	Speech Production	1072
Post-Hebbian Learning Algorithms	898	Speech Recognition Technology	1076
Potential Fields and Neural Networks	901	Spiking Neurons, Computation with	1080
Prefrontal Cortex in Temporal Organization of Action	905	Spinal Cord of Lamprey: Generation of Locomotor Patterns	1084
Principal Component Analysis	910	Statistical Mechanics of Generalization	1087
Probabilistic Regularization Methods for Low-Level Vision	913	Statistical Mechanics of Neural Networks	1090
Programmable Neurocomputing Systems	916	Statistical Mechanics of On-line Learning and Generalization	1095
Prosthetics, Motor Control	919	Statistical Parametric Mapping of Cortical Activity Patterns	1098
Prosthetics, Neural	923	Stereo Correspondence	1104
Prosthetics, Sensory Systems	926	Stochastic Approximation and Efficient Learning	1108
Pursuit Eye Movements	929	Stochastic Resonance	1112
Q-Learning for Robots	934	Structured Connectionist Models	1116
Radial Basis Function Networks	937	Support Vector Machines	1119
Rate Coding and Signal Processing	941	Synaptic Interactions	1126
Reaching Movements: Implications for Computational Models	945	Synaptic Noise and Chaos in Vertebrate Neurons	1130
Reactive Robotic Systems	949	Synaptic Transmission	1133
Reading	951	Synchronization, Binding and Expectancy	1136
Recurrent Networks: Learning Algorithms	955	Synfire Chains	1143
Recurrent Networks: Neurophysiological Modeling	960	Synthetic Functional Brain Mapping	1146
Reinforcement Learning	963	Systematicity of Generalizations in Connectionist Networks	1151
Reinforcement Learning in Motor Control	968	Temporal Dynamics of Biological Synapses	1156
Respiratory Rhythm Generation	972	Temporal Integration in Recurrent Microcircuits	1159
Retina	975	Temporal Pattern Processing	1163
Robot Arm Control	979	Temporal Sequences: Learning and Global Analysis	1167
Robot Learning	983	Tensor Voting and Visual Segmentation	1171
Robot Navigation	987	Thalamus	1176
Rodent Head Direction System	990	Universal Approximators	1180
Schema Theory	993	Unsupervised Learning with Global Objective Functions	1183
Scratch Reflex	999	Vapnik-Chervonenkis Dimension of Neural Networks	1188
Self-Organization and the Brain	1002	Vestibulo-Ocular Reflex	1192
Self-Organizing Feature Maps	1005	Visual Attention	1196
Semantic Networks	1010	Visual Cortex: Anatomical Structure and Models of Function	1202
Sensor Fusion	1014	Visual Course Control in Flies	1205
Sensorimotor Interactions and Central Pattern Generators	1016	Visual Scene Perception, Neurophysiology	1210
Sensorimotor Learning	1020	Visual Scene Segmentation	1215
Sensory Coding and Information Transmission	1023	Visuomotor Coordination in Frog and Toad	1219
Sequence Learning	1027	Visuomotor Coordination in Salamander	1225
Short-Term Memory	1030	Winner-Take-All Networks	1228
Silicon Neurons	1034	Ying-Yang Learning	1231
Simulated Annealing and Boltzmann Machines	1039		
Single-Cell Models	1044	Editorial Advisory Board	1239
Sleep Oscillations	1049	Contributors	1241
Somatosensory System	1053	Subject Index	1255
Somatotopy: Plasticity of Sensory Maps	1057		
Sound Localization and Binaural Processing	1061		
Sparse Coding in the Primate Cortex	1064		
Speech Processing: Psycholinguistics	1068		

Preface to the Second Edition

Like the first edition, which it replaces, this volume is inspired by two great questions: “How does the brain work?” and “How can we build intelligent machines?” As in the first edition, the heart of the book is a set of close to 300 articles in Part III which cover the whole spectrum of *Brain Theory and Neural Networks*. To help readers orient themselves with respect to this cornucopia, I have written Part I to provide the elementary background on the modeling of both brains and biological and artificial neural networks, and Part II to provide a series of road maps to help readers interested in a particular topic steer through the Part III articles on that topic. More on the motivation and structure of the book can be found in the Preface to the First Edition, which is reproduced after this. I also recommend reading the section “How to Use This Book”—one reader of the first edition who did not do so failed to realize that the articles in Part III were in alphabetical order, or that the Contributors list lets one locate each article written by a given author.

The reader new to the study of *Brain Theory and Neural Networks* will find it wise to read Part I for orientation before jumping into Part III, whereas more experienced readers will find most of Part I familiar. Many readers will simply turn to articles in Part III of particular interest at a given time. However, to help readers who seek a more systematic view of a particular subfield of *Brain Theory and Neural Networks*, Part II provides 22 Road Maps, each providing an essay linking most of the articles on a given topic. (I say “most” because the threshold is subjective for deciding when a particular article has more than a minor mention of the topic in a Road Map.) The Road Maps are organized into 8 groups in Part II as follows:

Grounding Models of Neurons and Networks

- Grounding Models of Neurons
- Grounding Models of Networks

Brain, Behavior, and Cognition

- Neuroethology and Evolution
- Mammalian Brain Regions
- Cognitive Neuroscience

Psychology, Linguistics, and Artificial Intelligence

- Psychology
- Linguistics and Speech Processing
- Artificial Intelligence

Biological Neurons and Networks

- Biological Neurons and Synapses
- Neural Plasticity
- Neural Coding
- Biological Networks

Dynamics and Learning in Artificial Networks

- Dynamic Systems
- Learning in Artificial Networks
- Computability and Complexity

Sensory Systems

- Vision
- Other Sensory Systems

Motor Systems

- Robotics and Control Theory
- Motor Pattern Generators
- Mammalian Motor Control

Applications, Implementations, and Analysis

Applications

Implementation and Analysis

The authors of the articles in Part III come from a broad spectrum of disciplines—such as biomedical engineering, cognitive science, computer science, electrical engineering, linguistics, mathematics, physics, neurology, neuroscience, and psychology—and have worked hard to make their articles accessible to readers across the spectrum. The utility of each article is enhanced by cross-references to other articles within the body of the article, and lists at the end of the article referring the reader to road maps, background material, and related reading.

To get some idea of how radically the new edition differs from the old, note that the new edition has 285 articles in Part III, as against the 266 articles of the first edition. Of the articles that appeared in the first edition, only 9 are reprinted unchanged. Some 135 have been updated (or even completely rewritten) by their original authors, and more than 30 have been written anew by new authors. In addition, there are over 100 articles on new topics. The primary shift of emphasis from the first edition has been to drastically reduce the number of articles on applications of artificial neural networks (from astronomy to steelmaking) and to greatly increase the coverage of models of fundamental neurobiology and neural network approaches to language, and to add the new papers which are now listed in the Road Maps on Cognitive Neuroscience, Neural Coding, and Other Sensory Systems (i.e., other than Vision, for which coverage has also been increased). Certainly, a number of the articles in the first edition remain worthy of reading in themselves, but the aim has been to make the new edition a self-contained introduction to brain theory and neural networks in all its current breadth and richness.

The new edition not only appears in print but also has its own web site.

Acknowledgments

My foremost acknowledgment is again to Prue Arbib, who served as Editorial Assistant during the long and arduous process of eliciting and assembling the many, many contributions to Part III. I thank the members of the Editorial Advisory Board, who helped update the list of articles from the first edition and focus the search for authors, and I thank these authors not only for their contributions to Part III but also for suggesting further topics and authors for the *Handbook*, in an ever-widening circle as work advanced on this new edition. I also owe a great debt to the hundreds of reviewers who so constructively contributed to the final polishing of the articles that now appear in Part III. Finally, I thank the staff of P. M. Gordon Associates and of The MIT Press for once again meeting the high standards of copy editing and book production that contributed so much to the success of the first edition.

Michael A. Arbib
Los Angeles and La Jolla
October 2002

Preface to the First Edition

This volume is inspired by two great questions: “How does the brain work?” and “How can we build intelligent machines?” It provides no simple, single answer to either question because no single answer, simple or otherwise, exists. However, in hundreds of articles it charts the immense progress made in recent years in answering many related, but far more specific, questions.

The term *neural networks* has been used for a century or more to describe the networks of biological neurons that constitute the nervous systems of animals, whether invertebrates or vertebrates. Since the 1940s, and especially since the 1980s, the term has been used for a technology of parallel computation in which the computing elements are “artificial neurons” loosely modeled on simple properties of biological neurons, usually with some adaptive capability to change the strengths of connections between the neurons.

Brain theory is centered on “computational neuroscience,” the use of computational techniques to model biological neural networks, but also includes attempts to understand the brain and its function through a variety of theoretical constructs and computer analogies. In fact, as the following pages reveal, much of brain theory is not about neural networks per se, but focuses on structural and functional “networks” whose units are in scales both coarser and finer than that of the neuron. Computer scientists, engineers, and physicists have analyzed and applied artificial neural networks inspired by the adaptive, parallel computing style of the brain, but this *Handbook* will also sample non-neural approaches to the design and analysis of “intelligent” machines. In between the biologists and the technologists are the connectionists. They use artificial neural networks in psychology and linguistics and make related contributions to artificial intelligence, using neuron-like units which interact “in the style of the brain” at a more abstract level than that of individual biological neurons.

Many texts have described limited aspects of one subfield or another of brain theory and neural networks, but no truly comprehensive overview is available. The aim of this *Handbook* is to fill that gap, presenting the entire range of the following topics: detailed models of single neurons; analysis of a wide variety of neurobiological systems; “connectionist” studies; mathematical analyses of abstract neural networks; and technological applications of adaptive, artificial neural networks and related methodologies. The excitement, and the frustration, of these topics is that they span such a broad range of disciplines, including mathematics, statistical physics and chemistry, neurology and neurobiology, and computer science and electrical engineering, as well as cognitive psychology, artificial intelligence, and philosophy. Much effort, therefore, has gone into making the book accessible to readers with varied backgrounds (an undergraduate education in one of the above areas, for example, or the frequent reading of related articles at the level of the *Scientific American*) while still providing a clear view of much of the recent specialized research.

The heart of the book comes in Part III, in which the breadth of brain theory and neural networks is sampled in 266 articles, presented in alphabetical order by title. Each article meets the following requirements:

1. It is authoritative within its own subfield, yet accessible to students and experts in a wide range of other fields.
2. It is comprehensive, yet short enough that its concepts can be acquired in a single sitting.
3. It includes a list of references, limited to 15, to give the reader a well-defined and selective list of places to go to initiate further study.
4. It is as self-contained as possible, while providing cross-references to allow readers to explore particular issues of related interest.

Despite the fourth requirement, some articles are more self-contained than others. Some articles can be read with almost no prior knowledge; some can be read with a rather general knowledge of a few key concepts; others require fairly detailed understanding of material covered in other articles. For example, many articles on applications will make sense only if one understands the “backpropagation” technique for training artificial neural networks; and a number of studies of neuronal function will make sense only if one has at least some idea of the Hodgkin-Huxley equation. Whenever appropriate, therefore, the articles include advice on background articles.

Parts I and II of the book provide a more general approach to helping readers orient themselves. Part I: Background presents a perspective on the “landscape” of brain theory and neural networks, including an exposition of the key concepts for viewing neural networks as dynamic, adaptive systems. Part II: Road Maps then provides an entrée into the many articles of Part III, with “road maps” for 23 different themes. The “Meta-Map,” which introduces Part II, groups these themes under eight general headings which, in and of themselves, give some sense of the sweep of the *Handbook*:

Connectionism: Psychology, Linguistics, and Artificial Intelligence
Dynamics, Self-Organization, and Cooperativity
Learning in Artificial Neural Networks
Applications and Implementations
Biological Neurons and Networks
Sensory Systems
Plasticity in Development and Learning
Motor Control

A more detailed view of the structure of the book is provided in the introductory section “How to Use this Book.” The aim is to ensure that readers will not only turn to the book to get good brief reviews of topics in their own specialty, but also will find many invitations to browse widely—finding parallels amongst different subfields, or simply enjoying the discovery of interesting topics far from familiar territory.

Acknowledgments

My foremost acknowledgment is to Prue Arbib, who served as Editorial Assistant during the long and arduous process of eliciting and assembling the many, many contributions to Part III; we both thank Paulina Tagle for her help with our work. The initial plan for the book was drawn up in 1991, and it benefited from the advice of a number of friends, especially George Adelman, who shared his experience as Editor of the *Encyclopedia of Neuroscience*. Refinement of the plan and the choice of publishers occupied the first few months of 1992, and I thank Fiona Stevens of The MIT Press for her support of the project from that time onward.

As can be imagined, the plan for a book like this has developed through a time-consuming process of constraint satisfaction. The first steps were to draw up a list of about 20 topic areas (similar to, but not identical with, the 23 areas surveyed in Part II), to populate these areas with a preliminary list of over 100 articles and possible authors, and to recruit the first members of the Editorial Advisory Board to help expand the list of articles and focus on the search for authors. A very satisfying number of authors invited in the first round accepted my invitation, and many of these added their voices to the Editorial Advisory Board in suggesting further topics and authors for the *Handbook*.

I was delighted, stimulated, and informed as I read the first drafts of the articles; but I have also been grateful for the fine spirit of cooperation with which the authors have responded to editorial comments and reviews. The resulting articles not only are authoritative and accessible in themselves, but also have been revised to match the overall style of the *Handbook* and to meet the needs of a broad readership. With this I express my sincere thanks to the editorial advisors, the authors, and the hundreds of reviewers who so

constructively contributed to the final polishing of the articles that now appear in Part III; to Doug Gordon and the copy editors and typesetters who transformed the diverse styles of the manuscripts into the style of the *Handbook*; and to the graduate students who helped so much with the proofreading.

Finally, I want to record a debt that did not reach my conscious awareness until well into the editing of this book. It is to Hiram Haydn, who for many years was editor of *The American Scholar*, which is published for general circulation by Phi Beta Kappa. In 1971 or so, Phi Beta Kappa conducted a competition to find authors to receive grants for books to be written, if memory serves aright, for the Bicentennial of the United States. I submitted an entry. Although I was not successful, Mr. Haydn, who had been a member of the jury, wrote to express his appreciation of that entry, and to invite me to write an article for the *Scholar*. What stays in my mind from the ensuing correspondence was the sympathetic way in which he helped me articulate the connections that were at best implicit in my draft, and find the right voice in which to “speak” with the readers of a publication so different from the usual scientific journal. I now realize that it is his example I have tried to follow as I have worked with these hundreds of authors in the quest to see the subject of brain theory and neural networks whole, and to share it with readers of diverse interests and backgrounds.

Michael A. Arbib
Los Angeles and La Jolla
January 1995

How to Use This Book

More than 90% of this book is taken up by Part III, which, in 285 separately authored articles, covers a vast range of topics in brain theory and neural networks, from language to motor control, and from the neurochemistry to the statistical mechanics of memory. Each article has been made as self-contained as possible, but the very breadth of topics means that few readers will be expert in a majority of them. To help the reader new to certain areas of the *Handbook*, I have prepared Part I: Background and Part II: Road Maps. The next few pages describe these aids to comprehension, as well as offering more information on the structure of articles in Part III.

Part I: Background: The Elements of Brain Theory and Neural Networks

Part I provides background material for readers new to computational neuroscience or theoretical approaches to neural networks considered as dynamic, adaptive systems. Section I.1, "Introducing the Neuron," conveys the basic properties of neurons and introduces several basic neural models. Section I.2, "Levels and Styles of Analysis," explains the interdisciplinary nexus in which the present study of brain theory and neural networks is located, with historical roots in cybernetics and with current work going back and forth between brain theory, artificial intelligence, and cognitive psychology. We also review the different levels of analysis involved, with schemas providing the functional units intermediate between an overall task and neural networks. Finally, Section I.3, "Dynamics and Adaptation in Neural Networks," provides a tutorial on the concepts essential for understanding neural networks as dynamic, adaptive systems. We close by stressing that the full understanding of the brain and the improved design of intelligent machines will require not only improvements in the learning methods presented in Section I.3, but also fuller understanding of architectures based on networks of networks, with initial structures well constrained for the task at hand.

Part II: Road Maps: A Guided Tour of Brain Theory and Neural Networks

The reader who wants to survey a major theme of brain theory and neural networks, rather than seeking articles in Part III one at a time, will find in Part II a set of 22 road maps that, among them, place every article in Part III in a thematic perspective. Section II.1 presents a Meta-Map, which briefly surveys all these themes, grouping them under eight general headings:

Grounding Models of Neurons and Networks

- Grounding Models of Neurons
- Grounding Models of Networks

Brain, Behavior, and Cognition

- Neuroethology and Evolution
- Mammalian Brain Regions
- Cognitive Neuroscience

Psychology, Linguistics, and Artificial Intelligence

- Psychology
- Linguistics and Speech Processing
- Artificial Intelligence

Biological Neurons and Networks

- Biological Neurons and Synapses
- Neural Plasticity
- Neural Coding
- Biological Networks

Dynamics and Learning in Artificial Networks

- Dynamic Systems
- Learning in Artificial Networks
- Computability and Complexity

Sensory Systems

- Vision
- Other Sensory Systems

Motor Systems

- Robotics and Control Theory
- Motor Pattern Generators
- Mammalian Motor Control

Applications, Implementations, and Analysis

- Applications
- Implementation and Analysis

This ordering of the themes has no special significance. It is simply one way to approach the richness of the *Handbook*, making it easy for you to identify one or two key road maps of special interest. By the same token, the order of articles in each of the 22 road maps that follow the Meta-Map is one among many such orderings. Each road map starts with an alphabetical listing of the articles most relevant to the current theme. The road map itself will provide suggestions for *interesting* traversals of articles, but this need not imply that an article provides *necessary* background for the articles it precedes.

Part III: Articles

Part III comprises 285 articles. These articles are arranged in alphabetical order, both to make it easier to find a specific topic (although a Subject Index is provided as well, and the alphabetical list of Contributors on page 1241 lists all the articles to which each author has contributed) and because a given article may be relevant to more than one of the themes of Part II, a fact that would be hidden were the article to be relegated to a specific section devoted to a single theme. Most of these articles assume some prior familiarity with neural networks, whether biological or artificial, and so the reader new to neural networks is encouraged to master the material in Part I before tackling Part III.

Most articles in Part III have the following structure: The introduction provides a non-technical overview of the material covered in the whole article, while the final section provides a discussion of key points, open questions, and linkages with other areas of brain theory and neural networks. The intervening sections may be more or less technical, depending on the nature of the topic, but the first and last sections should give most readers a basic appreciation of the topic, irrespective of such technicalities. The bibliography for each article contains about 15 references. People who find their favorite papers omitted from the list should blame my editorial decision, not the author's judgment. The style I chose for the *Handbook* was *not* to provide exhaustive coverage of research papers for the expert. Rather, references are there primarily to help readers who look for an *introduction* to the literature on the given topic, including background material, relevant review articles, and original research citations. In addition to formal references to the literature, each article contains numerous cross-references to other articles in the *Handbook*. These may occur either in the body of the article in the form THE TITLE OF THE ARTICLE IN SMALL CAPS, or at the end of the article, designated as “**Related Reading**.” In addition to suggestions for related reading, the reader will find, just prior to the list of references in each article, a mention of the **road map(s)** in which the article is discussed, as well as **background** material, when the article is more advanced.

In summary, turn directly to Part III when you need information on a specific topic. Read sections of Part I to gain a general perspective on the basic concepts of brain theory and neural networks. For an overview of some theme, read the Meta-Map in Part II to

choose road maps in Part II; read a road map to choose articles in Part III. A road map can also be used as an explicit guide for systematic study of the area under review. Then continue your exploration through further use of road maps, by following cross-references in Part III, by looking up terms of interest in the index, or simply by letting serendipity take its course as you browse through Part III at random.

Part I: Background

The Elements of Brain Theory and Neural Networks

Michael A. Arbib

How to Use Part I

Part I provides background material, summarizing a set of concepts established for the formal study of neurons and neural networks by 1986. As such, it is designed to hold few, if any, surprises for readers with a fair background in computational neuroscience or theoretical approaches to neural networks considered as dynamic, adaptive systems. Rather, Part I is designed for the many readers—be they neuroscience experimentalists, psychologists, philosophers, or technologists—who are sufficiently new to *brain theory and neural networks* that they can benefit from a compact overview of basic concepts prior to reading the road maps of Part II and the articles in Part III. Of course, much of what is covered in Part I is also covered at some length in the articles in Part III, and cross-references will steer the reader to these articles for alternative expositions and reviews of current research. In this exposition, as throughout the *Handbook*, we will move back and forth between *computational neuroscience*, where the emphasis is on modeling biological neurons, and *neural computing*, where the emphasis shifts back and forth between biological models and artificial neural networks based loosely on abstractions from biology, but driven more by technological utility than by biological considerations.

Section I.1, “Introducing the Neuron,” conveys the basic properties of neurons, receptors, and effectors, and then introduces several simple neural models, including the discrete-time McCulloch-Pitts model and the continuous-time leaky integrator model. References to Part III alert the reader to more detailed properties of neurons which are essential for the neuroscientist and provide interesting hints about future design features for the technologist.

Section I.2, “Levels and Styles of Analysis,” is designed to give the reader a feel for the interdisciplinary nexus in which the present study of brain theory and neural networks is located. The selection begins with a historical fragment which traces our federation of disciplines back to their roots in cybernetics, the study of control and communication in animals and machines. We look at the way in which the research addresses brains, machines, and minds, going

back and forth between brain theory, artificial intelligence, and cognitive psychology. We then review the different levels of analysis involved, whether we study brains or intelligent machines, and the use of schemas to provide intermediate functional units that bridge the gap between an overall task and the neural networks which implement it.

Section I.3, “Dynamics and Adaptation in Neural Networks,” provides a tutorial on the concepts essential for understanding neural networks as dynamic, adaptive systems. It introduces the basic dynamic systems concepts of stability, limit cycles, and chaos, and relates Hopfield nets to attractors and optimization. It then introduces a number of basic concepts concerning adaptation in neural nets, with discussions of pattern recognition, associative memory, Hebbian plasticity and network self-organization, perceptrons, network complexity, gradient descent and credit assignment, and backpropagation. This section, and with it Part I, closes with a cautionary note. The basic learning rules and adaptive architectures of neural networks have already illuminated a number of biological issues and led to useful technological applications. However, these networks must have their initial structure well constrained (whether by evolution or technological design) to yield approximate solutions to the system’s tasks—solutions that can then be efficiently and efficaciously shaped by experience. Moreover, the full understanding of the brain and the improved design of intelligent machines will require not only improvements in these learning methods and their initialization, but also a fuller understanding of architectures based on networks of networks. Cross-references to articles in Part III will set the reader on the path to this fuller understanding. Because Part I focuses on the basic concepts established for the formal study of neurons and neural networks by 1986, it differs hardly at all from Part I of the first edition of the *Handbook*. By contrast, Part II, which provides the road maps that guide readers through the radically updated Part III, has been completely rewritten for the present edition to reflect the latest research results.

I.1. Introducing the Neuron

We introduce the *neuron*. The dangerous word in the preceding sentence is *the*. In biology, there are radically different types of neurons in the human brain, and endless variations in neuron types of other species. In brain theory, the complexities of real neurons are abstracted in many ways to aid in understanding different aspects of neural network development, learning, or function. In *neural computing* (technology based on networks of “neuron-like” units), the artificial neurons are designed as variations on the abstractions of brain theory and are implemented in software, or VLSI or other media. There is no such thing as a “typical” neuron, yet this section will nonetheless present examples and models which provide a starting point, an essential set of key concepts, for the appreciation of the many variations on the theme of neurons and neural networks presented in Part III.

An analogy to the problem we face here might be to define *vehicle* for a handbook of transportation. A vehicle could be a car, a train, a plane, a rowboat, or a forklift truck. It might or might not carry people. The people could be crew or passengers, and so on. The problem would be to give a few key examples of form (such as car versus plane) and function (to carry people or goods, by land, air, or sea, etc.). Moreover, we would find interesting examples of co-evolution: for example, modern highway systems would

not have been created without the pressure of increasing car traffic; most features of cars are adapted to the existence of sealed roads, and some features (e.g., cruise control) are specifically adapted to good freeway conditions. Following a similar procedure, Part III offers diverse examples of neural form and function in both biology and technology.

Here, we start with the observation that a brain is made up of a network of cells called neurons, coupled to receptors and effectors. Neurons are intimately connected with glial cells, which provide support functions for neural networks. New empirical data show the importance of glia in regeneration of neural networks after damage and in maintaining the neurochemical milieu during normal operation. However, such data have had very little impact on neural modeling and so will not be considered further here. The input to the network of neurons is provided by *receptors*, which continually monitor changes in the external and internal environment. Cells called *motor neurons* (or *motoneurons*), governed by the activity of the neural network, control the movement of muscles and the secretion of glands. In between, an intricate network of neurons (a few hundred neurons in some simple creatures, hundreds of billions in a human brain) continually combines the signals from the receptors with signals encoding past experience to barrage the motor

neurons with signals that will yield adaptive interactions with the environment. In animals with backbones (vertebrates, including mammals in general and humans in particular), this network is called the *central nervous system* (CNS), and the brain constitutes the most headward part of this system, linked to the receptors and effectors of the body via the spinal cord. Invertebrate nervous systems (neural networks) provide astounding variations on the vertebrate theme, thanks to eons of divergent evolution. Thus, while the human brain may be the source of rich analogies for technologists in search of “artificial intelligence,” both invertebrates and vertebrates provide endless ideas for technologists designing neural networks for sensory processing, robot control, and a host of other applications. (A few of the relevant examples may be found in the Part II road maps, **Vision, Robotics and Control Theory, Motor Pattern Generators**, and **Neuroethology and Evolution**.)

The brain provides far more than a simple stimulus-response chain from receptors to effectors (although there are such reflex paths). Rather, the vast network of neurons is interconnected in loops and tangled skeins so that signals entering the net from the receptors interact there with the billions of signals already traversing the system, not only to yield the signals that control the effectors but also to modify the very properties of the network itself, so that future behavior will reflect prior experience.

The Diversity of Receptors

Rod and cone receptors in the eyes respond to light, hair cells in the ears respond to pressure, and other cells in the tongue and the mouth respond to subtle traces of chemicals. In addition to touch receptors, there are receptors in the skin that are responsive to movement or to temperature, or that signal painful stimuli. These external senses may be divided into two classes: (1) the proximity senses, such as touch and taste, which sense objects in contact with the body surface, and (2) the distance senses, such as vision and hearing, which let us sense objects distant from the body. Olfaction is somewhere in between, using chemical signals “right under our noses” to sense nonproximate objects. Moreover, even the proximate senses can yield information about nonproximate objects, as when we feel the wind or the heat of a fire. More generally, much of our appreciation of the world around us rests on the unconscious fusion of data from diverse sensory systems.

The appropriate activity of the effectors must depend on comparing where the system should be—the current target of an ongoing movement—with where it is now. Thus, in addition to the

external receptors, there are receptors that monitor the activity of muscles, tendons, and joints to provide a continual source of feedback about the tensions and lengths of muscles and the angles of the joints, as well as their velocities. The vestibular system in the head monitors gravity and accelerations. Here, the receptors are hair cells monitoring fluid motion. There are also receptors to monitor the chemical level of the bloodstream and the state of the heart and the intestines. Cells in the liver monitor glucose, while others in the kidney check water balance. Receptors in the *hypothalamus*, itself a part of the brain, also check the balance of water and sugar. The hypothalamus then integrates these diverse messages to direct behavior or other organs to restore the balance. If we stimulate the hypothalamus, an animal may drink copious quantities of water or eat enormous quantities of food, even though it is already well supplied; the brain has received a signal that water or food is lacking, and so it instructs the animal accordingly, irrespective of whatever contradictory signals may be coming from a distended stomach.

Basic Properties of Neurons

To understand the processes that intervene between receptors and effectors, we must have a closer look at “the” neuron. As already emphasized, *there is no such thing as a typical neuron*. However, we will summarize properties shared by many neurons. The “basic neuron” shown in Figure 1 is abstracted from a motor neuron of mammalian spinal cord. From the *soma* (cell body) protrudes a number of ramifying branches called *dendrites*; the soma and dendrites constitute the input surface of the neuron. There also extrudes from the cell body, at a point called the *axon hillock* (abutting the initial segment), a long fiber called the *axon*, whose branches form the *axonal arborization*. The tips of the branches of the axon, called *nerve terminals* or *boutons*, impinge on other neurons or on effectors. The locus of interaction between a bouton and the cell on which it impinges is called a *synapse*, and we say that the cell with the bouton *synapses upon* the cell with which the connection is made. In fact, axonal branches of some neurons can have many varicosities, corresponding to synapses, along their length, not just at the end of the branch.

We can imagine the flow of information as shown by the arrows in Figure 1. Although “conduction” can go in either direction on the axon, most synapses tend to “communicate” activity to the dendrites or soma of the cell they synapse upon, whence activity passes to the axon hillock and then down the axon to the terminal arbo-

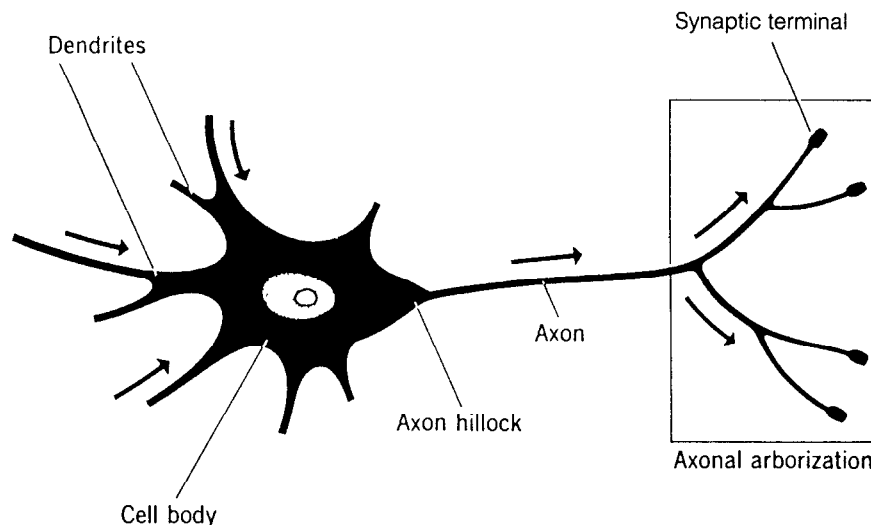


Figure 1. A “basic neuron” abstracted from a motor neuron of mammalian spinal cord. The dendrites and soma (cell body) constitute the major part of the input surface of the neuron. The axon is the “output line.” The tips of the branches of the axon form synapses upon other neurons or upon effectors (although synapses may occur along the branches of an axon as well as at the ends). (From Arbib, M. A., 1989, *The Metaphorical Brain 2: Neural Networks and Beyond*, New York: Wiley-Interscience, p. 52. Reproduced with permissions. Copyright © 1989 by John Wiley & Sons, Inc.)

rization. The axon can be very long indeed. For instance, the cell body of a neuron that controls the big toe lies in the spinal cord and thus has an axon that runs the complete length of the leg. We may contrast the immense length of the axon of such a neuron with the very small size of many of the neurons in our heads. For example, amacrine cells in the retina have branchings that cannot appropriately be labeled dendrites or axons, for they are short and may well communicate activity in either direction to serve as local modulators of the surrounding network. In fact, the propagation of signals in the “counter-direction” on dendrites away from the soma has in recent years been seen to play an important role in neuronal function, but this feature is not included in the account of the “basic neuron” given here (see DENDRITIC PROCESSING—titles in SMALL CAPS refer to articles in Part III).

To understand more about neuronal “communication,” we emphasize that the cell is enclosed by a membrane, across which there is a difference in electrical charge. If we change this potential difference between the inside and outside, the change can propagate in much the same passive way that heat is conducted down a rod of metal: a normal change in potential difference across the cell membrane can propagate in a passive way so that the change occurs later, and becomes smaller, the farther away we move from the site of the original change. This passive propagation is governed by the *cable equation*

$$\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2}$$

If the starting voltage at a point on the axon is V_0 , and no further conditions are imposed, the potential will decay exponentially, having value $V(x) = V_0 e^{-x/\lambda}$ at distance x from the starting point, where the length unit, the *length constant*, is the distance in which the potential changes by a factor of $1/e$. This length unit will differ from axon to axon. For “short” cells (such as the rods, cones, and bipolar cells of the retina), passive propagation suffices to signal a potential change from one end to the other; but if the axon is long, this mechanism is completely inadequate, since changes at one end will decay almost completely before reaching the other end. Fortunately, most nerve cells have the further property that if the change in potential difference is large enough (we say it exceeds a *threshold*), then in a cylindrical configuration such as the axon, a pulse can be generated that will actively propagate at full amplitude instead of fading passively.

If propagation of various potential differences on the dendrites and soma of a neuron yields a potential difference across the membrane at the axon hillock which exceeds a certain threshold, then a regenerative process is started: the electrical change at one place is enough to trigger this process at the next place, yielding a *spike* or *action potential*, an undiminishing pulse of potential difference propagating down the axon. After an impulse has propagated along the length of the axon, there is a short *refractory period* during which a new impulse cannot be propagated along the axon.

The propagation of action potentials is now very well understood. Briefly, the change in membrane potential is mediated by the flow of ions, especially sodium and potassium, across the membrane. Hodgkin and Huxley (1952) showed that the *conductance* of the membrane to sodium and potassium ions—the ease with which they flow across the membrane—depends on the transmembrane voltage. They developed elegant equations describing the voltage and time dependence of the sodium and potassium conductances. These equations (see the article AXONAL MODELING in Part III) have given us great insight into cellular function. Much mathematical research has gone into studying Hodgkin-Huxley-like equations, showing, for example, that neurons can support rhythmic pulse generation even without input (see OSCILLATORY AND BURSTING PROPERTIES OF NEURONS), and explicating trig-

gered long-distance propagation. Hodgkin and Huxley used curve fitting from experimental data to determine the terms for conductance change in their model. Subsequently, much research has probed the structure of complex molecules that form *channels* which selectively allow the passage of specific ions through the membrane (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION). This research has demonstrated how channel properties not only account for the terms in the Hodgkin-Huxley equation, but also underlie more complex dynamics which may allow even small patches of neural membrane to act like complex computing elements. At present, most artificial neurons used in applications are very simple indeed, and much future technology will exploit these “subneural subtleties.”

An impulse traveling along the axon from the axon hillock triggers new impulses in each of its branches (or *collaterals*), which in turn trigger impulses in their even finer branches. Vertebrate axons come in two varieties, myelinated and unmyelinated. The myelinated fibers are wrapped in a sheath of *myelin* (Schwann cells in the periphery, oligodendrocytes in the CNS—these are glial cells, and their role in axonal conduction is the primary role of glia considered in neural modeling to date). The small gaps between successive segments of the myelin sheath are called *nodes of Ranvier*. Instead of the somewhat slow active propagation down an unmyelinated fiber, the nerve impulse in a myelinated fiber jumps from node to node, thus speeding passage and reducing energy requirements (see AXONAL MODELING).

Surprisingly, at most synapses, the direct cause of the change in potential of the postsynaptic membrane is not electrical but chemical. When an impulse arrives at the presynaptic terminal, it causes the release of *transmitter* molecules (which have been stored in the bouton in little packets called vesicles) through the presynaptic membrane. The transmitter then diffuses across the very small *synaptic cleft* to the other side, where it binds to receptors on the postsynaptic membrane to change the conductance of the postsynaptic cell. The effect of the “classical” transmitters (later we shall talk of other kinds, the neuromodulators) is of two basic kinds: either *excitatory*, tending to move the potential difference across the postsynaptic membrane in the direction of the threshold (*depolarizing* the membrane), or *inhibitory*, tending to move the polarity away from the threshold (*hyperpolarizing* the membrane). There are some exceptional cell appositions that are so large or have such tight coupling (the so-called gap junctions) that the impulse affects the postsynaptic membrane without chemical mediation (see NEOCORTX: CHEMICAL AND ELECTRICAL SYNAPSES).

Most neural modeling to date focuses on the excitatory and inhibitory interactions that occur on a fast time scale (a millisecond, more or less), and most biological (as distinct from technological) models assume that all synapses from a neuron have the same “sign.” However, neurons may also secrete transmitters that modulate the function of a circuit on some quite extended time scale. Modeling that takes account of this *neuromodulation* (see SYNAPTIC INTERACTIONS and NEUROMODULATION IN INVERTEBRATE NERVOUS SYSTEMS) will become increasingly important in the future, since it allows cells to change their function, enabling a neural network to switch dramatically its overall mode of activity.

The excitatory or inhibitory effect of the transmitter released when an impulse arrives at a bouton generally causes a subthreshold change in the postsynaptic membrane. Nonetheless, the cooperative effect of many such subthreshold changes may yield a potential change at the axon hillock that exceeds threshold, and if this occurs at a time when the axon has passed the refractory period of its previous firing, then a new impulse will be fired down the axon.

Synapses can differ in shape, size, form, and effectiveness. The geometrical relationships between the different synapses impinging on the cell determine what patterns of synaptic activation will yield the appropriate temporal relationships to excite the cell (see

DENDRITIC PROCESSING). A highly simplified example (Figure 2) shows how the properties of nervous tissue just presented would indeed allow a simple neuron, by its very dendritic geometry, to compute some useful function (cf. Rall, 1964, p. 90). Consider a neuron with four dendrites, each receiving a single synapse from a visual receptor, so arranged that synapses A, B, C, and D (from left to right) are at increasing distances from the axon hillock. (This is not meant to be a model of a neuron in the retina of an actual organism; rather, it is designed to make vivid the potential richness of single neuron computations.) We assume that each receptor re-

acts to the passage of a spot of light above its surface by yielding a generator potential which yields, in the postsynaptic membrane, the same time course of depolarization. This time course is propagated passively, and the farther it is propagated, the later and the lower is its peak. If four inputs reached A, B, C, and D simultaneously, their effect may be less than the threshold required to trigger a spike there. However, if an input reaches D before one reaches C, and so on, in such a way that the peaks of the four resultant time courses at the axon hillock coincide, the total effect could well exceed threshold. This, then, is a cell that, although very

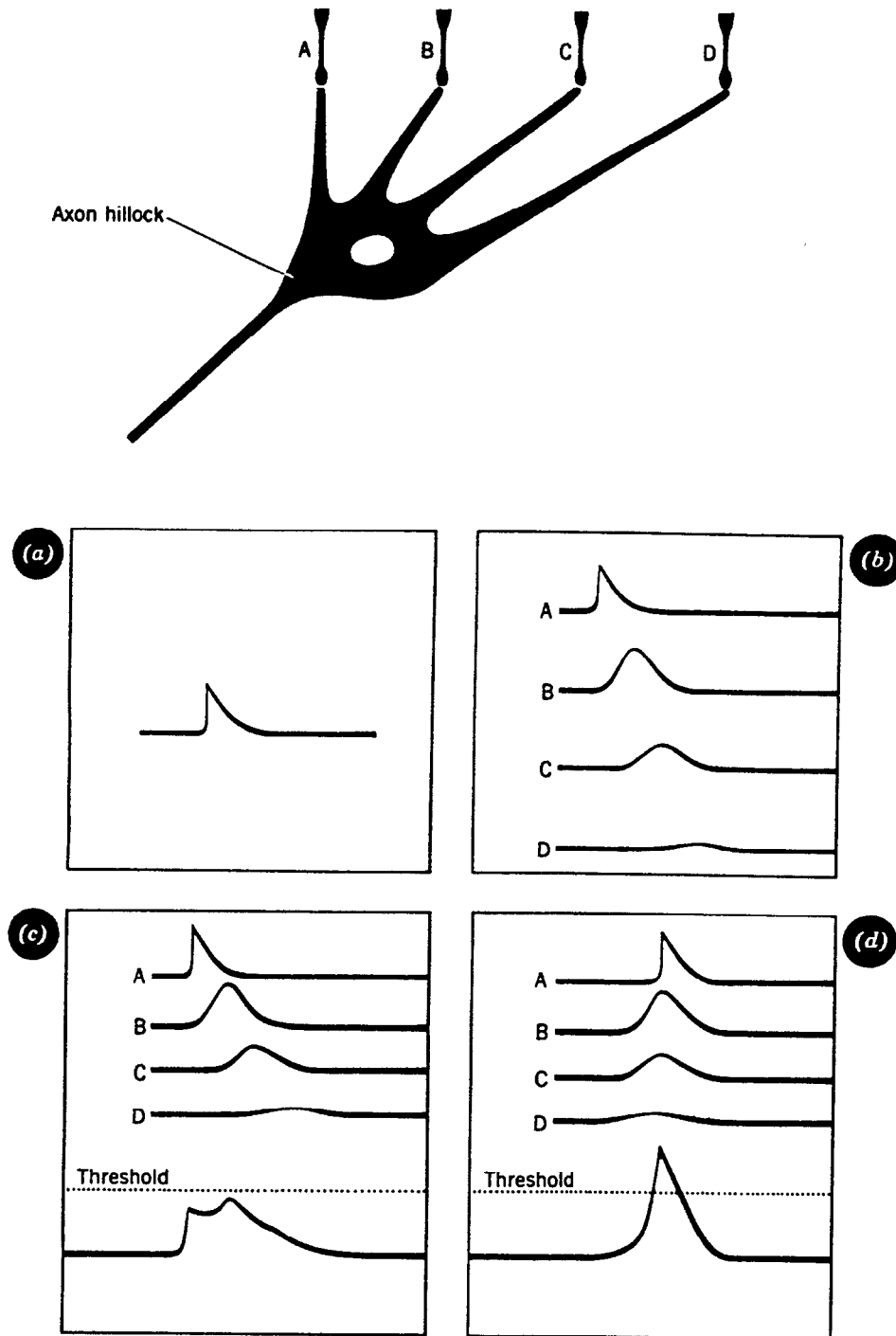


Figure 2. An example, conceived by Wilfrid Rall, of the subtleties that can be revealed by neural modeling when dendritic properties (in this case, length-dependent conduction time) are taken into account. As shown in Part C, the effect of simultaneously activating all inputs may be subthreshold, yet the cell may respond when inputs traverse the cell from right to left (D). (From Arbib, M. A., 1989, *The Metaphorical Brain 2: Neural Networks and Beyond*, New York: Wiley-Interscience, p. 60. Reproduced with permission. Copyright © 1989 by John Wiley & Sons, Inc.)

simple, can detect direction of motion across its input. It responds only if the spot of light is moving from right to left, and if the velocity of that motion falls within certain limits. Our cell will not respond to a stationary object, or one moving from left to right, because the asymmetry of placement of the dendrites on the cell body yields a preference for one direction of motion over others (for a more realistic account of biological mechanisms, see *DIRECTIONAL SELECTIVITY*). This simple example illustrates that the *form* (i.e., the geometry) of the cell can have a great impact on the *function* of the cell, and we thus speak of *form-function* relations. When we note that neurons in the human brain may have 10,000 or more synapses upon them, we can understand that the range of functions of single neurons is indeed immense.

Receptors and Effectors

On the “input side,” receptors share with neurons the property of generating potentials, which are transmitted to various synapses upon neurons. However, the input surface of a receptor does not receive synapses from other neurons, but can transduce environmental energy into changes in membrane potential, which may then propagate either actively or passively. (Visual receptors do not generate spikes; touch receptors in the body and limbs use spike trains to send their message to the spinal cord.) For instance, the rods and cones of the eye contain various pigments that react chemically to light in different frequency bands, and these chemical reactions, in turn, lead to local potential changes, called generator potentials, in the membrane. If the light falling on an array of rods and cones is appropriately patterned, then their potential changes will induce interneuron changes to, in turn, fire certain ganglion cells (retinal output neurons whose axons course toward the brain). Properties of the light pattern will thus be signaled farther into the nervous system as trains of impulses (see *RETINA*).

At the receptors, increasing the intensity of stimulation will increase the generator potential. If we go to the first level of neurons that generate pulses, the axons “reset” each time they fire a pulse and then have to get back to a state where the threshold and the input potential meet. The higher the generator potential, the shorter the time until they meet again, and thus the higher the frequency of the pulse. Thus, at the “input” it is a useful first approximation to say that intensity or quantity of stimulation is coded in terms of pulse frequency (more stimulus \approx more spikes), whereas the quality or type of stimulus is coded by different lines carrying signals from different types of receptors. As we leave the periphery and move toward more “computational” cells, we no longer have such simple relationships, but rather interactions of inhibitory cells and excitatory cells, with each inhibitory input moving a cell away from, and each excitatory input moving it toward, threshold.

To discuss the “output side,” we must first note that a muscle is made up of many thousands of muscle fibers. The motor neurons that control the muscle fibers lie in the spinal cord or the brainstem, whence their axons may have to travel vast distances (by neuronal standards) before synapsing upon the muscle fibers. The smallest functional entity on the output side is thus the *motor unit*, which consists of a motor neuron cell body, its axon, and the group of muscle fibers the axon influences.

A muscle fiber is like a neuron to the extent that it receives its input via a synapse from a motor neuron. However, the response of the muscle fiber to the spread of depolarization is to contract. Thus, the motor neurons which synapse upon the muscle fibers can determine, by the pattern of their impulses, the extent to which the whole muscle comprised of those fibers contracts, and can thus control movement. (Similar remarks apply to those cells that secrete various chemicals into the bloodstream or gut, or those that secrete sweat or tears.)

Synaptic activation at the *motor end-plate* (i.e., the synapse of a motor neuron upon a muscle fiber) yields a brief “twitch” of the muscle fiber. A low repetition rate of action potentials arriving at a motor end-plate causes a train of twitches, in each of which the mechanical response lasts longer than the action potential stimulus. As the frequency of excitation increases, a second action potential will arrive while the mechanical effect of the prior stimulus still persists. This causes a mechanical summation or fusion of contractions. Up to a point, the degree of summation increases as the stimulus interval becomes shorter, although the summation effect decreases as the interval between the stimuli approaches the refractory period of the muscle, and maximum tension occurs. This limiting response is called a *tetanus*. To increase the tension exerted by a muscle, it is then necessary to recruit more and more fibers to contract. For more delicate motions, such as those involving the fingers of primates, each motor neuron may control only a few muscle fibers. In other locations, such as the shoulder, one motor neuron alone may control thousands of muscle fibers. As descending signals in the spinal cord command a muscle to contract more and more, they do this by causing motor neurons with larger and larger thresholds to start firing. The result is that fairly small fibers are brought in first, and then larger and larger fibers are recruited. The result, known as Henneman’s Size Principle, is that at any stage, the increment of activation obtained by recruiting the next group of motor units involves about the same percentage of extra force being applied, aiding smoothness of movement (see *MOTOR-NEURON RECRUITMENT*).

Since there is no command that a neuron may send to a muscle fiber that will cause it to lengthen—all the neuron can do is stop sending it commands to contract—the muscles of an animal are usually arranged in pairs. The contraction of one member of the pair will then act around a pivot to cause the expansion of the other member of the pair. Thus, one set of muscles *extends* the elbow joint, while another set *flexes* the elbow joint. To extend the elbow joint, we do not signal the *flexors* to lengthen, we just stop signaling them to contract, and then they will be automatically lengthened as the *extensor* muscles contract. For convenience, we often label one set of muscles as the “prime mover” or *agonist*, and the opposing set as the *antagonist*. However, in such joints as the shoulder, which are not limited to one degree of freedom, many muscles, rather than an agonist-antagonist pair, participate. Most real movements involve many joints. For example, the wrist must be fixed, holding the hand in a position bent backward with respect to the forearm, for the hand to grip with its maximum power. *Synergists* are muscles that act together with the main muscles involved. A large group of muscles work together when one raises something with one’s finger. If more force is required, wrist muscles may also be called in; if still more force is required, arm muscles may be used. In any case, muscles all over the body are involved in maintaining posture.

Neural Models

Before presenting more realistic models of the neuron (see *PERSPECTIVE ON NEURON MODEL COMPLEXITY; SINGLE-CELL MODELS*), we focus on the work of McCulloch and Pitts (1943), which combined neurophysiology and mathematical logic, using the all-or-none property of neuron firing to model the neuron as a binary discrete-time element. They showed how excitation, inhibition, and threshold might be used to construct a wide variety of “neurons.” It was the first model to tie the study of neural nets squarely to the idea of computation in its modern sense. The basic idea is to divide time into units comparable to a refractory period so that, in each time period, at most one spike can be generated at the axon hillock of a given neuron. The McCulloch-Pitts neuron (Figure 3A) thus operates on a discrete-time scale, $t = 0, 1, 2, 3, \dots$, where the

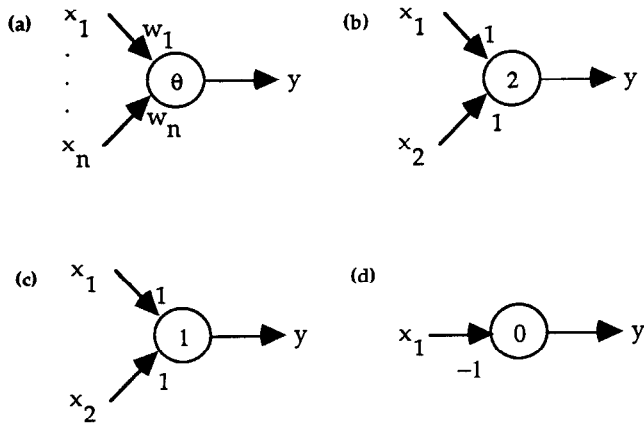


Figure 3. *a*, A McCulloch-Pitts neuron operating on a discrete-time scale. Each input has an attached weight w_i , and the neuron has a threshold θ . The neuron “fires” at time $t + 1$ just in case the weighted values of its inputs at time t is at least θ . *b*, Settings of weights and threshold for neurons that function as an AND gate (i.e., the output fires if x_1 and x_2 both fire). *c*, An OR gate (the output fires if x_1 or x_2 , or both fire). *d*, A NOT gate (the output fires if x_1 does NOT fire).

time unit is (in biology) on the order of a millisecond. We write $y(t) = 1$ if a spike does appear at time t , and $y(t) = 0$ if not. Each connection, or *synapse*, from the output of one neuron to the input of another has an attached *weight*. Let w_i be the weight on the i th connection onto a given neuron. We call the synapse *excitatory* if $w_i > 0$, and *inhibitory* if $w_i < 0$. We also associate a *threshold* θ with each neuron, and assume exactly one unit of delay in the effect of *all* presynaptic inputs on the cell’s output, so that a neuron “fires” (i.e., has value 1 on its output line) at time $t + 1$ if the weighted value of its inputs at time t is at least θ . Formally, if at time t the value of the i th input is $x_i(t)$ and the output one time step later is $y(t + 1)$, then

$$y(t + 1) = 1 \quad \text{if and only if} \quad \sum_i w_i x_i(t) \geq \theta$$

Parts *b* through *d* of Figure 3 show how weights and threshold can be set to yield neurons that realize the logical functions AND, OR, and NOT. As a result, McCulloch-Pitts neurons are sufficient to build networks that can function as the control circuitry for a computer carrying out computations of arbitrary complexity; this discovery played a crucial role in the development of automata theory and in the study of learning machines. Although the McCulloch-Pitts neuron no longer plays an active part in *computational neuroscience*, it is still widely used in *neural computing*, especially when it is generalized so that the input and output values can lie anywhere in the range $[0, 1]$ and the function $f(\sum_i w_i x_i(t))$, which yields $y(t + 1)$, is a continuously varying function rather than a step function. However, it is one thing to define model neurons with sufficient logical power to subserve any discrete computation; it is quite another to understand how the neurons in actual brains perform their tasks. More generally, the problem is to select just which units to model, and to decide how such units are to be represented. Thus, when we turn from neural computing to computational neuroscience, we must turn to more realistic models of neurons. On the other hand, we may say that neural computing cannot reach its full power without applying new mechanisms based on current and future study of biological neural networks (see the road map **Biological Neurons and Synapses**).

Modern brain theory no longer uses the binary model of the neuron, but instead uses continuous-time models that either rep-

resent the variation in average firing rate of the neuron or actually capture the time course of membrane potentials. It is only through such correlates of measurable brain activity that brain models can really feed back to biological experiments. Such models also require the brain theorist to know a great deal of detailed anatomy and physiology as well as behavioral data. Hodgkin and Huxley (1952) have shown us how much can be learned from analysis of membrane properties about the propagation of electrical activity along the axon: Rall (1964; cf. Figure 2) was a leader in showing that the study of membrane properties in a variety of connected “compartments” of membrane in dendrite, soma, and axon can help us understand small neural circuits, as in the OLFACTORY BULB (q.v.) or for DENDRITIC PROCESSING (q.v.). Nonetheless, in many cases, the complexity of compartmental analysis makes it more insightful to use a more lumped representation of the individual neuron if we are to assemble the model neurons to analyze large networks. A computer simulation of the response of a whole brain region which analyzed each component at the finest level of detail available would be too large to run on even a network of computers. In addition to the importance of detailed models of single neurons in themselves, such studies can also be used to fine-tune more economical models of neurons, which can then serve as the units in models of large networks, whether to model systems in the brain or to design artificial neural networks which exploit subtle neural capabilities.

We may determine units in the brain *physiologically*, e.g., by electrical recording, and *anatomically*, e.g., by staining. In many regions of the brain, we have an excellent correlation between physiological and anatomical units; that is, we know which anatomical entity yields which physiological response. Unfortunately, this is not always the case. We may have data on the electrophysiological correlates of animal behavior, and anatomical data as well, yet not know which specific cell, defined anatomically, yields an observed electrophysiological response. Another problem that we confront in modeling is that we have both too much and too little anatomical detail: too much in that there are many synapses that we cannot put into our model without overloading our capabilities for either mathematical analysis or computer simulation, and too little in that we often do not know which details of synaptology may determine the most important modes of behavior of a particular region of the brain. Judicious choices from available data, and judicious hypotheses concerning missing data, must thus be made in setting up a model, leading to the design of experiments whose results may either confirm these hypotheses or lead to their modification. An important point of good modeling methodology is thus to set up simulations in such a way that we can use different connectivity on different simulations, both to test alternative hypotheses and to respond to new data as they become available.

The simplest “realistic” model consonant with the above material is the *leaky integrator* model. Although some biological neurons communicate by the passive propagation (cable equation) of membrane potential down their (necessarily short) axons, most communicate by the active propagation of “spikes.” The generation and propagation of such spikes has been described in detail by the Hodgkin-Huxley equations. However, the leaky integrator model omits such details. It is a continuous-time model based on using the *firing rate* (e.g., the number of spikes traversing the axon in the most recent 20 ms) as a continuously varying *output* measure of the cell’s activity, in which the *internal state* of the neuron is described by a single variable, the membrane potential at the spike initiation zone. The firing rate is approximated by a simple, sigmoid function of the membrane potential. That is, we introduce a function σ of the membrane potential m such that $\sigma(m)$ increases from 0 to some maximum value as m increases from $-\infty$ to $+\infty$ (e.g., the sigmoidal function $k/[1 + \exp(-m/\theta)]$, increasing from 0 to its maximum k). Then the firing rate $M(t)$ of the cell is given by the

equation:

$$M(t) = \sigma(m(t))$$

The time evolution of the cell's membrane potential is given by a differential equation. Consider first the simple equation

$$\tau \frac{dm(t)}{dt} = -m(t) + h \quad (1)$$

We say that $m(t)$ is in an *equilibrium* if it does not change under the dynamics described by the differential equation. However, $dm(t)/dt = 0$ if and only if $m(t) = h$, so that h is the *unique* equilibrium of Equation 1. To get more information, we now integrate Equation 1 to get

$$m(t) = e^{-t/\tau} m(0) + (1 - e^{-t/\tau}) h$$

which tends to the *resting level* h with *time constant* τ with increasing t so long as τ is positive. We now add synaptic inputs to obtain

$$\tau \frac{dm(t)}{dt} = -m(t) + \sum_i w_i X_i(t) + h \quad (2)$$

where $X_i(t)$ is the firing rate at the i th input. Thus, an excitatory input ($w_i > 0$) will be such that increasing it will increase $dm(t)/dt$, while an inhibitory input ($w_i < 0$) will have the opposite effect. A neuron described by Equation 2 is called a *leaky integrator neuron*. This is because the equation

$$\tau \frac{dm(t)}{dt} = \sum_i w_i X_i(t) \quad (3)$$

would simply integrate the inputs with scaling constant τ ,

$$m(T) = m(0) + \frac{1}{\tau} \int_0^T \sum_i w_i X_i(t) dt \quad (4)$$

but the $-m(t)$ term in Equation 3 opposes this integration by a “leakage” of the potential $m(t)$ as it tries to return to its input-free equilibrium h .

It should be noted that, even at this simple level of modeling, there are alternative models. In the foregoing model, we have used subtractive inhibition. But there are inhibitory synapses which seem better described by *shunting* inhibition which, applied at a given point on a dendrite, serves to divide, rather than subtract from, the potential change passively propagating from more distal synapses. Again, the “lumped frequency” model cannot model the relative timing effects crucial to our motion detector example (see Figure 2). These might be approximated by introducing appropriate delay terms

$$\tau \frac{dm(t)}{dt} = -m(t) + \sum_i w_i X_i(t - t_i) + h$$

Another class of neuron models—spiking neurons, including integrate-and-fire neurons—are intermediate in complexity between leaky integrator models in which the output is the average firing rate (see RATE CODING AND SIGNAL PROCESSING) and detailed biophysical models in which the fine details of action potential generation are modeled using the Hodgkin-Huxley equation. In these intermediate models, the output is a spike whose timing is continuously variable as a result of cellular interactions, but the spike is represented simply by its time of occurrence, with no internal structure. For example, one may track the continuous variable (4), then generate a spike each time this quantity reaches threshold, while simultaneously resetting the integral to some baseline value (see INTEGRATE-AND-FIRE NEURONS AND NETWORKS).

Such models include the ability to transmit information very rapidly through small temporal differences between the spikes sent out by different neurons (see SPIKING NEURONS, COMPUTATION WITH).

All this reinforces the observation that there is no modeling approach that is automatically appropriate. Rather, we seek to find the simplest model adequate to address the complexity of a given range of problems. The articles in Part III of the *Handbook* will provide many examples of the diversity of neural models appropriate to different tasks.

More Detailed Properties of Neurons

In Section I.3, the only details we will add to the neuron models just presented will be various, relatively simple, rules of synaptic plasticity. This level of detail (though with many variations) will suffice for a fair range of models of biological neural networks, and for a range of current work on artificial neural networks (ANNs). The road map **Biological Neurons and Synapses** in Part II surveys a set of articles that demonstrate that biological neurons are vastly more complex than the present models suggest. Other road maps show the special structures revealed in “special-purpose” neural circuitry in different species of animals. Table 1 lists some of the relevant articles on such circuits, together with the specific animal types on which the studies were based. The point is that much is to be learned from features specific to many different types of nervous systems, as well as from studies in humans, monkeys, cats, and rats that focus on commonalities with the human nervous system.

An appreciation of this complexity is necessary for the computational neuroscientist wishing to address the increasingly detailed database of experimental neuroscience, but it should also prove important for the technologist looking ahead to the incorporation of new capabilities into the next generation of ANNs. Nonetheless, much can be accomplished with simple models, as we shall see in Section I.3.

Table 1. A Sampling of Articles Showing the Lessons to be Learned from the Study of Nervous Systems Very Different from Those of Humans

Crustacean Stomatogastric System	Crabs and lobsters
Development of Retinotectal Maps	Frogs
Echolocation: Cochleotopic and Computational Maps	Bats
Electrolocation	Electric fish
Half-Center Oscillators Underlying Rhythmic Movements	Various
Invertebrate Models of Learning	<i>Aplysia</i> and <i>Hermisenda</i>
Locomotion, Invertebrate	Various insects
Locust Flight: Components and Mechanisms in the Motor	Locusts
Motor Primitives	Frogs
Neuromodulation in Invertebrate Nervous Systems	Various
Oscillatory and Bursting Properties of Neurons	Various
Scratch Reflex	Turtles
Sound Localization and Binaural Processing	Owls
Spinal Cord of Lamprey: Generation of Locomotor Patterns	Lampreys
Visual Course Control in Flies	Flies
Visuomotor Coordination in Frog and Toad	Frogs and toads
Visuomotor Coordination in Salamander	Salamanders

1.2. Levels and Styles of Analysis

Many articles in this book show the benefits of interplay between biology and technology. Nonetheless, it is essential to distinguish between studying the brain and building an effective technology for intelligent systems and computation, and to distinguish among the various levels of investigation that exist (from the molecular to the system level) in these related, but by no means identical, disciplines. The present section provides a fuller sense of the disciplines that come together in brain theory and neural networks, and of the different levels of analysis involved in the study of complex biological and technological systems.

A Historical Fragment

Perhaps the simplest history of brain theory and neural networks would restrict itself to just three items: studies by McCulloch and Pitts (1943), Hebb (1949), and Rosenblatt (1958). These publications introduced the first model of neural networks as “computing machines,” the basic model of network self-organization, and the model of “learning with a teacher,” respectively. (Section I.3 provides a semitechnical introduction to this work and a key set of currently central ideas that build upon it.) The present historical fragment is designed to take us up to 1948, the year preceding the publication of Hebb’s book, to reveal our present federation of disciplines as the current incarnation of what emerged in the 1940s and is aptly summed up in the title of the book, *Cybernetics: Or Control and Communication in the Animal and the Machine* (Wiener, 1948). But whereas Wiener’s view of cybernetics was dominated by concepts of control and communication, our subject is dominated by notions of parallel and distributed computation, with special attention to learning in neural networks. On the other hand, notions of information and statistical mechanics championed by Wiener have reemerged as a strong strand in the study of neural networks today (see, e.g., the articles FEATURE ANALYSIS and STATISTICAL MECHANICS OF NEURAL NETWORKS in Part III). The articles in Part III will make abundantly clear how far we have come since 1948, and also how many problems remain. My intent in the present “fragment” is to enrich the reader’s understanding of current contributions by using a selective historical tour to place them in context.

Noting that the Greek word *cybernetics* (κυβερνητική) means the helmsman of a ship (cf. the Latin word *gubernator*, which gives us the word “governor” in English), Wiener (1948) used the term for a subject in which feedback played a central role. Feedback is the process whereby, e.g., the helmsman notes the “error,” the extent to which he is off course, and “feeds it back” to decide which way to move the rudder. We can see the importance of this concept in endowing automata (“self-moving” machines) with flexible behavior. Two hundred years earlier, in *L’Homme machine*, La Mettrie had suggested that such automata as the mechanical duck and flute player of Vaucanson indicated the possibility of one day building a mechanical man that could talk. While these clockwork automata were capable of surprisingly complex behavior, they lacked a crucial aspect of animal behavior, let alone human intelligence: they were unable to adapt to changing circumstances. In the following century, machines were built that could automatically counter disturbances to restore desired performance. Perhaps the best-known example of this is Watt’s governor for the steam engine, which would let off excess steam if the velocity of the engine became too great. This development led to Maxwell’s (1868) paper, “On Governors,” which laid the basis for both the theory of negative feedback and the study of system stability (both of which are discussed in Section I.3). Negative feedback was feedback in which the error (in Watt’s case, the amount by which actual velocity ex-

ceeded desired velocity) was used to counteract the error; stability occurred if this feedback was apportioned to reduce the error toward zero. Bernard (1878) brought these notions back to biology with his study of what Cannon (1939) would later dub *homeostasis*, observing that physiological processes often form circular chains of cause and effect that could counteract disturbances in such variables as body temperature, blood pressure, and glucose level in the blood. In fact, following publication of Wiener’s 1948 book, the Josiah Macy, Jr., Foundation conferences, in which many of the pioneers of cybernetics were involved, became referred to as the Cybernetics Group, with the proceedings entitled *Cybernetics: Circular Causal and Feedback Mechanisms in Biological and Social Systems*, (see Heims, 1991, for a history of the conferences and their participants).

The nineteenth century also saw major developments in the understanding of the brain. At an overall anatomical level, a major achievement was the understanding of localization in the cerebral cortex (see Young, 1970, for a history). Magendie and Bell had discovered that the dorsal roots of the spinal cord were sensory, carrying information from receptors in the body, while the ventral roots (on the belly side) were motor, carrying commands to the muscles. Fritsch and Hitzig, and then Ferrier, extended this principle to the brain proper, showing that the rear of the brain contains the primary receiving areas for vision, hearing, and touch, while the motor cortex is located in front of the central fissure. All this understanding of localization in the cerebral cortex led to the nineteenth century neurological doctrine, perhaps best exemplified in Lichtheim’s (1885) development of the insights of Broca and Wernicke into brain mechanisms of language, which viewed different mental “faculties” as being localized in different regions of the brain. Thus, neurological deficits were to be explained as much in terms of lesions of the connections linking two such regions as in terms of lesions to the regions themselves. We may also note a major precursor of the connectionism of this volume, where the connections are those between neuron-like units rather than anatomical regions: the associationist psychology of Alexander Bain (1868), who represented associations of ideas by the strengths of connections between “neurons” representing those ideas.

Around 1900, two major steps were taken in revealing the finer details of the brain. In Spain, Santiago Ramón y Cajal (e.g., 1906) gave us exquisite anatomical studies of many regions of the brain, revealing the particular structure of each as a network of neurons. In England, the physiological studies of Charles Sherrington (1906) on reflex behavior provided the basic physiological understanding of synapses, the junction points between the neurons. Somewhat later, in Russia, Ivan Pavlov (1927), extending associationist psychology and building on the Russian studies of reflexes by Sechenov in the 1860s, established the basic facts on the modifiability of reflexes by conditioning (see Fearing, 1930, for a historical review).

A very different setting of the scene for cybernetics came from work in mathematical logic in the 1930s. Kurt Gödel published his famous *Incompleteness Theorem* in 1931 (see Arbib, 1987, for a proof as well as a debunking of the claim that Gödel’s theorem sets limits on machine intelligence). The “formalist” program initiated by David Hilbert, which sought to place all mathematical truth within a single formal system, had reached its fullest expression in the *Principia Mathematica* of Whitehead and Russell. But Gödel showed that, if one used the approach offered in *Principia Mathematica* to set up consistent axioms for arithmetic and prove theorems by logical deduction from them, the theory *must* be incomplete, no matter which axioms (“knowledge base”) one started

with—there would be true statements of arithmetic that could not be deduced from the axioms.

Following Gödel's 1931 study, many mathematical logicians sought to formalize the notion of an effective procedure, of what could *and could not* be done by explicitly following an algorithm or set of rules. Kleene (1936) developed the theory of partial recursive functions; Turing (1936) developed his machines; Church (1941) developed the lambda calculus, the forerunner of McCarthy's list processing language, LISP, a one-time favorite of artificial intelligence (AI) workers; while Emil Post (1943) introduced systems for rewriting strings of symbols, of which Chomsky's early formalizations of grammars in 1959 were a special case. Fortunately, these methods proved to be equivalent. Whatever could be computed by one of these methods could be computed by any other method if it were equipped with a suitable "program." It thus came to be believed (Church's thesis) that if a function could be computed by any machine at all, it could be computed by each one of these methods.

Turing (1936) helped chart the limits of the computable with his notion of what is now called a *Turing machine*, a device that followed a fixed, finite set of instructions to read, write, and move upon a finite but indefinitely extendible tape, each square of which bore a symbol from some finite alphabet. As one of the ingredients of Church's thesis, Turing offered a "psychology of the computable," making plausible the claim that any effectively definable computation, that is, anything that a human could do in the way of symbolic manipulation by following a finite and completely explicit set of rules, could be carried out by such a machine equipped with a suitable program. Turing also provided the most famous example of a noncomputable problem, "the unsolvability of the Halting Problem." Let p be the numerical code for a Turing machine program, and let x be the code for the initial contents of a Turing machine's tape. Then the halting function $h(p, x) = 1$ if Turing machine p will eventually halt if started with data x ; otherwise it is 0. Turing showed that there was no "computer program" that could compute h .

And so we come to 1943, the key year for bringing together the notions of control mechanism and intelligent automata.

In "A Logical Calculus of the Ideas Immanent in Nervous Activity," McCulloch and Pitts (1943) united the studies of neurophysiology and mathematical logic. Their formal model of the neuron as a threshold logic unit (see Section I.1) built on the neuron doctrine of Ramón y Cajal and the excitatory and inhibitory synapses of Sherrington, using notation from the mathematical logic of Whitehead, Russell, and Carnap. McCulloch and Pitts provided the "physiology of the computable" by showing that the control box of any Turing machine, the essential formalization of symbolic computation, could be implemented by a network (with loops) of their formal neurons. The ideas of McCulloch and Pitts influenced John von Neumann and his colleagues when they defined the basic architecture of stored program computing. Thus, as electronic computers were built toward the end of World War II, it was understood that whatever they could do could be done by a network of neurons.

Craik's (1943) book, *The Nature of Explanation*, viewed the nervous system "as a calculating machine capable of modeling or paralleling external events," suggesting that the process of forming an "internal model" that paralleled the world is the basic feature of thought and explanation. In the same year, Rosenblueth, Wiener, and Bigelow published "Behavior, Purpose and Teleology." Engineers had noted that if feedback used in controlling the rudder of a ship were too brusque, the rudder would overshoot, compensatory feedback would yield a larger overshoot in the opposite direction, and so on and so on as the system wildly oscillated. Wiener and Bigelow asked Rosenblueth whether there was any corresponding pathological condition in humans and were given the example of intention tremor associated with an injured cerebellum. This evi-

dence for feedback within the human nervous system (see MOTOR CONTROL, BIOLOGICAL AND THEORETICAL) led the three scientists to advocate that neurophysiology move beyond the Sherringtonian view of the CNS as a reflex device adjusting itself in response to sensory inputs. Rather, setting reference values for feedback systems could provide the basis for analysis of the brain as a purposive system explicable only in terms of circular processes, that is, from nervous system to muscles to the external world and back again via receptors.

Such studies laid the basis for the emergence of cybernetics, which in turn gave birth to a number of distinct new disciplines, such as AI, biological control theory, cognitive psychology, and neural modeling, which each went their separate ways in the 1970s. The next subsection introduces a number of these disciplines and the relations between them; this analysis will continue in many articles in Part III of the *Handbook*.

Brains, Machines, and Minds

Brains. *Brain theory* comprises many different theories as to how the structures of the brain can subserve such diverse functions as perception, memory, control of movement, and higher mental function. As such, it includes both attempts to extend notions of computing, as well as applications of modern electronic computers to explore the performance of complex models. An example of the former is the study of *cooperative computation* between different structures in the brain which seeks to offer a new paradigm for computing that transcends classical notions associated with serial execution of symbolic programs. For the latter, *computational neuroscience* makes systematic use of mathematical analysis and computer simulation to provide ever better models of the structure and function of living brains, building on earlier work in both neural modeling and biological control theory.

Machines. *Artificial intelligence* studies how computers may be programmed to yield "intelligent" behavior without necessarily attempting to provide a correlation between structures in the program and structures in the brain. *Robotics* is related to AI but emphasizes the flexible control of machines (robots) which have receptors (e.g., television cameras) and effectors (e.g., wheels, legs, arms, grippers) that allow them to interact with the world.

Brain theory has spawned a companion field of *neural computing*, which involves the design of machines with circuitry inspired by, but which need not faithfully emulate, the neural networks of brains. Many technologists usurp the term "neural networks" for this latter field, but we will use it as an umbrella term which may, depending on context, describe biological nervous systems, models thereof, and the artificial networks which (sometimes at great remove) they inspire. When the emphasis is on "higher mental functions," neural computing may be seen as a new branch of AI (see the road map **Artificial Intelligence** in Part II), but it also contributes to robotics (especially to those robot designs inspired by analysis of animal behavior), and to a wide range of technologies, including those based on image analysis, signal processing, and control (see the road map **Applications**).

For the latter work, many people emphasize adaptive neural networks which, without specific programming, can adjust their connections through self-organization or to meet specifications given by some teacher. There are also significant contributions to the systematic design, rather than emergence through learning, of neural networks, especially for applications in low-level vision (such as stereopsis, optic flow, and shape-from-shading). However, complex problems cannot, in general, be solved by the tuning or the design of a single unstructured network. For example, robot control may integrate a variety of low-level vision networks with a set of competing and cooperating networks for motor control and its

planning. Brain theory and neural computing thus have to address the analysis and design, respectively, of networks of networks (see, e.g., **HYBRID CONNECTIONIST/SYMBOLIC SYSTEMS** and **MODULAR AND HIERARCHICAL LEARNING SYSTEMS**).

Minds. Here, I want to distinguish the brain from the mind (the realm of the “mental”). In great part, brain theory seeks to analyze how the brain guides the behaving organism in its interactions with the dynamic world around it, but much of the control of such interactions is not mental, and much of what is mental is subsymbolic and/or unconscious (see **PHILOSOPHICAL ISSUES IN BRAIN THEORY AND CONNECTIONISM** and **CONSCIOUSNESS, NEURAL MODELS OF**). Without offering a precise definition of “mental,” let me just say that many people can agree on examples of mental activity (perceiving a visual scene, reading, thinking, etc.) even if they take the diametrically opposite philosophical positions of dualism (mind and brain are separate) or monism (mind is a function of brain). They would then agree that some mental activity (e.g., contemplation) need not result in overt “interactions with the dynamic real world,” and that much of the brain’s activity (e.g., controlling normal breathing) is not mental. Face recognition seems to be a mental activity that we do not carry out through symbol manipulation. Indeed, even psychologists who reject Freud’s particular psychosexual theories accept his notion that much of our mental behavior is shaped by unconscious forces (for an assessment of Freud and an account of consciousness, see Arbib and Hesse, 1986).

Cognitive psychology attempts to explain the mind in terms of “information processing” (a notion which is continuing to change). It thus occupies a middle ground between brain theory and AI in which the model must explain psychological data (e.g., what tasks are hard for humans, people’s ability at memorization, the development of the child, patterns of human errors, etc.) but in which the units of the model need not correspond to actual brain structures. In the 1960s and 1970s, the majority of cognitive psychologists formulated their theories in terms of information theory and/or symbol manipulation, while theories of biological organization were ignored. However, workers in both AI and cognitive psychology now pay increasing attention to the cooperative computation paradigm. The term *connectionism* has come to be used for studies that model human thought and behavior in terms of parallel distributed networks of neuron-like units, with learning mediated by changes in strength of the connections between these elements (see **COGNITIVE MODELING: PSYCHOLOGY AND CONNECTIONISM**).

The study of brain theory and neural networks thus has a twofold aim: (1) to enhance our understanding of human thought and the neural basis of human and animal behavior (brain theory), and (2) to learn new strategies for building “intelligent” machines or adaptive robots (neural computing). In either case, we seek organizational principles that will help us understand how neurons (whether biological or artificial) can work together to yield complex patterns of behavior. *Brain theory* requires empirical data to shape and constrain modeling, but in return provides concepts and hypotheses to shape and constrain experimentation. In *neural computing*, the criterion for success is the design of a machine that can perform a task cheaply, reliably, and effectively, even if, in the process of making the best use of available (e.g., silicon) technology, the final design departs radically from the biological neural network that inspired it. It will be important in reading this *Handbook*, then, to be clear as to whether a particular study is an exercise in brain theory/computational neuroscience or in AI/neural computing. What will not be in doubt is that the influence of these subjects works both ways: not only can brain mechanisms inspire new technology, but new technologies provide metaphors to drive new theories of brain function. To this it must be added that most workers in ANNs know little of brain function, and relatively few neuroscientists have a deep understanding of brain theory or know much

of neural computing beyond the basic ideas of Hebbian plasticity and, perhaps, backpropagation (see Section I.3). However, the level of interchange has increased since the first edition of this *Handbook* appeared, and this new edition is designed to further increase the flow of information between these scientific communities.

Levels of Analysis

Whether the emphasis is on humans, animals, or machines, it becomes clear that we can seek insight at many different levels of analysis; from large information processing blocks down to the finest details of molecular structure. Much of psychology and linguistics looks at human behavior “from the outside,” whether studying overall competence or attending to details of performance. Neuropsychology relates behavior to the interaction of various brain regions. Neurophysiology studies the activity of neurons, both to understand the intrinsic properties of the neurons and to help understand their role in the subsystems dissected out by the neuropsychologist, such as networks for pattern recognition or for visuomotor coordination. Molecular and cell biology and biophysics correlate the structure and connectivity of the membranes and subcellular systems which constitute cells with the way these cells transform incoming patterns or subserve memory by changing function with repeated interactions.

These differing levels make it possible to focus individual research studies, but they are ill-defined, and a scientist who works on any one level needs to make occasional forays, both downward to find mechanisms for the functions studied, and upward to understand what role the studied function can play in the overall scheme of things. *Top-down* modeling starts from some overall behavior and explains it in terms of the interaction of high-level functional units, while *bottom-up* modeling starts from the interaction of individual neurons (or even smaller units) to explain network properties. It requires a judicious blend of the two to connect the clear overview of crucial questions to the hard data of neuroscience or, in the case of neural engineering, to the details of implementation. Most successful modeling will be purely bottom-up or top-down only in its initial stages, if at all—constraints on an initial top-down model will be given, for example, by the data on regional localization offered by the neurologist, or the circuit-cell-synapse studies of much current neuroscience.

We must now distinguish the brain’s computation from connectionist computation “in the *style* of the brain.” If a connectionist model succeeds in describing some psychological input/output behavior, it may become an important hypothesis that its internal structure is “real” (see **RECURRENT NETWORKS: NEUROPHYSIOLOGICAL MODELING**). In general, however, much additional work will be required to find and assimilate neurophysiological data to provide brain models in which the neurons are not mere formal units but actually represent biological neurons in the brain.

Much study of the brain is guided by evolutionary and comparative studies of animal behavior and brain function (cf. **EVOLUTION OF THE ANCESTRAL VERTEBRATE BRAIN** and related articles in the road map **Neuroethology and Evolution**). The information about the function of the human brain that is gained in the neurological clinic or during neurosurgery can thus be supplemented by humane experimentation on animals. (However, as evidenced by Table 1 of Section I.1, we can learn a great deal by studying the *differences*, as well as the similarities, between the brains of different species.) We learn by stimulating, recording from, or excising portions of an animal’s brain and seeing how the animal’s behavior changes. We may then compare such results with observations using such techniques as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) of the relative activity of different parts of the human brain during different tasks (see **IMAGING THE GRAMMATICAL BRAIN**, **IMAGING THE MOTOR**

BRAIN, and IMAGING THE VISUAL BRAIN). The grand aim of *cognitive neuroscience* (as neuropsychology has now become; see the **Cognitive Neuroscience** road map) is to use clinical data and brain imaging to form a high-level view of the involvement of various brain regions in human cognition, using single-cell activity recorded from animals engaged in analogous behaviors to suggest the neural networks underlying this involvement (see SYNTHETIC FUNCTIONAL BRAIN MAPPING). The catch, of course, is that the “analogous behaviors” of animals are not very analogous at all when it comes to such symbolic activities as language and reasoning. In Part III, we will see that “higher mental functions” tend to be modeled more in connectionist terms constrained (if at all) by psychological or psycholinguistic data (cf. the Part II road maps **Psychology** and **Linguistics and Speech Processing**), while the greatest successes in seeking the neural underpinnings of human behavior have come in areas such as vision, memory, and motor control, where we can make neural network models of animal analogues of human capabilities (cf. the road maps **Vision**, **Other Sensory Systems**, **Neural Plasticity**, **Biological Networks**, **Motor Pattern Generators**, and **Mammalian Motor Control**).

We also learn from the attempt to reproduce various aspects of human behavior in a robot, even though human action, memory, learning, and perception are far richer than those of any machine yet built or likely to be built in the near future (see BIOLOGICALLY INSPIRED ROBOTICS). Thus, when we suggest that the brain can be thought of in some ways as a (highly distributed) computer, we are not trying to reduce humans to the level of extant machines, but rather to understand ways in which machines give us insight into human attributes. This type of study has been referred to as *cybernetics*, extending the concept of Norbert Wiener, who, as we have seen, defined the subject as “the study of control and communication in man and machine.”

To the extent that they address “higher mental function,” the studies presented in this *Handbook* suggest that there is no single “thing” called *intelligence*, but rather a plexus of properties that, taken one at a time, may be little cause for admiration, but any sizable collection of which will yield behavior that we would label as intelligent. Turing (1950) argued that we would certainly regard a machine as intelligent if it could pass the following test: An experimenter sits in a room with two teletypes by which she conducts a “conversation” with two systems. One is a human, the other is a machine, but the experimenter is not told which is which. If, after asking many questions, she is likely to have much doubt about which is human and which is machine, we should, says Turing, concede intelligence to the machine. However, unless one dogmatically insists that being intelligent entails behaving in a human way, it is “harder” for a machine to pass this *Turing test* than to be intelligent. For instance, whereas a computer can answer problems in arithmetic quickly and correctly, a much more complex program would be required to ensure that it answered as slowly and erratically as a human. Turing’s aim was not to find a necessary set of conditions to ensure intelligence, but rather to devise a test which, if passed by a machine, would convince most skeptics that the machine had intelligence.

Schema Theory

The analysis of complex systems, whether they subserve natural or artificial intelligence, requires a coarser grain of analysis to complement that of neural networks. To make sense of the brain, we often divide it into functional systems—such as the motor system, the visual system, and so on—as well as into structural subsystems—from the spinal cord and the hippocampus to the various subdivisions of the prefrontal cortex. Similarly, in distributed AI (see MULTIAGENT SYSTEMS), the solution of a task may be distributed over a complex set of interacting *agents*, each with their

dedicated processors for handling the information available to them locally. Thus, both neuroscience and artificial intelligence require a language for expressing the distribution of function across units intermediate between overall function and the final units of analysis (e.g., neurons or simple instructions).

Since the “units of thought” or the subfunctions of a complex behavior may be quite high-level compared to the fine-grain computation of the myriad neurons in the human brain, SCHEMA THEORY (q.v.; see also Arbib, 1981; Arbib, Érdi, and Szentágothai, 1998, chap. 3) complements connectionism by providing a bridging language between functional description and neural networks. It is based on a theory of the concurrent activity of interacting functional units called *schemas*. Perceptual schemas are those used for perceptual analysis, while motor schemas are those which provide the control systems that can be coordinated to effect a wide variety of movement. Other schemas compete and cooperate to meld action, internal state, and perception in an ongoing action-perception cycle.

Figure 4A represents brain theory, while Figure 4B offers a similar but distinct picture for distributed AI. We may model the brain either functionally, analyzing some behavior in terms of interacting schemas, or structurally, through the interaction of anatomically defined units, such as brain regions (cf. the examples in the road map **Mammalian Brain Regions**) or substructures of these regions, such as layers or columns. In brain theory, we ultimately seek an explanation in terms of neural networks, since the neuron may be considered the basic unit of function as well as of structure, and much further work in computational neuroscience seeks to explain the complex functionality of real neurons in terms of “sub-neural” units, such as membrane compartments, channels, spines, and synapses. What makes the story more subtle is that, in general, a functional analysis proceeding “top-down” from some overall behavior need not map directly into a “bottom-up” analysis proceeding upward from the neural circuitry (brain theory) or basic set of processors (distributed AI), and that several iterations from the “middle out” may be required to bring the structural and functional accounts into consonance. Brain theory may then seek to replace an initially plausible schema analysis with one whose schemas may be constituted by an assemblage of schemas which can each be embodied in one structure (without denying that a given brain region may support the activity of multiple schemas). The schemas that serve as the functional units in our initial hypotheses about the decomposition of some overall function may well differ from the more refined hypotheses which provide an account of structural correlates as well. On the other hand, distributed AI may adopt any schema analysis that is technologically effective, and the schemas may be implemented in whatever medium is appropriate, whether as conventional computer programs, ANNs, or special-purpose devices. These different approaches then rest on effective design of VLSI “chips” or other computing materials (cf. the road map **Implementation and Analysis**).

For brain theory, the top-level schemas must be “large” enough to allow an analysis of behavior at or near the psychological level, yet also be subject to successive decomposition down to a level that may, in certain cases, be implemented in specific neural networks. We again distinguish a schema as a *functional* unit from a neural network as a *structural* unit. A given schema may be distributed across several neural networks; a given neural network may be involved in the implementation of several different schemas. The same will be true for relating connectionist units to single biological neurons. If there is to be a fuller rapprochement between connectionism and neuropsychology, it will be important to use a vocabulary (or context) that allows one to make the necessary distinctions between connectionist and biological neurons.

A top-down analysis (decomposing a function) may suggest that a certain schema is embedded in a certain part of the brain; we can then marshal the available data from anatomy and neurophysiology

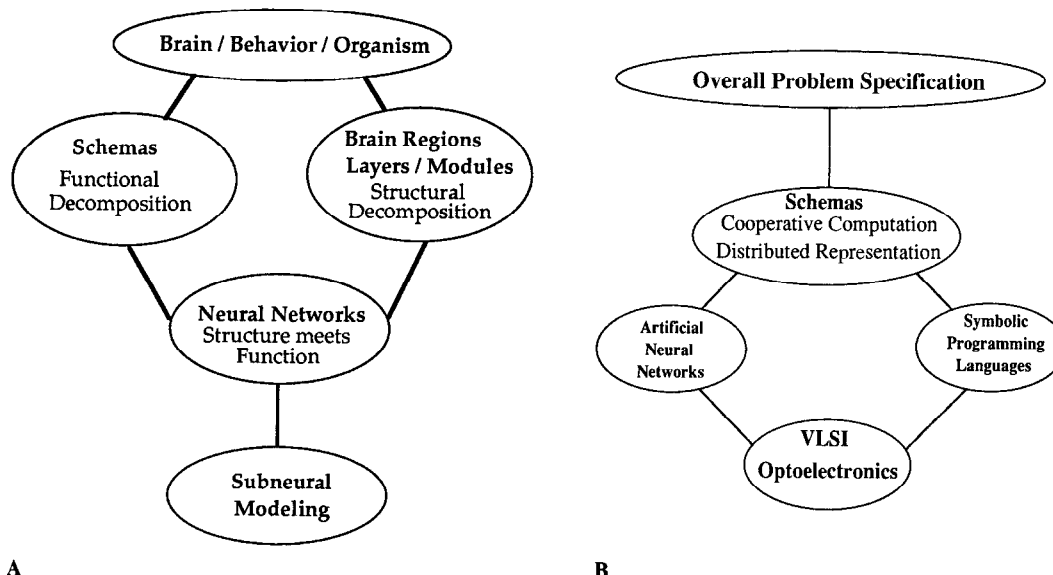


Figure 4. Views of level of analysis of brain and behavior (A) and a distributed technological system (B), highlighting the role of schemas as an intermediate level of functional analysis in each case.

to assess whether the circuitry can, indeed, subserve an instance of that schema. It often happens that the empirical data are inadequate. We then make hypotheses for experimental confirmation. Alternatively, bottom-up analysis of a brain region (assembling its constituents) may suggest that it subserves a different schema from that originally hypothesized, and we must then conduct a new top-down analysis in the light of these newfound constraints.

To illuminate the notion of experimental insight modifying an initial top-down analysis, we consider an example from *Rana computatrix*, a set of models of visuomotor coordination in the frog and toad (cf. VISUOMOTOR COORDINATION IN FROG AND TOAD). Frogs and toads snap at small moving objects and jump away from large ones (to oversimplify somewhat). Thus, a simple schema-model of the frog brain might simply postulate four schemas: two perceptual schemas (processes for recognizing objects or situations) and two motor schemas (processes for controlling some structured behavior). One perceptual schema would recognize small moving objects and activate a motor schema for approaching the prey; the other would recognize large moving objects and activate a motor schema for avoiding the predator. Lesion experiments can put such a model to the test if it is enhanced by hypotheses on the localization of each schema in the brain. It was thought that the tectum (a key visual region in the animal's midbrain) was the locus for recognizing small moving objects, while the pretectum (a region just in front of the tectum) was the locus for recognizing large moving objects. Based on these localization hypotheses, the model described would predict that an animal with a lesioned pretectum would be unresponsive to large objects, but would respond normally to small objects. However, the facts are quite different. A pretectum-lesioned toad will approach moving objects, both large and small, and does not exhibit avoidance behavior. This has led to a new schema model in which a perceptual schema to recognize large moving objects is still localized in the pretectum, but the tectum now contains a perceptual schema for *all* moving objects. We then add that activity of the pretectal schema not only triggers the avoidance motor schema but also inhibits approach. This new schema model still yields the normal behavior to large and small moving objects, but also fits the lesion data, since removal of the pretectum removes inhibition, meaning that the ani-

mal will now approach any moving object (Ewert and von Seelen, 1974).

We have thus seen how schemas may be used to provide falsifiable models of the brain, using lesion experiments to test schema models of behavior, and leading to new functional models that better match the structure of the brain. Note again that, in different species, the map from function to brain structure may be different, while in distributed AI the constraints are not those of analysis but rather those of design—namely, for a given function and a given set of processors, a schema decomposition must be found that will map most efficiently onto a network of processors of a certain kind.

While the brain may be considered a network of interacting “boxes” (anatomically distinguishable structures), there is no reason to expect each such box to mediate a single function that is well-defined from a behavioral standpoint. We have just seen that the frog tectum is implicated in both approach and (when modulated by pretectum) avoidance behavior. The language of schemas lets us express hypotheses about the various functions that the brain performs without assuming localization of any one function in any one region, but also allows us to express the way in which many regions participate in a given function, or a given region participates in many functions.

The style of *cooperative computation* (see COOPERATIVE PHENOMENA) exhibited in both schema theory and connectionism is far removed from serial computation and the symbol-based ideas that have dominated conventional AI. As we shall see in example after example in Part III, the brain has many specialized areas, each with a partial representation of the world. It is only through the interaction of these regions that the unity of behavior of the animal emerges, and the human is no different in this regard. The representation of the world is *the pattern of relationships between all its partial representations*. Much work in AI contributes to schema theory, even when it does not use this term. For example, Brooks (1986) builds robot controllers using layers made up of asynchronous modules that can be considered to be a version of schemas (see REACTIVE ROBOTIC SYSTEMS). This work shares with schema theory, with its mediation of action through a network of schemas, the point that no single, central, logical representation of the world

needs link perception and action. It is also useful to view cooperative computation as a social phenomenon. A schema is a self-contained computing agent (object) with the ability to communicate with other agents, and whose function is specified by some behavior. Whereas schema theory was motivated in great part by the study of interacting brain regions (other influences are reviewed

in SCHEMA THEORY), much early work in distributed AI was motivated by a social analogy in which the schemas were thought of as “agents” analogous to people interacting in a social setting to compete or cooperate in solving some overall problem, a theme elaborated on by Minsky (1985) and whose current status is reviewed in MULTIAGENT SYSTEMS.

I.3. Dynamics and Adaptation in Neural Networks

Section I.1 introduced a number of key concepts from the biological study of neurons, stressing the diversity of neurons both within the human CNS and across species. It presented several simple models of neurons, noting that computational neuroscience has gone on to produce more subtle and complicated neuronal models, while neural computing tends to use simple neurons augmented by “learning rules” for changing connection strengths on the basis of “experience.” The purpose of this section is to introduce two key approaches that dominate the modern study of neural networks: (1) the study of neural networks as dynamic systems (developed more fully in the road map **Dynamic Systems**), and (2) the study of neural networks as adaptive systems (see **Learning in Artificial Networks**). To make this section essentially self-contained, we start by recalling the definitions of the McCulloch-Pitts and leaky integrator neurons from Section I.1, but we do this in the context of a general, semiformal, introduction to dynamic systems.

Dynamic Systems

We motivate the notion of dynamic systems by considering how to abstract the interaction of an organism (or a machine) with its environment. The organism will be influenced by aspects of the current environment—the *inputs* to the organism—while the activity of the environment will be responsive in turn to aspects of the current activity of the organism, the *outputs* of the organism. The inputs and outputs that actually enter into a *theory* of the organism (or machine) are a small sampling of the flux of its interactions with the rest of the universe. There is essentially no limit to how many variables one could include in the analysis; a crucial task in any theory building is to pick the “right” variables.

Depending on the context, we will use the word *system* to denote either the physical reality (which we cannot know in its entirety) or the abstraction with which we approximate it. Inputs and outputs do not constitute a complete description of a system. We cannot predict how someone will answer a question unless we know her state of knowledge; nor can we tell how a computer will process its data unless we know the instructions controlling its computation. In short, we must include a description of the *internal state* of the system which determines what it will extract from its current stimulation in determining its current actions and modifying its internal state. Our abstraction of any real system contains five elements:

1. The set of *inputs*: those variables of the environment which we believe will affect the system behavior of interest to us.
2. The set of *outputs*: those variables of the system which we choose to observe, or which we believe will significantly affect the environment.
3. The set of *states*: those internal variables of the system (which may or may not also be output variables) which determine the relationship between input and output. Essentially, the state of a system is the system’s “internal residue of the past”: when we

know the state of a system, no further information about the past behavior of the system will enable us to refine predictions of the way in which future inputs and outputs of the system will be related.

4. The *state-transition function*: that function which determines how the state will change when the system obtains various inputs.
5. The *output function*: that function which determines what output the system will yield with a given input when in a given state.

Any system in which the state-transition function and output function uniquely determine the new state and output from a specification of the initial state and subsequent inputs is called a *deterministic* system. If, no matter how carefully we specify subsequent inputs to a system, we cannot specify exactly what will be the subsequent states and outputs, we say the system is *probabilistic* or *stochastic*. A stochastic treatment may be worthwhile, either because we are analyzing systems, which are “inescapably” stochastic (e.g., at the quantum level), or because we are analyzing macroscopic systems, which lend themselves to a stochastic description by ignoring “fine details” of microscopic variables. For example, it is usually more reasonable to describe a coin in terms of a 0.5 probability of coming up heads than to measure the initial placement of the coin on the finger and the thrust of the thumb in sufficient detail to determine whether the coin will come up heads or tails.

Continuous-Time Systems

In Newtonian mechanics, the state of the system comprises the positions of its components, which are directly observable, and their velocities, which can be estimated from the observed trajectory over a period of time. Time is continuous (i.e., characterized by the set \mathbb{R} of real numbers), and the way in which the state changes is described by a differential equation: classical mechanics provides the basic example of *continuous-time* systems in which the present state and input determine *the rate at which the state changes*. This requires that the input, output, and state spaces be continuous spaces in which such continuous changes can occur. Consider the simple example of a point mass undergoing rectilinear motion. At any time, its position $y(t)$ is the observable output of the system, and the force $u(t)$ acting upon it is the input applied to the system. Newton’s third law says that the force applied to the system equals the mass times the acceleration $\ddot{y}(t) = mu(t)$, where the acceleration $\ddot{y}(t)$ is the second derivative of $y(t)$. According to Newton’s laws, the state of the system is given by the position and velocity of the particle. We call the position-velocity pair, at any time, the *instantaneous state* $q(t)$ of the system. In fact, the earlier equation gives us enough information to deduce the rate of change $dq(t)/dt$ of this state. Using standard matrix formalism, we thus

have

$$\frac{d}{dt} \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} \dot{y}(t) \\ mu(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ m \end{bmatrix} u(t)$$

while

$$y(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} q(t)$$

This is an example of a *linear* system in which the rate of change of state depends linearly on the present state and input, and the present output depends linearly on the present state. That is, there are matrices F , G , and H such that

$$\frac{dq(t)}{dt} = Fq(t) + Gu(t); y(t) = Hq(t)$$

More generally, a physical system can be expressed by a pair of equations:

$$\begin{aligned} \frac{dq(t)}{dt} &= f(q(t), u(t)) \\ y(t) &= g(q(t)) \end{aligned}$$

The first expresses the rate of change $dq(t)/dt$ of the state as a function of both the state $q(t)$ and the input or control vector $u(t)$ applied at any time t ; the second reads the output from the current state.

We now present the definition of a *leaky integrator neuron* as a continuous-time system. The internal state of the neuron is its membrane potential, $m(t)$, and its output is the firing rate, $M(t)$. The *state transition function* of the cell is expressed as

$$\tau \frac{dm(t)}{dt} = -m(t) + \sum_i w_i X_i(t) + h \quad (1)$$

while the *output function* of the cell is given by the equation

$$M(t) = \sigma(m(t)) \quad (2)$$

Thus, if there are m inputs $X_i(t)$, $i = 1, \dots, m$, then the *input space* of the neuron is \mathbb{R}^m , with current value $(X_1(t), \dots, X_m(t))$, while the *state* and *output* spaces of the neuron both equal \mathbb{R} , with current values $m(t)$ and $M(t)$, respectively.

Let us now briefly (and semiformaly) see how a neural network comprised of leaky integrator neurons can also be seen as a continuous-time system in this sense. As typified in Figure 5, we characterize a neural network by selecting N neurons (each with specified input weights and resting potential) and by taking the axon of each neuron, which may be split into several branches

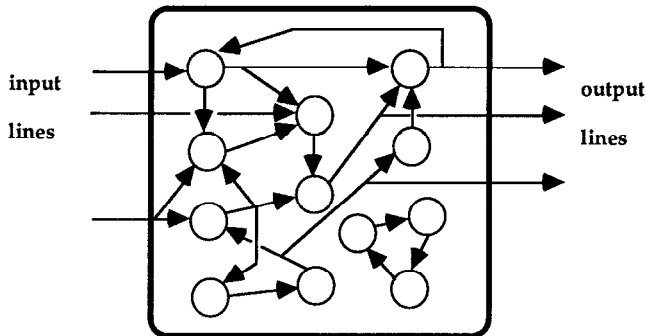


Figure 5. A neural network viewed as a system. The input at time t is the pattern of firing on the input lines, the output is the pattern of firing on the output lines; and the internal state is the vector of firing rates of all the neurons of the network.

carrying identical output signals, and either connecting each line to a unique input of another neuron or feeding it outside the net to provide one of the K network output lines. Then every input to a given neuron must be connected either to an output of another neuron or to one of the (possibly split) L input lines of the network. Thus the input set $X = \mathbb{R}^L$, the state set $Q = \mathbb{R}^N$, and the output set $Y = \mathbb{R}^K$. If the i th output line comes from the j th neuron, then the *output function* is determined by the fact that the i th component of the output at time t is the firing rate $M_j(t) = \sigma_j(M_j(t))$ of the j th neuron at time t . The state transition function for the neural network follows from the state transition functions of each of the N neurons

$$\tau \frac{dm_i(t)}{dt} = -m_i(t) + \sum_j w_{ij} X_j(t) + h_i \quad (3)$$

as soon as we specify whether $X_{ij}(t)$ is the output $M_j(t)$ of the k th neuron or the value $x_i(t)$ currently being applied on the i th input line of the overall network.

Discrete-Time Systems

In contrast to continuous-time systems, which *must* have continuous state spaces on which the differential equations for the state transition function can be defined, *discrete-time* systems may have either continuous or discrete state spaces. (A *discrete* state space is just a set with no specific metric or topological structure.) For example, a McCulloch-Pitts neuron is considered to operate on a discrete-time scale, $t = 0, 1, 2, 3, \dots$, and has connection weights w_i and threshold θ . If at time t the value of the i th input is $x_i(t)$, then the output one time step later, $y(t+1)$, equals 1 if and only if $\sum_i w_i x_i(t) \geq \theta$. If there are m inputs $(x_1(t), \dots, x_m(t))$, then, since inputs and outputs are binary, such a neuron has input set $= \{0, 1\}^m$, state set $=$ output set $\{0, 1\}$ (we treat the current state and output as being identical). On the other hand, the important learning scheme known as backpropagation (defined later) is based on neurons which operate on discrete time, but with both input and output taking continuous values in some range, say $[0, 1]$.

In computer science, an *automaton* is a discrete-time system with discrete input, output, and state spaces. Formally, we describe an automaton by the sets X , Y , and Q of inputs, outputs, and states, respectively, together with the *next-state function* $\delta: Q \times X \rightarrow Q$ and the *output function* $\beta: Q \rightarrow Y$. If the automaton is in state q and receives input x at time t , then its next state will be $\delta(q, x)$ and its next output will be $\beta(q)$. It should be clear that a McCulloch-Pitts neural network (i.e., a network like that shown in Figure 5, but a discrete-time network with each neuron a McCulloch-Pitts neuron) functions like a finite automaton, as each neuron changes state synchronously on each tick of the time scale $t = 0, 1, 2, 3, \dots$. Conversely, it can be shown (see Arbib, 1987; the result was essentially, though inscrutably, due to McCulloch and Pitts, 1943) that any finite automaton can be simulated by a suitable McCulloch-Pitts neural network.

Stability, Limit Cycles, and Chaos

With the previous discussion, we now have more than enough material to understand the crucial dynamic systems concept of *stability* and the related concepts of limit cycles and chaos (see COMPUTING WITH ATTRACTORS and CHAOS IN NEURAL SYSTEMS). We want to know what happens to an “unperturbed” system, i.e., one for which the input is held constant (possibly with some specific “null input,” usually denoted by 0, the “zero” input in X). An *equilibrium* is a state q in which the system can stay at rest, i.e., such that $\delta(q, 0) = q$ (discrete time) or $dq/dt = f(q, 0) = 0$ (continuous time). The study of stability is concerned with the issue of whether or not this rest point will be maintained in the face of slight disturbances. To see the variety of equilibria, we use the image of a sticky ball

rolling on the “hillside” of Figure 6. We say that point A on the “hillside” in this diagram is an *unstable equilibrium* because a slight displacement from A will tend to increase over time. Point B is in a region of *neutral equilibrium* because slight displacements will tend not to change further, while C is a point of *stable equilibrium*, since small displacements will tend to decrease over time. Note the word “small”: in a nonlinear system like that of Figure 6, a large displacement can move the ball from the *basin of attraction* of C (the set of states whose dynamics tends toward C) to another one. Clearly, the ball will not tend to return to C after a massive displacement that moves the ball to the far side of A’s hilltop.

Many nonlinear systems have another interesting property: they may exhibit *limit cycles*. These are closed trajectories in the state space, and thus may be thought of as “dynamic equilibria.” If the state of a system follows a limit cycle, we may also say it oscillates or exhibits periodic behavior. A limit cycle is *stable* if a small displacement will be reduced as the trajectory of the system comes closer and closer to the original limit cycle. By contrast, a limit cycle is *unstable* if such excursions do not die out. Research in nonlinear systems has also revealed what are called *strange attractors*. These are attractors which, unlike simple limit cycles, describe such complex paths through the state space that, although the system is deterministic, a path that approaches the strange attractor gives every appearance of being random. The point here is that very small differences in initial state may be amplified with the passage of time, so that differences that at first are not even noticeable will yield, in due course, states that are very different indeed. Such a trajectory has become the accepted mathematical model of *chaos*, and it is used to describe a number of physical phenomena, such as the onset of turbulence in a weather system, as well as a number of phenomena in biological systems (see CHAOS IN BIOLOGICAL SYSTEMS; CHAOS IN NEURAL SYSTEMS).

Hopfield Nets

Many authors have treated neural networks as dynamical systems, employing notions of equilibrium, stability, and so on, to classify their performance (see, e.g., Grossberg, 1967; Amari and Arbib, 1977; see also COMPUTING WITH ATTRACTORS). However, it was a paper by John Hopfield (1982) that was the catalyst in attracting the attention of many physicists to this field of study. In a McCulloch-Pitts network, every neuron processes its inputs to determine a new output at each time step. By contrast, a *Hopfield net* is a net of such units with (1) *symmetric weights* ($w_{ij} = w_{ji}$) and no self-connections ($w_{ii} = 0$), and (2) *asynchronous updating*. For instance, let s_i denote the state (0 or 1) of the i th unit. At each time

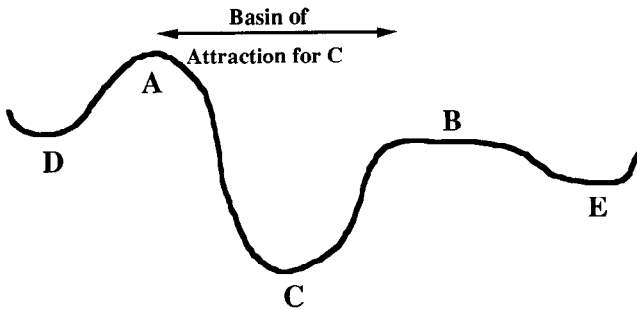


Figure 6. An energy landscape: For a ball rolling on the “hillside,” point A is an *unstable equilibrium*, point B lies in a region of *neutral equilibrium*, and point C is a point of *stable equilibrium*. Point C is called an attractor: the basin of attraction of C comprises all states whose dynamics tend toward C.

step, pick just one unit at random. If unit i is chosen, s_i takes the value 1 if and only if $\sum_j w_{ij}s_j \geq \theta_i$. Otherwise s_i is set to 0. Note that this is an *autonomous* (input-free) network: there are no inputs (although instead of considering θ_i as a threshold we may consider $-\theta_i$ as a constant input, also known as a bias).

Hopfield defined a measure called the *energy* for such a net (see ENERGY FUNCTIONALS FOR NEURAL NETWORKS)

$$E = -\frac{1}{2} \sum_{ij} s_i s_j w_{ij} + \sum_i s_i \theta_i \quad (1)$$

This is not the physical energy of the neural net but a mathematical quantity that, in some ways, does for neural dynamics what the potential energy does for Newtonian mechanics. In general, a mechanical system moves to a state of lower potential energy just as, in Figure 6, the ball tends to move downhill. Hopfield showed that his symmetrical networks with asynchronous updating had a similar property.

For example, if we pick a unit and the foregoing firing rule does not change its s_i , it will not change E . However, if s_i initially equals 0, and $\sum_j w_{ij}s_j \geq \theta_i$, then s_i goes from 0 to 1 with all other s_j constant, and the “energy gap,” or change in E , is given by

$$\begin{aligned} \Delta E &= -\frac{1}{2} \sum_j (w_{ij}s_j + w_{ji}s_j) + \theta_i \\ &= -\sum_j w_{ij}s_j + \theta_i, \text{ by symmetry} \\ &\leq 0 \text{ since } \sum_j w_{ij}s_j \geq \theta_i \end{aligned}$$

Similarly, if s_i initially equals 1, and $\sum_j w_{ij}s_j < \theta_i$, then s_i goes from 1 to 0 with all other s_j constant, and the energy gap is given by

$$\Delta E = \sum_j w_{ij}s_j - \theta_i < 0$$

In other words, with every asynchronous updating, we have $\Delta E \leq 0$. Hence the dynamics of the net tends to move E toward a minimum. We stress that there may be different such states—they are *local* minima, just as, in Figure 6, both D and E are local minima (each of them is lower than any “nearby” state) but not global minima (since C is lower than either of them). Global minimization is not guaranteed.

The expression just presented for ΔE depends on the symmetry condition, $w_{ij} = w_{ji}$, for without this condition, the expression would instead be $\Delta E = -(\frac{1}{2})\sum_j (w_{ij}s_j + w_{ji}s_j) + \theta_i$ and in this case, Hopfield’s updating rule need not yield a passage to energy minimum, but might instead yield a limit cycle, which could be useful in, e.g., controlling rhythmic behavior (see, e.g., RESPIRATORY RHYTHM GENERATION). In a control problem, a link w_{ij} might express the likelihood that the action represented by i would precede that represented by j , in which case $w_{ij} = w_{ji}$ is normally inappropriate.

The condition of *asynchronous* update is crucial, too. If we consider the simple “flip-flop” with $w_{12} = w_{21} = 1$ and $\theta_1 = \theta_2 = 0.5$, then the McCulloch-Pitts network will *oscillate* between the states (0, 1) and (1, 0) or will sit in the states (0, 0) or (1, 1); in other words, there is no guarantee that it will converge to an equilibrium. However, with $E = -(\frac{1}{2})\sum_{ij} s_i s_j w_{ij} + \sum_i s_i \theta_i$, we have $E(0, 0) = 0$, $E(0, 1) = E(1, 0) = 0.5$, and $E(1, 1) = 0$, and the Hopfield network will *converge* to the global minimum at either (0, 0) or (1, 1).

Hopfield also aroused much interest because he showed how a number of optimization problems could be “solved” using neural networks. (The quotes around “solved” acknowledge the fact that the state to which a neural network converges may represent a local, rather than a global, optimum of the corresponding optimization

problem.) Such networks were similar to the “constraint satisfaction” networks that had already been studied in the computer vision community. (In most vision algorithms—see, e.g., STEREO CORRESPONDENCE—constraints can be formulated in terms of symmetric weights, so that $w_{ij} = w_{ji}$ is appropriate.) The aim, given a “constraint satisfaction” problem, is to so choose weights for a neural network so that the energy E for that network is a measure of the overall constraint violation. A famous example is the Traveling Salesman Problem (TSP): There are n cities, with a road of length l_{ij} joining city i to city j . The salesman wishes to find a way to visit the cities that is optimal in two ways: each city is visited only once, and the total route is as short as possible. We express this as a constraint satisfaction network in the following way: Let the activity of neuron N_{ij} express the decision to go straight from city i to city j . The cost of this move is simply l_{ij} , and so the total “transportation cost” is $\sum_{ij} l_{ij} N_{ij}$. It is somewhat more challenging to express the cost of violating the “visit a city only once” criterion, but we can reexpress it by saying that, for city j , there is one and only one city i from which j is directly approached. Thus, $\sum_j (\sum_i N_{ij} - 1)^2 = 0$ just in case this constraint is satisfied; a non-zero value measures the extent to which this constraint is violated. This can then be mapped into the setting of weights and thresholds for a Hopfield network. Hopfield and Tank (1986) constructed chips for this network which do indeed settle very quickly to a local minimum of E . Unfortunately, there is no guarantee that this minimum is globally optimal. The article OPTIMIZATION, NEURAL presents this and a number of other neurally based approaches to optimization. The article SIMULATED ANNEALING AND BOLTZMANN MACHINES shows how noise may be added to “shake” a system out of a local minimum and let it settle into a global minimum. (Consider, for example, shaking that is strong enough to shake the ball from D to A , and thus into the basin of attraction of C , in Figure 6, but not strong enough to shake the ball back from C toward D .)

Adaptation in Dynamic Systems

In the previous discussion of neural networks as dynamic systems, the dynamics (i.e., the state transition function) has been fixed. However, just as humans and animals learn from experience, so do many important applications of ANNs depend on the ability of these networks to adapt to the task at hand by, e.g., changing the values of the synaptic weights to improve performance. We now introduce the general notion of an adaptive system as background to some of the most influential “learning rules” used in adaptive neural networks. The key motivation for using learning networks is that it may be too hard to program explicitly the behavior that one sees in a black box, but one may be able to drive a network by the actual input/output behavior of that box, or by some description of its trajectories, to cause it to adapt itself into a network which approximates that given behavior. However, as we will

stress at the end of this section, a learning algorithm may not solve a problem within a reasonable period of time unless the initial structure of the network is suitable.

Adaptive Control

A key problem of technology is to control a complex system so that it behaves in some desired way, whether getting a space probe on course to Mars or a steel mill to produce high-quality steel. A common situation that complicates this *control problem* is that the controlled system may not be known accurately; it may even change its character somewhat with time. For example, as fuel is depleted, the mass and moments of inertia of the probe may change in unpredicted ways. The *adaptation problem* involves determining, on the basis of interaction with a given system, an appropriate “model” of the system which the controller can use in solving the control problem.

Suppose we have available an *identification procedure* which can find an adequate parametric representation of the controlled system (see IDENTIFICATION AND CONTROL). Then, rather than build a controller specifically designed to control this one system, we may instead build a general-purpose controller which can accommodate to any reasonable set of parameters. The controller then uses the parameters which the identification procedure provides as the best estimate of the controlled system’s parameters at that time. If the identification procedure can make accurate estimates of the system’s parameters as quickly as they actually change, the controller will be able to act efficiently despite fluctuations in controlled system dynamics. The controller, when coupled to an identification procedure, is an *adaptive controller*; that is, it adapts its control strategy to changes in the dynamics of the controlled system. However, the use of an explicit identification procedure is only one way of building an adaptive controller. Adaptive neural nets may be used to build adaptive procedures which may directly modify the parameters in some control rule, or identify the system *inverse* so that desired outputs can be automatically transformed into the inputs that will achieve them. (See SENSORIMOTOR LEARNING for the distinction between forward and inverse models.)

Pattern Recognition

In the setup shown in Figure 7, the *preprocessor* extracts from the environment a set of “confidence levels” for various input features (see FEATURE ANALYSIS), with the result represented by a vector of d real numbers. In this formalization, any pattern x is represented by a point (x_1, x_2, \dots, x_d) in a d -dimensional Euclidean space \mathbb{R}^d called the *pattern space*. The pattern recognizer then takes the pattern and produces a response that may have one of K distinct values where there are K categories into which the patterns must be sorted; points in \mathbb{R}^d are thus grouped into at least K different sets (see CONCEPT LEARNING and PATTERN RECOGNITION). However, a

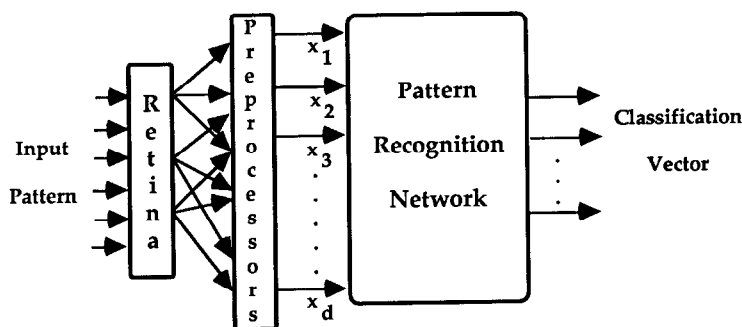


Figure 7. One strategy in pattern recognition is to precede the adaptive neural network by a fixed layer of “preprocessors” or “feature extractors” which replace the image by a finite vector for further processing. In other approaches, the functions defined by the early layers of the network may themselves be subject to training.

category might be represented in more than one region of \mathbb{R}^d . To take an example from visual pattern recognition (although the theory of pattern recognition networks applies to any classification of \mathbb{R}^d), a and A are members of the category of the first letter of the English alphabet, but they would be found in different connected regions of a pattern space. In such cases, it may be necessary to establish a hierarchical system involving a separate apparatus to recognize each subset, and a further system that recognizes that the subsets all belong to the same set (a related idea was originally developed by Selfridge, 1959; for adaptive versions, see MODULAR AND HIERARCHICAL LEARNING SYSTEMS). Here we avoid this problem by concentrating on the case in which the decision space is divided into exactly two connected regions.

We call a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ a *discriminant function* if the equation $f(x) = 0$ gives the *decision surface* separating two regions of a pattern space. A basic problem of pattern recognition is the specification of such a function. It is virtually impossible for humans to “read out” the function they use (not to mention *how* they use it) to classify patterns. Thus, a common strategy in pattern recognition is to provide a classification machine with an adjustable function and to “train” it with a set of patterns of known classification that are typical of those with which the machine must ultimately work. The function may be linear, quadratic, polynomial, or even more subtle yet, depending on the complexity and shape of the pattern space and the necessary discriminations. The experimenter chooses a class of functions with parameters which, it is hoped, will, with proper adjustment, yield a function that will successfully classify any given pattern. For example, the experimenter may decide to use a linear function of the form

$$f(x) = w_1x_1 + w_2x_2 + \dots + w_dx_d + w_{d+1}$$

(i.e., a McCulloch-Pitts neuron!) in a two-category pattern classifier. The equation $f(x) = 0$ gives a hyperplane as the decision surface, and training involves adjusting the coefficients ($w_1, w_2, \dots, w_d, w_{d+1}$) so that the decision surface produces an acceptable separation of the two classes. We say that two categories are *linearly separable* if an acceptable setting of such linear weights exists. Thus, pattern recognition poses (at least) the following challenges to neural networks:

(a) Find a “good” set of preprocessors. Competitive learning based on Hebbian plasticity (see COMPETITIVE LEARNING, as well as the following text) provides one way of finding such features by extracting statistically significant patterns from a set of input patterns. For example, if such a network were exposed to many, but only, letters of the Roman alphabet, then it would find that certain line segments and loops occurred repeatedly, even if there were no teacher to tell it how to classify the patterns.

(b) Given a set of preprocessors and a set of patterns which have already been classified, adjust the connections of a neural network so that it acts as an effective pattern recognizer. That is, its response to a preprocessed pattern should usually agree well with the classification provided by a teacher.

(c) Of course, if the neural network has multiple layers with adaptable synaptic weights, then the early layers can be thought of as preprocessors for the later layers, and we have a case of supervised, rather than Hebbian, formation of these “feature detectors”—emphasizing features which are not only statistically significant elements of the input patterns but which also serve to distinguish usefully to which class a pattern belongs.

Associative Memory

In pattern recognition, we associate a pattern with a “label” or “category.” Alternatively, an associative memory takes some “key” as input and returns some “associated recollection” as output (see ASSOCIATIVE NETWORKS). For example, given the sound of a word,

we may wish to recall its spelling. Given a misspelled word, we may wish to recall the correctly spelled word of which it is most plausibly a “degraded image.” There are two major approaches to the use of neural networks as associative memories:

In *nonrecurrent* neural networks, there are no loops (i.e., we cannot start at any neuron and “follow the arrows” to get back to that neuron). We use such a network by fixing the pattern of inputs as the key, and holding them steady. Since the absence of loops ensures that the input pattern uniquely determines the output pattern (after the new inputs have time to propagate their effects through the network), this uniquely determined output pattern is the recollection associated with the key.

In *recurrent* networks, the presence of loops implies that the input alone may not determine the output of the net, since this will also depend on the initial state of the network. Thus, recurrent networks are often used as associative memories in the following way. The inputs are only used transiently to establish the initial state of the neural network. After that, the network operates autonomously (i.e., uninfluenced by any inputs). If and when it reaches an equilibrium state, that state is read out as the recollection associated with the key.

In either case, the problem is to set the weights of the neural network so that it associates keys as accurately as possible with the appropriate recollections.

Learning Rules

Most learning rules in current models of “lumped neurons” (i.e., those that exclude detailed analysis of the fine structure of the neuron or the neurochemistry of neural plasticity) take the form of schemes for adjusting the synaptic weights, the “ws.” The two classic learning schemes for McCulloch-Pitts-type formal neurons are due to Hebb (see HEBBIAN SYNAPTIC PLASTICITY) and Rosenblatt (the perceptron, see PERCEPTRONS, ADALINES, AND BACKPROPAGATION), and we now introduce these in turn.

Hebbian Plasticity and Network Self-Organization

In Hebb’s (1949) learning scheme (see HEBBIAN SYNAPTIC PLASTICITY), the connection between two neurons is strengthened if both neurons fire at the same time. The simplest example of such a rule is to increase w_{ij} by the following amount:

$$\Delta w_{ij} = ky_i x_j$$

where synapse w_{ij} connects a presynaptic neuron with firing rate x_j to a postsynaptic neuron with firing rate y_i . The trouble with the original Hebb model is that every synapse will eventually get stronger and stronger until they all saturate, thus destroying any selectivity of association. Von der Malsburg’s (1973) solution was to normalize the synapses impinging on a given neuron. To accomplish this, one must first compute the Hebbian “update” $\Delta w_{ij} = ky_i x_j$ and then divide this by the total putative synaptic weights to get the final result which replaces w_{ij} by

$$\frac{w_{ij} + \Delta w_{ij}}{\sum_k (w_{kj} + \Delta w_{kj})}$$

where the summation k extends over all inputs to the neuron. This new rule not only increases the strengths of those synapses with inputs strongly correlated with the cell’s activity, but also decreases the synaptic strengths of other connections in which such correlations did not arise.

Von der Malsburg was motivated by the pattern recognition problem and was concerned with how individual cells in his network might come to be tuned so as to respond to one particular

input “feature” rather than another (see OCULAR DOMINANCE AND ORIENTATION COLUMNS for background as well as a review of more recent approaches). This exposed another problem with Hebb’s rule: a lot of nearby cells may, just by chance, all have initial random connectivity which makes them easily persuadable by the same stimulus; alternatively, the same pattern might occur many times before a new pattern is experienced by the network. In either case, many cells would become tuned to the same feature, with not enough cells left to learn important and distinctive features. To solve this, von der Malsburg introduced *lateral inhibition* into his model. In this connectivity pattern, activity in any one cell is distributed laterally to reduce (partially inhibit) the activity of nearby cells. This ensures that if one cell—call it A—were especially active, its connections to nearby cells would make them less active, and so make them less likely to learn, by Hebbian synaptic adjustment, those features that most excite A.

In summary, then, when the Hebbian rule is augmented by a normalization rule, it tends to “sharpen” a neuron’s predisposition “without a teacher,” getting its firing to become better and better correlated with a cluster of stimulus patterns. This performance is improved when there is some competition between neurons so that if one neuron becomes adept at responding to a pattern, it inhibits other neurons from doing so (COMPETITIVE LEARNING). Thus, the final set of input weights to the neuron depends both on the initial setting of the weights and on the pattern of clustering of the set of stimuli to which it is exposed (see DATA CLUSTERING AND LEARNING). Other “post-Hebbian” rules, motivated both by technological efficiency and by recent biological findings, are discussed in several articles in Part III, including HEBBIAN LEARNING AND NEURONAL REGULATION AND POST-HEBBIAN LEARNING ALGORITHMS.

In the adaptive architecture just described, the inputs are initially randomly connected to the cells of the processing layer. As a result, none of these cells is particularly good at pattern recognition. However, by sheer statistical fluctuation of the synaptic connections, one will be slightly better at responding to a particular pattern than others are; it will thus slightly strengthen those synapses which allow it to fire for that pattern and, through lateral inhibition, this will make it harder for cells initially less well tuned for that pattern to become tuned to it. Thus, without any teacher, this network automatically organizes itself so that each cell becomes tuned for an important cluster of information in the sensory inflow. This is a basic example of the kind of phenomenon treated in SELF-ORGANIZATION AND THE BRAIN.

Perceptrons

Perceptrons are neural nets that change with “experience,” using an *error-correction rule* designed to change the weights of each response unit when it makes erroneous responses to stimuli that are presented to the network. We refer to the judge of what is correct as the “teacher,” although this may be another neural network, or some environmental input, rather than a signal supplied by a human teacher in the usual schoolroom sense. Consider the case in which a set \mathbf{R} of input lines feeds a Pitts-McCulloch neural network whose neurons are called *associator units* and which in turn provide the input to a single McCulloch-Pitts neuron (called the *output unit* of the perceptron) with adjustable weights (w_1, \dots, w_d) and threshold θ . (In the case of visual pattern recognition, we think of \mathbf{R} as a rectangular “retina” onto which patterns may be projected.) A *simple perceptron* is one in which the associator units are not interconnected, *which means that it has no short-term memory*. (If such connections are present, the perceptron is called *cross-coupled*. A cross-coupled perceptron may have multiple layers and loops back from an “earlier” to a “later” layer.) If the associator units feed the pattern $x = (x_1, \dots, x_d)$ to the output unit, then the response of that unit will be to provide the pattern discrimination

with discriminant function $f(x) = w_1x_1 + \dots + w_dx_d - \theta$. In other words, the simple perceptron can only compute a *linearly separable* function of the pattern as provided by the associator units. The question asked by Rosenblatt (1958) and answered by many others since (cf. Nilsson, 1965) was, “Given a simple perceptron (i.e., only the synaptic weights of the output unit are adjustable), can we ‘train’ it to recognize a given linearly separable set of patterns by adjusting the ‘weights’ on various interconnections on the basis of feedback on whether or not the network classifies a pattern correctly?” The answer was “Yes: if the patterns are linearly separable, then there is a learning scheme which will eventually yield a satisfactory setting of the weights.” The best-known perceptron learning rule strengthens an active synapse if the effluent neuron fails to fire when it should have fired, and weakens an active synapse if the neuron fires when it should not have done so:

$$\Delta w_{ij} = k(Y_i - y_i)x_j$$

As before, synapse w_{ij} connects a presynaptic neuron with firing rate x_j to a postsynaptic neuron with firing rate y_i , but now Y_i is the “correct” output supplied by the “teacher.” (This is similar to the Widrow-Hoff [1960] least mean squares model of adaptive control; see PERCEPTRONS, ADALINES, AND BACKPROPAGATION.) Notice that the rule does change the response to x_j “in the right direction.” If the output is correct, $Y_i = y_i$ and there is no change, $\Delta w_{ij} = 0$. If the output is too small, then $Y_i - y_i > 0$, and the change in w_{ij} will add $\Delta w_{ij}x_j = k(Y_i - y_i)x_j > 0$ to the output unit’s response to (x_1, \dots, x_d) . Similarly, if the output is too large, then $Y_i - y_i < 0$, Δw_{ij} will add $k(Y_i - y_i)x_j < 0$ to the output unit’s response. Thus, there is a sense in which the new setting $w' = w + \Delta w$ classifies the input pattern x “more nearly correctly” than w does. Unfortunately, in classifying x “more correctly” we run the risk of classifying another pattern “less correctly.” However, the *perceptron convergence theorem* (see Arbib, 1987, pp. 66–69, for a proof) shows that Rosenblatt’s procedure does not yield an endless seesaw, but will eventually converge to a correct set of weights if one exists, albeit perhaps after many iterations through the set of trial patterns.

Network Complexity

The perceptron convergence theorem states that, if a linear separation exists, the perceptron error-correction scheme will find it. Minsky and Papert (1969) revived the study of perceptrons (although some AI workers thought they had killed it!) by responding to such results with questions like, “Your scheme works when a weighting scheme exists, but *when* does there exist such a setting of the weights?” More generally, “Given a pattern-recognition problem, how much of the retina must each associator unit ‘see’ if the network is to do its job?” Minsky and Papert studied when it was possible for a McCulloch-Pitts neuron (no matter how trained) to combine information in a single preprocessing layer to perform a given pattern recognition task, such as recognizing whether a pattern X of 1s on the retina (the other retinal units having output 0) is *connected*, that is, whether a path can be drawn from any 1 of X to another without going through any 0s. Another question was to determine whether X is of *odd parity*, i.e., whether X contains an odd number of 1s. The question is, “How many inputs are required for the preprocessing units of a simple perceptron to successfully implement f ?” We can get away with using a single element, computing an arbitrary Boolean function, and connecting it to all the units of the retina. So the question that really interests us is whether we can get away with a response unit connected to preprocessors, each of which receives inputs from a limited set of retinal units to make a global decision by synthesizing an array of local views.

We convey the flavor of Minsky and Papert’s approach by the example of XOR, the simple Boolean operation of addition modulo

2, also known as the exclusive-or. If we imagine the square with vertices (0, 0), (0, 1), (1, 1), and (1, 0) in the Cartesian plane, with (x_1, x_2) being labeled by $x_1 \oplus x_2$, we have 0s at one diagonally opposite pair of vertices and 1s at the other diagonally opposite pair of vertices. It is clear that there is no way of interposing a straight line such that the 1s lie on one side and the 0s lie on the other side. However, we shall prove it mathematically to gain insight into the techniques used by Minsky and Papert.

Consider the claim that we wish to prove wrong: that there actually exists a neuron with threshold θ and weights α and β such that $x_1 \oplus x_2 = 1$ if and only if $\alpha x_1 + \beta x_2 \geq \theta$. The crucial point is to note that the function of addition modulo 2 is symmetric; therefore, we must also have $x_1 \oplus x_2 = 1$ if and only if $\beta x_1 + \alpha x_2 \geq \theta$, and, so, adding together the two terms, we have $x_1 \oplus x_2 = 1$ if and only if $(\frac{1}{2})(\alpha + \beta)(x_1 + x_2) \geq \theta$. Writing $(\frac{1}{2})(\alpha + \beta)$ as γ , we see that we have reduced three putative parameters α , β , and θ to just two, namely γ and θ .

We now set $t = x_1 + x_2$ and look at the polynomial $\gamma t - \theta$. It is a degree 1 polynomial, but note: at $t = 0$, $\gamma t - \theta$ must be less than zero ($0 \oplus 0 = 0$); at $t = 1$, it is greater than or equal to zero ($0 \oplus 1 = 1 \oplus 0 = 1$); and at $t = 2$, it is again less than zero ($1 \oplus 1 = 0$). This is a contradiction—a polynomial of degree 1 cannot change sign from positive to negative more than once. We conclude that there is no such polynomial, and thus that there is no threshold element which will add modulo 2.

We now understand a general method used again and again by Minsky and Papert: start with a pattern-classification problem. Observe that certain symmetries leave it invariant. For instance, for the parity problem (is the number of active elements even or odd?), which includes the case of addition modulo 2 when the retina has only two units, any permutation of the points of the retina would leave the classification unchanged. Use this to reduce the number of parameters describing the circuit. Then lump items together to get a polynomial and examine actual patterns to put a lower bound on the degree of the polynomial, fixing things so that this degree bounds the number of inputs to the response unit of a simple perceptron.

Minsky and Papert provide many interesting theorems (for the proof of an illustrative sample, see Arbib, 1987, pp. 82–84). As just one example, we may note that they prove that the parity function requires preprocessors big enough to scan the whole retina if the preprocessors can only be followed by a single McCulloch-Pitts neuron. By contrast, to tell whether the number of active retinal inputs reaches a certain threshold only requires two inputs per neuron in the first layer. (For other complexity results, see the articles listed in the road map **Computability and Complexity**.)

Gradient Descent and Credit Assignment

The implication of the results on “network complexity” is clear: if we limit the complexity of the units in a neural network, then in general we will need many layers, rather than a single layer, if the network is to have any chance of being trained to realize many “interesting” functions. This conclusion motivates the study of training rules for multilayer perceptrons, of which the most widely used is *backpropagation*. Before describing this method, we first discuss two general notions of which it is an important exemplar: *gradient descent* and *credit assignment*.

In discussing Hopfield networks, we introduced the metaphor of an “energy landscape” (see Figure 6). The asynchronous updates move the state of the network (the vector of neural activity levels) “downhill,” tending toward a local energy minimum. Our task now is to realize that the metaphor works again on a far more abstract level when we consider learning. In learning, the dynamic variable is not the network state, but rather the vector of synaptic weights (or whatever other set of network parameters is adjusted by the

learning rules). We now conduct *gradient descent in weight space*. At each step, the weights are adjusted in such a way as to improve the performance of the network. (As in the case of the simple perceptron, the improvement is a “local” one based on the current situation. It is, in this case, a matter for computer simulation to prove that the cumulative effect of these small changes is a network which solves the overall problem.)

But how do we recognize which “direction” in weight space is “downhill”? Suppose success is achieved by a complex mechanism after operating over a considerable period of time (for example, when a chess-playing program wins a game). To what particular decisions made by what particular components should the success be attributed? And, if failure results, what decisions deserve blame? This is closely related to the problem known as the “mesa” or “plateau” problem (Minsky, 1961). The performance evaluation function available to a learning system may consist of large level regions in which gradient descent degenerates to exhaustive search, so that only a few of the situations obtainable by the learning system and its environment are known to be desirable, and these situations may occur rarely.

One aspect of this problem, then, is the *temporal* credit assignment problem. The utility of making a certain action may depend on the sequence of actions of which it is a part, and an indication of improved performance may not occur until the entire sequence has been completed. This problem was attacked successfully in Samuel’s (1959) learning program for playing checkers. The idea is to interpret predictions of future reward as rewarding events themselves. In other words, neutral stimulus events can themselves become reinforcing if they regularly occur before events that are intrinsically reinforcing. Such *temporal difference learning* (see REINFORCEMENT LEARNING) is like a process of erosion: the original uninformative mesa, where only a few sink holes allow gradient descent to a local minimum, is slowly replaced by broader valleys in which gradient descent may successfully proceed from many different places on the landscape.

Another aspect of credit assignment concerns structural factors. In the simple perceptron, only the weights to the output units are to be adjusted. This architecture can only support maps which are linearly separable as based on the patterns presented by the preprocessors, and we have seen that many interesting problems require preprocessing units of undue complexity to achieve linear separability. We thus need multiple layers of preprocessors, and, since one may not know a priori the appropriate set of preprocessors for a given problem, these units should be trainable too. This raises the question, “How does a neuron deeply embedded within a network ‘know’ what aspect of the outcome of an overall action was ‘its fault’?” This is the *structural* credit assignment problem. In the next section, we shall study the most widely used solution to this problem, called *backpropagation*, which propagates back to a hidden unit some measure of its responsibility.

Backpropagation is an “adaptive architecture”: it is not just a local rule for synaptic adjustment; it also takes into account the position of a neuron in the network to indicate how the neuron’s weights are to change. (In this sense, we may see the use of lateral inhibition to improve Hebbian learning as the first example of an adaptive architecture in these pages.) This adaptive architecture is an example of “neurally inspired” modeling, not modeling of actual brain structures; and there is no evidence that backpropagation represents actual brain mechanisms.

Backpropagation

The task of backpropagation is to train a *multilayer* (feedforward) *perceptron* (or MLP), a loop-free network which has its units arranged in layers, with a unit providing input only to units in the next layer of the sequence. The first layer comprises fixed input

units; there may then be several layers of trainable “hidden units” carrying an internal representation, and finally, there is the layer of output units, also trainable. (A simple perceptron then corresponds to the case in which we view the input units as fixed associator units, i.e., they deliver a preprocessed, rather than a “raw,” pattern which connect directly to the output units without any hidden units in between.) For what follows, it is crucial that each unit *not* be binary: it has both input and output taking continuous values in some range, say $[0, 1]$. The response is a sigmoidal function of the weighted sum. Thus, if a unit has inputs x_k with corresponding weights w_{ik} , the output x_i is given by $x_i = f_i(\sum w_{ik}x_k)$, where f_i is a sigmoidal function, say

$$f_i(x) = \frac{1}{1 + e^{-(x + \theta_i)}}$$

with θ_i being a bias or threshold for the unit.

The environment only evaluates the output units. We are given a training set of input patterns p and corresponding desired target patterns t^p for the output units. With o^p the actual output pattern elicited by input p , the aim is to adjust the weights in the network to minimize the error

$$E = \sum_{\text{patterns } p} \sum_{\text{output neurons } k} (t_k^p - o_k^p)^2$$

Rumelhart, Hinton, and Williams (1986) were among those who devised a formula for propagating back the gradient of this evaluation from a unit to its inputs. This process can continue by back-propagation through the entire net. The scheme seems to avoid many false minima. At each trial, we fix the input pattern p and consider the corresponding “restricted error”

$$E = \sum_k (t_k - o_k)^2$$

where k ranges over designated “output units.” The net has many units interconnected by weights w_{ij} . The learning rule is to change w_{ij} so as to reduce E by *gradient descent*:

$$\Delta w_{ij} = -\frac{\partial E}{\partial w_{ij}} = 2 \sum_k (t_k - o_k) \frac{\partial o_k}{\partial w_{ij}}$$

Consider a net divided into $m + 1$ layers, with nets in layer $g + 1$ receiving all their inputs from layer g ; with layer 0 comprising the input units; and layer m comprising the output units. If i is an output unit (remember, w_{ij} connects from j to i) then the only non-zero term in the last equation has $k = i$. Now $o_k = \sum w_{il}o_l$ where $w_{il} \neq 0$ only for o_l which are outputs from the previous layer. We thus have

$$\Delta w_{ij} = 2(t_i - o_i) \frac{\partial f_i\left(\sum w_{il}o_l\right)}{\partial w_{ij}} = 2(t_i - o_i)f'_i o_j$$

where f'_i is the derivative of the activation function evaluated at the activation level $in_i = \sum w_{il}o_l$ to unit i . Thus Δw_{ij} for an output unit i is proportional to $\delta_i o_j$, where $\delta_i = (t_i - o_i)f'_i$.

Next, suppose that i is a hidden unit whose output drives only output units:

$$\Delta w_{ij} = 2 \sum_k (t_k - o_k) \frac{\partial f_k\left(\sum w_{kl}o_l\right)}{\partial w_{ij}}$$

However, the only o_l that depends on w_{ij} is o_i , and so

$$\frac{\partial f_k\left(\sum w_{kl}o_l\right)}{\partial w_{ij}} = \frac{\partial f_k\left(\sum w_{kl}o_l\right)}{\partial o_i} \frac{\partial o_i}{\partial w_{ij}} = [f'_k w_{ki}] \cdot [f'_i o_j]$$

so that $\Delta w_{ij} = 2 \sum_k (t_k - o_k) [f'_k w_{ki}] \cdot [f'_i o_j]$.

Recalling that $\delta_k = (t_k - o_k)f'_k$ for an output unit k , we may rewrite this as

$$\Delta w_{ij} = 2 \left(\sum_k \delta_k w_{ki} \right) f'_i o_j$$

Thus, Δw_{ij} is proportional to $\delta_i o_j$, with $\delta_i = (\sum_k \delta_k w_{ki})f'_i$, where k runs over all units which receive unit i 's output. More generally, we can prove the following, by induction on how many layers back we must go to reach a unit:

Proposition. Consider a layered loop-free net with error $E = \sum_k (t_k - o_k)^2$, where k ranges over designated “output units,” and let the weights w_{ij} be changed according to the gradient descent rule

$$\Delta w_{ij} = -\partial E / \partial w_{ij} = 2 \sum_k (t_k - o_k) \frac{\partial o_k}{\partial w_{ij}}$$

Then the weights may be changed inductively, working back from the output units, by the rule

$$\Delta w_{ij} \text{ is proportional to } \delta_i o_j$$

where:

Basis Step: $\delta_i = (t_i - o_i)f'_i$ for an output unit.

Induction Step: If i is a hidden unit, and if δ_k is known for all units that receive unit i 's output, then $\delta_i = (\sum_k \delta_k w_{ki})f'_i$, where k runs over all units which receive unit i 's output.

Thus the “error signal” δ_i propagates back layer by layer from the output units. In $\sum_k \delta_k w_{ki}$, unit i receives error propagated back from a unit k to the extent to which i affects k . For output units, this is essentially the *delta rule* given by Widrow and Hoff (1960) (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION).

The theorem just presented tells us how to compute Δw_{ij} for gradient descent. It does not guarantee that the above step-size is appropriate to reach the minimum, nor does it guarantee that the minimum, if reached, is global. The backpropagation rule defined by this proposition is, thus, a heuristic rule, not one guaranteed to find a global minimum, but is still perhaps the most diversely used adaptive architecture. Many other approaches to learning, including some which are “neural-like” in at best a statistical sense, rather than being embedded in adaptive neural networks, may be found in the road map **Learning in Artificial Networks** (not just neural networks).

A Cautionary Note

The previous subsections have introduced a number of techniques that can be used to make neural networks more adaptive. In a typical training scenario, we are given a network N which, in response to the presentation of any x from some set of input patterns, will eventually settle down to produce a corresponding y from the set Y of the network's output patterns. A training set is then a sequence of pairs (x_k, y_k) from $X \times Y$, $1 \leq k \leq n$. The foregoing results say that, in many cases (and the bounds are not yet well defined), if we train the net with repeated presentations of the various (x_k, y_k) , it will converge to a set of connections which cause N to compute a function $f: X \rightarrow Y$ with the property that, over the set of k 's from 1 to n , the $f(x_k)$ “correlate fairly well” with the y_k . Of course, there are many other functions $g: X \rightarrow Y$ such that the $g(x_k)$ “correlate fairly well” with the y_k , and they may differ wildly on those “tests” x in X that do not equal an x_k in the training set. The view that one may simply present a trainable net with a few examples of solved problems, and it will then adjust its connections to be able to solve all problems of a given class, glosses over three main issues:

- Complexity:* Is the network complex enough to encode a solution method?

- (b) *Practicality*: Can the net achieve such a solution within a feasible period of time? and
- (c) *Efficacy*: How do we guarantee that the generalization achieved by the machine matches our conception of a useful solution?

Part III provides many “snapshots” of the research underway to develop answers to these problems (for the “state of play” see, for example, LEARNING AND GENERALIZATION: THEORETICAL BOUNDS; PAC LEARNING AND NEURAL NETWORKS; and VAPNIK-CHERVONENKIS DIMENSION OF NEURAL NETWORKS). Nonetheless, it is clear that these training techniques will work best when training is based on an adaptive architecture and an initial set of weights appropriate to the given problem. Future work on the neurally inspired design of intelligent systems will involve many domain-specific techniques for system design, such as those exemplified in the road maps **Vision** and **Robotics and Control Theory**, as well as general advances in adaptive architectures.

Envoi

With this, our tour of some of those basic landmarks of *Brain Theory and Neural Networks* established by 1986 is complete. I now invite each reader to follow the suggestions of the section “How to Use this Book” of the *Handbook* to begin exploring the riches of Part III, possibly with the guidance of a number of the road maps in Part II.

Acknowledgments. All of Part I is a lightly edited version of Part I as it appeared in the first edition of the *Handbook*. Section I.1 is based in large part on material contained in Section 2.3 of Arbib (1989), while Section I.3 is based on Sections 3.4 and 8.2.

References

- Amari, S., and Arbib, M. A., 1977, Competition and cooperation in neural nets, in *Systems Neuroscience* (J. Metzler, Ed.), New York: Academic Press, pp. 119–165.
- Arbib, M. A., 1981, Perceptual structures and distributed motor control, in *Handbook of Physiology—The Nervous System*, vol. II, *Motor Control* (V. B. Brooks, Ed.), Bethesda, MD: American Physiological Society, pp. 1449–1480.
- Arbib, M. A., 1987, *Brains, Machines, and Mathematics*, 2nd ed., New York: Springer-Verlag.
- Arbib, M. A., 1989, *The Metaphorical Brain 2: Neural Networks and Beyond*, New York: Wiley-Interscience.
- Arbib, M. A., Erdi, P., and Szentágothai, J., 1998, *Neural Organization: Structure, Function, and Dynamics*, Cambridge, MA: MIT Press.
- Arbib, M. A., and Hesse, M. B., 1986, *The Construction of Reality*, New York: Cambridge University Press.
- Bain, A., 1868, *The Senses and the Intellect*, 3rd ed.
- Bernard, C., 1878, *Leçons sur les phénomènes de la Vie*.
- Brooks, R. A., 1986, A robust layered control system for a mobile robot, *IEEE Robot. Automat.*, RA-2:14–23.
- Cannon, W. B., 1939, *The Wisdom of the Body*, New York: Norton.
- Chomsky, N., 1959, On certain formal properties of grammars, *Inform. Control*, 2:137–167.
- Church, A., 1941, *The Calculi of Lambda-Conversion*, Annals of Mathematics Studies 6, Princeton, NJ: Princeton University Press.
- Craik, K. J. W., 1943, *The Nature of Explanation*, New York: Cambridge University Press.
- Ewert, J.-P., and von Seelen, W., 1974, Neurobiologie und System-Theorie eines visuellen Muster-Erkennungsmechanismus bei Kroten, *Kybernetik*, 14:167–183.
- Fearing, F., 1930, *Reflex Action*, Baltimore: Williams and Wilkins.
- Gödel, K., 1931, Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme: I, *Monats. Math. Phys.*, 38:173–198.
- Grossberg, S., 1967, Nonlinear difference-differential equations in prediction and learning theory, *Proc. Natl. Acad. Sci. USA*, 58:1329–1334.
- Hebb, D. O., 1949, *The Organization of Behavior*, New York: Wiley.
- Heims, S. J., 1991, *The Cybernetics Group*, Cambridge, MA: MIT Press.
- Hodgkin, A. L., and Huxley, A. F., 1952, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol. Lond.*, 117:500–544.
- Hopfield, J., 1982, Neural networks and physical systems with emergent collective computational properties, *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.
- Hopfield, J. J., and Tank, D. W., 1986, Neural computation of decisions in optimization problems, *Biol. Cybern.*, 52:141–152.
- Kleene, S. C., 1936, General recursive functions of natural numbers, *Math. Ann.*, 112:727–742.
- La Mettrie, J., 1953, *Man a Machine* (trans. by G. Bussey from the French original of 1748), La Salle, IL: Open Court.
- Lichtheim, L., 1885, On aphasia, *Brain*, 7:433–484.
- Maxwell, J. C., 1868, On governors, *Proc. R. Soc. Lond.*, 16:270–283.
- McCulloch, W. S., and Pitts, W. H., 1943, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.*, 5:115–133.
- Minsky, M. L., 1961, Steps toward artificial intelligence, *Proc. IRE*, 49:8–30.
- Minsky, M. L., 1985, *The Society of Mind*, New York: Simon and Schuster.
- Minsky, M. L., and Papert, S., 1969, *Perceptrons: An Essay in Computational Geometry*, Cambridge, MA: MIT Press.
- Nilsson, N., 1965, *Learning Machines*, New York: McGraw-Hill.
- Pavlov, I. P., 1927, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (translated from the Russian by G. V. Anrep), New York: Oxford University Press.
- Post, E. L., 1943, Formal reductions of the general combinatorial decision problem, *Am. J. Math.*, 65:197–268.
- Rall, W., 1964, Theoretical significance of dendritic trees for neuronal input–output relations, in *Neural Theory and Modeling* (R. Reiss, Ed.), Stanford, CA: Stanford University Press, pp. 73–97.
- Ramón y Cajal, S., 1906, The structure and connexion of neurons, reprinted in *Nobel Lectures: Physiology or Medicine, 1901–1921*, New York: Elsevier, 1967, pp. 220–253.
- Rosenblatt, F., 1958, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.*, 65:386–408.
- Rosenbluth, A., Wiener, N., and Bigelow, J., 1943, Behavior, purpose and teleology, *Philos. Sci.*, 10:18–24.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986, Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1 (D. Rumelhart and J. McClelland, Eds.), Cambridge, MA: MIT Press/Bradford Books, pp. 318–362.
- Samuel, A. L., 1959, Some studies in machine learning using the game of checkers, *IBM J. Res. Dev.*, 3:210–229.
- Selfridge, O. G., 1959, Pandemonium: A paradigm for learning, in *Mechanisation of Thought Processes*, London: Her Majesty’s Stationery Office, pp. 511–531.
- Sherrington, C., 1906, *The Integrative Action of the Nervous System*, New York: Oxford University Press.
- Turing, A. M., 1936, On computable numbers with an application to the Entscheidungsproblem, *Proc. Lond. Math. Soc. (Series 2)*, 42:230–265.
- Turing, A. M., 1950, Computing machinery and intelligence, *Mind*, 59:433–460.
- von der Malsburg, C., 1973, Self-organization of orientation-sensitive cells in the striate cortex, *Kybernetik*, 14:85–100.
- Widrow, B., and Hoff, M. E., Jr., 1960, Adaptive switching circuits, in *1960 IRE WESCON Convention Record*, vol. 4, pp. 96–104.
- Wiener, N., 1948, *Cybernetics: Or Control and Communication in the Animal and the Machine*, New York: Technology Press and Wiley (2nd ed., Cambridge, MA: MIT Press, 1961).
- Young, R. M., 1970, *Mind, Brain and Adaptation in the Nineteenth Century: Cerebral Localization and Its Biological Context from Gall to Ferrier*, New York: Oxford University Press.

Part II: Road Maps

A Guided Tour of Brain Theory and Neural Networks

Michael A. Arbib

How to Use Part II

Part II provides a guided tour of our subject in the form of 22 **road maps**, each of which provides an overview of a single theme in brain theory and neural networks and offers a précis of Part III articles related to that theme. The road maps are grouped under eight general headings:

Grounding Models of Neurons and Networks
Brain, Behavior, and Cognition
Psychology, Linguistics, and Artificial Intelligence
Biological Neurons and Networks
Dynamics and Learning in Artificial Networks
Sensory Systems
Motor Systems
Applications, Implementations, and Analysis

Part II starts with the **meta-map** (Section II.1), which is designed to give some sense of the diversity yet interconnectedness of the themes taken up in this *Handbook* by quickly surveying the 22 different road maps. We then offer eight sections, one for each of

the above headings, that comprise the 22 road maps. In the road maps, we depart from the convention used elsewhere in this text whereby titles in capitals and small capitals are used for cross-references to all other articles. In the road maps, we reserve capitals and SMALL CAPITALS for articles on the tour, and we use titles in quotation marks to refer to related articles that are not primary to the current road map. We will use **boldface** type to refer to road maps and other major sections in Part II.

Every article in Part III occurs in at least one road map, and a few articles appear in two or even three road maps. Clearly, certain articles unequivocally have a place in a given road map, but as I considered articles that were less central to a given theme, my decisions on which articles to include became somewhat arbitrary. Thus, I invite you to read each road map to get a good overview of the main themes of each road map, and then continue your exploration by browsing Part III and using the articles listed under Related Reading and the index of the *Handbook* to add your own personal extensions to each map.

II.1. The Meta-Map

There is no one best path for the study of brain theory and neural networks, and you should use the meta-map simply to get a broad overview that will help you choose a path that is pleasing, or useful, to you.

Grounding Models of Neurons and Networks

Grounding Models of Neurons
 Grounding Models of Networks

The articles surveyed in these two road maps can be viewed as continuing the work of Part I, providing the reader with a basic understanding of the models of both biological and artificial neurons and neural networks that are developed in the 285 articles in Part III. The road maps will help each reader decide which of these articles provide necessary background for their own reading of the *Handbook*.

Brain, Behavior, and Cognition

Neuroethology and Evolution
 Mammalian Brain Regions
 Cognitive Neuroscience

The road map **Neuroethology and Evolution** places the following road map, **Mammalian Brain Regions**, in a dual perspective. First, by reviewing work on modeling neural mechanisms of the behavior of a variety of nonmammalian animals, it helps us understand the wealth of subtle neural computations available in other species, enriching our study of nervous systems that are closer to that of humans. When we focus on ethology (animal behavior), we often study the integration of perception and action, thus providing a useful complement to the many articles that focus on a subsystem in relative isolation. Second, by offering a number of articles on both biological and artificial evolution, we take the first steps in understanding the ways in which different neural architectures may emerge across many generations. Turning to the mammalian brain, we first look at **Mammalian Brain Regions**. We will also study

the role of these brain regions in other road maps as we analyze such functions as vision, memory, and motor control. We shall see that every such function involves the “cooperative computation” of a multiplicity of brain regions. However, **Mammalian Brain Regions** reviews those articles that focus on a single brain region and give some sense of how we model its contribution to key neural functions. The road map **Cognitive Neuroscience** then pays special attention to a range of human cognitive functions, including perception, action, memory, and language, with emphasis on the range of data now available from imaging of the active human brain and the challenges these data provide for modeling.

Psychology, Linguistics, and Artificial Intelligence

Psychology
 Linguistics and Speech Processing
 Artificial Intelligence

Our next three road maps—**Psychology, Linguistics and Speech Processing**, and **Artificial Intelligence**—are focused more on the effort to understand human psychology than on the need to understand the details of neurobiology. For example, the articles on **Psychology** may overlap those on **Cognitive Neuroscience**, but overall the emphasis shifts to “connectionist” models in which the “neurons” rarely correspond to the actual biological neurons of the human brain (the underlying structure). Rather, the driving idea is that the functioning of the human mind (the functional expression of the brain’s activity) is best explored through a parallel, adaptive processing methodology in which large populations of elements are simultaneously active, pass messages back and forth between each other, and can change the strength of their connections as they do so. This is in contrast to the serial computing methodology, which is based on the computing paradigm that was dominant from the 1940s through the 1970s and that now is complemented in mainstream computer science by work in grid-based computing, embedded systems, and teams of intelligent agents.

In short, connectionist approaches to psychology and linguistics use “neurons” that are more like the artificial neurons used to build

new applications for parallel processing than they are like the real neurons of the living brain.

In dividing this introduction to connectionism into three themes, I have first distinguished those aspects of connectionist psychology that relate to perception, memory, emotion, and other aspects of cognition in general from those specifically involved in connectionist linguistics before turning to artificial intelligence. The road map **Psychology** also contains articles that address philosophical issues in brain theory and connectionism, including the notion of consciousness, as well as articles that approach psychology from a developmental perspective. The road map **Linguistics and Speech Processing** presents connectionist models of human language performance as well as approaches (some more neural than others) to technologies for speech processing. The central idea in connectionist linguistics is that rich linguistic representations can emerge from the interaction of a relatively simple learning device and a structured linguistic environment, rather than requiring the details of grammar to be innate, captured in a genetically determined universal grammar. The road map **Artificial Intelligence** presents articles whose themes are similar to those in **Psychology** in what they explain, but are part of artificial intelligence (AI) because the attempt is to get a machine to exhibit some intelligent-like behavior, without necessarily meeting the constraints imposed by experimental psychology or psycholinguistics. “Classical” symbolic AI is contrasted with a number of methods in addition to the primary concentration on neural network approaches. The point is that, whereas brain theory seeks to know “how the brain does it,” AI must weigh the value of artificial neural networks (ANNs) as a powerful technology for parallel, adaptive computation against that of other technologies on the basis of efficacy in solving practical problems on available hardware. The reader will, of course, find a number of models that are of equal interest to psychologists and to AI researchers.

The articles gathered in these three road maps will not exhaust the scope of their subject matter, for at least two reasons. First, in addition to connectionist models of psychological phenomena, there are many biological models that embody genuine progress in relating the phenomena to known parts of the brain, perhaps even grounding a phenomenon in the behavior of identifiable classes of biological neurons. Second, while **Artificial Intelligence** will focus on broad thematic issues, a number of these also appear in applying neural networks in computer vision, speech recognition, and elsewhere using techniques elaborated in articles of the road map **Learning in Artificial Networks**.

Biological Neurons and Networks

Biological Neurons and Synapses
Neural Plasticity
Neural Coding
Biological Networks

The next four road maps, **Biological Neurons and Synapses**, **Neural Plasticity**, **Neural Coding**, and **Biological Networks**, are ones that, for many readers, may provide the appropriate entry point for the book as a whole, namely, an understanding of neural networks from a biological point of view. The road map **Biological Neurons and Synapses** gives us some sense of how sophisticated real biological neurons are, with each patch of membrane being itself a subtle electrochemical structure. An appreciation of this complexity is necessary for the computational neuroscientist wishing to address the increasingly detailed database of experimental neuroscience on how signals can be propagated, and how individual neurons interact with each other. But such complexity may also provide an eye opener for the technologist planning to incorporate new capabilities into the next generation of ANNs. The road map **Neural Plasticity** then charts from a biological point of view a variety

of specific mechanisms at the level of synapses, or even finer-grained molecular structures, which enable the changes in the strength of connections that underlie both learning and development. A number of such mechanisms have already implied a variety of learning rules for ANNs (see **Learning in Artificial Networks**), but they also include mechanisms that have not seen technological use. This road map includes articles that analyze mechanisms that underlie both development and regeneration of neural networks and learning in biological systems. However, I again stress to the reader that one may approach the road maps, and the articles in Part III of this *Handbook*, in many different orders, so that some readers may prefer to study the articles described in the road map **Learning in Artificial Networks** before or instead of studying those on neurobiological learning mechanisms.

Two more road maps round out our study of **Biological Neurons and Networks**. The simplest models of neurons either operate on a discrete-time scale or measure neural output by the continuous variation in firing rate. The road map **Neural Coding** examines the virtues of other alternatives, looking at both the possible gains in information that may follow from exploiting the exact timing of spikes (action potentials) as they travel along axonal branches from one neuron to many others, and the way in which signals that may be hard to discern from the firing of a single neuron may be reliably encoded by the activity of a whole population of neurons. We then turn to articles that chart a number of the basic architectures whereby biological neurons are combined into **Biological Networks**—although clearly, this is a topic expanded upon in many articles in Part III which are not explicitly presented in this road map.

Dynamics and Learning in Artificial Networks

Dynamic Systems
Learning in Artificial Networks
Computability and Complexity

The next three road maps—**Dynamic Systems**, **Learning in Artificial Networks**, and **Computability and Complexity**—provide a broad perspective on the dynamics of neural networks considered as general information processing structures rather than as models of a particular biological or psychological phenomenon or as solutions to specific technological problems. Our study of **Dynamic Systems** is grounded in studying the dynamics of a neural network with fixed inputs: does it settle down to an equilibrium state, and to what extent can that state be seen as the solution of some problem of optimization? Under what circumstances will the network exhibit a dynamic pattern of oscillatory behavior (a limit cycle), and under what circumstances will it undergo chaotic behavior (traversing what is known as a strange attractor)? This theme is expanded by the study of cooperative phenomena. In a gas or a magnet, we do not know the behavior of any single atom with precision, but we can infer the overall “cooperative” behavior—the pressure, volume, and temperature of a gas, or the overall magnetization of a magnet—through statistical methods, methods which even extend to the analyses of such dramatic phase transitions as that of a piece of iron from an unmagnetized lump to a magnet, or of a liquid to a gas. So, too, can statistical methods provide insight into the large-scale properties of neural nets, abstracting away from the detailed function of individual neurons, when our interest is in statistical patterns of behavior rather than the fine details of information processing. This leads us to the study of self-organization in neural networks, in which we ask for ways in which the interaction between elements in a neural network can lead to the spontaneous expression of pattern; whether this pattern is constituted by the pattern of activity of the individual neurons or by the pattern of synaptic connections which records earlier experience.

With this question of earlier experience, we have fully made the transition to the study of learning, and we turn to the road map which focuses on **Learning in Artificial Networks**, complementing the road map **Neural Plasticity**. (This replaces two road maps from the first edition—**Learning in Artificial Neural Networks, Deterministic**, and **Learning in Artificial Neural Networks, Statistical**—for two reasons: (1) the use of statistical methods in the study of learning in ANNs is so pervasive that the attempt to distinguish deterministic and statistical approaches to learning is not useful, and (2) the statistical analysis of learning in ANNs has spawned a variety of statistical methods that are less closely linked to neurobiological inspiration, and we wish these, too, to be included in our road map.) The study of **Computability and Complexity** then provides a rapprochement between neural networks and a number of ideas developed within the mainstream of computer science, especially those arising from the study of complexity of computational structures. Indeed, it takes us back to the very foundations of the theory of neural networks, in which the study of McCulloch-Pitts neurons built on earlier work on computability to inspire the later development of automata theory.

Sensory Systems

Vision

Other Sensory Systems

Vision has been the most widely studied of all sensory systems, both in brain theory and in applications and analysis of ANNs, and thus has a special road map of its own. **Other Sensory Systems**, treated at less length in the next road map, include audition, touch, and pain, as well a number of fascinating special systems such as electrolocation in electric fish and echolocation in bats.

Motor Systems

Robotics and Control Theory

Motor Pattern Generators

Mammalian Motor Control

The next set of road maps—**Robotics and Control Theory, Motor Pattern Generators**, and **Mammalian Motor Control**—addresses the control of movement by neural networks. In the study of **Robotics and Control Theory**, the adaptive properties of neural networks play a special role, enabling a control system, through experience, to become better and better suited to solve a given repertoire of control problems, guiding a system through a desired trajectory, whether through the use of feedback or feedforward. These general control strategies are exemplified in a number of different approaches to robot control. The articles in the road map **Motor Pattern Generators** focus on subconscious functions, such as breathing or locomotion, in vertebrates and on a wide variety of

pattern-generating activity in invertebrates. The reader may wish to turn back to the road map **Neuroethology and Evolution** for other studies in animal behavior (neuroethology) which show how sensory input, especially visual input, and motor behavior are integrated in a cycle of action and perception. **Mammalian Motor Control** places increased emphasis on the interaction between neural control and the kinematics or dynamics of limbs and eyes, and also looks at various forms of motor-related learning. In showing how the goals of movement can be achieved by a neural network through the time course of activity of motors or muscles, this road map overlaps some of the issues taken up in the more applications-oriented road map, **Robotics and Control Theory**. Much of the material on biological motor control is of general relevance, but the road map also includes articles on primate motor control that examine a variety of movements of the eyes, head, arm, and hand which are studied in a variety of mammals but are most fully expressed in primates and humans. Of course, as many readers will be prepared to notice by now, **Mammalian Motor Control** will, for some readers, be an excellent starting place for their study, since, by showing how visual and motor systems are integrated in a number of primate and human behaviors, it motivates the study of the specific neural network mechanisms required to achieve these behaviors.

Applications, Implementations, and Analysis

Applications

Implementation and Analysis

We then turn to a small set of **Applications** of neural networks, which include signal processing, speech recognition, and visual processing (but exclude the broader set of applications to astronomy, speech recognition, high-energy physics, steel making, telecommunications, etc., of the first edition, since *The Handbook of Neural Computation* [Oxford University Press, 1996] now provides a large set of articles on ANN applications). Since a neural network cannot be applied unless it is implemented, whether in software or hardware, we close with the road map **Implementation and Analysis**. The implementation methodologies include simulation on a general-purpose computer, emulation on specially designed neurocomputers, and implementation in a device built with electronic or photonic materials. As for analysis, we present articles in the nascent field of neuroinformatics which combines database methodology, visualization, modeling, and data analysis in an attempt to master the explosive growth of neuroscience data. (In Europe, the term *neuroinformatics* is used to encompass the full range of computational approaches to brain theory and neural networks. In the United States, some people use *neuroinformatics* to refer solely to the use of databases in neuroscience. Here we focus on the middle ground, where the analysis of data and the construction of models are brought together.)

II.2. Grounding Models of Neurons and Networks

The first two road maps expand the exposition of Part I by presenting basic models of neurons and networks that provide the building blocks for many of the articles in Part III.

Grounding Models of Neurons

AXONAL MODELING

DENDRITIC PROCESSING

HEBBIAN SYNAPTIC PLASTICITY

PERCEPTRONS, ADALINES, AND BACKPROPAGATION

PERSPECTIVE ON NEURON MODEL COMPLEXITY

REINFORCEMENT LEARNING

SINGLE-CELL MODELS

SPIKING NEURONS, COMPUTATION WITH

This road map introduces classes of neuron models of increasing complexity and attention to detail. The point is that much can be

learned even at high degrees of abstraction, while other phenomena can be understood only by attention to subtle details of neuronal function. The reader of this *Handbook* will find many articles exploring biological phenomena and technological applications at different levels of complexity. The implicit questions will always be, “Do all the details matter?” and “Is the model oversimplified?” The answers will depend both on the phenomena under question and on the current subtlety of experimental investigations. After introducing articles that present neuron models across the range of model complexity, the road map concludes with a brief look at the most widely analyzed forms of synaptic plasticity.

Classes of neuron models can be defined by how they treat the train of action potentials issued by a neuron (see the road map **Neural Coding**). Many models assume that information is carried in the average rate of pulses over a time much longer than a typical pulse width, with the occurrence times of particular pulses simply treated as jitter on an averaged analog signal. A neural model in such a theory might be a mathematical function which produces a real-valued output from its many real-valued inputs; that function could be linear or nonlinear, static or adaptive, and might be instantiated in analog silicon circuits or in digital software. Examples given of such models in **SINGLE-CELL MODELS** are the McCulloch-Pitts model, the perceptron model, Hopfield neurons, and polynomial neurons. However, some models assume that each single neural pulse carries reliable, precisely timed information. A neural model in such a theory fires only upon the exact coincidence of several input pulses, and quickly “forgets” when it last fired, so that it is always ready to fire upon another coincidence. The simplest such models are the integrate-and-fire models. The article concludes by briefly introducing the Hodgkin-Huxley model of squid axon, based on painstaking analysis (without benefit of electronic computers) of data from the squid giant axon, and then introduces modified single-point models, compartmental models, and computation with both passive dendrites and active dendrites. **SPIKING NEURONS, COMPUTATION WITH** provides more detail on those neuron models of intermediate complexity in which the output is a spike whose timing is continuously variable as a result of cellular interactions, providing a model of biological neurons that offers more details than firing rate models but without the details of biophysical models. The virtues of such models include the ability to transmit information very quickly through small temporal differences between the spikes sent out by different neurons. Information theory can be used to quantify how much more information about a stimulus can be extracted from spike trains if the precise timing is taken into account. Moreover, computing with spiking neurons may prove of benefit for technology.

AXONAL MODELING is centered on the Hodgkin and Huxley model, arguably the most successful model in all of computational neuroscience. The article shows how the Hodgkin-Huxley equations extend the cable equation to describe the ionic mechanisms underlying the initiation and propagation of action potentials. The vast majority of contemporary biophysical models use a mathematical formalism similar to that introduced by Hodgkin and Huxley, even though their model of the continuous, deterministic, and macroscopic permeability changes of the membrane was achieved without any knowledge of the underlying all-or-none, stochastic, and microscopic ionic channels. The article also describes the differences between myelinated and nonmyelinated axons; and briefly discusses the possible role of heavily branched axonal trees in information processing.

PERSPECTIVE ON NEURON MODEL COMPLEXITY then shows how this type of modeling might be extended to the whole cell. The key point is that one neuron with detailed modeling of dendrites (especially with nonuniform distributions of synapses and ion channels) can perform tasks that would require a network of many simple binary units to duplicate. The point is not to choose the most

complex model of a neuron but rather to seek an intermediate level of complexity which preserves the most significant distinctions between different “compartments” of the neuron (soma, various portions of the dendritic tree, etc.). The challenge is to demonstrate a useful computation or discrimination that can be accomplished with a particular choice of compartments in a neuron model, and then show that this useful capacity is lost when a coarser decomposition of the neuron is used. **DENDRITIC PROCESSING** especially emphasizes developments in compartmental modeling of dendrites, arguing that we are in the midst of a “dendritic revolution” that has yielded a much more fascinating picture of the electrical behavior and chemical properties of dendrites than one could have imagined only a few years ago. The dendritic membrane hosts a variety of nonlinear voltage-gated ion channels that endow dendrites with potentially powerful computing capabilities. Moreover, the classic view of dendrites as carrying information unidirectionally, from synapses to the soma, has been transformed: dendrites of many central neurons also carry information in the “backward” direction, via active propagation of the action potentials from the axon to the dendrites. These “reversed” signals can trigger plastic changes in the dendritic input synapses. Moreover, it is now known that the fine morphology as well as the electrical properties of dendrites change dynamically, in an activity-dependent manner.

If the most successful model in all of computational neuroscience is the Hodgkin-Huxley model, then the second most successful is Hebb’s model of “unsupervised” synaptic plasticity. The former was based on rigorous analysis of empirical data; the latter was initially the result of theoretical speculation on how synapses might behave if assemblies of cells were to work together to store and reconstitute thoughts and associations. **HEBBIAN SYNAPTIC PLASTICITY** notes that predictions derived from Hebb’s postulate can be generalized for different levels of integration (synaptic efficacy, functional coupling, adaptive change in behavior) by simply adjusting the variables derived from various measures of neural activity and the time-scale over which it operates. The article addresses five major issues: Should the definition of “Hebbian” plasticity refer to a simple positive correlational rule of learning, or are there biological justifications for including additional “pseudo-Hebbian” terms (such as synaptic depression due to disuse or competition) in a generalized phenomenological algorithm? What are the spatiotemporal constraints (e.g., input specificity, temporal associativity) that characterize the induction process? Do the predictions of Hebbian-based algorithms account for most forms of activity-dependent dynamics in synaptic transmission throughout phylogenesis? On which time-scales (perception, learning, epigenesis) and at which stage of development of the organism (embryonic, “critical” postnatal developmental periods, adulthood) are activity-dependent changes in functional links predicted by Hebb’s rule? Are there examples of correlation-based plasticity that contradict the predictions of Hebb’s postulate (termed anti-Hebbian modifications)? The article thus frames many important issues to be developed in the articles of the road map **Neural Plasticity** but that are also implicit, for example, in articles reviewed in the road maps **Psychology** and **Linguistics and Speech Processing**, in which Hebbian (and other) learning rules are used for “formal neurons” that are psychological abstractions rather than representation of real neurobiological neurons or even biological neuron pools. Two other articles serve to introduce the basic learning rules that have been most central in both biological analysis and connectionist modeling. Supervised learning adjusts the weights in an attempt to respond to explicit error signals provided by a “teacher,” which may be external, or another network in the same “brain.” This model was introduced in the perceptron model, which is reviewed in **PERCEPTRONS, ADALINES, AND BACKPROPAGATION** (of which more details in the next road map, **Grounding Models of Networks**). On the other hand, **REINFORCEMENT LEARNING** (of which

more details in the road map **Learning in Artificial Networks**) shows how networks can improve their performance when given general reinforcement (“that was good,” “that was bad”) by a critic, rather than the explicit error information offered by a teacher.

Grounding Models of Networks

ASSOCIATIVE NETWORKS
COMPUTING WITH ATTRACTORS
PERCEPTRONS, ADALINES, AND BACKPROPAGATION
RADIAL BASIS FUNCTION NETWORKS
SELF-ORGANIZING FEATURE MAPS
SPIKING NEURONS, COMPUTATION WITH

The mechanisms and implications of association—the linkage of information with other information—have a long history in psychology and philosophy. ASSOCIATIVE NETWORKS discusses association as realized in neural networks as well as association in the more traditional senses. Many neural networks are designed as pattern associators, which link an input pattern with the “correct” output pattern. Learning rules are designed to construct useful linkages between input and output patterns whether in feedforward neural network architectures or in a network whose units are recurrently interconnected. Special attention is given to the critical importance of data representation at all levels of network operation. PERCEPTRONS, ADALINES, AND BACKPROPAGATION introduces the perceptron rule and the LMS (least-mean-squares) algorithm for training feedforward networks with multiple adaptive elements, where each element can be seen as an adaptive linear combiner of its inputs followed by a nonlinearity which produces the output. It then presents the major extension provided by the backpropagation algorithm for training multilayer neural networks—which can be viewed as dividing the input space into regions bounded by hyperplanes, one for the thresholded output of each neuron of the output layer—and shows how this technique has been used to attack problems requiring neural networks with high degrees of nonlinearity and precision.

COMPUTING WITH ATTRACTORS shows how neural networks (often seen now as operating in continuous time) may be viewed as dynamic systems (a theme developed in great detail by the articles of the road map **Dynamic Systems**). This article describes how to compute with networks with feedback, with the input of a computation being set as an initial state for the system and the result read off a suitably chosen set of units when the network has “settled down.” The state a dynamical system settles into is called an *attractor*, so this paradigm is called *computing with attractors*. It is possible to settle down into an equilibrium state, or into periodic or even chaotic patterns of activity. (An interesting possibility, not considered in this article, is to perform computations based on the transient approach to the attractor, rather than on the basis of the attractor alone.) The Hopfield model for associative memory is used as the key example, showing its dynamic behavior as well as how the connections necessary to embed desired patterns can be

learned and how the paradigm can be extended to time-dependent attractors.

SELF-ORGANIZING FEATURE MAPS (SOFMs) introduces a famous version of competitive learning based on a layer of adaptive “neurons” that gradually develops into an array of feature detectors. The learning method is an augmented Hebbian method in which learning by the element most responsive to an input pattern is “shared” with its neighbors. The result is that the resulting “compressed image” of the (usually higher-dimensional) input space forms a “topographic map” in which distance relationships in the input space (expressing, e.g., pattern similarities) are approximately preserved as distance relationships between corresponding excitation sites in the map, while clusters of similar input patterns tend to become mapped to areas in the neural layer whose size varies in proportion to the frequency of the occurrence of their patterns. From a statistical point of view, the SOFM provides a nonlinear generalization of principal component analysis.

SPIKING NEURONS, COMPUTATION WITH discusses both the use of spiking neurons as a useful approximation to biological neurons and the study of networks of spiking neurons as a formal model of computation for which the assumptions need not be biological (see also “Integrate-and-Fire Neurons and Networks”). If the spiking neurons are not subject to significant amounts of noise, then one can carry out computations in networks of spiking neurons where every spike matters, and some finite network of spiking neurons can simulate a universal Turing machine. Spiking neurons can also be used as computational units that function like radial basis functions in the temporal domain. Another code uses the order of firing of different neurons as the relevant signal conveyed by these neurons. Firing rates of neurons in the cortex are relatively low, making it hard for the postsynaptic neuron to “read” the firing rate of a presynaptic neuron. However, networks of spiking neurons can carry out complex analog computations if the inputs of the computation are presented in terms of a space rate or population code.

The last article in this road map gives an example of the utility of studying networks in which the response properties of the individual units are designed not as abstractions from biological neurons, but rather because their response functions have mathematically desirable properties. A multilayer perceptron can be viewed as dividing the input space into regions bounded by hyperplanes, one for the thresholded output of each neuron of the output layer. RADIAL BASIS FUNCTION NETWORKS describes an alternative approach to decomposition of a pattern space into regions, describing the clusters of data points in the space as if they were generated according to an underlying probability density function. Thus the perceptron method concentrates on class boundaries, while the radial basis function approach focuses on regions where the data density is highest, constructing global approximations to functions using combinations of basis functions centered around weight vectors. The article shows that this approach not only has a range of useful theoretical properties but also is practically useful, having been applied efficiently to problems in discrimination, time-series prediction, and feature extraction.

II.3. Brain, Behavior, and Cognition

Neuroethology and Evolution

COMMAND NEURONS AND COMMAND SYSTEMS
CRUSTACEAN STOMATOGASTRIC SYSTEM
ECHOLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS
ELECTROLOCATION

EVOLUTION AND LEARNING IN NEURAL NETWORKS
EVOLUTION OF ARTIFICIAL NEURAL NETWORKS
EVOLUTION OF GENETIC NETWORKS
EVOLUTION OF THE ANCESTRAL VERTEBRATE BRAIN
HIPPOCAMPUS: SPATIAL MODELS

LANGUAGE EVOLUTION AND CHANGE
 LANGUAGE EVOLUTION: THE MIRROR SYSTEM HYPOTHESIS
 LOCOMOTION, VERTEBRATE
 LOCUST FLIGHT: COMPONENTS AND MECHANISMS IN THE MOTOR
 MOTOR PRIMITIVES
 NEUROETHOLOGY, COMPUTATIONAL
 OLFATORY CORTEX
 SCRATCH REFLEX
 SENSORIMOTOR INTERACTIONS AND CENTRAL PATTERN
 GENERATORS
 SOUND LOCALIZATION AND BINAURAL PROCESSING
 SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR
 PATTERNS
 VISUAL COURSE CONTROL IN FLIES
 VISUOMOTOR COORDINATION IN FROG AND TOAD
 VISUOMOTOR COORDINATION IN SALAMANDER

Many readers will come to the *Handbook* with one of two main motivations: to understand the human brain, or to explore the potential of ANNs as a technology for adaptive, parallel computation. The present road map emphasizes a third motivation: to study neural mechanisms in creatures very different from humans and their mammalian cousins—for the intrinsic interest of discovering the diverse neural architectures that abound in nature, for the suggestions these provide for future technology, and for the novel perspective on human brain mechanisms offered by seeking to place them in an evolutionary perspective.

Ethology is the study of animal behavior, in which our concern is with the circumstances under which a particular motor pattern will be deployed as an appropriate part of the animal's activity. Neuroethology, then, is the study of neural mechanisms underlying animal behavior. The emphasis is thus on an integrative, systems approach to the neuroscience of the animal being studied, as distinct from a reductionist approach to, for example, the neurochemistry of synaptic plasticity. Of course, a major aim of this *Handbook* is to create a context in which the reader can see both approaches to the study of nervous systems and ponder how best to integrate them. In particular, the reader will find many examples of the neuroethology of mammalian systems in a wide variety of other road maps, such as **Cognitive Neuroscience**, **Vision**, **Other Sensory Systems**, and **Mammalian Motor Control**. However, the present road map is designed to guide the reader to articles on a number of fascinating nonmammalian systems—as well as a few “exotic” mammalian systems—as a basis for a brief introduction of the evolutionary approach to biological and artificial neural networks.

NEUROETHOLOGY, COMPUTATIONAL suggests that *computational* neuroethology applies not only to animals but also to nonbiological autonomous agents, such as some types of robots and simulated embodied agents operating in virtual worlds (see also “Embodied Cognition”). The key element is the use of sophisticated computer-based simulation and visualization tools to study the neural control of behavior within the context of “agents” that are both embodied and situated within an environment. Other examples include specific neuroethological modeling directed toward specific animals (the computational frog *Rana computatrix* and the computational cockroach *Periplaneta computatrix*) and their implications for the rebirth of ideas first introduced by Grey Walter in his 1950s design of *Machina speculatrix* and later developed in the book *Vehicles* by Valentino Braitenberg.

If a certain interneuron is stimulated electrically in the brain of a marine slug, the animal then displays a species-specific escape swimming behavior, although no predator is present. If in a toad a certain portion of the optic tectum is stimulated in this manner, snapping behavior is triggered, although no prey is present. In both cases, a stimulus produces a rapid ballistic response. Such command functions provide the sensorimotor interface between sensory

pattern recognition and localization, on the one side, and motor pattern generation on the other. COMMAND NEURONS AND COMMAND SYSTEMS analyzes the extent to which a motor pattern generator (MPG) may be activated alone or in concert with others through perceptual stimuli mediated by a single “command neuron” (as in the marine slug) or by more diffuse “command systems” (as in the toad). Three articles then focus specifically on visuomotor coordination. VISUAL COURSE CONTROL IN FLIES explains the mechanisms underlying the extraction of retinal motion patterns in the fly, and their transformation into the appropriate motor activity. Rotatory large-field motion can signal unintended deviations from the fly's course and initiate a compensatory turn; image expansion can signal that the animal approaches an obstacle and initiates a landing or avoidance response; and discontinuities in the retinal motion field indicate nearby stationary or moving objects. Since many of the cells responsible for motion extraction are large and individually identifiable, the fly is quite amenable to an analysis of sensory processing. Similarly, the small number of muscles and motor neurons used to generate flight maneuvers facilitates studies of motor output. VISUOMOTOR COORDINATION IN SALAMANDER shows how low-level mechanisms add up to produce complicated behaviors, such as the devious approach of salamanders to their prey. Coarse coding models demonstrate how the location of an object may be encoded with high accuracy using only a few neurons with large, overlapping receptive fields. (This fits with the fact that the brains of salamanders are anatomically the simplest among vertebrates, containing only about 1 million neurons—frogs have up to 10 times and humans 10 million times as many neurons.) The models have been extended to the case where several objects are presented to the animal by linking a segmentation network and a winner-take-all-like object selection network to the coarse coding network in a biologically plausible way. Compensation of background movement, selection of an object, saccade generation, and approach and snapping behavior in salamanders have also been modeled successfully, in line with behavioral and neurobiological findings. Again, VISUOMOTOR COORDINATION IN FROG AND TOAD stresses that visuomotor integration implies a complex transformation of sensory data, since the same locus of retinal activation might release behavior directed toward the stimulus (as in prey catching) or toward another part of the visual field (as in predator avoidance). The article also shows how the efficacy of visual stimuli to release a response is determined by many factors, including the stimulus situation, the motivational state of the organism itself, and previous experience with the stimulus (learning and conditioning), and the physical condition of the animal's CNS (e.g., brain lesions). In addition, other types of sensory signals can modulate frogs' and toads' response to certain moving visual stimuli. For example, the efficacy of a visual stimulus may be greatly enhanced by the presence of prey odor.

MOTOR PRIMITIVES and SCRATCH REFLEX are two of the articles on nonmammalian animal behaviors that are described more fully in the road map **Motor Pattern Generators**. These articles examine the behavior elicited in frogs and turtles, respectively, by an irritant applied to the animal's skin. The former article examines the extent to which motor behaviors can be built up through a combination of a small set of basic elements; the latter emphasizes how the form of the scratch reflex changes dramatically, depending on the locus of the irritant. Other articles in the road map **Motor Pattern Generators** describe mechanisms underlying a variety of forms of locomotion (swimming, walking, flying).

SOUND LOCALIZATION AND BINAURAL PROCESSING uses data from owls, which are exquisitely skillful in using auditory signals to locate their prey, even in the dark, to anchor models which explain how information from the two ears is brought together to localize the source of a sound. The article focuses on the use of interaural time difference (ITD) as one way to estimate the azi-

muthal angle of a sound source. It describes one biological model (ITD detection in the barn owl's brainstem) and two psychological models. The underlying idea is that the brain attempts to match the sounds in the two ears by shifting one sound relative to the other, with the shift that produces the best match assumed to be the one that just balances the "real" ITD.

ECHOLLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS explores the highly specialized auditory system used by mustached bats to analyze the return signals from the biosonar pulses they emit for orientation and for hunting flying insects. Each biosonar pulse consists of a long constant-frequency (CF) component followed by a short frequency-modulated (FM) component. The CF components constitute an ideal signal for target detection and the measurement of target velocity (relative motion in a radial direction and wing beats of insects), whereas the short FM components are suited for ranging, localizing, and characterizing a target. The article shows how different parameters of echoes received by the bat carry different types of information about a target and how these may be structured in computational maps via parallel-hierarchical processing of different types of biosonar signals. These maps guide the bat's behavior. ELECTROLOCATION discusses another "exotic" sensory system related to behavior, this time the electrosensory systems of weakly electric fish. Animals with active electrosensory systems generate an electrical field around their body by means of an electrical organ located in the trunk and tail, and measure this field via electroreceptors embedded in the skin. Distortions of the electrical field due to animate or inanimate targets in the environment or signals generated by other fish provide inputs to the system, and several distinct behaviors can be linked to patterns of electrosensory input. The article focuses on progress in understanding electrolocation behavior and on the neural implementation of an adaptive filter that attenuates the effects of the fish's own movements.

We now turn to motor systems. CRUSTACEAN STOMATO-GASTRIC SYSTEM shows that work on the rhythmic motor patterns of the four areas of the crustacean stomach, the esophagus, cardiac sac, gastric mill, and pylorus, has identified four widely applicable properties. First, rhythmicity in these highly distributed networks depends on both network synaptic connectivity and slow active neuronal membrane properties. Second, modulatory influences can induce individual networks to produce multiple outputs, "switch" neurons between networks, or fuse individual networks into single larger networks. Third, modulatory neuron terminals receive network synaptic input. Modulatory inputs can be sculpted by network feedback and become integral parts of the networks they modulate. Fourth, network synaptic strengths can vary as a function of pattern cycle period and duty cycle.

The lamprey is a very primitive form of fish whose spinal cord supports a traveling wave of activity that yields the swimming movements of the animal's body, yet also persists ("fictive swimming") when the spinal cord is isolated from the body and kept alive in a dish. SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS reviews the data which ground a circuit diagram for the spinal cord circuitry, then shows how the lamprey locomotor network has been simulated. There are a number of neuromodulators present in the lamprey spinal cord that alter the output of the locomotor network. These substances, such as serotonin, dopamine, and tachykinins, offer good opportunities to test our knowledge of the locomotor system by combining the cellular and synaptic actions of the modulators into detailed network models. However, models that do not depend on details of individual cells have also proved useful in advancing our understanding of lamprey locomotion such as the control of turning. Other models probe the nature of the coupling among the rhythm generators, explaining how it may be that the speed of the head-to-tail propagation of the rhythmic activity down the spinal cord can vary with the speed of

swimming even though conduction delays in axons are fixed. LOCUST FLIGHT: COMPONENTS AND MECHANISMS IN THE MOTOR stresses that locust flight motor patterns are generated by an interactive mixture of the intrinsic properties of flight neurons, the operation of complex circuits, and phase-specific proprioceptive input. These mechanisms are subject to the concentrations of circulating neuromodulators and are also modulated according to the demands of a constantly changing sensory environment to produce adaptive behaviors. The system is flexible and able to operate despite severe ablations, and then to recover from these lesions. SENSORIMOTOR INTERACTIONS AND CENTRAL PATTERN GENERATORS analyzes basic properties of the biological systems performing sensorimotor integration. The article discusses both the impact of sensory information on central pattern generators and the less well-understood influence of motor systems on sensory activity. Interaction between motor and sensory systems is pervasive, from the first steps of sensory detection to the highest levels of processing. While there is no doubt that cortical systems contribute to sensorimotor integration, the article questions the view that motor cortex sends commands to a passively responsive spinal cord. Motor commands are only acted upon as spinal circuits integrate their intrinsic activity with all incoming information.

Turning to evolution, we find two classes of articles. We first look at those which focus on simulated evolution in ANNs, with emphasis on the role of evolution as an alternative learning mechanism to fit network parameters to yield a network better adapted to a given task. We then turn to articles more closely related to comparative and evolutionary neurobiology.

When neural networks are studied in the broader biological context of artificial life (i.e., the attempt to synthesize lifelike phenomena within computers and other artificial media), they are sometimes characterized by genotypes and viewed as members of evolving populations of networks in which genotypes are inherited from parents to offspring. EVOLUTION OF ARTIFICIAL NEURAL NETWORKS shows how ANNs can be evolved by using evolutionary algorithms (also known as genetic algorithms). An initial population of different artificial genotypes, each encoding the free parameters (e.g., the connection strengths and/or the architecture of the network and/or the learning rules) of a corresponding neural network, is created randomly. (An important challenge for future research is to study models in which the genotypes are more "biological" in nature, and less closely tied to direct description of the phenotype.) The population of networks is evaluated in order to determine the performance (fitness) of each individual network. The fittest networks are allowed to reproduce by generating copies of their genotypes, with the addition of changes introduced by genetic operators such as mutations (i.e., the random change of a few genes that are selected randomly) or crossover (i.e., the combination of parts of the genotype derived from two reproducing networks). This process is repeated for a number of generations until a network that satisfies the performance criterion set by the experimenter is obtained. LOCOMOTION, VERTEBRATE shows that the combination of neural models with biomechanical models has an important role to play in addressing the evolutionary challenge of seeing what modifications may have occurred in the locomotor circuits between the generation of traveling waves for swimming (the most ancestral vertebrates were close to the lamprey), the generation of standing waves for walking, and the generation of multiple gaits for quadruped locomotion, and on to biped locomotion. One example uses "genetic algorithms" to model the transition from a lamprey-like spinal cord that supports traveling waves to a salamander-like spinal cord that supports both traveling waves for swimming and "standing waves" for terrestrial locomotion, and then shows how vision may modulate spinal activity to yield locomotion toward a goal (see also VISUOMOTOR COORDINATION IN SALAMANDER).

EVOLUTION AND LEARNING IN NEURAL NETWORKS then extends the analysis of ANN evolution to networks that are able to adapt to the environment as a result of some form of lifetime learning. Where evolution is capable of capturing relatively slow environmental changes that might encompass several generations, learning allows an individual to adapt to environmental changes that are unpredictable at the generational level. Moreover, while evolution operates on the genotype, learning affects the phenotype, and phenotypic changes cannot directly modify the genotype. The article shows how ANNs subjected both to an evolutionary and a lifetime learning process have been studied to look at the advantages, in terms of performance, of combining two different adaptation techniques and also to help understand the role of the interaction between learning and evolution in natural organisms. Continuing this theme, LANGUAGE EVOLUTION AND CHANGE offers another style of “connectionist evolution,” placing a number of connectionist models of basic forms of language processing in an evolutionary perspective. In some cases, connectionist networks are used as simulated agents to study how social transmission via learning may give rise to the evolution of structured communication systems. In other cases, the specific properties of neural network learning are enlisted to help illuminate the constraints and processes that may have been involved in the evolution of language. The article surveys this connectionist research, starting from the emergence of early syntax, to the role of social interaction and constraints on network learning in the subsequent evolution of language, to linguistic change within existing languages.

With this we turn to the study of evolution in the sense of natural selection in biological systems, building on the insights of Charles Darwin. Since brains do not leave fossils, evolutionary work is more at the level of comparative neurobiology, looking at the nervous systems of currently extant species, then trying to build a “family tree” of possible ancestors. The idea is that we may gain deeper insights into the brains of animals of a given species if we can compare them with the brains of other species, make plausible inferences about the brain structure of their common ancestor, and then seek to relate differences between the current brains and the putative ancestral brains by relating these changes to the possible evolutionary pressures that caused each species to adapt to a specific range of environments. EVOLUTION OF THE ANCESTRAL VERTEBRATE BRAIN notes that efforts to understand how the evolving brain has adapted to specific environmental constraints are complicated because there are always several ways to implement a certain function within existing connections using molecular and cellular mechanisms. In any case, adult diversity is viewed as the outcome of divergent genetic developmental mechanisms. Thus, study of adult structures is aided by placing adult structures within their developmental history as structured by the genes that guide such development. The article introduces a possible prototype of the ancestral vertebrate brain, followed by a scenario for mechanisms that may have diversified the ancestral vertebrate brain. Evolution of the brainstem oculomotor system is used as a focal case study.

The study of gene expression patterns is playing an increasingly important role in the empirical study of brains and neurons, and the pace of innovation in this area has greatly accelerated with the publication of two maps of the human genome as well as genome maps for more and more other species. As of 2002, however, the impact of “genomic neuroscience” on computational neuroscience is still small. To help readers think about the promise of increasing this impact, we not only have the discussion in EVOLUTION OF THE ANCESTRAL VERTEBRATE BRAIN of how during development the CNS becomes polarized and then subdivides into compartments, each characterized by specific pattern of gene expression, but also a companion article, EVOLUTION OF GENETIC NETWORKS, which outlines some of the computational problems in modeling genetic

networks that can direct the establishment of a diversity of neuronal networks in the brain. Since neuronal networks are composed of a wide variety of different cell types, the final fate or end-stage of each cell type represents the outcome of a dynamic amalgamation of gene networks. Genetic networks not only determine the cell fate acquisition from the original stem cell, they also govern contact formation between the cell populations of a given neuronal network. There are intriguing parallels between the establishment and functioning of genetic networks with those of neuronal networks, which can range from simple (on-off switch) to complex. To give some sense of the complexity of organismic development, the article outlines how intracellular as well as cell-cell interactions modify the complexity of gene interactions involved in genetic networks to achieve an altered status of cell function and, ultimately, the connection alterations in the formation of neuronal networks.

OLFACTORY CORTEX describes how, during phylogeny, the paleocortex and archicortex develop in extent and complexity but retain their three-layered character, whereas neocortex emerges in mammals as a five- to six-layered structure. It stresses the evolutionary significance of the olfactory cortex and includes an account for brain theorists interested in principles of cortical organization of the early appearance of the olfactory cortex in phylogeny. Certainly, the cerebral cortex is a distinctive evolutionary feature of the mammalian brain (which does not mean that it is “better” than structures in other genera to which it may be more or less related), and the next articles give two perspectives on its structure. “Grasping Movements: Visuomotor Transformations” presents the interactions of visual areas of parietal cortex with the F5 area of premotor cortex in the monkey brain in serving the visual control of hand movements. The companion article, LANGUAGE EVOLUTION: THE MIRROR SYSTEM HYPOTHESIS, starts from the observations that monkey F5 contains a special set of “mirror neurons” active not only when the monkey performs a specific grasp, but also when the monkey sees others perform a similar task; that F5 is homologous to human Broca’s area, an area of cortex usually thought of as related to speech production; but that Broca’s area also seems to contain a mirror system for grasping. These facts are used to ground a new theory of the evolution of the human brain mechanisms that support language. It adds a neurological “missing link” to the long-held view that imitation and communication based on hand signs may have preceded the emergence of human mechanisms for extensive vocal communication. With this example to hand, the reader is invited to look through the book for articles that study specific brain mechanisms or specific behaviors in a number of species more or less related to the human. The challenge then is to chart what aspects are common to human brains and the brains more generally of primates, mammals, or even vertebrates; and then, having done so, to see what, if any, distinctive properties human brain and behavior possess. One can then seek an evolutionary account which illuminates these human capacities. For example, it is well known that the human hippocampus is crucial for the creation of episodic memories, our memories of episodes located in specific contexts of space and time (though these memories are eventually consolidated outside hippocampus). On the other hand, HIPPOCAMPUS: SPATIAL MODELS emphasizes the role of the hippocampus and related brain regions in building a map of spatial relations in the rat’s world. To what extent can we come to better understand human episodic memory by looking for the generalization from a spatial graph of the environment to one whose nodes are linked in both space and time?

Mammalian Brain Regions

AUDITORY CORTEX
AUDITORY PERIPHERY AND COCHLEAR NUCLEUS
BASAL GANGLIA

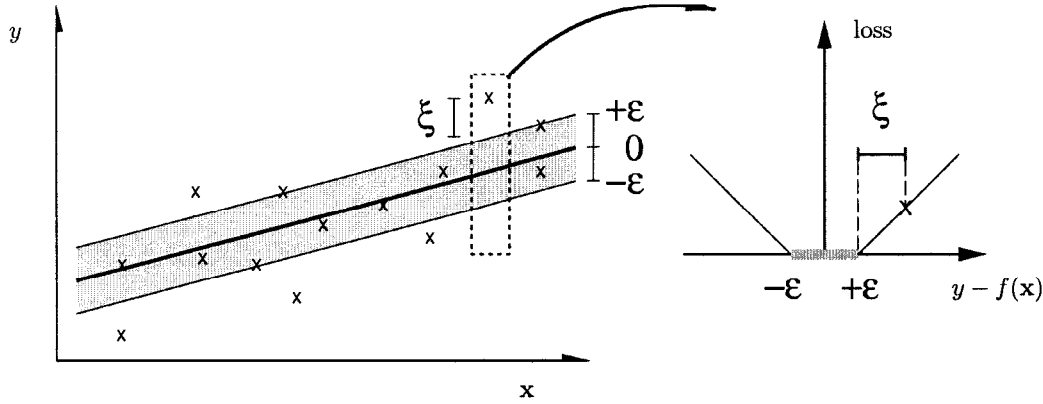


Figure 5. In support vector regression, a tube with radius ε is fitted to the data. The trade-off between model complexity and points lying outside of the tube (with positive slack variables ξ) is determined by minimizing Equa-

tion 39. (From Schölkopf, B., and Smola, A. J., 2002, *Learning with Kernels*, Cambridge, MA: MIT Press. Reprinted with permission.):

Kernel Principal Component Analysis

The kernel trick can be used to develop nonlinear generalizations of any algorithm that can be cast in terms of dot products, such as PRINCIPAL COMPONENT ANALYSIS (PCA) (q.v.).

PCA in feature space leads to an algorithm called *kernel PCA*. It is derived as follows. We wish to find eigenvectors \mathbf{v} and eigenvalues λ of the so-called *covariance matrix* \mathbf{C} in the feature space, where

$$\mathbf{C} := \frac{1}{m} \sum_{i=1}^m \Phi(x_i) \Phi(x_i)^\top \quad (43)$$

In the case when \mathcal{H} is very high-dimensional, the computational costs of doing this directly are prohibitive. Fortunately, one can show that all solutions to

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v} \quad (44)$$

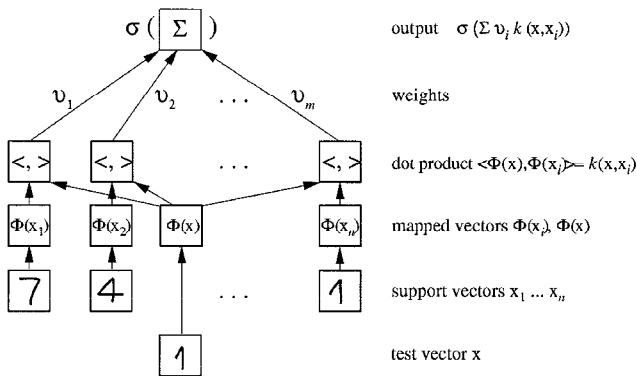


Figure 6. Architecture of SVMs and related kernel methods. The input x and the expansion patterns (SVs) x_i (we assume that we are dealing with handwritten digits) are nonlinearly mapped (by Φ) into a feature space \mathcal{H} where dot products are computed. Through the use of the kernel k , these two layers are in practice computed in one step. The results are linearly combined using weights v_i , found by solving a quadratic program (in pattern recognition, $v_i = y_i \alpha_i$; in regression estimation, $v_i = \alpha_i^* - \alpha_i$) or an eigenvalue problem (kernel PCA). The linear combination is fed into the function σ (in pattern recognition, $\sigma(x) = \text{sgn}(x + b)$; in regression estimation, $\sigma(x) = x + b$; in kernel PCA, $\sigma(x) = x$). (From Schölkopf, B., and Smola, A. J., 2002, *Learning with Kernels*, Cambridge, MA: MIT Press. Reprinted with permission.)

with $\lambda \neq 0$ must lie in the span of Φ -images of the training data. Thus, we may expand the solution \mathbf{v} as

$$\mathbf{v} = \sum_{i=1}^m \alpha_i \Phi(x_i) \quad (45)$$

thereby reducing the problem to that of finding the α_i . It turns out that this leads to a dual eigenvalue problem for the expansion coefficients,

$$m\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha} \quad (46)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ is normalized to satisfy $\|\boldsymbol{\alpha}\|^2 = 1/\lambda$.

To extract nonlinear features from a test point x , we compute the dot product between $\Phi(x)$ and the n th normalized eigenvector in feature space,

$$\langle \mathbf{v}^n, \Phi(x) \rangle = \sum_{i=1}^m \alpha_i^n k(x_i, x) \quad (47)$$

A toy example is given in Figure 7. As in the case of SVMs, the architecture can be visualized by Figure 6.

Implementation and Empirical Results

An initial weakness of SVMs was that the size of the quadratic programming problem scaled with the number of SVs. This was due to the fact that in Equation 32, the quadratic part contained at least all SVs—the common practice was to extract the SVs by going through the training data in chunks while regularly testing for the possibility that patterns initially not identified as SVs become SVs at a later stage. This procedure is referred to as *chunking*. Note that without chunking, the size of the matrix in the quadratic part of the objective function would be $m \times m$, where m is the number of all training examples.

What happens if we have a high-noise problem? In this case, many of the slack variables ξ_i become non-zero, and all the corresponding examples become SVs. For this case, decomposition algorithms were proposed, based on the observation that not only can we leave out the non-SV examples (the x_i with $\alpha_i = 0$) from the current chunk, but also some of the SVs, especially those that hit the upper boundary ($\alpha_i = C$). The chunks are usually dealt with using quadratic optimizers. Several public domain SV packages and optimizers are listed on the web page <http://www.kernel-machines.org>.

Modern SVM implementations made it possible to train on some rather large problems. Success stories include the 60,000 example

top of the layer. However, this discrepancy led to new experiments and related changes in the model which resulted in a good replication of the actual physiological data and required only feedforward excitation. The article continues by analyzing the anatomical substrates for orientation specificity and for surround modulation of visual responses, and concludes by discussing the origins of patterned anatomical connections. **VISUAL SCENE PERCEPTION** moves beyond V1 to chart the bifurcation of V1 output in monkeys and humans into a pathway that ascends to the parietal cortex (the dorsal “where/how” system involved in object location and setting of parameters for action) and a pathway that descends to inferotemporal cortex (the ventral “what” system involved in object recognition) (see also “Dissociations Between Visual Processing Modes”).

SOMATOSENSORY SYSTEM argues that the tactile stimulus representation changes from an original form (more or less isomorphic to the stimulus itself) to a completely distributed form (underlying perception) in a series of partial transformations in successive subcortical and cortical networks. At the level of primary somatosensory cortex, the neural image of the stimulus is sensitive to shape and temporal features of peripheral stimuli, rather than simply reflecting the overall intensity of local stimulation. The processing of somatosensory information is seen as modular on two different scales: macrocolumnar in terms of “segregates” such as the cortical barrels seen in rodent somatosensory cortex, each of which receives its principal input from one of the facial whiskers; and minicolumnar, with each minicolumn in a segregate receiving afferent connections from a unique subset of the thalamic neurons projecting to that segregate. The article argues that the causal factors involved in body/object interactions are represented by the pyramidal cells of somatosensory cortical areas in such a way that their ascending, lateral, and feedback connections develop an internal working model of mechanical interactions of the body with the outside world. Such an internal model can endow the somatosensory cortex with powerful interpretive and predictive capabilities that are crucial for haptic perception (i.e., tactile perception of proximal surroundings) and for control of object manipulation.

The auditory system is introduced in two articles. **AUDITORY PERIPHERY AND COCHLEAR NUCLEUS** spells out how the auditory periphery transforms a very high information rate acoustic stimulus into a series of lower information rate auditory nerve firings, with the incoming acoustic information split across hundreds of nerve fibers to avoid loss of information. The transformation involves complex mechanical-to-electrical transformations. The cochlear nucleus continues this process of parallelization by creating multiple representations of the original acoustic stimulus, with each representation presumably emphasizing different acoustic features that are fed to other brainstem structures, such as the superior olivary complex, the nuclei of the lateral lemniscus, and the inferior colliculus. These parallel pathways are believed to be specialized for the processing of different auditory features used for sound source classification and localization. From the inferior colliculus, auditory information is passed via the medial geniculate body in the thalamus to the auditory cortex. **AUDITORY CORTEX** stresses the crucial role that auditory cortex plays in the perception and localization of complex sounds. Although recent studies have expanded our knowledge of the neuroanatomical structure, the subdivisions, and the connectivities of all central auditory stages, relatively little is known about the functional organization of the central auditory system. Nevertheless, a few auditory tasks have been broadly accepted as vital for all mammals, such as sound localization, timbre recognition, and pitch perception. The article discusses a few of the functional and stimulus feature maps that have been found or postulated, and relates them to the more intuitive and better understood case of the echolocating bats (cf. “Echolocation: Cochleotopic and Computational Maps”).

The olfactory system is distinctive in that paths from periphery to cortex do not travel via a thalamic nucleus. The olfactory pathway begins with the olfactory receptor neurons in the nose, which project their axons to the olfactory bulb. The function of the olfactory bulb is to perform the initial stages of sensory processing of the olfactory signals before sending this information to the olfactory cortex. The study of the olfactory system offers prime examples of seeking a “basic circuit” that defines the irreducible minimum of neural components necessary for a model of the functions carried out by a region. **OLFACTORY BULB** offers examples of information processing without impulses and of output functions of dendrites (dendrodendritic synapses). The olfactory cortex is defined as the region of the cerebral cortex that receives direct connections from the olfactory bulb and is subdivided into several areas that are distinct in terms of details of cell types, lamination, and sites of output to the rest of the brain. The main area involved in olfactory perception is the piriform (also called prepyriform) cortex, which projects to the mediodorsal thalamus, which in turn projects to the frontal neocortex. This is often regarded as the main olfactory cortex, and is the subject of the article **OLFACTORY CORTEX**. Olfactory cortex is the earliest cortical region to differentiate in the evolution of the vertebrate forebrain and is the only region within the forebrain to receive direct sensory input. Models of olfactory cortex emphasize the importance of cortical dynamics, including the interactions of intrinsic excitatory and inhibitory circuits and the role of oscillatory potentials in the computations performed by the cortex.

We now introduce motor cortex, then turn to three systems related to motor control and to visuomotor coordination in mammals (cf. the road map **Mammalian Motor Control**): cortical areas involved in grasping, the basal ganglia, and cerebellum. **MOTOR CORTEX: CODING AND DECODING OF DIRECTIONAL OPERATIONS** spells out the relation between the direction of reaching and changes in neuronal activity that have been established for several brain areas, including the motor cortex. The cells involved each have a broad tuning function, the peak of which is viewed as the “preferred” direction of the cell. A movement in a particular direction will engage a whole population of cells. It is found that the weighted vector sum of their neuronal preferences is a “population vector” which points in (close to) the direction of the movement for discrete movements in 2D and 3D space. **GRASPING MOVEMENTS: VISUOMOTOR TRANSFORMATIONS** shows the tight coupling between (specific subregions of) parietal and premotor cortex in controlling grasping. The AIP region of inferior parietal lobe appears to play a fundamental role in extracting intrinsic visual properties (“affordances”) from the object for organizing grasping movements. The extracted visual information is then sent to the F5 region of premotor cortex, there activating neurons that code grip types congruent to the size, shape, and orientation of the object. In addition to visually activated neurons in AIP, there are AIP cells whose activity is linked to motor activity, possibly reflecting corollary discharges sent by F5 back to the parietal cortex. (For the possible relation of grasping to language, and the homology between F5 and Broca’s area, see “Language Evolution: The Mirror System Hypothesis.”)

The basal ganglia include the striatum, the globus pallidus, the substantia nigra, and the subthalamic nucleus. **BASAL GANGLIA** stresses that all of these structures are functionally subdivided into skeletomotor, oculomotor, associative, and limbic territories. The basal ganglia can be viewed as a family of loops, each taking its origin from a particular set of functionally related cortical fields, passing through the functionally corresponding portions of the basal ganglia, and returning to parts of those same cortical fields by way of specific zones in the dorsal thalamus. The article reviews models of the basal ganglia that attempt to incorporate appropriate anatomical or physiological data, but not those that use only generic

neural network architectures. Some models work at a comparatively low level of detail (membrane properties of individual neurons and microanatomical features) and restrict themselves to a single component of the basal ganglia nucleus; others work at the system level with the basal ganglia as a whole and with their interactions with related structures (e.g., thalamus and cortex). Since dopamine neurons discharge in relation to conditions involving the probability and imminence of behavioral reinforcement, dopamine neurons have been seen as playing a role in striatal information processing analogous to that of an "adaptive critic" in connectionist networks (cf. "Reinforcement Learning" and "Dopamine, Roles of").

The division of function between cerebellum and basal ganglia remains controversial. One view is that the basal ganglia play a role in determining when to initiate one phase of movement or another, and that the cerebellum adjusts the metrics of movement, tuning different movements and coordinating them into a graceful whole. *CEREBELLUM AND MOTOR CONTROL* reviews a number of models for cerebellar mechanisms underlying the learning of motor skills. Cerebellum can be decomposed into cerebellar nuclei and a cerebellar cortex. The only output cells of the cerebellar cortex are the Purkinje cells, and their only effect is to provide varying levels of inhibition on the cerebellar nuclei. Each Purkinje cell receives two types of input—a single climbing fiber, and many tens of thousands of parallel fibers. The most influential model of cerebellar cortex has been the Marr-Albus model of the formation of associative memories between particular patterns on parallel fiber inputs and Purkinje cell outputs, with the climbing fiber acting as "training signal." Later models place more emphasis on the relation between the cortex and nuclei, and on the way in which the subregions of this coupled cerebellar system can adapt and coordinate the activity of specific motor pattern generators. The plasticity of the cerebellum is approached from a different direction in *CEREBELLUM AND CONDITIONING*. Many experiments indicate that the cerebellum is involved in learning and performance of classically conditioned reflexes. The article reviews a number of models of the role of cerebellum in rabbit eye blink conditioning, providing a useful complement to models of the role of cerebellum in motor control.

The hippocampus has been implicated in a variety of memory functions, both as working memory and as basis for long-term memory. It was also the site for the discovery of long-term potentiation (LTP) in synapses (see "Hebbian Synaptic Plasticity"). Structurally, hippocampus is the simplest form of cortex. It contains one projection cell type, whose cell bodies are confined to a single layer, and receives inputs from all sensory systems and association areas. *HIPPOCAMPUS: SPATIAL MODELS* builds on the finding that single-unit recordings in freely moving rats have revealed "place cells" in subfields of the hippocampus whose firing is restricted to small portions of the rat's environment (the corresponding "place fields"). These data underlie the seminal idea of the hippocampus as a spatial map (cf. "Cognitive Maps"). The article reviews the data and describes some models of hippocampal place cells and of their role in circuits controlling the rat's navigation through its environment. *HIPPOCAMPAL RHYTHM GENERATION* provides data and models on theta and other rhythms as well as epileptic discharges, and also introduces the key cell types of the hippocampus and a number of interconnections between the hippocampus that seem to play a key role in the generation of these patterns of activity.

Finally, we turn to prefrontal cortex, the association cortex of the frontal lobes. It is one of the cortical regions to develop last and most in the course of both primate evolution and individual ontogeny. *PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION* suggests that the late morphological development of this cortex in both cases is related to its support of higher cognitive functions involving the capacity to execute novel and complex ac-

tions, which reaches its maximum in the adult human brain. The lateral region of prefrontal cortex is involved in the representation and temporal organization of sequential behavior. This article emphasizes the physiological functions of the lateral prefrontal cortex in the temporal organization of behavior. Temporal integration of sensory and motor information, through active short-term memory (working memory) and prospective set, supports the goal-directed performance of the perception-action cycle. This role extends to the temporal organization of higher cognitive operations, including, in the human, language and reasoning.

Cognitive Neuroscience

CORTICAL MEMORY

COVARIANCE STRUCTURAL EQUATION MODELING

EEG AND MEG ANALYSIS

EMOTIONAL CIRCUITS

EVENT-RELATED POTENTIALS

HEMISPHERIC INTERACTIONS AND SPECIALIZATION

IMAGING THE GRAMMATICAL BRAIN

IMAGING THE MOTOR BRAIN

IMAGING THE VISUAL BRAIN

IMITATION

LESIONED NETWORKS AS MODELS OF NEUROPSYCHOLOGICAL DEFICITS

NEUROLINGUISTICS

NEUROLOGICAL AND PSYCHIATRIC DISORDERS

NEUROPSYCHOLOGICAL IMPAIRMENTS

PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION SEQUENCE LEARNING

STATISTICAL PARAMETRIC MAPPING OF CORTICAL ACTIVITY PATTERNS

SYNTHETIC FUNCTIONAL BRAIN MAPPING

Cognitive neuroscience has been boosted tremendously in the last decade by the rapid development and increasing use of techniques to image the active human brain. The road map thus starts with several articles on ways of observing activity in the human brain and then examines various human cognitive functions.

The organization of large masses of neurons into synchronized waves of activity lies at the basis of phenomena such as the electroencephalogram (EEG) and evoked potentials, as well as the magnetoencephalogram (MEG). The EEG consists of the electrical activity of relatively large neuronal populations that can be recorded from the scalp, while the MEG can be recorded using very sensitive transducers arranged around the head. *EEG AND MEG ANALYSIS* reviews methods of quantitative analysis that have been applied to extract information from these signals, providing an indispensable tool for sleep and epilepsy research. Epilepsy is a neurological disorder characterized by the occurrence of seizures, sudden changes in neuronal activity that interfere with the normal functioning of neuronal networks, resulting in disturbances of sensory or motor activity and of the flow of consciousness. During an epileptic seizure, the neuronal network exhibits typical oscillations that usually propagate throughout the brain, involving progressively more brain systems. These oscillations are revealed in the EEG (see also "Hippocampal Rhythm Generation"). In general, the same brain sources account for the EEG and the MEG, with the reservation that the MEG reflects magnetic fields perpendicular to the skull that are caused by tangential current dipolar fields, whereas the EEG/MEG reflects both radial and tangential fields. This property can be used advantageously to disentangle radial sources lying in the convexity of cortical gyri from tangential sources lying in the sulci.

EVENT-RELATED POTENTIALS shows how cortical event-related potentials (ERPs) arise from synchronous interactions among large

numbers of participating neurons. These include dense local interactions involving excitatory pyramidal neurons and inhibitory interneurons, as well as long-range interactions mediated by axonal pathways in the white matter. Depending on the types of interaction that occur in a specific behavioral condition, cortical networks may display different states of synchrony, causing their ERPs to oscillate in different frequency bands, designated delta (0–4 Hz), theta (5–8 Hz), alpha (9–12 Hz), beta (13–30 Hz), and gamma (31–100 Hz). Depending on the location and size of the recording and reference electrodes, recorded cortical field potentials integrate neural activity over a range of spatial scales: from the intracortical local field potential (LFP) to the intracranial electrocorticogram (ECoG) to the extracranial electroencephalogram (EEG). ERP studies have shown that local cortical area networks are able to synchronize and desynchronize their activity rapidly with changes in cognitive state. When incorporated into ANNs, the result could be a metastable large-scale neural network design that recruits and excludes sub-networks according to their ability to reach consensual local patterns, with the ability to implement behavioral schemas and adapt to changing environmental conditions.

Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) provide means for seeing which brain regions are significantly more active in one task rather than another. Functional neuroimaging is generally used to make inferences about functional anatomy on the basis of evoked patterns of cortical activity. Functional anatomy involves an understanding of what each part of the brain does, and how different brain systems interact to support various sensorimotor and cognitive functions. Large-scale organization can be inferred from techniques that image the hemodynamic and metabolic sequelae of evoked neuronal responses. PET measures regional cerebral blood flow (rCBF) and fMRI measures oxygenation changes. Their spatial resolution is on the order of a few millimeters. Because PET uses radiotracers, its temporal resolution is limited to a minute or more by the half-life of the tracers employed. However, fMRI is limited only by the biophysical time constants of hemodynamic responses themselves (a few seconds).

STATISTICAL PARAMETRIC MAPPING OF CORTICAL ACTIVITY PATTERNS considers the neurobiological motivations for different designs and analyses of functional brain imaging studies, noting that the principles of *functional specialization* and *integration* serve as the motivation for most analyses. Statistical parametric mapping (SPM) is used to identify functionally specialized brain regions that respond selectively to experimental cognitive or sensorimotor changes, irrespective of changes elsewhere. SPM is a voxel-based approach (a voxel is a volume element of a 3D image) employing standard inferential statistics. SPM is a *mass-univariate* approach, in the sense that each data sequence, from every voxel, is treated as a univariate response. The massive numbers of voxels are analyzed in parallel, and dependencies among them are dealt with using random field theory (see “Markov Random Field Models in Image Processing”).

One approach to systems-level neural modeling aims at determining the network of brain regions mediating a specific cognitive task. This means finding the nodes of the network (i.e., the brain regions), and determining the task-dependent functional strengths of their interregional anatomical linkages. COVARIANCE STRUCTURAL EQUATION MODELING describes techniques applied to the correlations between PET- or fMRI-determined regional brain activities. These correlations are viewed as “functional connectivities.” They thus vary from task to task, as different patterns of excitation and inhibition are routed through the anatomical connections of these regions. Examples of questions that can be answered using this approach are: (1) As one learns a task, do the functional links between specific brain regions change their values? (2) In cases of similar performance, are the same brain networks

being used by normals and patients? The method is illustrated with studies of object and spatial vision showing cross-talk between the dorsal and ventral streams (see “Visual Scene Perception”), which implies that they need not be functionally independent. The article stresses the concept of a *neural context*, where the functional relevance of a particular region is determined by its interactions with other areas. Because the pattern of interactions with other connected areas differs from task to task, the resulting cognitive operations may vary within a single region as it engages in different tasks.

SYNTHETIC FUNCTIONAL BRAIN MAPPING analyzes ways in which models of neural networks grounded in primate neurophysiology can be used as the basis for predictions of the results of human brain imaging. This is crucial for furthering our understanding of the neural basis of behavior. Covariance structural equation modeling helps identify the nodes of the region-by-region network corresponding to a cognitive task, especially when there is little or no nonhuman data available (e.g., most language tasks). Synthetic functional brain mapping uses primate data to form hypotheses about the neural mechanisms whereby cognitive tasks are implemented in humans, with PET and fMRI data providing constraints on the possible ways in which these neural systems function. This is illustrated in relation to the mechanisms underlying saccadic eye movements and working memory.

The next three articles focus on what we are learning about vision, motor activity, and language from functional brain imaging. IMAGING THE VISUAL BRAIN addresses functional brain imaging of visual processes, with emphasis on limits in spatial and temporal resolution; constraints on subject participation; and trade-offs in experimental design. The articles focus on retinotopy, visual motion perception, visual object representation, and voluntary modulation of attention and visual imagery, emphasizing some of the areas where modeling and brain theory might be testable using current imaging tools. IMAGING THE MOTOR BRAIN shows that the behavioral form and context of a movement are important determinants of functional activity within cortical motor areas and the cerebellum, stressing that functional imaging of the human motor system requires one to study the interaction of neurological and cognitive processes with the biomechanical characteristics of the effectors. Multiple neural systems must interact to successfully perform motor tasks, encode relevant information for motor learning, and update behavioral performance in real time. The article discusses how evidence from functional imaging studies provides insight into motor automaticity as well as the role of internal models in movement. The article also discusses novel mathematical techniques that extend the scope of functional imaging experimentation.

IMAGING THE GRAMMATICAL BRAIN reviews brain imaging results that support the author’s view that linguistic rules are neurally real and form a constitutive element of the human language faculty. The focus is on linguistic combinations at the sentence level; but an analysis of cerebral representation of phonological units and of word meaning in its isolated and compositional aspects is provided as background. The study of brain mechanisms supporting language is further advanced in NEUROLINGUISTICS. Neurolinguistics began as the study of the language deficits occurring after brain injuries, and is rooted in the conceptual model of Broca’s aphasia, Wernicke’s aphasia, and other aphasic syndromes established over a hundred years ago. The article presents data and analyses for between-stage information flow, dynamics of within-stage processing, unitary representations and activation, and processing by constraint satisfaction. (For more background on these two articles, see the road map **Linguistics and Speech Processing**.)

PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION emphasizes the physiological functions of the lateral prefrontal cortex in the temporal organization of behavior, highlighting active

short-term memory (working memory) and prospective set. The two cooperate toward temporally integrating sensory and motor information by mediating cross-temporal contingencies of behavior (see also “Competitive Queuing for Planning and Serial Performance”). This temporal integration supports the goal-directed performance of the perception-action cycle. It is a role that extends to the temporal organization of higher cognitive operations, including language and reasoning in humans. **CORTICAL MEMORY** stresses that some components of memory are localized in discrete domains of cortex, while others are more widely distributed. It outlines a view of network memory in the neocortex that is supported by empirical evidence from neuropsychology, behavioral neurophysiology, and neuroimaging. Its essential features are the acquisition of memory by the formation and expansion of networks of neocortical neurons through changes in synaptic transmission; and the hierarchical organization of memory networks, with a hierarchy of networks in posterior cortex for perceptual memory and another in frontal cortex for executive memory. **SEQUENCE LEARNING** characterizes behavioral sequences in terms of their serial, temporal, and abstract structure, and analyzes the associated neural processing systems (see also “Temporal Pattern Processing”). Temporal structure is defined in terms of the durations of elements (and the possible pauses that separate them), and intuitively corresponds to the familiar notion of rhythm. Abstract structure is defined in terms of generative rules that describe relations between repeating elements within a sequence. Thus, the two sequences A-B-C-B-A-C and D-E-F-E-D-F are both generated from the same abstract structure 123-213. The article focuses on how the different dimensions of sequence structure can be encoded in neural systems, citing behavioral studies in different patient and control groups and related simulation studies. A recurrent network for manipulating abstract structural relations is implemented in a distributed network that potentially includes the perisylvian cortex in and around Broca’s area. It is argued that both transfer of sequence knowledge between domains and abstract rule representation are likely to be neurophysiological realities.

Complementing sequence learning is the study of imitation, the ability to recognize and reproduce others’ actions. Imitation is also related to fundamental capabilities for social cognition such as the recognition of conspecifics, the attribution of others’ intentions, and the ability to deceive and to manipulate others’ states of mind. **IMITATION** bridges between biology and engineering, reviewing the cognitive and neural processes behind the different forms of imitation seen in animals and showing how studies of biological processes influence the design of robot controllers and computational algorithms. Theoretical models have been proposed to, e.g., distinguish between purely associative imitation (low-level) and sequential imitation (high-level). It is argued that modeling of imitation will lead to a better understanding of the neural mechanisms at the basis of social cognition and will offer new perspectives on the evolution of animal abilities for social representation (see “Language Evolution: The Mirror System Hypothesis” for more on evolution and imitation).

EMOTIONAL CIRCUITS stresses the distinction between emotional experiences and the underlying processes that lead to emotional experiences. (See also “Motivation” for a discussion of the motivated or goal-directed behaviors that are often accompanied by emotion or affect.) The article is grounded in studies of how the brain detects and evaluates emotional stimuli and how, on the basis of such evaluations, appropriate responses are produced, treating emotion as a function that allows the organism to respond in an adaptive manner to challenges in the environment rather than being inextricably compounded with the subjective experience of emotion. The amygdala is shown to play a major role in the evaluation process. It is argued that fearful stimuli follow two main routes. The fast route involves the thalamo-amygdala pathway and responds best to simple stimulus features, while the slow route in-

volves the thalamo-cortical-amygdala pathway and carries more complex features (such as context). The expression of fear is mediated by the outputs of the amygdala to brainstem and hypothalamus, while the experience of fear involves the prefrontal cortex.

One cerebral hemisphere may perform better than the other for such diverse tasks as language, handedness, visuospatial processing, emotion and its facial expression, olfaction, and attention. Behavioral lateralization has not only been demonstrated in people, but also in rodents, birds, primates, and other animals in areas such as vocalization and motor preferences. Many anatomical, biochemical, and physiological asymmetries exist in the brain, but it is generally unclear which, if any, of these asymmetries actually contribute to hemispheric specialization. Pathways such as the corpus callosum connecting the hemispheres appear to mediate both excitatory and longer-term inhibitory interactions between the hemispheres. **HEMISPHERIC INTERACTIONS AND SPECIALIZATION** first considers models of hemispheric interactions that do not incorporate hemispheric differences, and conversely, models examining the effects of hemispheric differences that do not incorporate hemispheric interactions. It then looks in more detail at recent studies demonstrating how both hemispheric interactions and differences influence the emergence of lateralization in models where lateralization is not initially present.

As we already saw in, e.g., **NEUROLINGUISTICS**, cognitive neuropsychology uses neurological data on the performance of brain-damaged patients to constrain models of normal cognitive function. **LESIONED NETWORKS AS MODELS OF NEUROPSYCHOLOGICAL DEFICITS** surveys how connectionist techniques have been employed to model the operation and interaction of “modules” inferred from the neurological data. The advantage over “box-and-arrow” models is that removing neurons or connections in connectionist models leads to natural analogues of real brain damage. Moreover, such models let one explore the possibility that processing is actually more distributed and interactive than the older models implied. The article discusses the effects of simulated lesioning on various models, constructed either as feedforward networks or as attractor networks, paying special attention to the misleading artifacts that may arise when large brains are modeled by small ANNs. Continuing with this theme, **NEUROPSYCHOLOGICAL IMPAIRMENTS** cautions that the inferences that link a neuropsychological impairment to a particular theory in cognitive neuroscience are not as direct as one might at first assume. The brain is a distributed and highly interactive system, such that local damage to one part can unleash new modes of functioning in the remaining parts of the system. The article emphasizes neural network models of cognition and the brain that provide a framework for reasoning about the effects of local lesions in distributed, interactive systems. In many cases a model’s behavior after lesioning is somewhat counterintuitive and so can lead to very different interpretations regarding the nature of the normal system. A model of neglect dyslexia shows how an impairment in a prelexical attentional process could nevertheless show a lexicality effect. Prosopagnosia is an impairment of face recognition that can occur relatively independently of impairments in object recognition. The behavior of some prosopagnosic patients seems to suggest that that recognition and awareness depend on dissociable and distinct brain systems. However, a model of covert face recognition demonstrates how dissociation may occur without separate systems. **NEUROLOGICAL AND PSYCHIATRIC DISORDERS** shows how neural modeling may be harnessed to investigate the pathogenesis and potential treatment of brain disorders by studying the relation between the “microscopic” pathological alterations of the underlying neural networks and the “macroscopic” functional and behavioral disease manifestations that characterize the network’s function. The article reviews computational studies of the neurological disorders of Alzheimer’s disease, Parkinson’s disease, and stroke, and the psychiatric disorders of schizophrenia and affective disorders.

II.4. Psychology, Linguistics, and Artificial Intelligence

Psychology

ANALOGY-BASED REASONING AND METAPHOR
 ASSOCIATIVE NETWORKS
 COGNITIVE DEVELOPMENT
 COGNITIVE MAPS
 COGNITIVE MODELING: PSYCHOLOGY AND CONNECTIONISM
 COMPOSITIONALITY IN NEURAL SYSTEMS
 CONCEPT LEARNING
 CONDITIONING
 CONSCIOUSNESS, NEURAL MODELS OF
 DEVELOPMENTAL DISORDERS
 EMBODIED COGNITION
 EMOTIONAL CIRCUITS
 FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM
 MOTIVATION
 PHILOSOPHICAL ISSUES IN BRAIN THEORY AND CONNECTIONISM
 SCHEMA THEORY
 SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS

Much classical psychology was grounded in notions of association—of ideas, or of stimulus and response—which were well developed in the philosophy of Hume, but with roots going back as far as Aristotle. ASSOCIATIVE NETWORKS shows how these old ideas gain new power because neural networks can provide mechanisms for the formation of associations that automatically yield many further properties. One of these is that neural networks will in many cases have similar responses to similar inputs, a property that is exploited in the study of ANALOGY-BASED REASONING AND METAPHOR. Analogy and metaphor have been characterized as comparison processes that permit one domain to be seen in terms of another. Indeed, many of the advantages suggested for connectionist models—representation completion, similarity-based generalization, graceful degradation, and learning—also apply to analogy, yet analogical processing poses significant challenges for connectionist models. Analogy and metaphor involve structured pattern matching, structured pattern completion, and a focus on common *relational structure* rather than on common object descriptions. The article analyzes current connectionist models of analogy and metaphor in terms of representations and associated processes, not in terms of brain function. Challenges for future research include building analogical models that can preserve structural relations over incrementally extended analogies and that can be used as components of a broader cognitive system such as one that would perform problem solving. Indeed, people continually deal with composite structures whether they result from aggregation of symbols in a natural language into syllables, words, and sentences or aggregation of visual features into contour and regions, objects, and complete scenes. COMPOSITIONALITY IN NEURAL SYSTEMS addresses the question of what sort of neural dynamics allows composite structures to emerge, with the grouping and binding of parts into interpretable wholes. To this day it is still disputed whether ANNs are capable of adequately handling compositional data, and if so, which type of network is most suitable. Basic results have been obtained with simple recurrent networks, but some researchers argue that more complicated dynamics (see, e.g., “Synchronization, Binding and Expectancy”) or dynamics similar to classical symbolic processing mechanisms are necessary for successful modeling of compositionality. In a related vein, SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS presents the current “state of play” for Fodor and Pylyshyn’s critique of connectionist architecture. They claimed that human cog-

nitive abilities “come in clumps” (i.e., the abilities are systematically related), and that this systematic relationship does not hold in connectionist networks. The present article examines claims and counterclaims concerning the idea that learning in connectionist architectures can engender systematicity, with special attention paid to studies based on simple recurrent networks (SRNs) and recursive auto-associative memory (RAAM). The conclusion is that, for now, evidence for systematicity in such simple networks is rather limited. (One may ponder the fact that animal brains are vastly more complex than a single SRN or RAAM.)

The “units of thought” afforded by connectionist “neurons” are quite high level compared to the fine-grain computation of the myriad neurons in the human brain, and their properties may hence be closer to those of entire neural networks than to single biological neurons. Moreover, future connectionist accounts of cognition will certainly involve the coordination of connectionist modules (see, e.g., “Hybrid Connectionist/Symbolic Systems”). SCHEMA THEORY complements neuroscience’s well-established terminology for levels of structural analysis (brain region, neuron, synapse) with a functional vocabulary, a framework for analysis of behavior with no necessary commitment to hypotheses on the localization of each schema (unit of functional analysis), but which can be linked to a structural analysis whenever appropriate. The article focuses on two issues: structuring perceptual and motor schemas to provide an action-oriented account of behavior and cognition (as relevant to the roboticist as the ethologist), and how schemas describing animal behavior may be mapped to interacting regions of the brain. Schema-based modeling becomes part of neuroscience when constrained by data provided by, e.g., human brain mapping, studies of the effects of brain lesions, or neurophysiology. The resulting model may constitute an adequate explanation in itself or may provide the framework for modeling at the level of neural networks or below. Such a neural schema theory provides a functional/structural decomposition, in strong contrast to models that employ learning rules to train a single neural network to respond as specified by some training set.

Connectionism can apply many different types of ANN techniques to explain psychological phenomena, and the article COGNITIVE MODELING: PSYCHOLOGY AND CONNECTIONISM places a sample of these in perspective. The general idea is that much of psychology is better understood in terms of parallel networks of adaptive units than in terms of serial symbol processing, and that connectionism gains much of its power from using very simple units with explicit learning rules. The article points out that connectionist models of cognition can be used both to model cognitive processes and to simulate the performance of tasks and that, unlike many traditional computational models, they are not explicitly programmed by the investigator. However, important aspects of the performance of a connectionist net are controlled by the researcher, so that the achievement of a good fit to the psychological data depends both on the way in which analogs to the data are derived and on the results of “extensional programming,” such as decisions about the selection and presentation of training data. The article also notes the work of “cognitive connectionists,” whose computational experiments have demonstrated the ability of connectionist representations to provide a promisingly different account of important characteristics of cognition (compositionality and systematicity), previously assumed to be the exclusive province of the classical symbolic tradition. PHILOSOPHICAL ISSUES IN BRAIN THEORY AND CONNECTIONISM asks the following questions: (1) Do neural systems exploit classical compositional and systematic representations, distributed representations, or no representations at all? (2) How do results emerging from neuroscience help constrain

cognitive scientific models? (3) In what ways might embodiment, action, and dynamics matter for understanding the mind and the brain? There is a growing emphasis on the computational economies afforded by real-world action and the way larger structures (of agents and artifacts) both scaffold and transform the shape of individual reason. However, rather than seeing representations as opposed to interactive dynamics, the article advocates a broader vision of the inner representational resources themselves, stressing the benefits of converging influences from robotics, systems-level neuroscience, cognitive psychology, evolutionary theory, AI, and philosophical analysis. This philosophical theme is further developed in *CONSCIOUSNESS, NEURAL MODELS OF*, which reviews the basic ways in which consciousness has been defined, relevant neuropsychological data, and preliminary progress in neural modeling. Among the characteristics needed for consciousness are temporal duration, attentional focus, binding, bodily inputs, salience, past experience, and inner perspective. Brain imaging, as well as insights into single-cell activity and the effects of brain deficits, is leading to a clearer picture of the neural correlates of consciousness. The article presents a specific attention control model of the emergence of awareness in which experience of the prereflective self is identified with the corollary discharge of the attention movement control signal. This signal is posited to reside briefly in its buffer until the arrival of the associated attended input activation at its own buffer. The article concludes by reviewing other neural models of consciousness.

Much of the early work on ANNs was inspired by the problem of "Pattern Recognition" (q.v.). *CONCEPT LEARNING* provides a general introduction to recent work, placing such ideas in a psychological perspective. Concepts are mental representations of kinds of objects, events, or ideas. The article focuses on learning mental representations of new concepts from experience and how mental representations of concepts are used to make categorization decisions and other kinds of judgments. The article reviews five types of concept learning models: rule models, prototype models, exemplar models, mixed models, and neuroscientific models. The mechanisms discussed briefly here are developed at greater length in many articles in the road map *Learning in Artificial Networks*. The psychology of concept learning receives special application in the study of *FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM*, which relates connectionist approaches to face recognition to psychological theories for the subtasks of representing faces and retrieving them from memory, comparing human and model performance along these dimensions.

Many of the concepts of connectionist psychology are strongly related to work in behaviorism, but neural networks provide a stronger "internal structure" than stimulus-response probabilities. Connectionist research has enriched a number of concepts that seemed "anticognitive" by embedding them in mechanisms, namely, neural nets, which can both support internal states and yield stimulus-response pairs as part of a general input-output map. This is shown in *CONDITIONING*. During conditioning, animals modify their behavior as a consequence of their experience of the contingencies between environmental events. This article presents formal theories and neural network models that have been proposed to describe classical and operant conditioning. During *classical conditioning*, animals change their behavior as a result of the contingencies between the conditioned stimulus (CS) and the unconditioned stimulus (US). Contingencies may vary from very simple to extremely complex ones. For example, in Pavlov's proverbial experiment, dogs were exposed to the sound of a bell (CS) followed by food (US). At the beginning of training, animals salivated (generated an unconditioned response, UR) only when the US was presented. With an increasing number of CS-US pairings, CS presentations elicited a conditioned response (CR). The article discusses variations in the effectiveness of the CS, the US, and the CS and

US together, as well as attentional models. During *operant (or instrumental) conditioning*, animals change their behavior as a result of a triple contingency between its responses (R), discriminative stimuli (SD), and the reinforcer (US). Animals are exposed to the US in a relatively close temporal relationship with the SD and R. As in "Reinforcement Learning" (q.v.), during operant conditioning animals learn by trial and error from feedback that evaluates their behavior but does not indicate the correct behavior. The article discusses positive reinforcement and negative reinforcement. Such ideas are further developed in *COGNITIVE MAPS*. Tolman introduced the notion of a *cognitive map* to explain animals' capacity for place learning, latent learning, detours, and shortcuts. In some models, Tolman's vicarious trial-and-error behavior has been regarded as reflecting the animal's comparison of different expectancies: at choice points, animals make a decision after sampling the intensity of the activation elicited by the various alternative paths. Other models still use Tolman's stimulus-approach view and assume that animals approach the place with the strongest appetitive activation, thereby performing a gradient ascent toward the goal. In addition to storing the representation of the environment in the terms of the contiguity between places, cognitive maps can store information about differences in height and the type of terrain between adjacent places, contain a priori knowledge of the space to be explored, distinguish between roads taken and those not taken, and keep track of which places have been examined. Neural networks with more than two layers can also be used to represent both the contiguity between places and the relative position of those places. Hierarchical cognitive maps can represent the environment at multiple levels. In contrast to their nonhierarchical counterparts, hierarchical maps can plan navigation in large environment, use a smaller number of connections in their networks, and have shorter decision times.

Learning in neural nets can be either supervised or unsupervised, and supervision can be in terms of a specific error signal or some general reinforcement. However, in real animals, these signals seem to have some "heat" to them, which brings us to the issues of motivation and emotion. Motivated or goal-directed behaviors are sets of motor actions that direct an animal toward a particular goal object. Interaction with the goal either promotes the survival of an individual or maintains the species. Motivated behaviors include sleep/wake, ingestive, reproductive, thermoregulatory, and aggressive/defensive behaviors (see also "Pain Networks"). They are often accompanied by emotion or affect. Given the difficulty of defining the terms *drive*, *instinct*, and *motivation* with respect to the neural substrates of behavior, *MOTIVATION* adopts a neural systems approach that discusses what and how particular parts of the brain contribute to the expression of behaviors that have a motivated character. The approach is based on Hullian incentive models of motivation, where the probability of a particular behavior depends on the integration of information from systems that control circadian timing and regulate arousal state, inputs derived from interosensory information that encode internal state (e.g., hydration state, plasma glucose, leptin, etc.), modulatory hormonal inputs such as gonadal steroids that mediate sexual behavior, and inputs derived from classic sensory modalities. *EMOTIONAL CIRCUITS* analyzes the nature of emotion, emphasizing its role in behavior rather than the subjective feelings that accompany human emotions, then examines the role of brain structures such as the amygdala, the interaction of body and cognitive states, and the status of neural modeling. The expression of fear is seen as mediated by the outputs of the amygdala to lower brain centers (brainstem, hypothalamus), while the experience of fear involves the prefrontal cortex.

Finally, we turn to development, a theme of special concern in connectionist linguistics (see the next road map). *COGNITIVE DEVELOPMENT* reviews connectionist models of the origins of knowledge, the mechanisms of change, and the task-dependent nature of

developing knowledge across a variety of domains. In each case, the models provided explicit instantiations and controlled tests of specific theories of development, and allowed the exploration of complex, emergent phenomena. However, most connectionist models are “fed” their input patterns regardless of what they output, whereas even very young children shape their environments based on how they behave. Moreover, most connectionist models are designed for and tested on a single task within a single domain, whereas children face a multitude of tasks across a range of domains each day. Capturing such features of development will require future models to take in a variety of types of information and learn how to perform successfully across a number of tasks. **DEVELOPMENTAL DISORDERS** uses the comparison of different abnormal phenotypes to explore further the modeling of the developing mind/brain. The article reviews recent examples of connectionist models of developmental disorders. Autism is a developmental disorder characterized primarily by deficits in social interaction, communication, and imagination, but also by a range of secondary deficits. One hypothesis suggests that these structural deficits are consistent with too few neurons in some brain areas, such as the cerebellum, and too many neurons in other areas, such as the amygdala and hippocampus. This grounds a simple connectionist model trained on categorization tasks linking such differences in neuro-computational constraints to some of the secondary deficits found in autism. Other models relate disordered feature maps or hidden unit numbers to higher-level cognitive deficits that characterize autism. Developmental dyslexia has been modeled by changing parameters in models of the normal processes of reading. Another model captures some features of specific language impairment, specifically the difficulty of affected patients in learning rule-based inflectional morphology in verbs, using an attractor network mapping between semantic codes and phonological codes. The article also reports new empirical findings on Williams syndrome patients which reveal a deficit in generalizing knowledge of inflectional patterns to novel forms. Alterations in the initial computational constraints of a connectionist model of past tense development are shown to account for some of the patterns seen in such data, demonstrating how different computational constraints interact in the process of development. Connectionist models thus provide a powerful tool with which to investigate the role of initial computational constraints in determining the trajectory of both typical and atypical development, ensuring that selective deficits in developmental disorders are seen in terms of the outcome of the developmental process itself.

Linguistics and Speech Processing

CONSTITUENCY AND RECURSION IN LANGUAGE
CONVOLUTIONAL NETWORKS FOR IMAGES, SPEECH, AND TIME
SERIES

HIDDEN MARKOV MODELS

IMAGING THE GRAMMATICAL BRAIN

LANGUAGE ACQUISITION

LANGUAGE EVOLUTION AND CHANGE

LANGUAGE EVOLUTION: THE MIRROR SYSTEM HYPOTHESIS

LANGUAGE PROCESSING

MOTOR THEORIES OF PERCEPTION

NEUROLINGUISTICS

OPTIMALITY THEORY IN LINGUISTICS

PAST TENSE LEARNING

READING

SPEECH PROCESSING: PSYCHOLINGUISTICS

SPEECH PRODUCTION

SPEECH RECOGNITION TECHNOLOGY

The traditional grounding of linguistics is in grammar, a systematic set of rules for structuring the sentences of a particular language.

Much modern work in linguistics has been dominated by the ideas of Noam Chomsky, who placed the notion of grammar in a mathematical framework. His ideas have gone through successive stages in which the formulation of grammars has changed radically. However, two themes have remained stable in the “generative linguistics” that has grown from his work:

- There is a *universal grammar* which defines what makes a language human, and each human language has a grammar that is simply a parametric variation of the universal grammar.
- Language is too complicated for a child to learn from scratch; instead a child has universal grammar as an innate mental capacity. When the child hears example sentences of a language, they set parameters in the universal grammar so that the child can then acquire the grammar of the particular language.

Connectionist linguistics attacks this reasoning on two fronts:

- It says that language processing is better understood in terms of connectionist processing, which, as a performance model (i.e., a model of behavior, as distinct from a competence model, which gives a static representation of a body of knowledge), can give an account of errors as well as regularities in language use.
- It notes that connectionism has powerful learning tools that Chomsky has chosen to ignore. With those tools, connectionism can model how children could acquire language on the basis of far less specific mental structures than those posited in universal grammar.

LANGUAGE PROCESSING reviews many application of connectionist modeling. Despite the insights gained into syntactic structure across languages, the formal study of language has revealed relatively little about learning and development. Thus, as we shall see later in this road map, the connectionist program for understanding language has concentrated on the process of *change*, exploring topics such as language development, language breakdown, the dynamics of representation in complex systems which themselves may be receiving changing input, and even the evolution of language. The article briefly reviews models of lexical processing (reading single words, recognizing spoken words, and word production) as well as higher-level processing. It concludes that there has been important progress in many areas of connectionist-based research into language processing, and this modeling influences both psychological and neuropsychological experimentation and observation. However, it concedes that the major debates on top-down feedback, on the capacity of connectionist models to capture the productivity and systematicity of human language, and on the degree of modularity in language processing remain to be settled.

CONSTITUENCY AND RECURSION IN LANGUAGE then provides more detail on connectionist approaches to syntax. Words group together to form coherent building blocks, *constituents*, within a sentence, so that “The girl liked a boy” decomposes into “the girl” and “liked a boy,” forming a subject noun phrase (NP) and a verb phrase (VP), respectively. In linguistics, grammar rules such as Sentence $S \rightarrow NP VP$ determine how constituents can be put together to form sentences. To capture the full generativity of human language, *recursion* needs to be introduced into the grammar. For example, if we add the rules $NP \rightarrow (det) N(PP)$ (noun with optional determiner and prepositional phrase) and $PP \rightarrow Preposition NP$, then the rules are recursive, because in this case, NP can invoke rules that eventually call for another instance of NP. This article discusses how constituency and recursion may fit into a connectionist framework, and the possible implications this may have for linguistics and psycholinguistics.

LANGUAGE ACQUISITION presents models used by developmental connectionists to support the claim that rich linguistic represen-

tations can emerge from the interaction of a relatively simple learning device and a structured linguistic environment. The article reviews connectionist models of lexical development, inflectional morphology, and syntax acquisition, stressing that these models use similar learning algorithms to solve diverse linguistic problems. PAST TENSE LEARNING then presents issues in word morphology as a backdrop for a detailed discussion of the prime debate between a rule-based and a connectionist account of language processing, over the forming of regular and irregular past tenses of verbs in English. The dual mechanism model—use the general rule “add-ed” unless an irregular past tense is found in a table of exceptions—was opposed by the view that all past tenses, even for regular verbs, are formed by a connectionist network. The article concludes that most researchers now agree that the mental processing of irregular inflections is not rule governed but rather works much like a connectionist network. Certainly, rules provide an intuitively appealing explanation for regular behavior. Indeed, people are clearly able to consciously identify regularities and describe them with explicit rules that can then be deliberately followed, but this does not imply that a neural encoding of these rules, rather than a connectionist network which yields rule-like behavior, is the better account of “mental reality.” The matter is subtle because the brain is composed of neurons. Thus the issue is not “Does the brain’s language processing use neural networks?” but whether or not the activity of those networks is best described as explicitly encoding a set of rules.

READING covers connectionist models of reading and associated processes, including the reading disorder known as dyslexia. Where a skilled reader can recognize many thousands of printed words, each in a fraction of a second, with no noticeable effort, a dyslexic child may need great effort to recognize a printed word as a particular word. Most connectionist networks for reading are models of word recognition. However, word recognition is more than an analytic letter-by-letter process that translates spelling into phonology, and so the synthetic-analytic debate provides the organizing theme for this article. The authors argue that, rather than see modeling word recognition as a distinct, separable component of reading, it may be better to investigate more integrative, nonlinear iterative network models. However, SPEECH PROCESSING: PSYCHOLINGUISTICS reviews attempts to capture psycholinguistic data using connectionist models, with the primary focus on speech segmentation and word recognition. This article analyzes how far the problem of segmenting speech into words occurs independently of word recognition; considers the interplay of connectionist models of word recognition with empirical research and theory; and assesses the gap that remains between psycholinguistic studies of speech processing and modeling of the human brain. Although data from neuropsychology and functional imaging are becoming increasingly important (see IMAGING THE GRAMMATICAL BRAIN and NEUROLINGUISTICS), the main empirical constraints on psycholinguistic models are derived from laboratory studies of human language processing that are unrelated to neural data. The article suggests that connectionist modeling helps bridge the gulf between psycholinguistics and neuroscience by employing computational models that embody at least some of the computational principles of the brain.

IMAGING THE GRAMMATICAL BRAIN notes that there is little agreement on the best way to analyze language. Contrary to the connectionist approach (see, e.g., PAST TENSE LEARNING), the author sees inventories of combinatorial rules, and stores of complex objects of several types over which these rules operate, as being at the core of language. The “language faculty,” in this view, inheres in a cerebrally represented knowledge base (rule system) and in algorithms that instantiate it. It is divided into levels for the identification and segmentation of speech sounds (universal phonetics), a system that enables the concatenation of phonetic units into se-

quences (phonology), then into words (morphology, where word structure is computed), sentences (syntax), and meaning (lexical and compositional semantics). The article reviews results emanating from brain imaging that support the neural reality of linguistic rules as a constitutive element of the human language faculty. The focus is on linguistic combinations at the sentence level, but an analysis of cerebral representation of phonological units and of word meaning in its isolated and compositional aspects is provided as background. The study of brain mechanisms supporting language is further advanced in NEUROLINGUISTICS. Neurolinguistics began as the study of the language deficits occurring after brain injuries and is rooted in the conceptual model of Broca’s aphasia, Wernicke’s aphasia, and other aphasic syndromes established over a hundred years ago. However, thanks to recent research, critical details are now seen differently, and finer details have been added. Speech and language are now recognized as the products of interacting dynamic systems, with major implications for modeling normal and abnormal performance and for understanding their neural substrates. The article analyzes between-stage information flow, dynamics of within-stage processing, unitary representations and activation, and processing by constraint satisfaction. How the cognitive elements (nodes) of psychological theorizing correspond to actual neuronal activity is not known for certain. However, the article suggests that the attractor states that can occur in recurrent networks are viable candidates for behaving as nodes. Indeed, many modeling efforts in neurolinguistics have been concerned with the consequences of relatively large-scale assumptions about stages and connections (see “Lesioned Networks as Models of Neuropsychological Deficits”).

On the output side, SPEECH PRODUCTION focuses on work in motor control, dynamical systems and neural networks, and linguistics that is critical to understanding the functional architecture and characteristics of the speech production system. The central point is that spoken word forms are not unstructured wholes but rather are composed from a limited inventory of phonological units that have no independent meaning but that can be (relatively freely) combined and organized in the construction of word forms. The production of speech by the lips, tongue, vocal folds, velum, and respiratory system can thus be understood as arising from choreographed linguistic action units. However, when phonological units are made manifest in word and sentence production, their spatiotemporal realization by the articulatory system, and consequent acoustic character presented to the auditory system, is highly variable and context dependent. The speech production system is sometimes viewed as having two components, one (traditionally referred to as phonology) concerned with categorical and linguistically contrastive information, and the other concerned with gradient, noncontrastive information (traditionally referred to as phonetics). However, current work in connectionist and dynamical systems models blurs this dichotomy. MOTOR THEORIES OF PERCEPTION reviews reasons why speech scientists have doubted the claim that the speech motor system participates in speech perception and then argues against such doubts, showing that the theory accrues credibility when it is set in the larger context of investigations of perception, action, and their coupling. The mirror neurons in primates (see LANGUAGE EVOLUTION: THE MIRROR SYSTEM HYPOTHESIS) are seen as providing an existence proof of neuronal perceptuomotor couplings. The article further argues that, although the motor theory of speech perception was motivated by requirements of speaking and listening, real-world functional perception-action coupling is central to the “design” of animals more generally.

We have already contrasted connectionism with rule-based frameworks that account for linguistic patterns through the sequential application of transformations to lexical entries. OPTIMALITY THEORY IN LINGUISTICS introduces optimality theory (OT) as

a framework for linguistic analysis that has largely supplanted rule-based frameworks within phonology; it has also been applied to syntax and semantics, though not as widely. Generation of utterances in OT involves two functions, *Gen* and *Eval*. *Gen* takes an input and returns a (possibly infinite) set of output candidates. Some candidates might be identical to the input, others modified somewhat, others unrecognizable. *Eval* chooses the candidate that best satisfies a set of ranked constraints; this optimal candidate becomes the output. The constraints can conflict, so the constraints' ranking, which differs from language to language, determines the outcome. One language might eliminate consonant clusters by deleting consonants; another might retain all input consonants. OT was partly inspired by neural networks, employing as it does the ideas of optimization, parallel evaluation, competition, and soft, conflicting constraints. OT can be implemented in a neural network with constraints that are implemented as connection weights. The network implements a Lyapunov function that maximizes "harmony" ($\sum_{ij} a_i w_{ij} a_j$; the sum, for all pairs i, j of neurons, of the product of the neurons' activations and their connection weight). Hierarchically structured representations (e.g., consonants and vowels grouped into syllables) can be represented as matrices of neurons, where each matrix is the tensor product of a vector for a linguistic unit and a vector for its position in the hierarchy.

An approach to language that emphasizes the learning processes of each new speaker rather than the existence of a set of immutable rules shared by all humans seems well equipped to approach the issue of how a language changes from generation to generation. Computational modeling has been used to test competing theories about specific aspects of language evolution under controlled circumstances. Connectionist networks have been used as simulated agents to study how social transmission via learning may give rise to the evolution of structured communication systems. In other cases, properties of neural network learning are enlisted to help illuminate the constraints and processes that may have been involved in the evolution of language. *LANGUAGE EVOLUTION AND CHANGE* surveys this connectionist research, starting from the emergence of early syntax and continuing on to the role of social interaction and constraints on network learning in subsequent evolution of language. It also discusses linguistic change within existing languages, showing how the inherent generalization ability of neural networks makes certain errors in language transmission from one generation to the next more likely than others. (However, such models say more about the simplification of grammars than about how language complexity arises in the first place.) Where this article stresses computational efficacy of various models proposed for the emergence of features characteristic of current human languages, *LANGUAGE EVOLUTION: THE MIRROR SYSTEM HYPOTHESIS* focuses on brain mechanisms shared by humans with other primates, and seeks to explain how these generic mechanisms might have become specialized during hominid evolution to support language. It is argued that imitation and pantomime provide a crucial bridging capability between general primate capabilities for action recognition and the language readiness of the human brain.

At present, the state of play may be summarized as follows: generative linguistics has shown how to provide grammatical rules that explain many subtle sentence constructions of English and many other languages, revealing commonalities and differences between languages, with the differences in some cases being reduced to very elegant and compact formulations in terms of general rules with parametric variations. However, in offering the notion of universal grammar as the substrate for language acquisition, generative linguistics ignores issues of learning that must, in any case, be faced in explaining how children acquire the large and idiosyncratic vocabulary of their native tongue. Connectionist linguistics, on the other hand, has made great strides in bringing learning to the center, not only showing how specific language skills (e.g., use of the past

tense) may be acquired, but also providing insight into psycholinguistics, the study of language behavior. However, connectionist linguistics still faces two major hurdles: it lacks the systematic overview of language provided by generative linguistics, and little progress has been made in developing a neurolinguistic theory of the contributions of specific brain regions to language capabilities. It is one thing to train an ANN to yield a convincing model of performance on the past tense; it is quite another to offer an account of how this skill interfaces with all the other aspects of language, and what neural substrates are necessary for their acquisition by the human child.

The remaining articles look at speech processing from a technological perspective rather than in relation to human psycholinguistic data. *SPEECH RECOGNITION TECHNOLOGY* introduces the way computer systems that transcribe speech waveforms into words rely on digital signal processing and statistical modeling methods to analyze and model the speech signal. Although commercial technology is typically not based on connectionist methods, neural network processing is commonly seen as a promising alternative to some of the current algorithms, and the article focuses on speech recognizers that process large-vocabulary continuous speech and that use multilayer feedforward neural networks. Traditional speech recognition systems follow a hierarchical architecture. A grammar specifies the sentences allowed by the application. (Alternatively, for very large vocabulary systems, a statistical language model may be used to define the probabilities of various word sequences in the domain of application.) Each word allowed by the grammar is listed in a dictionary that specifies its possible pronunciations in terms of sequences of phonemes which are further decomposed into smaller units whose acoustic realizations are represented by statistical acoustic models. When a speech waveform is input to a recognizer, it is first processed by a front-end unit that extracts a sequence of observations, or "features," from the raw signal. This sequence of observations is then decoded into the sequence of speech units whose acoustic models best fit the observations and that respect the constraints imposed by the dictionary and language model. Hidden Markov models (HMMs) have been an essential part of the toolkit for continuous speech recognition, as well as other complex temporal pattern recognition problems such as cursive (handwritten) text recognition, time-series prediction, and biological sequence analysis. *HIDDEN MARKOV MODELS* describes the use of deterministic and stochastic finite state automata for sequence processing, with special attention to HMMs as tools for the processing of complex piecewise stationary sequences. It also describes a few applications of ANNs to further improve these methods. HMMs allow complex learning problems to be solved by assuming that the sequential pattern can be decomposed into piecewise stationary segments, with each stationary segment parameterized in terms of a stochastic function. The HMM is called "hidden" because there is an underlying stochastic process (i.e., the sequence of states) that is not directly observable but that nonetheless affects the observed sequence of events. *CONVOLUTIONAL NETWORKS FOR IMAGES, SPEECH, AND TIME SERIES* shows how shift invariance is obtained in convolutional networks by forcing the replication of weight configurations across space. This takes the topology of the input into account, enabling such networks to force the extraction of local features by restricting the receptive fields of hidden units to be local, and enforcing a built-in invariance with respect to translations, or local distortions of the inputs.

Artificial Intelligence

ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS
BAYESIAN NETWORKS
COMPETITIVE QUEUING FOR PLANNING AND SERIAL
PERFORMANCE

COMPOSITIONALITY IN NEURAL SYSTEMS
 CONNECTIONIST AND SYMBOLIC REPRESENTATIONS
 DECISION SUPPORT SYSTEMS AND EXPERT SYSTEMS
 DYNAMIC LINK ARCHITECTURE
 GRAPHICAL MODELS: PARAMETER LEARNING
 GRAPHICAL MODELS: PROBABILISTIC INFERENCE
 GRAPHICAL MODELS: STRUCTURE LEARNING
 HYBRID CONNECTIONIST/SYMBOLIC SYSTEMS
 MEMORY-BASED REASONING
 MULTIAGENT SYSTEMS
 SCHEMA THEORY
 SEMANTIC NETWORKS
 STRUCTURED CONNECTIONIST MODELS
 SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS

In the 1950s, the precursors of today's fields of artificial intelligence and neural networks were still subsumed under the general heading of *cybernetics*. Much of the work in the 1960s sought to distance artificial intelligence (AI) from its cybernetic roots, emphasizing models of, e.g., logical inference, game playing, and problem solving that were based on explicit symbolic representations manipulated by serial computer programs. However, work in computer vision and in robotics (discussed in the road maps **Vision** and **Robotics and Control Theory**, respectively) showed that this distinction was never entirely convincing, since these were areas of AI that made use of parallel computation and numerical transformations. For a while, a case could be made that the use of parallelism might be appropriate for peripheral sensing and motor control but not for the "central" processes involved in "real" intelligence. However, work from at least the mid-1970s onward has made this fallback position untenable. For example, in the HEARSAY system, speech understanding was achieved not by serial manipulation of symbolic structures but by the action (implicitly distributed, though in the 1970s still implemented on a serial computer) of knowledge sources (what we would now call "agents") to update numerical confidence levels of multiple hypotheses distributed across a set of "levels" in a data structure known as a blackboard. MULTIAGENT SYSTEMS introduces the methodology that has grown out of such beginnings. What constitutes an "individual" can be highly subjective: an individual to one researcher can, to another, be a complex distributed system comprised of finer-grained agents. Research in brain theory has dealt with different levels, from neurons to brain regions to humans whereas AI work in multi-agent systems has focused on coarse-grained levels of individuality and interaction, where the goal is to draw upon sociological, political, and economic insights. The article is designed to survey enough of this work on multi-agent systems to foster comparisons between the ANN, brain theory, and multi-agent approaches. A crucial notion is that agents either have or learn models of the agents with which they interact. These models allow agents to avoid dealing with malicious or broken agents. Agents may even build nested models of the other agents that include an agent's models of other agents, and so on. By using their models of each other, the agents loosely organize themselves into self-reinforcing communities of trust, avoiding unproductive future interactions with other agents. In another branch of AI, work on *expert systems*—information systems that represent expert knowledge for a particular problem area as a set of rules, and that perform inferences when new data are entered—provided an important application success in which numerical confidence values played a role, but with the emphasis still on manipulation of hypotheses through the serial application of explicit rules. As shown in DECISION SUPPORT SYSTEMS AND EXPERT SYSTEMS, we now see many cases in which the application of separate rules is replaced by transformations effected in parallel by (trainable) neural net-

works. A *decision system* is either a decision support system or an expert system in the classic AI sense. The article reviews results on connectionist-based decision systems. In particular, trainable knowledge-based neural networks can be used to accumulate both knowledge (rules) and data, building adaptive decision systems with incremental, on-line learning.

As the general overview article ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS makes clear, there are many problems for which the (not necessarily serial) manipulation of symbolic structures can still outperform connectionist approaches, at least with today's software running on today's hardware. Nonetheless, if we define AI by the range of problems it is to solve—or the "packets of intelligence" it is to implement—then it is no longer useful to define it in opposition to connectionism. In general, the technologist facing a specific problem should choose between, or should combine, connectionist and symbolic approaches on the basis of efficacy, not ideology. On occasion, for rhetorical purposes, authors will use the term AI for a serial symbolic methodology distinct from connectionism. However, we will generally use it in an extended sense of a technology that seeks to realize aspects of intelligence in machines by whatever methods work best. The term *symbolic AI* will then be used for the "classical" approach. The article examines the relative merits of symbolic AI systems and neural networks, and ways of attempting to bridge between the two. In brain theory, everything, whether symbolic or not, is, in the final analysis, implemented in a neural network. But even here, an analysis of the brain will often best be conducted in terms of interacting subsystems that are not all fully explicated in neural network terms. SCHEMA THEORY complements neuroscience's well-established terminology for levels of structural analysis (brain region, neuron, synapse) with a framework for analysis of behavior with no necessary commitment to hypotheses on the localization of each schema (unit of functional analysis), but which can be linked to a structural analysis whenever appropriate. The article focuses on two issues: structuring perceptual and motor schemas to provide an action-oriented account of behavior and cognition (as relevant to the roboticist as the ethologist), and how schemas describing animal behavior may be mapped to interacting regions of the brain. Schema-based modeling becomes part of neuroscience when constrained by data provided by, e.g., human brain mapping, studies of the effects of brain lesions, or neurophysiology. The resulting model may constitute an adequate explanation in itself or may provide the framework for modeling at the level of neural networks or below. Such a neural schema theory provides a functional/structural decomposition, in strong contrast to models that employ learning rules to train a single, otherwise undifferentiated, neural network to respond as specified by some training set. HYBRID CONNECTIONIST/SYMBOLIC SYSTEMS reviews work on hybrid systems that integrate neural (ANN) and symbolic processes. Cognitive processes are not homogeneous, and so some are best captured by symbolic models and others by connectionist models. Correspondingly, from a technological viewpoint, AI systems for practical applications can benefit greatly from a proper combination of different techniques combining, e.g., symbolic models (for capturing explicit knowledge) and connectionist models (for capturing implicit knowledge).

Use of the term *systematicity* in relation to connectionist networks originated with Fodor and Pylyshyn's critique of connectionist architecture. They claimed that human cognitive abilities are systematically related in a way that does not hold in connectionist networks, unlike formal systems akin to propositional logic. SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS starts by noting that this critique made no reference to learning-based generalization, and then proceeds to examine claims and counterclaims concerning the claim that learning in connectionist architectures can engender systematicity. Special attention is paid

to studies based on simple recurrent networks (SRNs) and recursive auto-associative memory (RAAM). The article suggests that, for now, evidence for systematicity in such simple networks is rather limited. Perhaps this is not so surprising, given that there is little evidence of systematicity in most animals, and animal brains are vastly more complex than SRNs or RAAMs. Compare “Language Evolution: The Mirror System Hypothesis” for a discussion of how evolution may have shaped the human brain to extend capabilities shared with other species to yield novel human cognitive abilities.

The notion of representation plays a central role in AI. As discussed in *SEMANTIC NETWORKS*, one classic form of representation in AI is the *semantic network*, in which nodes represent concepts and links represent relations between them. Semantic networks were originally developed for couching “semantic” information, either in the psychologist’s sense of static information about concepts or in the semanticist’s sense of the meanings of natural language sentences. However, they are also used as a general knowledge representation tool. The more elaborate types of semantic networks are similar in their representational abilities to sophisticated forms of symbolic logic. The article discusses various ways of implementing or emulating semantic networks in neural networks, and of forming hybrid semantic network-neural network systems. *STRUCTURED CONNECTIONIST MODELS* emphasizes those neural networks in which the translation from symbolic to neural is fairly direct: nodes become “neurons,” but now processing is done by neural interactions rather than by an “inference engine” acting on a passive representation. At the other extreme, certain neural networks (connectionist, rather than biological) may transform input “questions” to output “answers” via the distributed activity of neurons whose firing conditions have no direct relationship to the concepts that might normally arise in a logical analysis of the problem (cf. “Past Tense Learning”). In the fully distributed version of the latter approach, each “item” (concept or mental object) is represented as a pattern of activity distributed over a common pool of nodes. However, if “John” and “Mary,” for example, are represented as patterns of activity over the entire network such that each node in the network has a specific value in the patterns for “John” and “Mary,” respectively, then how can the network represent “John” and “Mary” at the same time? To address such problems, the structured approach often employs small clusters of nodes that act as “focal” nodes for concepts and provide access to more elaborate structures that make up the detailed encoding of concepts (cf. “Localized Versus Distributed Representations”). The discussion of these varying styles of representation is continued in *CONNECTIONIST AND SYMBOLIC REPRESENTATIONS*. In symbolic representations, the heart of mathematics and many models of cognition, symbols are meaningless entities to which arbitrary significance may be assigned. Composing ordered tuples from symbols and other tuples allows us to create an infinitude of complex structures from a finite set of tokens and combination rules. Inference in the symbolic framework is founded on structural comparison and rule-governed manipulation of these objects. However, AI makes extensive use of nondeductive reasoning methods. Symbolists have moved to more complex formalizations of cognitive processes, using heuristic and unsound inference rules. Connectionists explore a radical alternative: that cognitive processes are mere epiphenomena of a completely different type of underlying system, whose operations can never be adequately formalized in symbolic language. The article examines representation and processing issues in the connectionist move from classical discrete, set-theoretic semantics to a continuous, statistical, vector-based semantics.

In symbolic AI, two concepts can be linked by providing a pointer between them. In a neural net, the problem of “binding” the two patterns of activity that represent the concepts is a more subtle one, and several models address the use of rapidly changing synaptic strengths to provide temporary “assemblages” of currently

related data. This theme is developed not only in *STRUCTURED CONNECTIONIST MODELS*, but also in the articles *COMPOSITIONALITY IN NEURAL SYSTEMS* (how can inferences about a structure be based on the way it is composed of various elements?), and “Object Structure, Visual Processing” (combining visual elements of an object into a recognizable whole). *DYNAMIC LINK ARCHITECTURE*, the basic methodology, views the brain’s data structure as a graph composed of nodes connected by links. Both units and links bear activity variables changing on the rapid time scale of fractions of a second. The nodes play the role of symbolic elements. The intensity of activity measures the degree to which a node is active in a given time interval, signifying the degree to which the meaning of the node is alive in the mind of the animal, while correlations of activity between nodes quantify the degree to which the signal of one node is related to that of others. The strength of links can change on two time scales, represented by two variables called temporary weight and permanent weight. The permanent weight corresponds to the usual synaptic weight, can change on the slow time scale of learning, and represents permanent memory. The temporary weight can change on the same time scale as the node activity—it is what makes the link dynamic. On this view, dynamic links constitute the glue by which higher data structures are built up from more elementary ones.

Complementing the theme of representation in symbolic AI has been that of planning, going from (representations of) the current state and some desired state to a sequence of operations that will transform the former to the latter. *COMPETITIVE QUEUING FOR PLANNING AND SERIAL PERFORMANCE* presents neural network studies based on two assumptions: that more than one plan representation can be simultaneously active in a planning layer, and that which plan to enact next is chosen as the most active plan representation by a competition in a second neural layer. Once a plan wins the competition and is used to initiate a response, its representation is deleted from the field of competitors in the planning layer, and the competition is re-run. This iteration allows the two-layer network to transform an initial activity distribution across plan representations into a serial performance. Such models provide a very different basis for control of serial behavior than that given by recurrent neural networks. The article suggests that such a system was probably an ancient invention in the evolution of animals yet may still serve as a viable core for the highest levels of planning and skilled sequencing exhibited by humans.

The final articles in this road map are not on neural nets per se, but instead provide related methods that add to the array of techniques extending AI beyond the serial, rule-based approach. *BAYESIAN NETWORKS* provides an explicit method for following chains of probabilistic inference such as those appropriate to expert systems, extending Bayes’s rule for updating probabilities in the light of new evidence. The nodes in a Bayesian network represent propositional variables of interest and the links represent informational or causal dependencies among the variables. The dependencies are quantified by conditional probabilities for each node given its parents in the network. The network supports the computation of the probabilities of any subset of variables given evidence about any other subset, and the reasoning processes can operate on Bayesian networks by propagating information in any direction. *GRAPHICAL MODELS: PROBABILISTIC INFERENCE* introduces the graphical models framework, which has made it possible to understand the relationships among a wide variety of network-based approaches to computation, and in particular to understand many neural network algorithms and architectures as instances of a broader probabilistic methodology. Graphical models use graphs to represent and manipulate joint probability distributions. The graph underlying a graphical model may be directed, in which case the model is often referred to as a belief network or a Bayesian network, or the graph may be undirected, in which case the model

is generally referred to as a Markov random field. The articles **GRAPHICAL MODELS: STRUCTURE LEARNING** and **GRAPHICAL MODELS: PARAMETER LEARNING** present learning algorithms that build on these inference algorithms and allow parameters and structures to be estimated from data. (A fuller précis of the three articles on graphical models can be found in the road map **Learning in Artificial Networks**.) Finally, **MEMORY-BASED REASONING** applies massively parallel computing to answer questions about a new situation by searching for data on the most similar stored instances. Memory-based reasoning (MBR) refers to a family of nearest-neighbor-like methods for making decisions or classifications. Where nearest-neighbor methods generally use a simple overlap distance metric, MBR uses variants of the value distance metric.

MBR and neural nets form decision surfaces differently, and so will perform differently. MBR can become arbitrarily accurate if large numbers of cases are available, and if these cases are well behaved and properly categorized, whereas neural nets cannot respond well to isolated cases but tend to be good at smooth extrapolation. For each article reviewed in this paragraph, the reader may ponder whether these methods are alternatives to connectionist AI, or whether they can contribute to the emergence of a technologically efficacious hybrid. As stated before, where brain theory seeks to know “how the brain does it,” AI must weigh the value of ANNs as a powerful technology for parallel, adaptive computation against that of other technologies on the basis of efficacy in solving practical problems on available hardware.

II.5. Biological Neurons and Networks

Biological Neurons and Synapses

ACTIVITY-DEPENDENT REGULATION OF NEURONAL
CONDUCTANCES
AXONAL MODELING
BIOPHYSICAL MECHANISMS IN NEURONAL MODELING
BIOPHYSICAL MOSAIC OF THE NEURON
DENDRITIC PROCESSING
DENDRITIC SPINES
DIFFUSION MODELS OF NEURON ACTIVITY
ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION
NEOCORTEX: BASIC NEURON TYPES
NEOCORTEX: CHEMICAL AND ELECTRICAL SYNAPSES
OSCILLATORY AND BURSTING PROPERTIES OF NEURONS
PERSPECTIVE ON NEURON MODEL COMPLEXITY
SINGLE-CELL MODELS
SYNAPTIC INTERACTIONS
SYNAPTIC NOISE AND CHAOS IN VERTEBRATE NEURONS
SYNAPTIC TRANSMISSION
TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES

Nearly all the articles in the road maps **Psychology, Linguistics and Speech Processing**, and **Artificial Intelligence** discuss networks made of very simple neurons describable by a single internal variable, either binary or real-valued (the “membrane potential”) and that communicate with other neurons by a simple (generally nonlinear) function of that variable, sometimes referred to as the firing rate. Incoming signals are usually summed linearly via “synaptic weights,” and these weights in turn may be adjusted by simple learning rules, such as the Hebbian rule, the perceptron rule, or a reinforcement learning rule. Such simplifications remain valuable both for technological application of ANNs and for approximate models of large biological networks. Nonetheless, biological neurons are vastly more complex than these single-compartment models suggest. An appreciation of this complexity is necessary for the computational neuroscientist wishing to address the increasingly detailed database of experimental neuroscience. It is also important for the technologist looking ahead to the incorporation of new capabilities into the next generation of ANNs.

The neocortex is functionally parcellated into vertical columns (~0.5 mm in diameter) traversing all six layers. These columns have no obvious anatomical boundaries, and the topographic mapping of afferent and efferent pathways probably determines their locations and dimensions as well as their functions. **NEOCORTEX: BASIC NEURON TYPES** shows that these apparently stereotypical microcircuits are composed of a daunting variety of precisely and

intricately interconnected neurons and argues that this neuronal diversification may provide a foundation for maximizing the computational abilities of the neocortex. All anatomical cell types can display multiple discharge patterns and molecular expression profiles. Different cell types are synaptically interconnected according to complex organizational principles to form intricate stereotypical microcircuits. The article challenges neural network modelers to incorporate and account for this cellular diversity and the role of different cells in the computational capability of cortical microcircuits. **NEOCORTEX: CHEMICAL AND ELECTRICAL SYNAPSES** summarizes the diverse functional properties of synapses in neocortex. These synapses tend to be small, but their structure and biochemistry are complex. Both chemical and electrical synapses exist in neocortex. *Chemical synapses* are the “usual synapses” of neural network models, and are far more abundant. They use a chemical neurotransmitter that is packaged presynaptically into vesicles, released in quantized (vesicle-multiple) amounts, and binds to post-synaptic receptors that either open an ion channel directly (voltage-dependent ion channels) or modulate the channel through an intracellular molecule that links the activated receptor to the opening or closing of the channel. The latter molecule is called a “second messenger,” to contrast it with the case in which the transmitter itself provides a “primary message” that acts directly on the channel, in this case called “ligand-gated.” Second-messenger-based synaptic interaction occurs on a slower time scale than ligand-gated interaction and is called *neuromodulation*, since it may modulate the behavior of the postsynaptic neuron over a time scale of seconds or minutes rather than milliseconds (cf. “Neuromodulation in Invertebrate Nervous Systems” and “Neuromodulation in Mammalian Nervous Systems”). The essential element of an *electrical synapse* is a protein called a connexin; 12 connexins form a single intercytoplasmic ion channel, and a cluster of such channels constitutes a gap junction. Electrical synapses provide a direct pathway that allows ionic current or small organic molecules to flow from the cytoplasm of one cell to that of another. Short-term dynamics allow synapses to serve as temporal filters of neural activity. Long-term synaptic plasticity provides specific, localized substrates for various forms of memory. Modulation of synaptic function by neurotransmitters (see “Neuromodulation in Mammalian Nervous Systems”) provides a mechanism for globally altering the properties of a neural circuit during changes of behavioral state. Each of these functions has diverse forms that vary between synapses, depending on their site within the cortical circuit (and elsewhere in the brain).

PERSPECTIVE ON NEURON MODEL COMPLEXITY discusses the wide range of model complexity, from very simple to rather complex neuron models. Which model to choose depends, in each case,

on the context, such as how much information we already have about the neurons under consideration and what questions we wish to answer. The use of more realistic neuron models when seeking functional insights into biological nervous systems does not mean choosing the most complex model, at least in the sense of including all known anatomical and physiological details. Rather, the key is to preserve the most significant distinctions between regions (soma, proximal dendritic, distal dendritic, etc.), using “compartmental modeling,” whereby one compartment represents each functionally distinct region. *SINGLE-CELL MODELS* starts by reviewing the “simple” models of Part I (the McCulloch-Pitts, perceptron, and Hopfield models) and the slightly more complex polynomial neuron. It then turns to more realistic biophysical models, most of which are explored in further detail in this road map. These include the Hodgkin-Huxley model of squid axon, integrate-and-fire models, modified single-point models, cable and compartmental models, and models of synaptic conductances.

Before turning to a detailed analysis of mechanisms of neuronal function, we first consider an article that offers a high-level view of the neuron, but this time a stochastic one. Most nerve cells encode their output as a series of action potentials, or spikes, that originate at or close to the cell body and propagate down the axon at constant velocity and amplitude. *DIFFUSION MODELS OF NEURON ACTIVITY* studies the membrane potential of a single neuron as engaged in a stochastic process that will eventually bring it to the threshold for spike initiation. This leads to the first-passage-time problem, inferring the distribution of neuronal spiking based on the “first passage” of the membrane potential from its resting value to threshold. In addition to using stochastic differential equations, the article shows how the Wiener and Ornstein-Uhlenbeck neuronal models can be obtained as the limit of a Markov process with discrete state spaces. Besides these models, characterized by additive noise terms appearing in the corresponding stochastic differential equations, the article also reviews diffusion models with multiplicative noise, showing that these can be used not only for the description of steady-state firing under constant stimulation, but also for effects of periodic stimulation.

Now for the details of neuronal function. The ionic mechanisms underlying the initiation and propagation of action potentials were elucidated in the squid giant axon by a number of workers, most notably Hodgkin and Huxley. Variations on the Hodgkin-Huxley equation underlie the vast majority of contemporary biophysical models. *AXONAL MODELING* describes this model and its assumptions, introduces the two classes of axons (myelinated and non-myelinated) found in most animals, and concludes by briefly commenting on the possible functions of axonal branching in information processing. The Hodgkin-Huxley equation was brilliantly inferred from detailed experiments on conduction of nerve impulses. Much research since then has revealed that the basis for these equations is provided by “channels,” structures built from a few macromolecules and embedded in the neuron which, in a voltage-dependent way, can selectively allow different ions to pass through the cell membrane to change the neuron’s membrane potential. Similarly, channels (also known in this case as receptors) in the postsynaptic membrane can respond to neurotransmitters, chemicals released from the presynaptic membrane, to change the neuron’s local membrane potential in response to presynaptic input. These changes, local to the synapse, must propagate down the dendrites and across the cell body to help determine whether or not the axon will “pass threshold” and generate an action potential. *ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION* notes that channels not only produce action potentials but can set a particular firing pattern, latency, rhythm, or oscillation for the firing of these spikes. Each neuronal class is endowed with a different set of channels, and the diversity of channels between different types of neurons explains the functional classes of neurons found in the brain. Some

neurons fire spontaneously, some show adaptation, some fire in bursts, and so on. Therefore, a channel-based cellular physiology is relevant to questions about the role of different brain regions in overall function.

Biophysically detailed compartmental models of single neurons typically aim to quantitatively reproduce membrane voltages and currents in response to some sort of “synaptic” input. We may think of them as “Hodgkin-Huxley-Rall” models, based on the hypothesis of the neuron as a dynamical system of nonlinear membrane channels distributed over an electrotonic cable skeleton. Such models can incorporate as much biophysical detail as desired (or practical), but in general, all include some explicit assortment of voltage-dependent and transmitter-gated (synaptic) membrane channels. *BIOPHYSICAL MECHANISMS IN NEURONAL MODELING* first presents general issues regarding model formulations and data interpretation. It then describes the modeling of various features of Hodgkin-Huxley-Rall models, including Hodgkin-Huxley and Markov kinetic descriptions of voltage- and second-messenger-dependent ion channels as well as methods for describing intracellular calcium dynamics and the associated buffer systems and membrane pumps. The models for each of these mechanisms are at an intermediate level of biophysical detail, appropriate for describing macroscopic variables (e.g., membrane currents, ionic concentrations) on the scale of the entire cell or anatomical compartments thereof. Similar models of synaptic mechanisms are covered in *SYNAPTIC INTERACTIONS*, which provides kinetic models of how synaptic currents arise from ion channels whose opening and closing are controlled (gated) directly or indirectly by the release of neurotransmitter. The article compares several models of synaptic interaction, focusing on simple models based on the kinetics of postsynaptic receptors, and shows how these models capture the time courses of postsynaptic currents of several types of synaptic responses, as well as synaptic summation, saturation, and desensitization.

The membrane potential of central neurons undergoes synaptic noise, fluctuations that depend on both the summed firing of action potentials by neurons presynaptic to the investigated cell and the spontaneous release of transmitter. *SYNAPTIC NOISE AND CHAOS IN VERTEBRATE NEURONS* argues that, despite its random appearance, synaptic noise may be a true signal associated with neural coding, possibly a chaotic one. In addition to reviewing tools for detecting chaotic behavior, the article pays special attention to Mauthner cells, a pair of identified neurons in the hindbrain of teleost fishes. When the fish is subjected to an unexpected stimulus, one of the cells triggers an escape reaction. Their excitability is controlled by powerful inhibitory presynaptic interneurons that continuously generate an intense synaptic noise. While it is still an open question whether this synaptic noise exhibits deterministic chaos or is truly random, it is worth stressing that the “noise” has adaptive value for the fish: the variability along output pathways introduces uncertainty in the expression of the reflex, and therefore enhances the fish’s success in evading predators.

TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES complements the many studies of synaptic plasticity in the *Handbook* that focus on long-term changes in synaptic strength by showing how synaptic function can be profoundly influenced by activity over time scales of milliseconds to seconds. Synapses that exhibit such short-term plasticity are powerful computational elements that can have profound impact on cortical circuits (cf. “Dynamic Link Architecture”). Short-term plasticity includes both synaptic depression and a number of components of short-term enhancement (facilitation, augmentation, and posttetanic potentiation) acting over increasingly longer periods of time. Synaptic facilitation appears to result from enhanced transmitter release due to elevated presynaptic calcium levels, while depression is believed to result, in part, from depletion of a readily releasable pool of vesicles. Depression ap-

pears to be a particularly prominent feature of transmission at excitatory synapses onto pyramidal cells. In addition to having complex short-term dynamics, synapses are stochastic, and it is argued that constructive roles for unreliable transmission become apparent when short-term plasticity is considered in connection with stochastic transmission, with synapses acting as stochastic temporal filters of their presynaptic spike trains. Indeed, SYNAPTIC TRANSMISSION is concerned with the uncertainties introduced by noise and their relation to synaptic plasticity. The probability that a single activated synapse will release neurotransmitter has a broad distribution, well fitted by a gamma function, with a mean near 0.3. The dynamic regulation of synaptic strength depends on a complicated set of mechanisms that record the history of synaptic use over many time scales, and serve to filter the incoming spike train in a way that reflects the past use of the synapse. The article provides equations which describe how synaptic use determines the number of vesicles available for release, and for the release probability in turn.

OSCILLATORY AND BURSTING PROPERTIES OF NEURONS offers a dynamic systems analysis of the linkage between a fascinating variety of endogenous oscillations (neuronal rhythms) and appropriate sets of channels. However, membrane potential oscillations with apparently similar characteristics can be generated by different ionic mechanisms, and a given cell type may display several different firing patterns under different neuromodulatory conditions. Here, membrane dynamics are described by coupled differential equations, the behavior modes by attractors (cf. "Computing with Attractors"), and the transitions between modes by bifurcations. The rest state is represented by a time-independent steady state, and repetitive firing is represented by a limit cycle. ("Silicon Neurons" shows how such differential equations can be directly mapped into an electronic circuit built using analog VLSI, to allow real-time exploration of the behavior of quite realistic neural models.)

Roughly a dozen different types of ion channels contribute to the membrane conductance of a typical neuron. ACTIVITY-DEPENDENT REGULATION OF NEURONAL CONDUCTANCES takes as its starting point the fact that the electrical characteristics of a neuron depend on the number of channels of each type active within the membrane and on how these channels are distributed over the surface of the cell. A complex array of biochemical processes controls the number and distribution of ion channels by constructing and transporting channels, modulating their properties, and inserting them into and removing them from the neuron's membrane. The point to note here is that channels are small groupings of large molecules, and they are assembled on the basis of genetic instructions in the cell nucleus. Thus, changing which genes are active (i.e., regulating gene expression) can change the set of channels in a cell, and thus the characteristics of the cell. In fact, electrical activity in the cell can affect a range of processes, from activity-induced gene expression to activity-dependent modulation of assembled ion channels. Channel synthesis, insertion, and modulation are much slower than the usual voltage- and ligand-dependent processes that open and close channels. Thus, consideration of activity-dependent regulation of conductances introduces a dynamics acting on a new, slower time scale into neuronal modeling, a feedback mechanism linking a neuron's electrical characteristics to its activity. A similar theme is developed in BIOPHYSICAL MOSAIC OF THE NEURON, which is structured around the metaphor of the mosaic neuron. A mosaic is a collection of discrete parts, each with unique properties, that are fitted together in such a way that an image emerges from the whole in a nonobvious way. Similarly, the neuronal membrane is packed with a diversity of receptors and ion channels and other proteins with a recognizable distribution. In addition, the cytoplasm is not just water with ions, but a mosaic of interacting molecular systems that can directly affect the functional properties of membrane proteins. The argument is that, just as a

mosaic painting provokes perception of a complete image out of a maze of individually diversified tiles, so a given neuron performs a well-defined computational role that depends not only on the network of cells in which it is embedded, but also to a large extent on the dynamic distribution of macromolecules throughout the cell.

DENDRITIC PROCESSING focuses on dendrites as electrical input-output devices that operate on a time scale range of several to a few hundred milliseconds. (See "Dendritic Learning" for modeling of the plasticity of dendritic function and the assertion that the concept of "overall connection strength between two neurons" is ill-defined, since it is the distribution of synapses in relation to dendritic geometry that proves crucial.) The input to a dendrite consists of temporal patterns of synaptic inputs spatially distributed over the dendritic surface, whereas the output is (except, for example, in the case of dendrodendritic interactions) an ionic current delivered to the soma for transformation there, via a threshold mechanism, to a train of action potentials at the axon. The article discusses how the morphology, electrical properties, and synaptic inputs of dendrites interact to perform their input-output operation. It uses cable theory and compartmental modeling to model the spread of electric current in dendritic trees. The variety of excitable (voltage-gated) channels that are found in many types of dendrites enrich the computational capabilities of neurons, with interaction proceeding in both directions, away from and toward the soma. Computer modeling methods for neurons offer numerical methods for solving the equations describing branched cables. DENDRITIC SPINES are short appendages found on the dendrites of many different cell types. They are composed of a bulbous "head" connected to the dendrite by a thin "stem." An excitatory synapse is usually found on the spine head, and some spines also have a second, usually inhibitory, synapse located on or near the spine stem. Models in which the spine is represented as a passive electrical circuit show that the large resistance of a thin spine stem can attenuate a synaptic input delivered to the spine head. Other models address calcium diffusion and plasticity in spines. Current research focuses on the hypothesis that the spine stem provides a diffusional resistance that allows calcium to become concentrated in the spine head and calcium-dependent reactions to be localized to the synapse. This could be very important for plasticity changes, such as those that occur with long-term potentiation.

Neural Plasticity

AXONAL PATH FINDING
CEREBELLUM AND CONDITIONING
CEREBELLUM AND MOTOR CONTROL
CEREBELLUM: NEURAL PLASTICITY
CONDITIONING
DENDRITIC LEARNING
DEVELOPMENT OF RETINOTECTAL MAPS
DYNAMIC LINK ARCHITECTURE
HABITUATION
HEBBIAN LEARNING AND NEURONAL REGULATION
HEBBIAN SYNAPTIC PLASTICITY
INFORMATION THEORY AND VISUAL PLASTICITY
INVERTEBRATE MODELS OF LEARNING: *APLYSIA* AND
HERMISSENDA
NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK
MODELS
OCULAR DOMINANCE AND ORIENTATION COLUMNS
POST-HEBBIAN LEARNING ALGORITHMS
SHORT-TERM MEMORY
SOMATOTOPY: PLASTICITY OF SENSORY MAPS
TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES

Most studies of learning in ANNs involve a variety of learning rules, inspired in great part by the psychological hypotheses of

Hebb and Rosenblatt (cf. Section I.3) about ways in which synaptic connections may change their strength as a result of experience. In recent years, much progress has been made in tracing the processes that underlie the plasticity of synapses of biological neurons. The present road map samples this research together with related modeling. Although the emphasis will be on synaptic plasticity, several articles stress the role of axonal growth in forming new connections, and the road map closes with an article suggesting that changes in location of synapses may be just as important as changes in synaptic strength.

Hebb's idea was that a synapse (what we would now call a Hebbian synapse) strengthens when the presynaptic and postsynaptic elements tend to be coactive. The plausibility of this hypothesis has been enhanced by the neurophysiological discovery of a synaptic phenomenon in the hippocampus known as long-term potentiation (LTP), which is induced by a Hebbian mechanism. Hebb's postulate has received various modifications to address, e.g., the saturation problem.

HEBBIAN SYNAPTIC PLASTICITY shows that a variety of experimental networks ranging from the abdominal ganglion in the invertebrate *Aplysia* to visual cortex and the CA1 region of hippocampus offer converging validation of Hebb's postulate on strengthening synapses by (more or less) coincident presynaptic and postsynaptic activity. In these networks, similar algorithms of potentiation can be implemented using different cascades of second messengers triggered by activation of synaptic and/or voltage-dependent conductances. Most cellular data supporting Hebb's predictions have been derived from electrophysiological measurements of composite postsynaptic potentials or synaptic currents, or of short-latency peaks in cross-correlograms, which cannot always be interpreted simply at the synaptic level. The basic conclusion of these experiments is that covariance between pre- and postsynaptic activity upregulates and downregulates the "effective" connectivity between pairs of functionally coupled cells. The article thus suggests that what changes according to a correlational rule is not so much the efficacy of transmission at a given synapse, but rather a more general coupling term mixing the influence of polysynaptic excitatory and inhibitory circuits linking the two cells, modulated by the diffuse network background activation. Replacing this composite interaction by a single coupling term defines an ideal Hebbian synapse.

The crucial role played in the CA1 form of LTP by channels called NMDA receptors in the synapses is further explained in NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK MODELS. NMDA receptors are subtypes of receptors for the excitatory neurotransmitter glutamate and are involved in diverse physiological as well as pathological processes. They mediate a relatively "slow" excitatory postsynaptic potential, and act as coincidence detectors of presynaptic and postsynaptic activity. The interactions between the slow NMDA-mediated and fast AMPA-mediated currents provide the basis for a range of dynamic properties that contribute to diverse neuronal processes. NMDA receptors have attracted much interest in neuroscience because of their role in learning and memory. Their ability to act as coincidence detectors make them an ideal molecular device for producing Hebbian synapses. The article reviews data related to the biological characteristics of NMDA receptors and models that have been used to describe their function in isolated membrane patches, in neurons, and in complex circuits.

A classic problem with Hebb's original rule is that it only strengthens synapses. But this means that all synapses would eventually saturate, depriving the cell of its pattern separation ability. A number of biologically inspired responses to this problem are described in the next two articles. HEBBIAN LEARNING AND NEURONAL REGULATION stresses that, for both computational and biological reasons, Hebbian plasticity will involve many synapses of

the same neuron. Biologically, synaptic interactions are inevitable as synapses compete for the finite resources of a single neuron. Computationally, neuron-specific modifications of synaptic efficacies are required in order to obtain efficient learning, or to faithfully model biological systems. Hence neuronal regulation, a process modulating all synapses of a postsynaptic neuron, is a general phenomenon that complements Hebbian learning. The article shows that neuronal regulation may answer important questions, such as: What bounds the positive feedback loop of Hebbian learning and guarantees some normalization of the synaptic efficacies of a neuron? How can a neuron acquire specificity to particular inputs without being prewired? How can memories be maintained throughout life while synapses suffer degradation due to metabolic turnover? In unsupervised learning, neuronal regulation allows for competition between the various synapses on a neuron and leads to normalization of their synaptic efficacies. In supervised learning, neuronal regulation improves the capacity of associative memory models and can be used to guarantee the maintenance of biological memory systems. Our basic tour of Hebbian learning concludes with POST-HEBBIAN LEARNING ALGORITHMS. This article starts by observing that Hebb's original postulate was a verbally described phenomenological rule, without specification of detailed mechanisms. Subsequent work has shown the computational usefulness of many variations of the original learning rule. This article presents background material on conditioning, neural development, and physiologically realistic cellular-level learning phenomena as a prelude to a review of several families of rules providing computational implementations of Hebbian-inspired rules.

CEREBELLUM AND MOTOR CONTROL reviews a number of models for cerebellar mechanisms underlying the learning of motor skills. Cerebellum can be decomposed into cerebellar nuclei and a cerebellar cortex. The only output cells of the cerebellar cortex are the Purkinje cells, and their only effect is to provide varying levels of inhibition on the cerebellar nuclei. Each Purkinje cell receives two types of input: a single climbing fiber, and many tens of thousands of parallel fibers. The most influential model of cerebellar cortex has been the Marr-Albus model of the formation of associative memories between particular patterns on parallel fiber inputs and Purkinje cell outputs, with the climbing fiber acting as "training signal." Later models place more emphasis on the relation between the cortex and nuclei, and on the way in which the subregions of this coupled cerebellar system can adapt and coordinate the activity of specific motor pattern generators. The plasticity of the cerebellum is approached from a different direction in CEREBELLUM AND CONDITIONING. Many experiments indicate that the cerebellum is involved in learning and performance of classically conditioned reflexes; the present article reviews a number of models of the role of cerebellum in rabbit eyelid conditioning. (A more general perspective on conditioning is given in CONDITIONING and described more fully in the road map **Psychology**, which describes several formal theories and neural network models for classical and operant conditioning.) Inspired by the Marr-Albus hypothesis, neurophysiological research eventually showed that coincidence of climbing fiber and parallel fiber activity on a Purkinje cell led to long-term depression (LTD) of the synapse from parallel fiber to Purkinje cell. CEREBELLUM: NEURAL PLASTICITY offers readers an exhaustive overview of the data on the neurochemical mechanisms underlying this form of plasticity. The authors conclude that the timing conditions for LTD induction may account for the temporal specificity of cerebellar motor learning, and suggest that an important future development in the field will be to study developmental aspects of LTD in relation to acquisition of motor skills. However, the article cites only one model of LTD. It is clear that there are immense challenges to neural modelers in exploring the implications of the plethora of neurochemical interactions swirling about this single class of synaptic plasticity and, by implication,

the variety of different mechanisms expressed elsewhere in the nervous system.

There is now strong evidence for a process of short-term memory (STM) involved in performing tasks requiring temporary storage and manipulation of information to guide appropriate actions. SHORT-TERM MEMORY addresses three issues: What are the different types of STM traces? How do intrinsic and synaptic mechanisms contribute to the formation of STM traces? How do STM traces translate into long-term memory representation of temporal sequences? The stress is on the computational mechanisms underlying these processes, with the suggestion that these mechanisms may well underlie a wide variety of seemingly different biological processes. The article examines both the short-term preservation of patterns of neural firing in a circuit and ways in which short-term maintained activity may be transferred into long-term memory traces.

There is no hard and fast line between the cellular mechanisms underlying the development of the nervous system and those involved in learning. Nonetheless, the former emphasizes the questions of how one part of the brain comes to be connected to another and how overall patterns of connectivity are formed, while the latter tends to regard the connections as in place, and asks how their strengths can be modified to improve the network's performance. Studies of regeneration—the reforming of connections after damage to neurons or cell tracts—are thus associated more with developmental mechanisms than with learning per se. Another significant area of research that complements development is that of aging, but there is still too little work relating aging to neural modeling.

Study of the regeneration of retinotopic eye-brain maps in frogs (i.e., neighboring points in the frog retina map, in a one-to-many fashion, to neighboring points in the optic tectum) has been one of the most fruitful areas for theory-experiment interaction in neuroscience. Following optic nerve section, optic nerve fibers tended to regenerate connections with those target neurons to which they were connected before surgery, even after eye rotation. This suggests that each cell in both retina and tectum has a unique chemical marker signaling 2D location, and that retinal axons seek out tectal cells with the same positional information. However, in experiments in which lesions were made in goldfish retina or tectum, it was found that topographic maps regenerated in conformance with whatever new boundary conditions were created by the lesions; e.g., the remaining half of a retina would eventually connect in a retinotopic way to the whole of the tectum, rather than just to the half to which it was originally connected. Although there is wide variation between species in the degree of order existing in the optic nerve, it is almost always the case that the final map in the tectum is ordered to a greater extent than is the optic nerve. Theory and experiment paint a subtle view in which genetics sets a framework for development, but the final pattern of connections depends both on boundary conditions and on patterns of cellular activity. This view is now paradigmatic for our understanding of how patterns of neural connectivity are determined. The development of such maps appears to proceed in two stages: the first involves axon guidance independent of neural activity; the second involves the refinement of initially crude patterns of connections by processes dependent on neural activity. AXONAL PATH FINDING focuses on the former events, while DEVELOPMENT OF RETINOTECTAL MAPS discusses the latter. Understanding the molecular basis of retinotectal map formation has been transformed since the appearance of the first edition of the *Handbook* by discoveries centering on ephrins and the corresponding Eph receptors. The Eph/ephrins come in two families, A and B, with the A family important for mapping along the rostral-caudal axis of the tectum, while the B family may be important for mapping along the dorsal-ventral axis. Most models of development of retinotectal maps take synaptic strengths as their

primary variable between arrays of retinal and tectal locations, with initial synaptic strengths then updated according to rules that depend in various ways on correlated activity, competition for tectal space, molecular gradients, and fiber-fiber interactions. However, actual movement or branching of axons to find their correct targets is rarely considered. Thus, future computational models of retinotectal map formation should take into account data on Eph receptors and ephrin ligands, data on the guidance of retinal axons that enter the tectum by ectopic routes, and the results of retinal and tectal ablation and transplantation experiments. Up to now, the great majority of theoretical work in the neural network tradition has focused on changes in synaptic strengths within a fixed connectational architecture, but how axons chart their initial path toward the correct target structure has generally not been addressed. AXONAL PATH FINDING reviews recent experimental work addressing how retinal ganglion cell axons find the optic disk, how they then exit the retina, why they grow toward the optic chiasm, why some then cross at the midline while others do not, and so on—a body of knowledge that now has the potential to be framed and interpreted in terms of theoretical models. Whereas work in neural networks has usually focused on processes such as synaptic plasticity that are dependent on neural activity, models for axon guidance must generally be phrased in terms of activity-independent mechanisms, particularly guidance by molecular gradients. Many fundamental questions remain unresolved, for which theoretical models have the potential to make an important contribution. What is the minimum gradient steepness detectable by a growth cone, and how does this vary with the properties of the receptor-ligand interaction and the internal state of the growth cone? How is a graded difference in receptor binding internally converted into a signal for directed movement? And, how do axons integrate multiple cues?

OCULAR DOMINANCE AND ORIENTATION COLUMNS studies two issues that go beyond basic map formation to provide further insight into activity-dependent development. When cells in layer IVc of visual cortex are tested to see which eye drives them more strongly, it is found that ocular dominance takes the form of a zebra-stripe-like pattern of alternating dominance. Model and experiment support the view that the stripes are not genetically specified but instead form through network self-organization. Another classic example is the formation of orientation specificity. A number of models are reviewed in light of current data, both theoretical analysis based on the idea that leading eigenvectors dominate (cf. “Pattern Formation, Biological” and “Pattern Formation, Neural”) and computer simulations.

TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES complements the many studies of synaptic plasticity in the *Handbook* that focus on long-term changes in synaptic strength by showing the importance of fast synaptic changes over time scales of milliseconds to seconds. Short-term plasticity includes both synaptic depression and a number of components of short-term enhancement (facilitation, augmentation, and posttetanic potentiation) acting over increasingly longer periods of time. In addition to having complex short-term dynamics, synapses are stochastic (see “Synaptic Transmission”), and it is argued that constructive roles for unreliable transmission become apparent when short-term plasticity is considered in connection with stochastic transmission, with synapses acting as stochastic temporal filters of their presynaptic spike trains. DYNAMIC LINK ARCHITECTURE develops the theme of fast synaptic changes at the level of network function, viewing the brain's data structure as a graph composed of nodes connected by links whose strength can change on two time scales, represented by two variables called temporary weight and permanent weight. The permanent weight corresponds to the usual synaptic weight, can change on the slow time scale of learning, and represents permanent memory. The temporary weight can change on the same time scale as the node activity, providing the dynamic links that, according to

this model, constitute the glue by which higher data structures are built up from more elementary ones.

INFORMATION THEORY AND VISUAL PLASTICITY demonstrates some features of information theory that are relevant to the relaying of information in cortex and presents cases in which information theory led people to seek methods for Gaussianizing the input distribution and, in other cases, to seek learning goals for non-Gaussian distributions. The MDL principle (see “Minimum Description Length Analysis”) was presented as a learning goal which takes into account the complexity of the decoding network. In particular, the article connects entropy-based methods, projection pursuit, and extraction of simple cells in visual cortex.

As can be seen from the above, neural network models of development and regeneration have been dominated by studies of the visual system. The next article, however, takes us to the somatosensory system. Research in the past decade has demonstrated plastic changes at all levels of the adult somatosensory system in a wide range of mammalian species. Changes in the relative levels of sensory stimulation as a result of experience or injury produce modifications in sensory maps. SOMATOTOPY: PLASTICITY OF SENSORY MAPS discusses which features of somatotopic maps change and under what conditions, the mechanisms that may account for these changes, and the functional consequences of sensory map changes.

Just as the giant squid axon provided invaluable insights into the active properties of neural membrane summarized in the Hodgkin-Huxley equation, so have invertebrates provided many insights into other basic mechanisms (see “Neuromodulation in Invertebrate Nervous Systems” and “Crustacean Stomatogastric System” for two examples). INVERTEBRATE MODELS OF LEARNING: *APLYSIA* AND *HERMISSENDA* does the same for basic learning mechanisms. A ganglion (localized neural network) of these invertebrates can control a variety of different behaviors, yet a given behavior such as a withdrawal response may be mediated by 100 neurons or less. Moreover, many neurons are relatively large and can be uniquely identified, functional properties of an individual cell can be related to a specific behavior, and changes in cellular properties during learning can be related to specific changes in behavior. Biophysical and molecular events underlying the changes in cellular properties can then be determined and mathematically modeled. The present article illustrates this with studies of two gastropod mollusks: associative and nonassociative modifications of defensive siphon and tail withdrawal reflexes in *Aplysia* and associative learning in *Hermisenda*.

HABITUATION describes one of the simplest forms of learning, the progressive decrement in a behavioral response with repeated presentations of the eliciting stimulus, and reveals the complexity in this apparent simplicity. This article reviews the fundamental characteristics of habituation and describes experimental preparations in which the neural basis of habituation has been examined as well as attempts to model habituation. Experimental studies have identified at least two important neural mechanisms of habituation, homosynaptic depression within the reflex circuit and extrinsic descending modulatory input. A number of systems are put forward as good candidates for future modeling. Habituation of defensive reflexes was among the first types of learning explained successfully at the cellular level. Habituation in the crayfish tail-flip reflex, due to both afferent depression as well as descending inhibition, offers the opportunity to analyze the interaction and cooperativity of mechanisms intrinsic and extrinsic to the reflex circuit. The nematode *C. elegans* offers the possibility of a genetic analysis of habituation.

As shown in “Dendritic Processing,” dendrites are highly complex structures, both anatomically and physiologically, and are the principal substrates for information processing within the neuron. DENDRITIC LEARNING assesses the consequences of axodendritic

structural plasticity for learning and memory, countering the view that neural plasticity is limited to the strengthening and weakening of existing synaptic connections. In particular, the article supports the view that long-term storage may involve the correlation-based sorting of synaptic contacts onto the many separate dendrites of a target neuron. In the models offered in this article, the output of the cell represents the sum of a moderately large set of separately thresholded dendritic subunits, so that a single neuron as modeled here is equivalent to a conventional ANN built from two layers of point neurons. As a result, the concept of “overall connection strength between two neurons” is no longer well defined, for it is the distribution of synapses in relation to dendritic geometry that proves crucial.

Neural Coding

ADAPTIVE SPIKE CODING

INTEGRATE-AND-FIRE NEURONS AND NETWORKS

LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS

MOTOR CORTEX: CODING AND DECODING OF DIRECTIONAL

OPERATIONS

OPTIMAL SENSORY ENCODING

POPULATION CODES

RATE CODING AND SIGNAL PROCESSING

SENSORY CODING AND INFORMATION TRANSMISSION

SPARSE CODING IN THE PRIMATE CORTEX

SYNCHRONIZATION, BINDING AND EXPECTANCY

SYNFIRE CHAINS

In the McCulloch-Pitts neuron, the output is binary, generated on a discrete-time scale; at the other extreme, the Hodgkin-Huxley equations can create a dazzling array of patterns of axonal activity in which the shape as well as the timing of each spike is continuously variable. In between, we have models such as the leaky integrator model, in which only the rate of firing of a cell is significant, while in the spiking neuron model the timing but not the shape of spikes is continuously variable. This raises the question of how sensory inputs and motor outputs, let alone “thoughts” and other less mental intervening variables, are coded in neural activity. In answering this question, we must not only seek to understand the significance of the firing pattern of an individual neuron but also probe how variables may be encoded in patterns of firing distributed across a whole population of neurons.

Retinotopic feature maps are the norm near the visual periphery and up into the early stages of the visual cortex. Here, the firing of a cell peaks for stimuli that fall on a specific patch of the retina and also for a specific feature. Perhaps the most famous example of this is provided by the simple cells discovered in visual cortex by Hubel and Wiesel, which are edge-sensitive cells tuned both for the retinal position and orientation of the edge. In such studies, the cell is characterized by its firing rate during presentation of the stimulus. Similar results are seen for other feature types (see “Feature Analysis”) and other sensory systems. The issue of how other information may be coded by activity in the nervous systems of animals is addressed in a number of articles. LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS asks whether the final neural encoding of visual recognition of one’s grandmother, say, involves neurons that respond selectively to “grandmother”—so-called “grandmother cells”—or whether the sight of grandmother is never made explicit at the single neuron level, with the representation instead distributed across a large number of cells, none of which responds selectively to “grandmother” alone. Few neuroscientists argue that individual neurons might explicitly represent particular objects, but many connectionists have used localist representations to model phenomena that include word and letter perception, although they generally insist that the units in their models are not real neurons. The article examines neurophysiological evidence

that both distributed and local coding are used in high-order visual areas and then goes “against the stream” by forwarding computational reasons for preferring representations that are more localist in some parts of the brain, before examining how work on temporal coding schemes has changed the nature of the local versus distributed debate. *SPARSE CODING IN THE PRIMATE CORTX* marshals theoretical reasons and experimental evidence suggesting that the brain adopts a compromise between distributed and local representations that is often referred to as *sparse coding*. This thesis is illustrated with data on object recognition and face recognition in inferotemporal cortex (the “what” pathway) in monkey.

Perhaps the best-known example of motor coding is that described in *MOTOR CORTX: CODING AND DECODING OF DIRECTIONAL OPERATIONS* for the relation between the direction of reaching and changes in neuronal activity that have been established for several brain areas, including the motor cortex. The cells involved each have a broad tuning function the peak of which is considered to be the “preferred” direction of the cell. A movement in a particular direction will engage a whole population of cells. It is found that, during discrete movements in 2D and 3D space, the weighted vector sum of these neuronal preferences is a “population vector” which points in (close to) the direction of the movement. Such examples underlie the more general analysis given in *POPULATION CODES*. Population codes are computationally appealing both because the overlap among the neurons’ tuning curves allows precise encoding of values that fall between the peaks of two adjacent tuning curves and because many cortical functions, such as sensorimotor transformations, can be easily modeled with population codes. The article focuses on decoding, or reading out, population codes. Neuronal responses are noisy, leading to the need for good estimators for the encoded variables. The article reviews the various estimators that have been proposed, and considers their neuronal implementations. Moreover, there are cases where it is reasonable to assume that population activity codes for more than just a single value, and could even code for a whole probability distribution. The goal of decoding is then to recover an estimate of this probability distribution.

INTEGRATE-AND-FIRE NEURONS AND NETWORKS shows how these models offer potential principles of coding and dynamics. At the single neuron level, it is shown that coherent input is more efficient than incoherent spikes in driving a postsynaptic neuron. Questions discussed for homogeneous populations include conditions under which it is possible, in the absence of an external stimulus, to stabilize a population of spiking neurons at a reasonable level of spontaneous activity, and the relation of frequency of collective oscillations to neuronal parameters, and how rapidly population activity responds to changes in the input. An extension to mixed excitatory/inhibitory populations as found in the cortex is also discussed. *SYNCHRONIZATION, BINDING AND EXPECTANCY* argues that the “binding” of cells that correspond to features of a given visual object may exploit another dimension of cellular firing, namely, the phase at which a cell fires within some overall rhythm of firing. The article presents data consistent with the proposal that the synchronization of responses on a time scale of milliseconds provides an efficient mechanism for response selection and binding of population responses. Synchronization also increases the saliency of responses because it allows for effective spatial summation in the population of neurons receiving convergent input from synchronized input cells. *SYNFIRE CHAINS* were introduced to account for the appearance of precise firing sequences with long interspike delays, dealing with the ways in which such chains might be generated, activity propagation along the chain, how synfire chains can be used to compute, and how they might be detected in electrophysiological recordings. A synfire chain is composed of many pools (or layers) of neurons connected in a feedforward fashion. In a random network with moderate connectivity, many synfire chains can be found by chance, but such ran-

dom synfire chains may not function reproducibly unless the synaptic connections are strengthened by some appropriate learning rule. A given neuron can participate in more than one synfire chain. The extent to which such repeated membership can take place without compromising reproducibility is known as the *memory capacity* of synfire chains. Synfire chains may be considered a special case of the “cell assembly” suggested by Hebb. However, in Hebb’s concepts the cell assembly was a network with multiple feedback connections, whereas the synfire chain is a feedforward net. This allows for much faster computations by synfire chains. While noting that there have also been criticisms of the theory, the article argues that classical anatomy and physiology of the cortex sustain the idea that activity may be organized in synfire chains and that one can create compositional systems from synfire chains.

RATE CODING AND SIGNAL PROCESSING investigates ways in which the sequence of spike occurrence times may encode the information that a neuron communicates to its targets. Spike trains are often quite variable under seemingly identical stimulation conditions. Does this variability carry information about the stimulus? The term *rate coding* is applied in situations where the precise timing of spikes is thought not to play a significant role in carrying sensory information. The article analyzes the sensory information conveyed by two types of rate codes, mean firing rate codes and instantaneous firing rate codes, by adapting classical methods of statistical signal processing to the analysis of neuronal spike trains. While focusing on various examples of rate coding, such as that of neurons of weakly electric fish sensitive to electrical field amplitude, the article also notes cases in which spike timing plays a crucial role.

Recent years have seen an increasing number of quantitative studies of neuronal coding based on Shannon’s information theory, in which the “information” or “entropy” of a message is a purely statistical measure based on the probability of the message within an ensemble: the less likely the message is to occur, the greater its information content. *SENSORY CODING AND INFORMATION TRANSMISSION* reviews two recent approaches to measuring transmitted information. The first is based on direct estimation of the spike train entropies in terms of which transmitted information is defined; the second is based on an expansion to second order in the length of the spike trains. The meaning of any signal that we receive from our environment is modulated by the context within which it appears. *ADAPTIVE SPIKE CODING* explores the analysis of “context” as the statistical ensemble in which the signal is embedded. Interpreting a message requires both registering the signal itself and knowing something about this statistical ensemble. The relevant temporal or spatial ensemble depends on the task. Information theoretically, representations that appropriately take into account the statistical properties of the incoming signal are more efficient (see *OPTIMAL SENSORY ENCODING* and “Information Theory and Visual Plasticity”). The article focuses on neural adaptation, reversible change in the response properties of neurons on short time scales. Since the first observations of adaptation in spiking neurons, it had been suggested that adaptation serves a useful function for information processing, preventing a neuron from continuing to transmit redundant information, viewing both the filtering and the threshold function of a neuron as adaptive functions of the input that may implement the goal of increasing information transmission. Issues include adaptation to the stimulus distribution, with the information about the ensemble read off from the statistics of spike time differences; the separation of different time scales in adaptation; and adaptation of receptive fields. The article also explores the role of calcium and of channel dynamics in providing adaptation mechanisms.

OPTIMAL SENSORY ENCODING focuses on the visual system, seeking to understand what type of data encoding for signals passing from retina to cerebral cortex could reduce the data rate without significant information loss, exploiting the fact that nearby image

pixels tend to convey similar signals and thus carry redundant information. One strategy is to transform the original redundant signal (e.g., in photoreceptors) to nonredundant signals in the retinal ganglion cells or cortical neurons, as in the Infomax proposal. The article presents different coding schemes with different advantages. The retinal code has the advantage of small and identical receptive field (RF) shapes, involving shorter neural wiring and easier specifications. The cortical multiscale code is preferred when invariance is needed for objects moving in depth. Again, whereas the Infomax principle applies well to explain the RFs of the more numerous class of retinal ganglion cells, the P cells in monkeys or X cells in cats, another class of ganglion cells, M cells in monkeys or Y cells in cats, have RFs that are relatively larger, color unselective, and tuned to higher temporal frequencies. These M cells do not extract the maximum information possible (Infomax) about the input but can serve to extract the information as quickly as possible. It is argued that information theory is more likely to find its application in the early stages of the sensory processing, before information is selected or discriminated for any specific cognitive task, and that optimal sensory coding in later stages of sensory pathways will depend on cognitive tasks that require applications of alternative theories.

Biological Networks

CORTICAL HEBBIAN MODULES
CORTICAL POPULATION DYNAMICS AND PSYCHOPHYSICS
DOPAMINE, ROLES OF
HIPPOCAMPAL RHYTHM GENERATION
INTEGRATE-AND-FIRE NEURONS AND NETWORKS
LAYERED COMPUTATION IN NEURAL NETWORKS
NEUROMODULATION IN INVERTEBRATE NERVOUS SYSTEMS
NEUROMODULATION IN MAMMALIAN NERVOUS SYSTEMS
RECURRENT NETWORKS: NEUROPHYSIOLOGICAL MODELING
SLEEP OSCILLATIONS
TEMPORAL INTEGRATION IN RECURRENT MICROCIRCUITS

We turn now to studies of biological neural networks, a study complemented by articles in the road map **Mammalian Brain Regions** and in other road maps on sensory systems, memory, and motor control.

CORTICAL HEBBIAN MODULES models the activity seen in cortical networks during the delay period following the presentation of the stimulus in a delay match-to-sample or delay eye-movement task. The rates observed are in the range of about 10–20 spikes/s, with the subset of neurons that sustain elevated rates being selective of the sample stimulus and concentrated in localized columns in associative cortex. The article shows how to model these selective activity distributions through the autonomous local dynamics in the column. The model presents neural elements and synaptic structures that can reproduce the observed neuronal spike dynamics; showing how Hebbian synaptic dynamics can give rise, in a process of training, to a synaptic structure in the local module capable of sustaining selective activity during the delay period. The mathematical framework for the analysis is provided by the mean field theory of statistical mechanics.

LAYERED COMPUTATION IN NEURAL NETWORKS abstracts from the biology to present a general framework for modeling computations performed in layered structures (which occur in many parts of the vertebrate and invertebrate brain, including the optic tectum, the avian visual wulst, and the cephalopod optic lobe, as well as the mammalian cerebral cortex). A general formalism is presented for the connectivity between layers and the dynamics of typical units of each layer. Information processing capabilities of neural layers include filter operations; lateral cooperativity and competition that can be used in, e.g., stereo vision and winner-take-all;

topographic mapping that underlies the allocation of cortical neurons to different parts of the visual field (fovea/periphery), or the processing of optic flow patterns; and feature maps and population coding, which may be applied both to sensory systems and to “motor fields” of neurons so that the flow of activity in motor areas can predict initiated movements. In a related vein, CORTICAL POPULATION DYNAMICS AND PSYCHOPHYSICS describes cortical population dynamics in the form of structurally simple differential equations for the neurons’ firing activities, using a model class introduced by Wilson and Cowan. The Wilson-Cowan model is powerful enough to reproduce a variety of cortical phenomena and captures the dynamics of neuronal populations seen in a variety of experiments, yet simple enough to allow for analytical treatment that yields an understanding of the mechanisms leading to the observed behavior. The model is applied here to explain dynamical properties of the primate visual system on different levels, reaching from single neuron properties like selectivity for the orientation of a stimulus up to higher cognitive functions related to the binding and processing of stimulus features in psychophysical discrimination experiments.

HIPPOCAMPAL RHYTHM GENERATION notes that global brain states in both normal and pathological situations may be associated with spontaneous rhythmic activities of large populations of neurons. This article presents data and models on the main such states associated with the hippocampus: the two main normally occurring states—the theta rhythm with the associated gamma oscillation, and the irregular sharp waves (SPW) with the associated high-frequency (ripple) oscillation—and a pathological brain state associated with epileptic seizures. Several different modeling strategies are compared in studying rhythmicity in the hippocampal CA3 region.

SLEEP OSCILLATIONS analyzes cortical and thalamic networks at multiple levels, from molecules to single neurons to large neuronal assemblies, with techniques ranging from intracellular recordings to computer simulations, to illuminate the generation, modulation, and function of brain oscillations. Sleep is characterized by synchronized events in billions of synaptically coupled neurons in thalamocortical systems. The early stage of quiescent sleep is associated with EEG spindle waves, which occur at a frequency of 7 to 14 Hz; as sleep deepens, waves with slower frequencies appear on the EEG. The other sleep state, associated with rapid eye movements (REM sleep) and dreaming, is characterized by abolition of low-frequency oscillations and an increase in cellular excitability, very much like wakefulness, although motor output is markedly inhibited. Activation of a series of neuromodulatory transmitter systems during arousal blocks low-frequency oscillations, induces fast rhythms, and allows the brain to recover full responsiveness.

It is a truism that similarity of input-output behavior is no guarantee of similarity of internal function in two neural networks. In particular, a recurrent neural network trained by backpropagation to mimic some biological function may have little internal resemblance to the neural networks responsible for that function in the living brain. Nonetheless, RECURRENT NETWORKS: NEUROPHYSIOLOGICAL MODELING demonstrates that dynamic recurrent network models (see “Recurrent Networks: Learning Algorithms” for the formal background) can provide useful tools to help systems neurophysiologists understand the neural mechanisms mediating behavior. Biological experiments typically involve bits of the system; neural network models provide a method of generating working models of the complete system. Confidence in such models is increased if they not only simulate dynamic sensorimotor behavior but also incorporate anatomically appropriate connectivity. The utility of such models is illustrated in the analysis of four types of biological function: oscillating networks, primate target tracking, short-term memory tasks, and the construction of neural integrators.

As is evident in the road map **Biological Neurons and Synapses**, not all neurons are alike: they show a rich variety of conductances that endow them with different functional properties. These

properties and hence the collective activity of interacting groups of neurons are not fixed, but are instead subject to modulation. The term *neuromodulation* usually refers to the effect of neurochemicals such as acetylcholine, dopamine, norepinephrine, and serotonin, and other substances, including neuropeptides. By contrast with the rapid transmission of information through the nervous system by excitatory and inhibitory synaptic potentials, neuromodulators primarily activate receptor proteins, which do not contain an ion channel (metabotropic receptors). These receptors in turn activate enzymes, which change the internal concentration of substances called second messengers. Second messengers cause slower and longer-lasting changes in the physiological properties of neurons, resulting in changes in the processing characteristics of the neural circuit. *NEUROMODULATION IN INVERTEBRATE NERVOUS SYSTEMS* stresses that the sensory information an animal needs depends on a number of factors, including its activity patterns and motivational state. The modulation of the sensitivities of many sensory receptors is shown for a stretch receptor in crustaceans. Modulators can activate, terminate, or modify rhythmic pattern-generating networks. One example of such “polymorphism” is that neuromodulation can reconfigure the same network to produce either escape swimming or reflexive withdrawal in the nudibranch mollusk *Tritonia*. Mechanisms and sites of neuromodulation include alteration of intrinsic properties of neurons, alteration of synaptic efficacy by neuromodulators, and modulation of neuromuscular junctions and muscles. All this makes clear the subtlety of neuronal function that must be addressed by computational neuroscience and that may inspire the design of a new generation of artificial neurons. Turning to the mammalian brain, we find that the anatomical distribution of fibers releasing neuromodulatory substances in the brain is usually very diffuse, with the activity of a small number of neuromodulatory neurons influencing the functional properties of broad regions of the brain. *NEUROMODULATION IN MAMMALIAN NERVOUS SYSTEMS* starts by summarizing physiological effects of neuromodulation, including effects on resting membrane potential of pyramidal cells and interneurons, spike frequency adaptation, synaptic transmission, and long-term potentiation. It is stressed that the effect of a neurochemical is receptor dependent: a single neuromodulator such as serotonin can have dramatically different effects on different neurons, depending on the type of receptor it activates. Indeed, a chemical may function as a neurotransmitter for one receptor and as a neuromodulator for another. The second half of the article reviews neural network models that help us understand how neuromodulatory effects that appear small at the single neuron level may have a significant effect on dynamical properties when distributed throughout a network. The article reviews several different models of the function of modulatory influences in neural circuits, including noradrenergic modulation of attentional processes (strangely, noradrenergic neurons are those sensitive to norepinephrine), dopaminergic modulation (by dopamine) of working memory, cholinergic modulation (by acetylcholine) of input versus internal processing, and modulation of oscillatory dynamics in cortex and thalamus. *DOPAMINE, ROLES*

OF then focuses specifically on roles of dopamine in both neuromodulation and in synaptic plasticity. Dopamine is a neuromodulator that originates from small groups of neurons in the ventral tegmental area, the substantia nigra, and in the diencephalon. Dopaminergic projections are in general very diffuse and reach large portions of the brain. The time scales of dopamine actions are diverse, from a few hundred milliseconds to several hours. The article focuses on the mesencephalic dopamine centers because they are the most studied, and because they are thought to be involved in diseases such as Tourette’s syndrome, schizophrenia, Parkinson’s disease, Huntington’s disease, drug addiction, and depression. These centers are also involved in such normal brain functions as working memory, reinforcement learning, and attention. The article discusses the biophysical effects of dopamine, how dopamine levels influence working memory, the ways in which dopamine responses resemble the reward prediction signal of the temporal difference model of reinforcement learning, and the role of dopamine in allocation of attention.

INTEGRATE-AND-FIRE NEURONS AND NETWORKS presents relatively simple models that take account of the fact that most biological neurons communicate by action potentials, or spikes (see also “Spiking Neurons, Computation with”). In contrast to the standard neuron model used in ANNs, integrate-and-fire neurons do not rely on a temporal average over the pulses. Instead, the pulsed nature of the neuronal signal is taken into account and considered as potentially relevant for coding and information processing. However, integrate-and-fire models do not explicitly describe the form of an action potential. Integrate-and-fire and similar spiking neuron models are phenomenological descriptions on an intermediate level of detail. Compared to other single-cell models, they allow coding principles to be discussed in a transparent manner. Moreover, the dynamics in networks of integrate-and-fire neurons can be analyzed mathematically, and large systems with thousands of neurons can be simulated rather efficiently.

TEMPORAL INTEGRATION IN RECURRENT MICROCIRCUITS hypothesizes that the ability of neural computation in behaving organisms to produce a response at any time that depends appropriately on earlier sensory inputs and internal states rests on a common principle by which neural microcircuits operate in different cortical areas and species. The article argues that, while tapped delay lines, finite state machines, and attractor neural networks are suitable for modeling specific tasks, they appear to be incompatible with results from neuroanatomy (highly recurrent diverse circuitry) and neurophysiology (fast transient dynamics of firing activity with few attractor states). The authors thus view the transient dynamics of neural microcircuits as the main carrier of information about past inputs, from which specific information needed for a variety of different tasks can be read out in parallel and at any time by different readout neurons. This approach leads to computer models of generic recurrent circuits of integrate-and-fire neurons for tasks that require temporal integration of inputs and, it is argued, provides a new conceptual framework for the experimental investigation of neural microcircuits and larger neural systems.

II.6. Dynamics and Learning in Artificial Networks

Dynamic Systems

AMPLIFICATION, ATTENUATION, AND INTEGRATION
CANONICAL NEURAL MODELS
CHAINS OF OSCILLATORS IN MOTOR AND SENSORY SYSTEMS
CHAOS IN BIOLOGICAL SYSTEMS

CHAOS IN NEURAL SYSTEMS
COLLECTIVE BEHAVIOR OF COUPLED OSCILLATORS
COMPUTING WITH ATTRACTORS
COOPERATIVE PHENOMENA
DYNAMICS AND BIFURCATION IN NEURAL NETS
DYNAMICS OF ASSOCIATION AND RECALL

ENERGY FUNCTIONALS FOR NEURAL NETWORKS
 OPTIMIZATION, NEURAL
 PATTERN FORMATION, BIOLOGICAL
 PATTERN FORMATION, NEURAL
 PHASE-PLANE ANALYSIS OF NEURAL NETS
 SELF-ORGANIZATION AND THE BRAIN
 SHORT-TERM MEMORY
 STATISTICAL MECHANICS OF NEURAL NETWORKS
 STOCHASTIC RESONANCE
 WINNER-TAKE-ALL NETWORKS

Much interest in ANNs has been based on the use of trainable feedforward networks as universal approximators for functions $f: X \rightarrow Y$ from the input space X to the output space Y . However, their provenance was more general. The founding paper of Pitts and McCulloch established the result that, by the mid-1950s, could be rephrased as saying that any finite automaton could be simulated by a network of McCulloch-Pitts neurons. A *finite automaton* is a discrete-time dynamic system; that is, on some suitable time scale, it specifies the next state $q(t+1)$ as a function $\delta(q(t), x(t))$ of the current state and input (for articles related to automata and theory of computation, see the road map **Computability and Complexity**). But a neuron can be modeled as a continuous-time system (as in a leaky integrator neuron with the membrane potential as the state variable). A network of continuous-time neurons can then be considered as a continuous-time system with the rate of change of the state (which could, for example, be a vector whose elements are the membrane potentials of the individual neurons) defined as a function $\dot{q}(t) = f(q(t), x(t))$ of the current state and input. When the input is held constant, the network (whether discrete- or continuous-time) may be analyzed by dynamical systems theory. **COMPUTING WITH ATTRACTORS** shows some of the benefits of such an approach. In particular, a net with internal loops may go to equilibrium (providing a state from which the answer to some problem may be read out), enter a limit cycle (undergoing repetitive oscillations which are useful in control of movement, and in other situations in which a “clock cycle” is of value), or exhibit chaotic behavior (acting in an apparently random way, even though it is deterministic). In particular, the article builds on the notion of a Hopfield network. Hopfield contributed much to the resurgence of interest in neural networks in the 1980s by associating an “energy function” with a network, showing that if only one neuron changed state at a time, a symmetrically connected net would settle to a local minimum of the energy, and that many optimization problems could be mapped to energy functions for symmetric neural nets. **ENERGY FUNCTIONALS FOR NEURAL NETWORKS** uses the notion of Lyapunov function from the dynamical study of ordinary differential equations to show how the definition of energy function and the conditions for convergence to a local minimum can be broadened considerably. (Of course, a network undergoing limit cycles or chaos will not have an energy function that is minimized in this sense.) **OPTIMIZATION, NEURAL** shows that this property can be exploited to solve combinatorial optimization problems that require a more or less exhaustive search to achieve exact solutions, with a computational effort growing exponentially or worse with system size. The article shows that ANN methods can provide heuristic methods that yield reasonably good approximate solutions. Recurrent network methods based on deterministic annealing use an interpolating continuous (analog) space, allowing for shortcuts to good solutions (compare “Simulated Annealing and Boltzmann Machines”). The key to the approach offered here is the technique of mean-field approximation from statistical mechanics. While early neural optimizations were confined to problems encodable with a quadratic energy in terms of a set of binary variables, in the past decade the method has been extended to deal with more general problem types, both in terms of variable types and energy

functions, and has evolved to a general-purpose heuristic for combinatorial optimization.

DYNAMICS AND BIFURCATION IN NEURAL NETS notes that the powerful qualitative and geometric tools of dynamical systems theory are most useful when the behavior of interest is stationary in the sense that the inputs are at most time or space periodic. It then shows how to analyze what kind of behavior we can expect over the long run for a given neural network. In ANNs, the final state may represent the recognition of an input pattern, the segmentation of an image, or any number of machine computations. The stationary states of biological neural networks may correspond to cognitive decisions (e.g., binding via synchronous oscillations) or to pathological behavior such as seizures and hallucinations. Another important issue that is addressed by dynamical systems theory is how the qualitative dynamics depends on parameters. The qualitative change of a dynamical system as a parameter is changed is the subject of bifurcation theory, which studies the appearance and disappearance of branches of solutions to a given set of equations as some parameters vary. This article shows how to use these techniques to understand how the behavior of neural nets depends on both the parameters and the initial states of the network. **PHASE-PLANE ANALYSIS OF NEURAL NETS** complements the study of bifurcations with a technique for studying the qualitative behavior of small systems of interacting neural networks whose neurons are, essentially, leaky integrator neurons. A complete analysis of such networks is impossible but when there are at most two variables involved, a fairly complete description can be given. The article introduces this qualitative theory of differential equations in the plane, analyzing two-neuron networks that consist of two excitatory cells, two inhibitory cells, or an excitatory and inhibitory cell. While planar systems may seem to be a rather extreme simplification, it is argued that in some local cortical circuits we can view the simple planar system as representing a population of coupled excitatory and inhibitory neurons. Computational methods are a very powerful adjunct to this type of analysis. The article concludes with comments on numerical methods and software.

CANONICAL NEURAL MODELS starts from the observation that various models of the same neural structure could produce different results. It thus shows how to derive results that can be observed in a class or a family of models. To exemplify the utility of considering families of neural models instead of a single model, the article shows how to reduce an entire family of Hodgkin-Huxley-type models to a *canonical model*. A model is canonical for a family if there is a continuous change of variables that transforms any other model from the family into this one. As an example, a canonical phase model is presented for a family of weakly coupled oscillators. The change of variables does not have to be invertible, so the canonical model is usually lower-dimensional, simple, and tractable, and yet retains many important features of the family. For example, if the canonical model has multiple attractors, then each member of the family has multiple attractors.

Chaotic phenomena, in which a deterministic law generates complicated, nonperiodic, and unpredictable behavior, exist in many real-world systems and mathematical models. Chaos has many intriguing characteristics, such as sensitive dependence on initial conditions. **CHAOS IN BIOLOGICAL SYSTEMS** provides a view of the appearance of this phenomenon of “deterministic randomness” in a variety of models of physical and biological systems. Features used in assessing time series for chaotic behavior include the power spectrum, dimension, Lyapunov exponent, and Poincaré map. Examples are given from ion channels through cellular activity to complex networks, and “dynamical disease” is characterized by qualitative changes in dynamics in biological control systems. However, the high dimensions of biological systems and the environmental fluctuations that lead to nonstationarity make convincing demonstration of chaos in vivo (as opposed to computer

models) a difficult matter. CHAOS IN NEURAL SYSTEMS looks at chaos in the dynamics of axons, neurons, and networks. An open issue is to understand the significance, if any, of observed fluctuations that appear chaotic. Does a neuron function well despite fluctuations in the timing between spikes, or are the irregularities essential to its task? And if the irregularities are essential to the task, is there any reason to expect that deterministic (chaotic) irregularities would be better than random ones? The vexing question of whether chaos adds functionality to neural networks is still open (see also "Synaptic Noise and Chaos in Vertebrate Neurons"). STOCHASTIC RESONANCE is a nonlinear phenomenon whereby the addition of a random process, or "noise," to a weak incoming signal can enhance the probability that it will be detected by a system. Information about the signal transmitted through the system is also enhanced. The information content or detectability of the signal is degraded for noise intensities that are either smaller or larger than some optimal value. Stochastic resonance has been demonstrated at several levels in biology, from ion channels in cell membranes to animal and human cognition, perception, and, ultimately, behavior.

PATTERN FORMATION, BIOLOGICAL presents a general methodology, based on analysis of the largest eigenvalue, for tracing the asymptotic behavior of a dynamical system, and applies it to the problem of biological pattern formation. Turing originally considered the problem of how animal coat patterns develop. He suggested that chemical markers in the skin comprise a system of diffusion-coupled chemical reactions among substances called morphogens. Turing showed that in a two-component reaction-diffusion system, a state of uniform chemical concentration can undergo a diffusion-driven instability, leading to the formation of a spatially inhomogeneous state. In population biology, patchiness in population densities is the norm rather than the exception. In developmental biology, groups of previously identical cells follow different developmental pathways, depending on their position, to yield the rich spectrum of mammalian coat patterns and the patterns found in fishes, reptiles, mollusks, and butterflies. The article closes with a mechanical model of the process of angiogenesis (genesis of the blood supply) and network formation of endothelial cells in the extracellular matrix, as well as a new approach for predicting brain tumor growth. PATTERN FORMATION, NEURAL shows that the Turing mechanism for spontaneous pattern formation plays an important role in studying two key questions on the large-scale functional and anatomical structure of cortex: How did the structure develop? What forms of spontaneous and stimulus-driven neural dynamics are generated by such a cortical structure? In the neural context, interactions are mediated not by molecular diffusion but by long-range axonal connections. This neural version of the Turing instability has been applied to many problems concerning the dynamics and development of cortex. In the former case, pattern formation occurs in neural activity; in the latter it occurs in synaptic weights. In most cases there exists some underlying symmetry in the model that plays a crucial role in the selection and stability of the resulting patterns.

Complementing this theme of pattern formation, SELF-ORGANIZATION AND THE BRAIN contrasts the algorithmic division of labor between programmer and computer in most current man-made computers with the view of the brain as a dynamical system in which ordered structures arise by processes of self-organization. It argues that, whereas the theory of self-organization has so far focused on the establishment of static structures, the nervous system is concerned with the generation of purposeful, nested processes evolving in time. However, if a self-organizing system is to create the appropriate patterns, quite a few control parameters in a system must all be put in the right ballpark. The article argues that, in view of the variability of the physiological state of the nervous system, evolution must have developed general mechanisms to ac-

tively and autonomously regulate its systems such as to produce interesting self-organized processes and states. The process of brain organization is seen as a cascade of steps, each one taking place within the boundary conditions established by the previous one, but the theory of such cascades is still nonexistent, posing massive challenges for future research. COOPERATIVE PHENOMENA offers a related perspective, developing what has been a major theme in physics for the last century: statistical mechanics, which shows how, for example, to average out the individual variations in position and velocity of the myriad molecules in a gas to understand the relationship between pressure, volume, and temperature, or to see how variations in temperature can yield dramatic phase transitions, such as from ice to water or from water to steam. The article places these ideas in a general setting, stressing the notion of an *order parameter* (such as temperature in the previous example) that describes the macroscopic order of the system and whose variation can yield qualitative changes in system behavior. Unlike a control parameter, which is a quantity imposed on the system from the outside, an order parameter is established by the system itself via self-organization. The argument is mainly presented at a general level, but the article concludes by briefly examining cooperative phenomena in neuroscience, including pattern formation (see also PATTERN FORMATION, BIOLOGICAL), EEG, MEG, movement coordination, and hallucinations (see also PATTERN FORMATION, NEURAL).

STATISTICAL MECHANICS OF NEURAL NETWORKS introduces the reader to some of the basic methods of statistical mechanics and shows that they can be applied to systems made up of large numbers of (formal) neurons. Statistical mechanics has studied magnets as lattices with an atomic magnet (modeled as, e.g., a spin that can be up or down) at each lattice point, and this has led to the statistical analysis of neural networks as "spin glasses," where firing and nonfiring correspond to "spin up" and "spin down," respectively. It has also led to the study of "Markov Random Field Models in Image Processing," in which the initial information at each lattice site represents some local features of the raw image, while the final state allows one to read off a processed image.

COLLECTIVE BEHAVIOR OF COUPLED OSCILLATORS explains the use of phase models (here, the phase is the phase of an oscillation, not the type of phase whose transition is studied in statistical mechanics) to help understand how temporal coherence arises over populations of densely interconnected oscillators, even when their frequencies are randomly distributed. The phase oscillator model for neural populations exemplifies the idea that certain aspects of brain functions seem largely independent of the neurophysiological details of the individual neurons while trying to recover phase information, i.e., the kind of information encoded in the form of specific temporal structures of the sequence of neuronal spikings. The article reviews the collective behavior of coupled oscillators using the phase model and assuming all-to-all type interconnections. Despite this simplification, a great variety of collective behaviors is exhibited. Special attention is given to the onset and persistence of collective oscillation in frequency-distributed systems, splitting of the population into a few subgroups (clustering), and the more complex collective behavior called slow switching. Collections of oscillators that send signals to one another can phase lock, with many patterns of phase differences. CHAINS OF OSCILLATORS IN MOTOR AND SENSORY SYSTEMS discusses a set of examples that illustrate how those phases emerge from the oscillator interactions. Much of the work was motivated by spatiotemporal patterns in networks of neurons that govern undulatory locomotion. The original experimental preparation to which this work was applied is the lamprey central pattern generator (CPG) for locomotion, but the mathematics is considerably more general. The article discusses several motor systems, then turns to the procerebral lobe of *Limax*, the common garden slug, to illustrate chains of oscillators

in a sensory system. Since the details of the oscillators often are not known and difficult to obtain, the object of the mathematics is to find the consequences of what is known, and to generate sharper questions to motivate further experimentation.

AMPLIFICATION, ATTENUATION, AND INTEGRATION focuses on the computational role of the recurrent connections in networks of leaky integrator neurons. Setting the transfer function $f(u)$ to be simply u in the network equations yields a linear network that can be completely analyzed using the tools of linear systems theory. The article describes the properties of linear networks and gives some examples of their application to neural modeling. In this framework, it is shown how recurrent synaptic connectivity can either attenuate or speed up responses; both effects can occur simultaneously in the same network. Besides amplification and attenuation, a linear network can also carry out temporal integration, in the sense of Newtonian calculus, when the strength of feedback is precisely tuned for an eigenmode, so that its gain and time constant diverge to infinity. Finally, it is noted that the linear computations of amplification, attenuation, and integration can be ascribed to a number of brain areas.

WINNER-TAKE-ALL NETWORKS presents a number of designs for neural networks that solve the following problem: given a number of networks, each of which provides as output some “confidence measure,” find in a distributed manner the network whose output is strongest. Two important variants of winner-take-all are k -winner-take-all, where the k largest inputs are identified, and softmax, which consists of assigning each input a weight so that all weights sum to 1 and the largest input receives the biggest weight. The article first describes softmax and shows how winner-take-all can be derived as a limiting case; it then describes how they can both be derived from probabilistic, or energy function, formulations; and it closes with a discussion of VLSI and biological mechanisms. “Modular and Hierarchical Learning Systems” addresses a somewhat related topic: Given a complex problem, find a set of networks, each of which provides an approximate solution in some region of the state space, together with a gating network that can combine these approximations to yield a globally satisfactory solution (i.e., blend the “good” solutions rather than extract the “best” solution).

DYNAMICS OF ASSOCIATION AND RECALL uses dynamical studies to analyze the pattern recall process and its relation with the choice of initial state, the properties of stored patterns, noise level, and network architecture. For large networks and in global recall processes, the strategy is to derive dynamical laws at a *macroscopic* level (i.e., dependent on many neuron states). The challenge is to find the smallest set of macroscopic quantities which will obey closed deterministic equations in the limit of an infinitely large network. The article focuses on simple Hopfield-type models, but closes with a discussion of some variations and generalizations. SHORT-TERM MEMORY asks: What are the different types of STM traces? How do intrinsic and synaptic mechanisms contribute to the formation of STM traces? How do STM traces translate into long-term memory representation of temporal sequences? The stress is on computational mechanisms underlying these processes with the suggestion that these mechanisms may well underlie a wide variety of seemingly different biological processes. The article examines both the short-term preservation of patterns of neural firing in a circuit and ways in which short-term maintained activity may be transferred into long-term memory traces.

Learning in Artificial Networks

ADAPTIVE RESONANCE THEORY
ASSOCIATIVE NETWORKS
BACKPROPAGATION: GENERAL PRINCIPLES
BAYESIAN METHODS AND NEURAL NETWORKS

BAYESIAN NETWORKS
COMPETITIVE LEARNING
CONVOLUTIONAL NETWORKS FOR IMAGES, SPEECH, AND TIME SERIES
DATA CLUSTERING AND LEARNING
DYNAMICS OF ASSOCIATION AND RECALL
ENSEMBLE LEARNING
EVOLUTION AND LEARNING IN NEURAL NETWORKS
EVOLUTION OF ARTIFICIAL NEURAL NETWORKS
GAUSSIAN PROCESSES
GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS
GRAPHICAL MODELS: PARAMETER LEARNING
GRAPHICAL MODELS: PROBABILISTIC INFERENCE
GRAPHICAL MODELS: STRUCTURE LEARNING
HELMHOLTZ MACHINES AND SLEEP-WAKE LEARNING
HIDDEN MARKOV MODELS
INDEPENDENT COMPONENT ANALYSIS
LEARNING AND GENERALIZATION: THEORETICAL BOUNDS
LEARNING NETWORK TOPOLOGY
LEARNING AND STATISTICAL INFERENCE
LEARNING VECTOR QUANTIZATION
MINIMUM DESCRIPTION LENGTH ANALYSIS
MODEL VALIDATION
MODULAR AND HIERARCHICAL LEARNING SYSTEMS
NEOCOGNITRON: A MODEL FOR VISUAL PATTERN RECOGNITION
NEUROMANIFOLDS AND INFORMATION GEOMETRY
PATTERN RECOGNITION
PERCEPTRONS, ADALINES, AND BACKPROPAGATION
PRINCIPAL COMPONENT ANALYSIS
RADIAL BASIS FUNCTION NETWORKS
RECURRENT NETWORKS: LEARNING ALGORITHMS
REINFORCEMENT LEARNING
SELF-ORGANIZING FEATURE MAPS
SIMULATED ANNEALING AND BOLTZMANN MACHINES
STATISTICAL MECHANICS OF GENERALIZATION
STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION
STOCHASTIC APPROXIMATION AND EFFICIENT LEARNING
SUPPORT VECTOR MACHINES
TEMPORAL PATTERN PROCESSING
TEMPORAL SEQUENCES: LEARNING AND GLOBAL ANALYSIS
UNIVERSAL APPROXIMATORS
UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS
YING-YANG LEARNING

The majority of articles in this road map deal with learning in artificial neural networks. Nonetheless, the road map is titled “Learning in Artificial Networks” to emphasize the inclusion of a body of research on statistical inference and learning that can be seen either as generalizing neural networks or as analyzing other forms of networks, such as Bayesian networks and graphical models.

The fundamental difference between a system that learns and one that merely memorizes is that the learning system generalizes to unseen examples. Much of our concern is with supervised learning, getting a network to behave in a way that successfully approximates some specified pattern of behavior or input-output relationship. In particular, much emphasis has been placed on feedforward networks which have no loops, so that the output of the net depends on its input alone, since there is then no internal state defined by reverberating activity. The most direct form of this is a synaptic matrix, a one-layer neural network for which input lines directly drive the output neurons and a “supervised Hebbian” rule sets synapses so that the network will exhibit specified input-output pairs in its response repertoire. This is addressed in ASSO-

CIATIVE NETWORKS, which notes the problems that arise if the input patterns (the “keys” for associations) are not orthogonal vectors. Association also extends to recurrent networks, but in such systems of “dynamic memories” (e.g., Hopfield networks) there are no external inputs as such. Rather the “input” is the initial state of the network, and the “output” is the “attractor” or equilibrium state to which the network then settles. For neurons whose output is a sigmoid function of the linear combination of their inputs, the memory capacity of the associative memory is approximately $0.15n$, where n is the number of neurons in the net. Unfortunately, such an “attractor network” memory model has many spurious memories, i.e., equilibria other than the memorized patterns, and there is no way to decide whether a recalled pattern was memorized or not. DYNAMICS OF ASSOCIATION AND RECALL (see the road map **Dynamic Systems** for more details) shows how to move away from microscopic equations at the level of individual neurons to derive dynamical laws at a macroscopic level that characterize association and recall in Hopfield-type networks (with some discussion of variations and generalizations).

Historically, the earliest forms of supervised learning involved changing synaptic weights to oppose the error in a neuron with a binary output (the perceptron error-correction rule), or to minimize the sum of squares of errors of output neurons in a network with real-valued outputs (the Widrow-Hoff rule). This work is charted in PERCEPTRONS, ADALINES, AND BACKPROPAGATION, which also charts the extension of these classic ideas to multilayered networks. In multilayered networks, there is the *structural credit assignment problem*: When an error is made at the output of a network, how is credit (or blame) to be assigned to neurons deep within the network? One of the most popular techniques is called backpropagation, whereby the error of output units is propagated back to yield estimates of how much a given “hidden unit” contributed to the output error. These estimates are used in the adjustment of synaptic weights to these units within the network. BACKPROPAGATION: GENERAL PRINCIPLES places this idea in a broader framework by providing an overview of contributions that enrich our understanding of the pros and cons (such as “plateaus”) of this adaptive architecture. It also assesses the biological plausibility of backpropagation.

The underlying theoretical grounding is that, given any function $f: X \rightarrow Y$ for which X and Y are codable as input and output patterns of a neural network, then, as shown in UNIVERSAL APPROXIMATORS, f can be approximated arbitrarily well by a feedforward network with one layer of hidden units. The catch, of course, is that many, many hidden units may be required for a close fit. It is thus often treated as an empirical question whether there exists a sufficiently good approximation achievable in principle by a network of a given size—an approximation that a given learning rule may or may not find (it may, for example, get stuck in a local optimum rather than a global one). Gradient descent methods have also been extended to adapt the synaptic weights of recurrent networks. The backpropagation algorithm for feedforward networks has been successfully applied to a wide range of problems, but what can be implemented by a feedforward network is just a static mapping of the input vectors. However, to model dynamical functions of brains or machines, one must use a system capable of storing internal states and implementing complex dynamics. RECURRENT NETWORKS: LEARNING ALGORITHMS presents, then, learning algorithms for recurrent neural networks that have feedback connections and time delays. In a recurrent network, the state of the system can be encoded in the activity pattern of the units, and a wide variety of dynamical behaviors can be programmed by the connection weights. A popular subclass of recurrent networks consists of those with symmetric connection weights. In this case, the network dynamics is guaranteed to converge to a minimum of some “energy” function (see “Energy Functionals for Neural Networks” and

“Computing with Attractors”). However, steady-state solutions are only a limited portion of the capabilities of recurrent networks. For example, they can transform an input sequence into a distinct output sequence, and they can serve as a nonlinear filter, a nonlinear controller, or a finite-state machine. This article reviews the learning algorithms for training recurrent networks, with the main focus on supervised learning algorithms. (See “Recurrent Networks: Neurophysiological Modeling” for the use of such networks in modeling biological neural circuitry.)

One useful perspective for supervised learning views learning as hill-climbing in weight space, so that each “experience” adjusts the synaptic weights of the network to climb (or descend) a metaphorical hill for which “height” at a particular point in “weight space” corresponds to some measure of the performance of the network (or the organism or robot of which it is a part). When the aim is to minimize this measure, the learning process is then an example of what mathematicians call *gradient descent*. The term *reinforcement* comes from studies of animal learning in experimental psychology, where it refers to the occurrence of an event, in the proper relation to a response, that tends to increase the probability that the response will occur again in the same situation. REINFORCEMENT LEARNING describes a form of “semisupervised” learning where the network is not provided with an explicit form of error at each time step but rather receives only generalized reinforcement (“you’re doing well”; “that was bad!”), which yields little immediate indication of how any neuron should change its behavior. Moreover, the reinforcement is intermittent, thus raising the temporal credit assignment problem (see also “Reinforcement Learning in Motor Control”): How is an action at one time to be credited for positive reinforcement at a later time? The solution is to build an “adaptive critic” that learns to evaluate actions of the network on the basis of how often they occur on a path leading to positive or negative reinforcement. Methods for this assessment of future expected reinforcement include temporal difference (TD) learning and Q-learning (see “Q-Learning for Robots”). Current reinforcement learning research includes parameterized function approximation methods; understanding how exploratory behavior is best introduced and controlled; learning under conditions in which the environment state cannot be fully observed; introducing various forms of abstraction such as temporally extended actions and hierarchy; and relating computational reinforcement learning theories to brain reward mechanisms (see “Dopamine, Roles of”).

The task par excellence for supervised learning is pattern recognition—the problem of classifying objects, often represented as vectors or as strings of symbols, into categories. Historically, the field of pattern recognition started with early efforts in neural networks (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION). While neural networks played a less central role in pattern recognition for some years, recent progress has made them the method of choice for many applications. As PATTERN RECOGNITION demonstrates, properly designed multilayer networks can learn complex mappings in high-dimensional spaces without requiring complicated hand-crafted feature extractors. To rely more on learning, and less on detailed engineering of feature extractors, it is crucial to tailor the network architecture to the task, incorporating prior knowledge to be able to learn complex tasks without requiring excessively large networks and training sets. ENSEMBLE LEARNING describes algorithms that, rather than finding one best hypothesis to explain the data, construct a set (sometimes called a committee or ensemble) of hypotheses and then have those hypotheses vote to classify new patterns. Ensemble methods are often much more accurate than any single hypothesis. For example, the representational problem arises when the hypothesis space does not contain any hypotheses that are good approximations to the true decision function f . In some cases, a weighted sum of hypotheses expands the space of functions that can be represented. Hence, by taking a

weighted vote of hypotheses, the learning algorithm may be able to form a more accurate approximation to f . The bulk of research into ensemble methods has focused on constructing ensembles of decision trees. The article introduces the techniques of bagging and boosting, among others, and analyzes their relative merits under different conditions.

Many specific architectures have been developed to solve particular types of learning problem. ADAPTIVE RESONANCE THEORY (ART) bases learning on internal expectations. A pattern matching process compares an external input with the internal memory of various coded patterns. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the same or incorporate new information from matched portions of the current input. When the external world fails to match an ART network's expectations or predictions, a search process selects a new category, representing a new hypothesis about what is important in the present environment.

The neocognitron (see NEOCOGNITRON: A MODEL FOR VISUAL PATTERN RECOGNITION) was developed as a neural network model for visual pattern recognition that addresses the specific question, "How can a pattern be recognized despite variations in size and position?" by using a multilayer architecture in which local features are replicated in many different scales and locations. More generally, as shown in CONVOLUTIONAL NETWORKS FOR IMAGES, SPEECH, AND TIME SERIES, shift invariance in convolutional networks is obtained by forcing the replication of weight configurations across space. Moreover, the topology of the input is taken into account, enabling such networks to force the extraction of local features by restricting the receptive fields of hidden units to be local, and enforcing a built-in invariance with respect to translations, or local distortions of the inputs. The idea of connecting units to local receptive fields on the input goes back to the perceptron in the early 1960s, and was almost simultaneous with Hubel and Wiesel's discovery of locally sensitive, orientation-selective neurons in the cat's visual system.

Just as a polynomial of too high a degree is not useful for curve fitting, a network that is too large will fail to generalize well, and will require longer training times. Smaller networks, with fewer free parameters, enforce a smoothness constraint on the function found. For best performance, it is therefore desirable to find the smallest network that will "fit" the training data. To create a neural network, a designer typically fixes a network topology and uses training data to tune its parameters such as connection weights. The designer, however, often does not have enough knowledge to specify the ideal topology. It is thus desirable to learn the topology from training data as well. LEARNING NETWORK TOPOLOGY reviews algorithms that adjust network topology, adding neurons and removing neurons during the learning process, to arrive at a network appropriate to a given task. For topology learning, a bias is added to prefer smaller models. It is often found that this bias produces a neural network that has better generalization and is more interpretable. This framework is applied to learning the topologies of both feedforward neural networks and BAYESIAN NETWORKS. In Bayesian networks, all the nodes of the network are given and set, and one searches for a topology by adding or deleting links.

Many articles in the *Handbook* emphasize situations where, e.g., learning from examples is stochastic in the sense that examples are randomly generated and the network behavior is thus to be analyzed from a statistical point of view. Statistical estimation identifies the mechanism underlying stochastic phenomena. LEARNING AND STATISTICAL INFERENCE studies learning by using such statistical notions as Fisher information, Bayesian loss, and sequential estimation, as well as the Expectation-Maximization (EM) algorithm for estimating hidden variables. Nonlinear neurodynamics, learning,

and self-organization are seen as adding new concepts to statistical science. The article examines the dynamical behaviors of a learning network under a general loss criterion. The behavior of learning curves is related to neural network complexity to elucidate the discrepancy between training and generalization errors. This perspective is further developed in NEUROMANIFOLDS AND INFORMATION GEOMETRY. A neural network is specified by its architecture and a number of parameters such as synaptic weights and thresholds. Any neural network of this architecture is specified by a point in the parameter space. Learning takes place in the parameter space and a learning process is represented by a trajectory. The article presents the approach of information geometry which sees the geometrical structure of the parameter space as given by a Riemannian manifold. The article shows how dynamical behaviors of neural learning on these "neuromanifolds" are related to the underlying geometrical structures, using multilayer perceptrons and Boltzmann machines as examples.

GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS sets forth the essential relationship between multivariate function estimation in a statistical context and supervised machine learning. Given a training set consisting of (input, output) pairs (x_i, y_i) , the task is to construct a map that generalizes well in that, given a new value of x , the map will provide a reasonable prediction for the hitherto unobserved output associated with this x . Regularization simplifies the problem by applying constraints to the construction of the map that reduce the generalization error (see also "Probabilistic Regularization Methods for Low-Level Vision"). Ideally, these constraints embody a priori information concerning the true relationship between input and output, though various ad hoc constraints have sometimes been shown to work well in practice. Feedforward neural nets, radial basis functions, and various forms of splines all provide regularized or regularizable methods for estimating "smooth" functions of several variables from a given training set. Which method to use depends on the particular nature of the underlying but unknown "truth," the nature of any prior information that might be available about this "truth," the nature of any noise in the data, the ability of the experimenter to choose the various smoothing or regularization parameters well, and so on.

MODULAR AND HIERARCHICAL LEARNING SYSTEMS solves a complex learning problem by dividing it into a set of subproblems. In the context of supervised learning, modular architectures arise when the data can be described by a collection of functions, each of which works well over a relatively local region of the input space, allocating different modules to different regions of the space. The challenge is that, in general, the learner is not provided with prior knowledge of the partitioning of the input space. To solve this, a "gating network" can learn which module to "listen to" in different situations. The learning algorithms described here solve the credit assignment problem by computing a set of posterior probabilities that can be thought of as estimating the utility of different modules in different parts of the input space. An EM algorithm (cf. LEARNING AND STATISTICAL INFERENCE), an alternative to gradient methods, can be derived for estimating the parameters of both the modular system and its extension to hierarchical architectures. The latter arise when we assume that the data are well described by a multiresolution model in which regions are divided recursively into subregions.

BAYESIAN METHODS AND NEURAL NETWORKS shows how to apply Bayes's rule for the use of probabilities to quantify inferences about hypotheses from given data. The idea is to take the predictions $p(d|h_i)$ made by alternative models h_i about data d , and the prior probabilities of the models $p(h_i)$, and obtain the posterior probabilities $p(h_i|d)$ of the models given the data, using Bayes's rule in the form $p(h_i|d) = p(d|h_i)p(h_i)/p(d)$. To apply this to neural networks, regard a supervised neural network as a nonlinear param-

eterized mapping from an input x to an output $y = y(x; w)$, which depends continuously on the “weights” parameter w . The idea is to choose w from a weight space with some given probability distribution $p(w)$ so as to maximize the likelihood of the nets yielding the given set of (input, output) observations. The Bayesian framework deals with uncertainty in a natural, consistent manner by combining prior beliefs about which models are appropriate with how likely each model would be to have generated the data. This results in an elegant, general framework for fitting models to data that, however, may be compromised by computational difficulties in carrying out the ideal procedure. There are many approximate Bayesian implementations, using methods such as sampling, perturbation techniques, and variational methods. In the case of models linear in their parameters, Bayesian neural networks are closely related to GAUSSIAN PROCESSES (q.v.), where many of the computational difficulties of dealing with more general stochastic nonlinear systems can be avoided. Traditionally, neural networks are graphical representations of functions in which the computations at each node are deterministic. By contrast, networks in which nodes represent stochastic variables are called graphical models (see BAYESIAN NETWORKS and GRAPHICAL MODELS: PROBABILISTIC INFERENCE).

RADIAL BASIS FUNCTION NETWORKS applies Bayesian methods to the case where the approximation to the given $y = y(x; w)$ is based on a network using combinations of “radial basis” functions, each of which is “centered” around a weight vector w , so that the response to input x depends on some measure of “distance” of x from w , rather than on the dot product $w \cdot x = \sum_i w_i x_i$ as in many formal neurons. The distribution of the w ’s may be determined by some form of clustering (see DATA CLUSTERING AND LEARNING). Further learning adjusts the connection strengths to a neuron whose outputs give an estimate of, e.g., the posterior probability $p(c|x)$ that class c is present given the observation (network input) x . However, it is easier to model other related aspects of the data, such as the unconditional distribution of the data $p(x)$ and the likelihood of the data, $p(x|c)$, and then recreate the posterior from these quantities according to Bayes’s rule, $p(c|x) = p(c)p(x|c)/p(x)$.

GAUSSIAN PROCESSES continues the Bayesian approach to neural network learning, placing a prior probability distribution over possible functions and then letting the observed data “sculpt” this prior into a posterior using the available data. One can place a prior distribution $P(w)$ on the weights w of a neural network to induce a prior over functions $P(y(x;w))$ but the computations required to make predictions are not easy, owing to nonlinearities in the system. A Gaussian process, defined by a mean function and covariance matrix, can prove useful as a way of specifying a prior directly over function space—it is often simpler to do this than to work with priors over parameters. Gaussian processes are probably the simplest kind of function space prior that one can consider, being a generalization of finite-dimensional Gaussian distributions over vectors. A Gaussian process is defined by a mean function (which we shall usually take to be identically zero), and a covariance function $C(x, x')$ which indicates how correlated the value of the function y is at x and x' . This function encodes our assumptions about the problem (e.g., that the function is smooth and continuous) and will influence the quality of the predictions. The article shows how to use Gaussian processes for classification problems, and describes how data can be used to adapt the covariance function to the given prediction problem.

MINIMUM DESCRIPTION LENGTH ANALYSIS shows how ideas relating to minimum description length (MDL) have been applied to neural networks, emphasizing the direct relationship between MDL and Bayesian model selection methods. The classic MDL approach defined the information in a binary string to be the length of the shortest program with which a general-purpose computer could generate the string. The Bayes bridge is obtained by replacing the Bayesian goal of inferring the “most likely” model M from a set

of observations by minimizing the length of an encoded message which describe M as well as the data D expressed in term of M . MDL and Bayesian methods both formalize Occam’s razor in that a complex network is preferred only if its predictions are sufficiently more accurate.

UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS makes the point that even unsupervised learning involves an *implicit* training signal based on the network’s ability to predict its own input, or on some more general measure of the quality of its internal representation. The main problem in unsupervised learning research is then seen as the formulation of a performance measure or cost function for the learning to generate this internal supervisory signal. The cost function is also known as an objective function, since it sets the objective for the learning process. The article reviews three types of unsupervised neural network learning procedures: information-preserving algorithms, density estimation techniques, and invariance-based learning procedures. The first method is based on the preservation of mutual information $I_{x,y} = H(x) - H(x|y)$ between the input vector x and output vector y , where $H(x)$ is the entropy of random variable x and $H(x|y)$ is the entropy of the conditional distribution of x given y . The second approach is to assume a priori a class of models that constrains the general form of the probability density function and then to search for the particular model parameters defining the density function (or mixture of density functions) most likely to have generated the observed data (cf. the earlier discussion of Bayesian methods). Finally, invariance-based learning extracts higher-order features and builds more abstract representations. Once again, the approach is to make constraining assumptions about the structure that is being sought, and to build these constraints into the network’s architecture and/or objective function to develop more efficient, specialized learning procedures.

The Bayesian articles stress the “global” statistical idea of “find the weights which, according to given probability distributions maximize some expectation” as distinct from the deterministic idea of adjusting the weights at each time step to provide a local increment in performance on the current input. However, gradient descent provides an important tool for finding the weight settings which decrease some stochastic expectation of error, too. STOCHASTIC APPROXIMATION AND EFFICIENT LEARNING shows that gradient descent has a long tradition in the literature of stochastic approximation. Any stochastic process that can be interpreted as minimizing a cost function based on noisy gradient measurements in a sequential, recursive manner may be considered to be a stochastic approximation. “Sequential” means that each estimate of the location of a minimum is used to make a new observation, which in turn immediately leads to a new estimate; “recursive” means that the estimates depend on past gradient measurements only through a fixed number of scalar statistics. Such on-line algorithms are useful because they enjoy significant performance advantages for large-scale learning problems. The article describes their properties using stochastic approximation theory as a very broad framework, and provides a brief overview of newer insights obtained using information geometry (see NEUROMANIFOLDS AND INFORMATION GEOMETRY) and replica calculations (see STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION).

In order to understand the performance of learning machines, and to gain insight that helps to design better ones, it is helpful to have theoretical bounds on the generalization ability of the machines. The determination of such bounds is the subject of LEARNING AND GENERALIZATION: THEORETICAL BOUNDS. Here it is necessary to formalize the learning problem and turn the question of how well a machine generalizes into a mathematical question. The article adopts the formalization used in statistical learning theory, which is shown to include both pattern recognition and function learning. The road map **Computability and Complexity** gives

more information on this and related articles, such as “PAC Learning and Neural Networks” and “Vapnik-Chervonenkis Dimension of Neural Networks,” which offer bounds on the performance of learning methods. SUPPORT VECTOR MACHINES addresses the (binary) pattern recognition problem of learning theory: given two classes of objects, to assign a new object to one of the two classes. Trying to find the best classifier involves notions of similarity in the set X of inputs. Support vector machines (SVMs) build a decision function as a kernel expansion corresponding to a separating hyperplane in a feature space. SVMs rest on methods for the selection of the patterns on which the kernels are centered and in the choice of weights that are placed on the individual kernels in the decision function. SVMs and other kernel methods have a number of advantages compared to classical neural network approaches, such as the absence of spurious local minima in the optimization procedure, the need to tune only a few parameters, and modularity in the design. Kernel methods connect similarity measures, nonlinearities, and data representations in linear spaces where simple geometric algorithms are performed.

The passage of the “energy” of a Hopfield network to a local minimum can be construed as a means for solving an optimization problem. The catch is the word “local” in local minimum—the solution may be the best in the neighborhood, yet far better solutions may be located elsewhere. One resolution of this is described in SIMULATED ANNEALING AND BOLTZMANN MACHINES. At the expense of great increases in time to convergence, simulated annealing escapes local minima by adding noise, which is then gradually reduced (“lowering the temperature”). The initially high temperature (i.e., noise level) stops the system from getting trapped in “high valleys” of the energy landscape, the lowering of temperature allows optimization to occur in the “deepest valley” once it has been found. The Boltzmann machine then applies this method to design a class of neural networks. These machines use stochastic computing elements to extend discrete Hopfield networks in two ways: they replace the deterministic, asynchronous dynamics of Hopfield networks with a randomized local search dynamics, and they replace the Hebbian learning rule with a more powerful stochastic learning algorithm.

Turning from neural networks to another form of network structure, BAYESIAN NETWORKS (as distinct from BAYESIAN METHODS AND NEURAL NETWORKS) provides an explicit method for following chains of probabilistic inference such as those appropriate to expert systems, extending the Bayes’s rule for updating probabilities in the light of new evidence. The nodes in a Bayesian network represent propositional variables of interest and the links represent informational or causal dependencies among the variables. The dependencies are quantified by conditional probabilities for each node, given its parents in the network. The network supports the computation of the probabilities of any subset of variables, given evidence about any other subset, and the reasoning processes can operate on Bayesian networks by propagating information in any direction. HELMHOLTZ MACHINES AND SLEEP-WAKE LEARNING starts by observing that since unsupervised learning is largely concerned with finding structure among sets of input patterns, it is important to take advantage of cases in which the input patterns are generated in a systematic way, thus forming a manifold that has many fewer dimensions than the space of all possible activation patterns. The Helmholtz machine is an analysis-by-synthesis model. The key idea is to have an imperfect generative model train a better analysis or recognition model, and an imperfect recognition model train a better generative model. The generative model for the Helmholtz machine is a structured belief network (i.e., Bayesian network) that is viewed as a model for hierarchical top-down connections in the cortex. New inputs are analyzed in an approximate fashion using a second structured belief network (called the recognition model), which is viewed as a model for the standard,

bottom-up connections in cortex. The generative and recognition models are learned from data in two phases. In the *wake phase*, the recognition model is used to estimate the underlying generators for a particular input pattern, and then the generative model is altered so that those generators are more likely to have produced the input that is actually observed. In the *sleep phase*, the generative model fantasizes inputs by choosing particular generators stochastically, and then the recognition model is altered so that it is more likely to report those particular generators if the fantasized input were actually to be observed. YING-YANG LEARNING further develops this notion of simultaneously building up two pathways, a bottom-up pathway for encoding a pattern in the observation space into its representation in a representation space, and a top-down pathway for decoding or reconstructing a pattern from an inner representation back to a pattern in the observation space. The theory of Bayesian Ying-Yang harmony learning formulates the two-pathway approach in a general statistical framework, modeling the two pathways via two complementary Bayesian representations of the joint distribution on the observation space and representation space. The article shows how a number of major learning problems and methods can be seen as special cases of this unified perspective. Moreover, the ability of Ying-Yang learning for regularization and model selection is placed in an information-theoretic perspective.

GRAPHICAL MODELS: PROBABILISTIC INFERENCE introduces a generalization of Bayesian networks. The graphical models framework provides a clean mathematical formalism that has made it possible to understand the relationships among a wide variety of network-based approaches to computation, and in particular to understand many neural network algorithms and architectures as instances of a broader probabilistic methodology. Graphical models use graphs to represent and manipulate joint probability distributions. The graph underlying a graphical model may be directed, in which case the model is often referred to as a belief network or a Bayesian network, or the graph may be undirected, in which case the model is generally referred to as a Markov random field. A graphical model has both a structural component, encoded by the pattern of edges in the graph, and a parametric component, encoded by numerical “potentials” associated with sets of edges in the graph. General inference algorithms allow statistical quantities (such as likelihoods and conditional probabilities) and information-theoretic quantities (such as mutual information and conditional entropies) to be computed efficiently. The article closes by noting that many neural network architectures are special cases of the general graphical model formalism, both representationally and algorithmically. Special cases of graphical models include essentially all models of unsupervised learning, as well as Boltzmann machines, mixtures of experts, and radial basis function networks, while many other neural networks, including the classical multilayer perceptron, can be profitably analyzed from the viewpoint of graphical models. The next two articles present learning algorithms that build on these inference algorithms and allow parameters and structures to be estimated from data. GRAPHICAL MODELS: PARAMETER LEARNING discusses the learning of parameters for a fixed graphical model. As noted, each node in the graph represents a random variable, while the edges in the graph represent the qualitative dependencies between the variables; the absence of an edge between two nodes means that any statistical dependency between these two variables is mediated via some other variable or set of variables. The quantitative dependencies between variables that are connected via edges are specified via parameterized conditional distributions, or more generally nonnegative “potential functions.” The pattern of edges is the structure of the graph, while the parameters of the potential functions are parameters of the graph. The present article assumes that the structure of the graph is given, and shows how to then learn the parameters of the graph from data. GRAPHICAL MODELS: STRUCTURE LEARNING turns to the simul-

taneous learning of parameters and structure. Real-world applications of such learning abound, the example presented being an analysis of data regarding factors that influence the intention of high school students to attend college. For simplicity, the article focuses on directed-acyclic graphical models, but the basic principles thus defined can be applied more generally. The Bayesian approach is emphasized, and then several common non-Bayesian approaches are mentioned briefly.

COMPETITIVE LEARNING is a form of unsupervised learning in which each input pattern comes, through learning, to be associated with the activity of one or at most a few neurons, leading to sparse representations of data that are easy to decode. Competitive learning algorithms employ some sort of competition between neurons in the same layer via lateral connections. This competition limits the set of neurons to be affected in a given learning trial. Hard competition allows the final activity of only one neuron, the strongest one to start with, whereas in soft competition the activity of the lateral neurons does not necessarily drive all but one to zero. One form of competitive learning algorithm can be described as an application of a successful single-neuron learning algorithm in a network with lateral connections between adjacent neurons. The lateral connections are needed so that each neuron can be inhibited from adapting to a feature of the data already captured by other neurons. A second family of algorithms uses the competition between neurons for improving, sharpening, or even forming the features extracted from the data by each single neuron. DATA CLUSTERING AND LEARNING emphasizes the related idea of data clustering, discovering, and emphasizing structure that is hidden in a data set (e.g., the pronounced similarity of groups of data vectors) in an unsupervised fashion. There is a delicate trade-off: not to superimpose too much structure, and yet not to overlook structure. The choice of data representation predetermines what kind of cluster structures can be discovered in the data. Formulating the search for clusters as an optimization problem then supports validation of clustering results by checking that the cluster structures found in a data set vary little from one data set to a second data set generated by the same data source. The two tasks of clustering, density estimation and data compression, are tightly related by the fact that the correct identification of the probability model of the source yields the best code for data compression. PRINCIPAL COMPONENT ANALYSIS shows how, in data compression applications like image or speech coding, a distribution of input vectors may be economically encoded, with small expected values of the distortions, in terms of eigenvectors of the largest eigenvalues of the correlation matrix that describes the distribution of these patterns (these eigenvectors are the “principal components”). However, it is usually not possible to find the eigenvectors on-line. The ideal solution is then replaced by a neural network learning rule embodying a constrained optimization problem that converges to the solution given by the principal components. INDEPENDENT COMPONENT ANALYSIS (ICA) is a linear transform of multivariate data designed to make components of the resulting random vector as statistically independent (factorial) as possible. In signal processing it is used to attack the problem of the blind separation of sources, for example of audio signals that have been mixed together by an unknown process (the “cocktail party effect”). In the area of neural networks and brain theory, it is an example of an information-theoretic unsupervised learning algorithm. When an ICA network is trained on an ensemble of natural images, it learns localized-oriented receptive fields qualitatively similar to those found in area V1 of mammalian visual cortex. ICA has been used to decompose multivariate brain data into components that help us understand task-related spatial and temporal brain dynamics. Thus the same neural network algorithm is being used both as an explanation of brain properties and as a method of probing the brain. Where principal component analysis (PCA) uses second-order statistics (the covariance matrix)

to remove correlations between the elements of a vector, ICA uses statistics of all orders. PCA attempts to decorrelate the outputs, while ICA attempts to make the outputs statistically independent. The most widely used adaptive, on-line method for ICA is also the most “neural-network-like” and is the one described in the body of this article.

SELF-ORGANIZING FEATURE MAPS introduces the self-organizing feature map (SOFM or SOM; also known as a Kohonen map), a nonlinear method by which features can be obtained with an unsupervised learning process. It is based on a layer of adaptive “neurons” that gradually develops into an array of feature detectors. The linking of input signals to response locations in the map can be viewed as a nonlinear projection from a signal or input space to the (usually) 2D map layer. The learning method is an augmented Hebbian method in which learning by the element most responsive to an input pattern is “shared” with its neighbors. The result is that the resulting “compressed image” of the (usually higher-dimensional) input space has the property of a topographic map that reflects important metric and statistical properties of the input signal distribution: distance relationships in the input space (expressing, e.g., pattern similarities) are approximately preserved as distance relationships between corresponding excitation sites in the map, and clusters of similar input patterns tend to become mapped to areas of the neural array whose size varies in proportion to the frequency of the occurrence of their patterns. This resembles in many ways the structure of topographic feature maps found in many brain areas, for which the SOFM offers a neural model that bridges the gap between microscopic adaptation rules postulated at the single neuron or synapse level and the formation of experimentally better accessible, macroscopic patterns of feature selectivity in neural layers. From a statistical point of view, the SOFM provides a nonlinear generalization of principal component analysis and has proved valuable in many application contexts.

In order to give a quantitative answer to the question of how well the trained network will be able to classify an input that it has not seen before, it is common to assume that all inputs, both from the training set and the test set, are produced independently and at random. Clearly, the generalization error depends on the specific algorithm that was used during the training, and its calculation requires knowledge of the network weights generated by the learning process. In general, these weights will be complicated functions of the examples, and an explicit form will not be available in most cases. The methods of statistical mechanics provide an approach to this problem, which often enables an exact calculation of learning curves in the limit of a very large network. In the statistical mechanics approach one studies the ensemble of all networks that implement the same set of input/output examples to a given accuracy. In this way the typical generalization behavior of a neural network (in contrast to the worst or optimal behavior) can be described. We thus turn to two articles that apply the methods introduced in the article “Statistical Mechanics of Neural Networks”: STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION emphasizes on-line learning in which training examples are dealt with one at a time, while STATISTICAL MECHANICS OF GENERALIZATION emphasizes off-line or memory-based methods, where learning is guided by the minimization of a cost function as averaged over the whole training set. From a statistical physics point of view, the distinction is between systems that can be thought of as being in a state of thermal equilibrium (off-line \approx on-equilibrium) and away-from-equilibrium situations where the network is not allowed to extract all possible information from a set of examples (on-line \approx off-equilibrium). While on-line learning is an intrinsically stochastic process, the restriction to large networks, together with assumptions about the statistical properties of the inputs, permits a concise description of the dynamics in terms of coupled ordinary differential equations. These deterministic

equations govern the average evolution of quantities that completely define the macroscopic state of the ANN. The average is taken with respect to the data, which is straightforward if the presented examples are statistically independent. The probability that the network will make a mistake on the new input defines its generalization error for a given training set. Its average over many realizations of the training set, as a function of the number of examples, gives the so-called learning curve. Calculation of the learning curve requires knowledge of the network weights generated by the learning process, for which an explicit form will not be available in most cases. The methods of statistical mechanics provide an approach to this problem, in many cases yielding an exact calculation of learning curves in the “thermodynamic limit” of a very large network in which the network size increases in proportion to the number of training examples, while the statistical or information-theoretic approach is applicable to the learning curve of a medium-size network (cf. *LEARNING AND STATISTICAL INFERENCE*).

MODEL VALIDATION shows how the data analyst tries to infer a “model” that summarizes functional dependencies that may be observed in a given set of empirical data. A good model fit should reproduce the behavior of the studied system in the parameter range to be explained by the model study. Model complexity has to be controlled to avoid both missing essential features of the system (underfitting) and adapting to irrelevant fluctuations in the data (overfitting). Model validation provides the crucial step in modeling between model synthesis and analysis, assessing how appropriate the model is to gain insight into the real-world system. Model validation can make use of bounds of the VC type (cf. “Vapnik-Chervonenkis Dimension of Neural Networks”), which usually contain a complexity term that accounts for the flexibility of the hypothesis class and a fitting term that measures the contraction of measure due to the large number of samples. It is shown how these terms can be controlled either by numerical methods like cross-validation and bootstrap or by analytical techniques from computational learning theory. The trade-off between model complexity and goodness of fit and its relation to the computational complexity of learning remains a deep challenge for research.

HIDDEN MARKOV MODELS describes the use of deterministic and stochastic finite state automata for sequence processing, with special attention to hidden Markov models as tools for the processing of complex piecewise stationary sequences. It also describes a few applications of ANNs to further improve these methods. HMMs allow complex sequential learning problems to be solved by assuming that the sequential pattern can be decomposed into piecewise stationary segments, with each stationary segment parameterized in terms of a stochastic function. The HMM is called “hidden” because there is an underlying stochastic process (i.e., the sequence of states) that is not observable but that affects the observed sequence of events.

TEMPORAL PATTERN PROCESSING notes that time is embodied in a temporal pattern in two different ways: the temporal order among the components of a sequence and the temporal duration of the elements (see also “Sequence Learning”). A sequence is defined as *complex* if it contains repetitions of the same subsequence, and otherwise is *simple*. For the generation of complex sequences, the correct successor can be determined only by knowing components prior to the current one. We refer to the prior subsequence required to determine the current component as the *context* of the component. Temporal processing requires that a neural network have a capacity of short-term memory (STM) in order to maintain a component for some time. Time warping is challenging because we would like to have invariance over limited warping, but dramatic change in relative duration must be recognized differently. Another fundamental ability of human information processing is chunking, which, in the context of temporal processing, means that frequently encountered and meaningful subsequences organize into chunks

that form basic units for further chunking at a higher level. *TEMPORAL SEQUENCES: LEARNING AND GLOBAL ANALYSIS* studies how elementary pattern sequences may be represented in neural structures at a low architectural and computational cost, seeking to understand mechanisms to memorize spatiotemporal associations in a robust fashion within model neural networks. The article focuses on formal neural networks where the interplay between neural and synaptic dynamics and, in particular, the role of transmission delays can be analyzed using methods from nonlinear dynamics and statistical mechanics. Among the questions studied are how to train a network so that its limit cycles will resemble taught sequences. Such simplified systems are necessarily caricatures of biological structures yet suggest aspects that are important for more elaborate approaches to real neural systems.

EVOLUTION OF ARTIFICIAL NEURAL NETWORKS adds another temporal dimension to the biological process of adaptation, namely, that of evolution. Rather than adapt the weights of a single network to solve a problem in the network’s “lifetime,” the evolutionary approach applies the methodology of genetic algorithms to evolve a population of neural networks over several generations so that the population becomes better and better suited to some computational ecology. *EVOLUTION AND LEARNING IN NEURAL NETWORKS* extends this selection of networks on the basis of the result of their adaptation to the environment through lifetime learning. The article shows how studies of ANNs that are subjected both to an evolutionary and a lifetime learning process have been conducted to look at the advantages, in terms of performance, of combining two different adaptation techniques or to help understand the role of the interaction between learning and evolution in natural organisms.

Computability and Complexity

ANALOG NEURAL NETS: COMPUTATIONAL POWER
LEARNING AND GENERALIZATION: THEORETICAL BOUNDS
NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY
PAC LEARNING AND NEURAL NETWORKS
UNIVERSAL APPROXIMATORS
VAPNIK-CHERVONENKIS DIMENSION OF NEURAL NETWORKS

The 1930s saw the definition of an abstract notion of computability when it was discovered that the set of functions on the natural numbers, $f: \mathbb{N} \rightarrow \mathbb{N}$, computable by a Turing machine (an abstraction from following a finite set of rules to calculate on a finite but extendible tape, each square of which could hold one of a fixed set of symbols), lambda functions (which later came to be better known as functions computable by programs written in LISP), and general recursive functions (a class of functions obtained from very simple numerical functions by repeated application of composition, minimization, etc.), were identical. As general-purpose electronic computers were developed and used in the 1940s and 1950s, it was firmly established that these *computable functions* were precisely the functions that could be computed by such computers with suitable programs, provided there were no limitations on computer memory or computation time. This set the stage for the development of complexity theory in the 1960s and beyond: to chart the different subsets of the computable functions that would be obtained when restrictions were placed on computing resources.

Many classification or pattern recognition tasks can be formulated as mappings between subsets of multidimensional vector spaces by using a suitable coding of inputs and outputs, and many types of feedforward networks are *universal* in the sense that, given enough adjustable synaptic weights, they can approximate any mapping between subsets of Euclidean spaces. *UNIVERSAL APPROXIMATORS* surveys recent developments in the mathematical theory of feedforward networks and includes proofs of the universal approximation capabilities of perceptron and radial basis function

networks with general activation and radial functions, and provides estimates of rates of approximation. The article also characterizes sets of multivariable functions that can be approximated without the “curse of dimensionality,” which is an exponentially fast scaling of the number of parameters with the number of variables.

NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY explores what happens when the discrete operations of conventional automata theory are replaced by a computing model in which operations on real numbers are treated as basic. Whereas classical automata describe digital machines, neural models frequently require a framework of analog computation defined on a continuous phase space, with a dynamics characterized by the existence of real constants that influence the macroscopic behavior of the system. Moreover, unlike the flow in digital computation, analog models do not include local discontinuities. Neural networks with real weights are more powerful than traditional models of computation in that they can compute more functions within given time bounds. However, the practicality of an approach based on infinite precision real operations remains to be seen. Nonetheless, the new attention to real numbers has renewed complexity theory and introduced many open problems in computational learning theory and neural network theory. The article thus pays special attention to analog computation in the presence of noise. ANALOG NEURAL NETS: COMPUTATIONAL POWER then analyzes the exact and approximate representational power of feedforward and recurrent neural nets with synchronous update, with a brief discussion of networks of spiking neurons and their relation to sigmoidal nets. Learning complexity increases with increasing representational power of the underlying neural model and care has to be exercised to strike a balance between representational power on the one hand and learning complexity on the other. However, the emphasis of the article is on representational power, i.e., on what can be represented with networks using a given set of activation functions, rather than on learning complexity. Splines (i.e., piecewise polynomial functions) have turned out to be powerful approximators, and they are used here as the benchmark class of activation functions. Much attention is given to studying the properties that a class of activation functions needs to reach the approximation power of splines.

PAC LEARNING AND NEURAL NETWORKS discusses the “probably approximately correct” (PAC) learning paradigm as it applies to ANNs. Roughly speaking, if a large enough sample of randomly drawn training examples is presented, then it should be

likely that, after learning, the neural network will classify most other randomly drawn examples correctly. The PAC model formalizes the terms “likely” and “most.” The two main issues in PAC learning theory are how many training examples should be presented, and whether learning can be achieved using a fast algorithm. These are known, respectively, as the *sample complexity* and *computational complexity* problems. PAC learning makes use of the Vapnik-Chervonenkis dimension (VC-dimension) as a combinatorial parameter that measures the “expressive power” of a family of functions. This parameter is described more fully in VAPNIK-CHERVONENKIS DIMENSION OF NEURAL NETWORKS. Bounds for the VC-dimension of a neural net N provide estimates for the number of random examples that are needed to train N so that it has good generalization properties (i.e., so that the error of N on new examples from the same distribution is very small, with probability very close to 1). Typically, the VC-dimension for a class of networks grows polynomially (in many cases, between linearly and quadratically) with the number of adjustable parameters of the neural network. In particular, if the number of training examples is large compared to the VC-dimension, the network’s performance on training data is a reliable indication of its future performance on subsequent data. The bounds on training set size tend to be large, since they provide generalization guarantees simultaneously for any probability distribution on the examples and for any training algorithm that minimizes disagreement on the training examples. Tighter bounds are available for some special distributions and specific training algorithms. This theme is further developed in LEARNING AND GENERALIZATION: THEORETICAL BOUNDS in relation to three learning problems: pattern recognition, regression estimation, and density estimation. Because of the looseness of its bounds as well as the difficulty of evaluating them, VC theory was until recently largely neglected by practitioners. This has changed markedly with the development of support vector machines. Using nonlinear similarity measures, referred to as kernels, one can reduce a large class of learning algorithms to linear algorithms in an associated feature space. For the linear algorithms, a VC analysis can be carried out, identifying precisely the factors that need to be controlled to achieve high generalization ability in a variety of learning tasks. “Support Vector Machines” casts these factors into a convex optimization framework, leading to efficient and mathematically well-founded algorithms that have been shown to produce state-of-the-art results on a large variety of problems.

II.7. Sensory Systems

Vision

ADAPTIVE RESONANCE THEORY
COLLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL
COLOR PERCEPTION
CONTOUR AND SURFACE PERCEPTION
CORTICAL POPULATION DYNAMICS AND PSYCHOPHYSICS
DIRECTIONAL SELECTIVITY
DISSOCIATIONS BETWEEN VISUAL PROCESSING MODES
DYNAMIC LINK ARCHITECTURE
DYNAMIC REMAPPING
FACE RECOGNITION: NEUROPHYSIOLOGY AND NEURAL TECHNOLOGY
FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM
FAST VISUAL PROCESSING

FEATURE ANALYSIS

GABOR WAVELETS AND STATISTICAL PATTERN RECOGNITION
GLOBAL VISUAL PATTERN EXTRACTION
IMAGING THE VISUAL BRAIN
INFORMATION THEORY AND VISUAL PLASTICITY
KALMAN FILTERING: NEURAL IMPLICATIONS
LAMINAR CORTICAL ARCHITECTURE IN VISUAL PERCEPTION
MARKOV RANDOM FIELD MODELS IN IMAGE PROCESSING
MOTION PERCEPTION: ELEMENTARY MECHANISMS
MOTION PERCEPTION: NAVIGATION
NEOCOGNITRON: A MODEL FOR VISUAL PATTERN RECOGNITION
OBJECT RECOGNITION
OBJECT RECOGNITION, NEUROPHYSIOLOGY
OBJECT STRUCTURE, VISUAL PROCESSING
OCULAR DOMINANCE AND ORIENTATION COLUMNS
ORIENTATION SELECTIVITY

PERCEPTION OF THREE-DIMENSIONAL STRUCTURE
 PROBABILISTIC REGULARIZATION METHODS FOR LOW-LEVEL
 VISION
 PURSUIT EYE MOVEMENTS
 RETINA
 SENSOR FUSION
 STEREO CORRESPONDENCE
 SYNCHRONIZATION, BINDING AND EXPECTANCY
 TENSOR VOTING AND VISUAL SEGMENTATION
 VISUAL ATTENTION
 VISUAL CORTEX: ANATOMICAL STRUCTURE AND MODELS OF
 FUNCTION
 VISUAL SCENE PERCEPTION
 VISUAL SCENE SEGMENTATION

The topic of **Vision** has provided one of the most fertile fields of investigation both for brain theorists and for technologists constructing ANNs. Six articles in the road map **Mammalian Brain Regions**—RETINA, COLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL, “Thalamus,” VISUAL CORTEX: ANATOMICAL STRUCTURE AND MODELS OF FUNCTION, and VISUAL SCENE PERCEPTION—introduce various brain regions associated with vision. It is important to emphasize the role of “active vision” in gaining information relevant for animals and robots considered as real-time perception-action systems. This is a theme that is further developed in the road maps **Neuroethology and Evolution** and **Mammalian Motor Control**. Nonetheless, many articles in the present road map will analyze vision as the process of discovering from images what is present in the world: we may see active vision as more like the mode of vision employed by the “where/how” system described in VISUAL SCENE PERCEPTION, whereas “passive” vision may be closer to the role of the “what” pathway. DISSOCIATIONS BETWEEN VISUAL PROCESSING MODES explores the notion that the visual system has two kinds of jobs to do. One is to support visual cognition, the other is to drive visually guided behavior. Qualitative information about location may be adequate for cognition, but the sensorimotor function needs quantitative egocentrically calibrated spatial information to guide motor acts. The article reviews evidence from neurophysiology, neurological analysis of patients, and psychophysics that the two systems should be modeled as separate maps of visual space rather than as a single visual representation with two readouts. Moreover, spatial information can flow from the cognitive to the sensorimotor representation, but not in the other direction.

However, even “passive” vision is not so passive, since attentional mechanisms are constantly moving the eyes to foveate on items of particular relevance to the current interests of the organism. COLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL reviews the role of the superior colliculus in the control of gaze shifts (combined eye-head movements) and its possible involvement in the control of eye movements in 3D space (direction and depth). During attentive fixation, the “Vestibulo-Ocular Reflex” (VOR) and slow vergence maintain binocular foveal fixation to correct for body movements. When the task requires inspection of an eccentric stimulus, a complex synergy of coordinated movements comes into play. Such refixations typically involve a rapid combined eye-head movement (saccadic gaze shift) and often require binocular adjustment in depth (vergence). By virtue of its topographical organization, the superior colliculus has become a key area for experimental and modeling approaches to the question of how sensory signals can be transformed into goal-directed movements. Interestingly, the superior colliculus is not driven by visual input alone. Auditory and somatosensory cues are transformed to register with the visual map in the colliculus for the control of saccades. SENSOR FUSION picks up this theme of ways in which sensory information can be brought together in the brains of diverse

animals (snakes, cats, monkeys, and humans) and surveys biologically inspired technological implementations (such as the use of infrared to enhance vision). PURSUIT EYE MOVEMENTS takes us from saccadic “jumps” to those smooth eye movements involved in following a moving target. Current models of pursuit include “image motion” models, “target velocity” models, and models that address the role of prediction in pursuit. These models make no explicit reference to the neural structures that might be responsible, but the article analyzes the neural pathways for pursuit, stressing the importance of both visual areas of the cerebral cortex and oculomotor regions of the cerebellum.

The RETINA, the outpost of the brain that contains both light-sensitive receptors and several layers of neurons that “preprocess” these responses, transforms visual signals in a multitude of ways to code properties of the visual world such as contrast, color, and motion. The article suggests that much of the retina’s signal coding and structural detail is derived from the need to optimally amplify the signal and eliminate noise. But retinal circuitry is diverse. The exact details are probably related to the ecological niche occupied by the organism. In mammals, the retinal output branches into two pathways, the collicular pathway and the geniculostriate pathway. The destination of the former is the midbrain region known as the superior colliculus, discussed above. VISUAL CORTEX: ANATOMICAL STRUCTURE AND MODELS OF FUNCTION reviews features of the microcircuitry of the target of the geniculostriate pathway, the primary visual cortex (area V1), and discusses the physiological properties of cells in its different laminae. It then outlines several hypotheses as to how the anatomical structure and connections might serve the functional organization of the region. For example, a connectionist model of layer IVc of V1 demonstrated that the gradient of change in properties of the layer could indeed be replicated using dendritic overlap through the lower two-thirds of the IVc layer. However, it was insufficient to explain the continuous and sharply increasing field size and contrast sensitivity observable near the top of the layer. The article shows how this discrepancy led to new experiments and related changes in the model which resulted in a good replication of the actual physiological data and required only feedforward excitation. The article goes on to analyze the anatomical substrates for orientation specificity and for surround modulation of visual responses, and concludes by discussing the origins of patterned anatomical connections. OCULAR DOMINANCE AND ORIENTATION COLUMNS discusses further properties of cells in layer IVc of V1. When these cells are tested to see which eye drives them more strongly, it is found that ocular dominance takes the form of a zebra-stripe-like pattern of alternating dominance. Within this high-level organization are “hypercolumns” devoted to a particular retinotopic region of visual space, each hypercolumn being further refined into columns whose cells are best responsive to edges of the same specific orientation. The article also presents models for how these structures might form through self-organization during development. The article reviews data on the orientation specificity of cells of V1 and their columnar organization, and offers models for the way in which development may yield such features of cortical structure. GABOR WAVELETS AND STATISTICAL PATTERN RECOGNITION shows how the response properties of many cells in primary visual cortex may be better described by what are called “Gabor wavelets” than as simple edge detectors. Each Gabor wavelet responds best to patterns of a given spatial frequency and orientation within a given neighborhood. The article relates this notion to both biology and technology. The detection of edge information from within a visual scene is an essential component of visual processing. This processing is believed to be initiated in the primary visual cortex, where individual neurons are known to act as feature detectors of the orientation of edges within the visual scene. Individual neurons can have an *orientation preference* (which states that neuron’s preferred orientation of the

angle of edges) and *orientation selectivity* (which measures the neuron's sensitivity as a detector of orientation). **ORIENTATION SELECTIVITY** focuses on mechanisms of orientation selectivity in the visual cortex, arguing that the orientation preference of each neuron and the orderly orientation preference map in cortex are likely to be consequences of a pattern of feedforward convergence. However, the selectivity observed in steady-state and orientation dynamics experiments cannot be achieved by a purely feedforward model. Corticocortical inhibition is a crucial ingredient in the emergence of orientation selectivity in the visual cortex, while the relative importance of corticocortical excitation in enhancing orientation selectivity is still under investigation but appears to be more significant for the function of complex cells than for simple cells in V1. Moving beyond the orientation features of primary visual cortex, **INFORMATION THEORY AND VISUAL PLASTICITY** demonstrates some aspects of information theory that are relevant to relaying information in cortex and connects entropy-based methods, projection pursuit, and extraction of simple cells in visual cortex. **FEATURE ANALYSIS** offers a more general view of the characterization of visual features based on the redundancy of the visual signal and the transformation of the signal as it passes along the visual pathway. Describing a particular cell as an "x detector" implies that the cell responds when and only when that particular feature is present (e.g., an edge detector responds only in the presence of an edge), but the article argues that describing cells in the early visual system as "detectors" of any type of feature is misleading. Features are useful for describing natural images because the latter have massive informational redundancy. Image space itself is too vast to search directly. Feature analysis depends on the proposition that the search for particular objects can be concentrated in a subspace of image space, the feature space. Localized receptive fields in primary visual cortex provide the primitive basis set for the feature space of vision. These form the basis for the elaboration of neurons responding selectively to geometrical features in area TE of the inferotemporal cortex (IT), and these in turn form the basis for object recognition in different but overlapping areas of IT.

Given that cells in the early stages of the visual system, at least, provide a distributed (more or less retinotopic) set of "features" (in some suitably general sense, given the above caution, of patterns that yield the best response rather than patterns that yield the only response), the issue arises of how those features that correspond to a single object in the visual scene are bound together. **CONTOUR AND SURFACE PERCEPTION** introduces parallel interacting subsystems that follow complementary processing strategies. Boundary formation proceeds by spatially linking oriented contrast measures along smooth contour patterns, while perceptual surface attributes, such as lightness or texture, are derived from local ratio measures of image contrast of regions lying within contours. Mechanisms of both subsystems mutually interact to resolve initial ambiguities and to generate coherent representations of surface layout. Representations of intrinsic scene characteristics are constrained in terms of the consistency of the set of solutions, which often involve smoothness assumptions for correlated feature estimates. These consistency constraints are typically based on the laws of physical image generation. The article reviews fundamental approaches to computation of intrinsic scene characteristics and various neural models of boundary and surface computation. Each model involves lateral propagation of signals to interpolate and smooth sparse estimates.

ADAPTIVE RESONANCE THEORY (ART) bases learning on internal expectations. A pattern matching process (both for visual patterns and in other domains) compares an external input with the internal memory code for various patterns. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the

same or incorporate new information from matched portions of the current input. When the external world fails to match an ART network's expectations or predictions, a search process selects a new category, representing a new hypothesis about what is important in the present environment. **LAMINAR CORTICAL ARCHITECTURE IN VISUAL PERCEPTION** uses the LAMINART model (an extension of ART) to propose functional roles for cortical layers in visual perception. Neocortex has an intricate design that exhibits a characteristic organization into six distinct cortical layers, but few models have addressed the functional utility of the laminar organization itself in the control of behavior. LAMINART integrates data about visual perception and neuroscience for such processes as preattentive grouping and attention. It is suggested that the functional roles for cortical layers proposed here—binding together distributed cortical data through a combination of bottom-up adaptive filtering and horizontal associations, and modulating it with top-down attention—generalize, with appropriate specializations, to other forms of sensory and cognitive processing.

CORTICAL POPULATION DYNAMICS AND PSYCHOPHYSICS models cortical population dynamics to explain dynamical properties of the primate visual system on different levels, reaching from single neuron properties like selectivity for the orientation of a stimulus up to higher cognitive functions related to the binding and processing of stimulus features in psychophysical discrimination experiments. On the other hand, **SYNCHRONIZATION, BINDING AND EXPECTANCY** argues that the "binding" of cells that correspond to a given visual object may exploit another dimension of cellular firing, namely, the phase at which a cell fires within some overall rhythm of firing. The article presents data consistent with the proposal that the synchronization of responses on a time scale of milliseconds provides an efficient mechanism for response selection and binding of population responses. Synchronization also increases the saliency of responses because it allows for effective spatial summation in the population of neurons receiving convergent input from synchronized input cells. **VISUAL SCENE SEGMENTATION** tackles the segmentation of a visual scene into a set of coherent patterns corresponding to objects. Objects appear in a natural scene as the grouping of similar sensory features and the segregation of dissimilar ones. Studies in visual perception, in particular *Gestalt* psychology, have uncovered a number of principles for perceptual organization, such as proximity, similarity, connectedness, and relatedness in memory. Scene segmentation requires neural networks to address the binding problem. The temporal correlation approach is to encode the binding by the correlation of temporal activities of feature-detecting cells. A special form of temporal correlation is *oscillatory correlation*, where the basic units are neural oscillators. The article first reviews non-oscillatory approaches in scene segmentation, and then turns to oscillatory approaches. The temporal correlation approach is further developed in **DYNAMIC LINK ARCHITECTURE**, which views the brain's data structure as a graph composed of nodes connected by links, where both units and links bear activity variables changing on the rapid time scale of fractions of a second. The nodes play the role of symbolic elements. Dynamic links constitute the glue by which higher data structures are built up from more elementary ones.

Beyond the basic issue of how the visual scene is segmented (how visual elements are grouped) into possibly meaningful wholes lies the question of determining for a region so determined its color, motion, distance, shape, etc. These issues are addressed in the next set of articles. **COLOR PERCEPTION** stresses that color is not a local property inferred from the wavelength of light hitting a patch of retina but is a property of regions of space that depends both on the light they reflect and on the surrounding context. Our visual system "recreates" the world in the form of boundaries that contain surfaces, and color perception involves the perception of aspects

of these surfaces. Matching surfaces with the same reflectance properties in different parts of the visual scene or under different illuminants are the two problems of color constancy. In addition, wavelength signals can be used in the course of perceiving form or motion independent of their role in the subjective experience of color. **DIRECTIONAL SELECTIVITY** first reviews models of retinal direction selectivity (which contributes to oculomotor responses rather than motion perception). Older models depend on the way in which amacrine and other cells of the retina are connected to the ganglion cells, the retinal output cells. A newer model is based on the directionality of synaptic interactions on the dendrites of amacrine cells, involving a spatial asymmetry in the inputs and outputs of a dendrodendritic synapse, and its shunting inhibition. It is argued that development of this latter mechanism might involve Hebbian processes driven by spontaneous activity and light. Cortical directional selectivity (which does contribute to motion perception as well as the control of eye movements) involves many cortical regions. Directionally sensitive cells in primary visual cortex (V1) project to middle temporal cortex (MT) where directional selectivity becomes more complex, MT cells typically having larger receptive fields. From MT, the motion pathway projects to middle superior temporal cortex. Cortical directional selectivity has been modeled in three manners: as a spatially asymmetric excitatory drive followed by multiplication or squaring, via a spatially asymmetric nonlinear inhibitory drive, and through a spatially asymmetric linear inhibitory drive followed by positive feedback. This selectivity might involve Hebbian processes driven by spontaneous activity and binocular interactions. The issues in this article have some overlap with those presented in **MOTION PERCEPTION: ELEMENTARY MECHANISMS**, which emphasizes measurement of the direction and speed of movement of features in the 2D image linking successive views to infer *optic flow*, which is the pattern of image velocities that is projected onto the retina. The article discusses the cortical correlates of these various representations. **MOTION PERCEPTION: NAVIGATION** shows how, when an observer moves through the world, the optic flow can inform him about his own motion through space and about the 3D structure and motion of objects in the scene. This information is essential for tasks such as the visual guidance of locomotion through the environment and the manipulation and recognition of objects. This article focuses on the recovery of observer motion from optic flow. It includes strategies for detecting moving objects and avoiding collisions, discusses how optic flow may be used to control actions, and describes the neural mechanisms underlying heading perception. **GLOBAL VISUAL PATTERN EXTRACTION** continues the study of neural mechanisms which mediate between the extraction of local edge and contour information by orientation-selective simple cells in primary visual cortex (V1) and the high levels of cortical form vision in inferior temporal cortex (IT), where many neurons are sensitive to complex global patterns, including objects and faces. The ventral form vision pathway includes at least areas V1, V2, V4, TEO, and TE (the highest level of IT), raising the question of what processes occur at these intervening stages to transform local V1 orientation information into global pattern representations. Essentially the same question may be posed in cortical motion processing along the dorsal pathway comprising V1, V2, MT, MST, and higher parietal areas. V1 neurons extract only local motion vectors perpendicular to moving edge segments, while MST neurons are sensitive to complex optic flow patterns, including expansion. This article suggests answers to these analogous questions about transitions from local to global processing in both motion and form vision by focusing on intermediate levels of these two pathways, mainly V4 and MST.

PERCEPTION OF THREE-DIMENSIONAL STRUCTURE reviews various computational models for inferring an object's 3D structure from different types of optical information, such as shading, tex-

ture, motion, and stereo, and examines how the performance of these models compares with the capabilities and limitations of human observers in judging different aspects of 3D structure under varying viewing conditions. In particular, stereoscopic vision exploits the fact that points in a 3D scene will in general project to different positions in the images formed in the left and right eyes. The differences in these positions are termed *disparities*. The stereo *correspondence problem* is to identify which points in a pair of stereo images correspond to a single point in 3D space. Solving this problem allows the stereo pair to be mapped into a single representation, called a disparity map, that makes explicit the disparities of various points common to both images, thus revealing the distance of various visual elements from the observer. Depth perception is then completed by determining depth values for all points in the images. **STEREO CORRESPONDENCE** notes that various constraints have been used to help determine which features on the two eyes should be matched in inferring depth. These include compatibility of matching primitives, cohesivity, uniqueness, figural continuity, and the ordering constraint. Various neural network stereo correspondence algorithms are then reviewed, and the problems of surface discontinuities and uncorrelated points, and of transparency, are addressed. The article also reviews neurophysiological studies of disparity mechanisms.

A more abstract approach to the correspondence problem, from the perspective of computer vision rather than psychology or neurophysiology, is offered in **TENSOR VOTING AND VISUAL SEGMENTATION**. In 3D, as we have seen, surfaces are inferred from binocular images by obtaining depth hypotheses for points and/or edges. In image sequence analysis, the estimation of motion and shape starts with local measurements of feature correspondences, which gives noisy data for the subsequent computation of scene information. Hence, any salient structure estimator must be able to handle the presence of multiple structures and their interaction in the presence of noisy data. This article analyzes approaches to address early to midlevel vision problems, emphasizing the tensor voting methodology for the robust inference of multiple salient structures such as junctions, curves, regions, and surfaces from any combination of points, curve elements, and surface patch element inputs in 2D and 3D. The article describes two regularization formalisms, one that imposes certain physical constraints so that the search space can be constrained and algorithmically tractable, and another using a Bayesian formalism to transform an ill-posed problem into one of functional optimization.

PROBABILISTIC REGULARIZATION METHODS FOR LOW-LEVEL VISION offers regularization theory (cf. "Generalization and Regularization in Nonlinear Learning Systems") as a general mathematical framework to deal with the fact that the problem of inferring 3D structure from 2D images is *ill-posed*: there are many spatial configurations compatible with a given 2D image or set (motion sequence, stereo pair, etc.) of images. The issue then becomes to find which spatial configuration is most probable. We have already seen a number of constraints associated with stereo vision. Deterministic regularization theory defines a "cost function," which combines a measure of how close a spatial configuration comes to yielding the given image (set) with a measure of the extent to which the configuration violates the constraints, and then seeks that configuration which minimizes this cost. The present article emphasizes a more general probabilistic approach in which the "actual" field f and the observed field g are considered as realizations of random fields, with the reconstruction of f understood as an estimation problem. **MARKOV RANDOM FIELD MODELS IN IMAGE PROCESSING** views the task of image modeling as being one of finding an adequate representation of the intensity distribution of a given image. What is adequate often depends on the task at hand. The general properties of the local spatiotemporal structure of images or image sequences are characterized by a Mar-

kov random field (MRF) in which the probability distribution for the image intensity and a further set of other attributes (edges, texture, and region labels) at a particular location are conditioned on values in a neighborhood of pixels (picture elements or image points). The observed quantities are usually noisy, blurred images. The article presents five steps of MRF image modeling within a Bayesian estimation/inference paradigm, and provides a number of examples. Particular attention is paid to maximum a posteriori (MAP) estimates. MRF image models have proved versatile enough to be applied to image and texture synthesis, image restoration, flow field segmentation, and surface reconstruction.

KALMAN FILTERING; NEURAL IMPLICATIONS introduces Kalman filtering, which, under linear and Gaussian conditions, produces a recursive estimate of the hidden state of a dynamic system, i.e., one that is updated with each subsequent (noisy) measurement of the observed system. The article shows how Kalman filtering provides insight into visual recognition and the role of the cerebellum in motor control. In particular, it presents a hierarchically organized neural network for visual recognition, with each intermediate level of the hierarchy receiving two kinds of information: bottom-up information from the preceding level, and top-down information from the higher level. For its implementation, the model uses a multiscale estimation algorithm that may be viewed as a hierarchical form of the extended Kalman filter that is used to simultaneously learn the feedforward, feedback, and prediction parameters of the model on the basis of visual experiences in a dynamic environment. The resulting adaptive process involves a fast dynamic state-estimation process that allows the dynamic model to anticipate incoming stimuli, as well as a slow Hebbian learning process that provides for synaptic weight adjustments in the model.

IMAGING THE VISUAL BRAIN addresses functional brain imaging of visual processes, with emphasis on limits in spatial and temporal resolution, constraints on subject participation, and trade-offs in experimental design. The articles focus on retinotopy, visual motion perception and visual object representation, and voluntary modulation of attention and visual imagery, emphasizing some of the areas where modeling and brain theory might be testable using current imaging tools. **VISUAL ATTENTION** offers data and hypotheses for cortical mechanisms to overtly and covertly shift attention (i.e., with and without eye movements). Attention guides where to look next based on both bottom-up (image-based) and top-down (task-dependent) cues—and indeed, the anatomy of the visual system includes extensive feedback connections from later stages and horizontal connections within each layer. Vision appears to rely on sophisticated interactions between coarse, massively parallel, full-field preattentive analysis systems and the more detailed, circumscribed, and sequential attentional analysis system. The articles focus on the brain area involved in visual attention and then analyze a variety of relevant mechanisms. Yet, having stressed the way in which we normally take a number of shifts of attention to fully take in the details of a visual scene, it is intriguing to learn how much can be absorbed in a single fixation. **FAST VISUAL PROCESSING** notes that much information can be extracted from briefly glimpsed scenes, even at presentation rates of around 10 frames/s, a technique known as rapid sequential visual presentation (RSVP). Since interspike intervals for neurons are seldom shorter than 5 ms, the underlying algorithms should involve no more than about 20 sequential, though massively parallel, steps. There is an important distinction in neural computation between feedforward processing models and those with recurrent connections that allow feedback and iterative processing. Pure feedforward models (e.g., multilayer perceptrons, MLPs) can operate very quickly in parallel hardware. The article argues that even in systems that use extensive recurrent connections, the fastest behavioral responses may essentially depend on a single feedforward processing wave. It looks at how detailed measurements of processing speed can be combined with

anatomical and physiological constraints to constrain models of how the brain performs such computations.

There is a vast literature on pattern recognition in neural networks (see, for example, “Pattern Recognition” and “Concept Learning”). Here we discuss articles on face recognition and object recognition. The recognition of other individuals, and in particular the recognition of faces, is a major prerequisite for human social interaction and indeed has been shown to employ specific brain mechanisms. The ability to recognize people from their faces is part of a spectrum of related skills that include face segmentation (i.e., finding faces in a scene or image) and estimation of the pose, direction of gaze, and the person’s emotional state. **FACE RECOGNITION: NEUROPHYSIOLOGY AND NEURAL TECHNOLOGY** starts with a review of relevant neurophysiology. Brain injury can lead to prosopagnosia, the loss of ability to recognize individual faces, while leaving intact the ability to recognize general objects. Single-unit recordings in the IT cortex of macaque monkeys have revealed neurons with a high responsiveness to the presence of a face, an individual, or the expression on the face, and neural models for face recognition are reviewed in relation to such data. The article then focuses on computational theories that are inspired by neural ideas (see **DYNAMIC LINK ARCHITECTURE**; **GABOR WAVELETS AND STATISTICAL PATTERN RECOGNITION**) but that find their justification in the construction of successful computer systems for the recognition of human faces even when the gallery of possible faces is very large indeed. **FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM** provides a brief history of connectionist approaches to face recognition and surveys the broad range of tasks to which these models have been applied. The article relates the models to psychological theories for the subtasks of representing faces and retrieving them from memory, comparing human and model performance along these dimensions.

OBJECT RECOGNITION focuses on models of viewpoint-invariant object recognition that are constrained by psychological data on human object recognition. It presents three main approaches to object recognition—invariant based, model based, and appearance based—and analyzes the strengths of each of these in a framework of decision complexity, noting the trade-off between representations that emphasize invariance and those designed for discriminability. The analysis shows that it is unlikely for a single form of representation to satisfy all kinds of object recognition tasks a human or other visual animal may encounter. The article thus argues that a key ingredient in a comprehensive brain theory for object recognition is a computational framework that allows on-demand selection or adaptation of representations based on the current task and proposes a simple “first past the post” scheme (a temporal winner-take-all scheme) for self-selecting the most appropriate level of abstraction, given a finite set of available representations along a visual processing pathway.

OBJECT STRUCTURE, VISUAL PROCESSING emphasizes structure-processing tasks that call for separate treatment of various fragments of the visual stimulus, each of which spans only a fraction of the visual extent of the object or scene under consideration. Examples of structural tasks include recognition of part-part similarities, and identifying a region in an object toward which an action can be directed. After discussing object form processing in computer vision and relevant neurophysiological data on primate vision, the article focuses on two neuromorphic models of visual structure processing. The **JIM** model implements a recognition-by-components scenario based on geons (“geometrical elements,” which are generalized cylinders formed by moving a cross-section along a possibly curved axis). The **Chorus of Fragments** model exploits both the “what” and the “where” streams of visual cortex to recognize fragments no matter what their position, but then uses their approximate spatial relationships to see whether they together form cues for the recognition of an object. In particular, then, it

avoids the binding problem of explicitly linking neural activity related to a specific object as a prerequisite to analysis of that object's characteristics. (By contrast, **SYNCHRONIZATION, BINDING AND EXPECTANCY** argues that the brain does solve the binding problem, and does so by synchronization of neural firing for those neurons related to a single object.)

OBJECT RECOGNITION, NEUROPHYSIOLOGY reviews some theoretical approaches to object recognition in the context of mainly neurophysiological evidence. It also considers briefly the analysis of visual scenes. Scene analysis is relevant to object recognition because scenes may themselves be recognized initially at a holistic, object-like level, providing a context or "gist" that influences the speed and accuracy of recognition of the constituent objects. The article proposes that object recognition is based on a distributed, view-based representation in which objects are recognized on the basis of multiple, 2D-feature-selective neurons. Specialist cells appear to play a role in associating such feature combinations into certain nontrivial image transformations, coding for a certain percentage of all stimuli in a largely view-invariant manner. The article offers evidence that a convergent hierarchy is used to build invariant representations over several stages, and that at each stage lateral competitive processes are at work between the neurons. It is argued that the association of views of objects observed over the course of time could play a key role in building up object representations. The review focuses mainly on the "what" stream of IT cortex, seen as the center of object recognition. **VISUAL SCENE PERCEPTION** also brings in the "where/how" stream of parietal cortex as it analyzes how mechanisms that integrate schemas for recognition of different objects into the perception of some overall scene may be linked to the distributed planning of action. It also presents recent neurophysiology suggesting how the context of a natural scene may modify the response properties of neurons responsive to visual features. The article compares three approaches—the slide-box metaphor, short-term memory in the **VISIONS** system, and the visuospatial scratchpad—for creating a theory of how the visual perception of objects may be integrated with the perception of spatial layout. The first two stress a schema-theoretic approach, while the latter is strongly tied to visual neurophysiology and modeling in terms of quasi-neural attractor networks. The aim is to open the way to future research that will embed the study of visual scene perception in an action-oriented integration of IT and parietal visual systems.

Other Sensory Systems

AUDITORY CORTEX
AUDITORY PERIPHERY AND COCHLEAR NUCLEUS
AUDITORY SCENE ANALYSIS
ECHOLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS
ELECTROLOCATION
OLFACTORY BULB
OLFACTORY CORTEX
PAIN NETWORKS
PROSTHETICS, SENSORY SYSTEMS
SENSOR FUSION
SOMATOSENSORY SYSTEM
SOMATOTOPY: PLASTICITY OF SENSORY MAPS
SOUND LOCALIZATION AND BINAURAL PROCESSING

Here we analyze sensory systems other than vision—e.g., touch, audition, and pain. Moreover, when one sense cannot provide all the necessary information, complementary observations may be provided by another sense. For example, touch complements vision in placing a peg in a hole when the effector occludes the agent's view. Also, senses may offer competing observations, such as the competition between vision and the vestibular system in maintain-

ing balance (and its occasional side effect of seasickness). Another type of interplay between the senses is the use of information extracted by one sense to focus the attention of another sense, coordinating the two, as in audition cueing vision. **SENSOR FUSION** explores a number of ways sensory information is brought together in the brains of diverse animals (snakes, cats, monkeys, humans) and surveys biologically inspired technological implementations (such as the use of infrared to enhance vision). (See also "Collicular Visuomotor Transformations for Gaze Control" for an important example of sensor fusion—the transformation of auditory and somatosensory cues into a visual map for the control of rapid eye movements.)

The road map **Mammalian Brain Regions** introduced a number of regions linked to sensory systems other than vision, but we will now meet a number of related and additional topics as well. **SOMATOSENSORY SYSTEM** shows how the somatosensory system changes the tactile stimulus representation from a form more or less isomorphic to the stimulus to a completely distributed form in a series of partial transformations in successive subcortical and cortical networks. It further argues that the causal factors involved in body/object interactions are explicitly represented by an internal model in the pyramidal cells of somatosensory cortex that is crucial for haptic perception of proximal surroundings and for control of object manipulation. Somatotopy, a dominant feature of subdivisions of the somatosensory system, is defined by a topographic representation, or map, in the brain of sensory receptors on the body surface. **SOMATOTOPY: PLASTICITY OF SENSORY MAPS** shows that these orderly representations of cutaneous receptors in the spinal cord, lower brainstem, thalamus, and neocortex represent both the peripheral distribution of receptors and dynamic aspects of brain function. The article reviews evidence for somatosensory plasticity involving cortical reorganization after peripheral injury and as a result of training. The article analyzes the features of somatotopic maps that change, the contribution of subcortical changes to cortical plasticity, the mechanisms involved, and the functional consequences of sensory map changes. An important issue is the relation between the plasticity of the sensory and motor systems.

PAIN NETWORKS adds a new dimension to bodily sensation. The pain system encodes information on the intensity, location, and dynamics of tissue-threatening stimuli but differs from other sensory systems in its "emotional-motivational" factors (see also "Motivation"). In the pain system, these factors strongly modulate the relation between stimulus and felt response. At one extreme is allodynia, a state in which the slightest touch with a cotton wisp is agonizing. People display wide individual and trial-to-trial variability in the amount of pain reported following administration of calibrated noxious stimuli; pain sensation is subject to ongoing modulation by a complex of extrinsic (stimulus-generated) and intrinsic (CNS-generated) state variables. The article spells out how these act in the CNS as well as the periphery.

AUDITORY PERIPHERY AND COCHLEAR NUCLEUS spells out how the auditory periphery parcels out acoustic stimulus across hundreds of nerve fibers, and how the cochlear nucleus continues this process by creating multiple representations of the original acoustic stimulus. The article emphasizes monaural signal processing, whereas **SOUND LOCALIZATION AND BINAURAL PROCESSING** shows how information from the two ears is brought together. The article focuses on the use of interaural time difference (ITD) as one way to estimate the azimuthal angle of a sound source. It describes one biological model (ITD detection in the barn owl's brainstem) and two psychological models. The underlying idea is that the brain attempts to match the sounds in the two ears by shifting one sound relative to the other, with the shift that produces the best match assumed to be the one that just balances the "real" ITD. **AUDITORY CORTEX** stresses the crucial role that auditory cortex plays in the perception and localization of complex sounds, examining auditory

tasks vital for all mammals, such as sound localization, timbre recognition, and pitch perception. **AUDITORY SCENE ANALYSIS** discusses how the auditory system parses the acoustic mixture that reaches the ears of an animal to segregate a targeted sound source from the background of other sounds. The first stage, segmentation, decomposes the acoustic mixture into its constituent components. In the second stage, acoustic components that are likely to have arisen from the same environmental event are grouped, forming a perceptual representation (stream) that describes a single sound source. At the physiological level, segmentation corresponds (at least in part) to peripheral auditory processing, which performs a frequency analysis of the acoustic input, whereas the physiological substrate of auditory grouping is much less well understood. The article focuses on models that are at least physiologically plausible, while noting that other models of auditory scene analysis adopt a more abstract information processing perspective.

ECHOLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS provides us with a more detailed understanding of the auditory system in a very special class of mammals, the bats. Mustached bats emit echolocation (ultrasonic) pulses for navigation and for hunting flying insects. On the basis of the echo, prey must be detected and distinguished from the background clutter of vegetation, characterized as appropriate for consumption, and localized in space for orientation and prey capture. The bats emit ultrasonic pulses that consist of a long constant-frequency component followed by a short frequency-modulated component. Each pulse-echo combination provides a discrete sample of the continuously changing auditory scene. The auditory network contains two key design features: neurons that are sensitive to combinations of pulse and echo components, and computational maps that represent systematic changes in echo parameters to extract the relevant information.

Electrolocation is another sense that helps the animal locate itself in its world, but this time the animals are electric fishes rather than bats, and the signals are electrical rather than auditory. **ELECTROLOCATION** relates its topic to the general issue of mechanisms that facilitate the processing of relevant signals while rejecting noise, and of attentional processes that select which stimuli are to be attended to. Weakly electric fish generate an electrical field around their body and measure this field via electroreceptors embedded in the skin to “electrolocate” animate or inanimate targets in the environment. The article emphasizes a widespread but poorly understood characteristic of sensory processing circuits, namely, the presence of massive descending or feedback connections by which higher centers presumably modulate the operation of lower centers. Not only are response gain and receptive field

organization controlled by these descending connections, but there are adaptive filtering mechanisms that can reject stimuli that otherwise might mask critical functions. This use of stored sensory expectations for the cancellation or perhaps the identification of specific input patterns may yield insights into diverse neural circuits, including the cochlear nuclei and the cerebellum, in other species.

Two articles introduce data and models for the olfactory system (see also the road map **Mammalian Brain Regions**). **OLFACTORY BULB** describes the special circuitry involved in basic preprocessing, while **OLFACTORY CORTEX** presents a dynamical systems analysis of further olfactory processing. The olfactory bulb receives input from the sensory neurons in the olfactory epithelium and sends its outputs to the olfactory cortex, among other brain regions. The bulb was one of the first regions of the brain for which compartmental models of neurons were constructed, which led to some of the first computational models of functional microcircuits. **OLFACTORY BULB** gives an overview of olfactory bulb cells and circuits, current ideas about the computational functions of the bulb, and modeling studies to investigate these functions. The olfactory cortex is defined as the region of the cerebral cortex that receives direct connections from the olfactory bulb. It is the earliest cortical region to differentiate in the evolution of the vertebrate forebrain and the only region within the forebrain to receive direct sensory input. Moreover, the olfactory cortex has the simplest organization among the main types of cerebral cortex. **OLFACTORY CORTEX** thus views it as a model for understanding basic principles underlying cortical organization.

Finally, a very different view of sensory systems is provided by **PROSTHETICS, SENSORY SYSTEMS**, which discusses how information collected by electronic sensors may be delivered directly to the nervous system by electrical stimulation. After assessing the amenability of all sensory modalities (hearing, vision, touch, proprioception, balance, smell, and taste), the article focuses on auditory and visual prostheses. The great success story has been with cochlear implants. Here the article reviews improved temporospatial representations of speech sounds, combined electrical and acoustic stimulation in patients with residual hearing, and psychophysical correlates of performance variability. Since a prosthesis does not necessarily match natural neural encoding of a stimulus, the success of the prosthesis depends in part on the plasticity of the human brain as it remaps to accommodate this new class of signals. For example, the success of cochlear implants rests in part on the ability of auditory cortex to remap itself in a similar fashion to the remapping of somatosensory cortex described in **SOMATOTOPY: PLASTICITY OF SENSORY MAPS**.

II.8. Motor Systems

Robotics and Control Theory

ARM AND HAND MOVEMENT CONTROL
 BIOLOGICALLY INSPIRED ROBOTICS
 IDENTIFICATION AND CONTROL
 MOTOR CONTROL, BIOLOGICAL AND THEORETICAL
 POTENTIAL FIELDS AND NEURAL NETWORKS
 Q-LEARNING FOR ROBOTS
 REACTIVE ROBOTIC SYSTEMS
 REINFORCEMENT LEARNING IN MOTOR CONTROL
 ROBOT ARM CONTROL
 ROBOT LEARNING

ROBOT NAVIGATION SENSORIMOTOR LEARNING

As noted in the “Historical Fragment” section of Part I, the interchange between biology and technology that characterizes the study of neural networks is an outgrowth of work in *cybernetics* in the 1940s. One of the keys to cybernetics was control (the other was communication of the kind studied in information theory). It is thus appropriate that control theory should have become a major application area for neural networks as well as being a key concept of brain theory. The objective of control is to influence the behavior of a dynamical system in some desired fashion. The latter includes

maintaining the outputs of systems at constant values (regulation) or forcing them to follow prescribed time functions (tracking). Maintaining the altitude of an aircraft or the glucose level in the blood at constant values are examples of regulation; controlling a rocket to follow a given trajectory is an example of tracking. **MOTOR CONTROL, BIOLOGICAL AND THEORETICAL** sets forth the basic cybernetic concepts. A motor control system acts by sending motor commands to a controlled object, often referred to as “the plant,” which in turn acts on the local environment. The plant or the environment has one or more variables which the controller attempts to regulate. If the controller bases its actions on signals which are not affected by the plant output, it is said to be a feedforward controller. If the controller bases its actions on a comparison between desired behavior and the controlled variables, it is a feedback controller. “Motor Pattern Generation” provides a related perspective (see the road map **Motor Pattern Generators**).

The major advantage of negative feedback control, in which the controller seeks constantly to cancel the feedback error, is that it is a very simple, robust strategy that operates well without exact knowledge of the controlled object, and despite internal or external disturbances. The advantage of feedforward control is that it can, in the ideal case, give perfect performance with no error between the reference and the controlled variable. The main disadvantages are the practical difficulties in developing an accurate controller, and the lack of corrections for unexpected disturbances. **IDENTIFICATION AND CONTROL** explores the major strategy for developing an accurate controller, namely to “identify” the plant as belonging to (or more precisely, being well approximated by) a system obtained from a general family of systems by setting a key set of parameters (e.g., the coefficients in the matrices of a linear system). By coupling a controller to an identification procedure, one obtains an adaptive controller that can handle an unknown plant even if its dynamics are (slowly) changing. In both biology and many technological applications, nonlinearities and uncertainties play a major role, and linear approximations are not satisfactory. The article presents research using neural networks to handle these nonlinearities and examines the theoretical assumptions that have to be made when such networks are used as identifiers and controllers.

REINFORCEMENT LEARNING IN MOTOR CONTROL recalls the general theory introduced in “Reinforcement Learning” and proceeds to note its utility in motor control. Many motor skills are attained in the absence of explicit feedback about muscle contractions or joint angles. In contrast to supervised learning, such learning depends on “reinforcement” (or evaluative feedback; it need not involve pleasure or pain), which tells the learner whether or not, and possibly by how much, its behavior has improved, or provides an indication of success or failure. Instead of trying to match a standard of correctness, a reinforcement learning system tries to maximize the goodness of behavior as indicated by evaluative feedback. To do this, it has to actively try alternatives, compare the resulting evaluations, and use some kind of selection mechanism to guide behavior toward the better alternatives. **Q-LEARNING FOR ROBOTS** applies reinforcement learning techniques to robot control. Q-learning does not require a model of the robot-world interaction, and it uses learning examples in the form of triplets (situation, action, Q-value), where the Q-value is the *utility* of executing the action in the situation. Q-learning involves three different functions, *evaluation*, *memorization*, and *updating*. Heuristically adapted Q-learning has proved successful in applications such as obstacle avoidance, wall following, go-to-the-nest, etc., using neural-based implementations such as multilayer perceptrons trained with backpropagation, or self-organizing maps.

SENSORIMOTOR LEARNING explains how neural nets can acquire “models” of some desired sensorimotor transformation. A *forward model* is a representation of the transformation from motor commands to movements, in other words, a model of the controlled

object. An *inverse model* is a representation of the transformation from desired movements to motor commands, and so can be used as the controller for the controlled object. The managing of multiple models, each with their own range of applicability in given tasks, is given special attention. **ROBOT LEARNING** focuses on learning robot control, the process of acquiring a sensorimotor control strategy for a particular movement task and movement system. The article offers a formal framework within which to discuss robot learning in terms of the different methods that have been suggested for the learning of control policies, such as learning the control policy directly, learning the control policy in a modular way, indirect learning of control policies, imitation learning, and learning of motor control components. The article also reviews specific function approximation problems in robot learning, including neural network approaches. **ROBOT ARM CONTROL** addresses related issues concerning the availability of precise mappings from physical space or sensor space to joint space or motor space. Robot arm controllers are usually hierarchically structured from the lowest level of servomotors to the highest levels of trajectory generation and task supervision. In each case an actual motion is made to follow as closely as possible a commanded motion through the use of feedback. The difference lies in the coordinate systems used at each level. At least four coordinate spaces can be distinguished: the task space (used to specify tasks, possibly in terms of sensor readings), the workspace (6D Cartesian coordinates defining a position and orientation of the end-effector), the joint space (intrinsic coordinates determining a robot configuration), and the actuator space (in which actual motions are commanded). Correlational procedures carry out feature discovery or clustering and are often used to represent a given state space in a compact and topology-preserving manner, using procedures such as those described in “Self-Organizing Feature Maps.” Error-minimization procedures require explicit data on input-output pairs; their goal is to build a mapping from inputs to outputs that generalizes adequately using, e.g., the least-mean-squares (LMS) rule and backpropagation. In between both extremes lie procedures that use reinforcement learning to build a mapping that maximizes reward. **ARM AND HAND MOVEMENT CONTROL** discusses some of the most prominent regularities of arm and hand control, and examines computational and neural network models designed to explain them. The analysis reveals an interesting competition between explanations sought on the neural, biomechanical, perceptual, and computational levels that has created its share of controversy. Whereas some topics, such as internal model control, have gained solid grounding, the importance of the dynamic properties of the musculoskeletal system in facilitating motor control, the role of real-time perceptual modulation of motor control, and the balance between dynamical systems models versus optimal control-based models are still seen as offering many open questions.

BIOLOGICALLY INSPIRED ROBOTICS describes how modern robotics may learn from the way organisms are constructed biologically and how this creates adaptive behaviors. (I cannot resist noting here the acronym introduced by R. I. Damper, R. L. B. French, and T. W. Scutt, 2000, *ARBIB: An Autonomous Robot Based on Inspiration from Biology, Robotics and Autonomous Systems*, 31:247–274.) Research on autonomous robots based on inspiration from biology ranges from modeling animal sensors in hardware to guiding robots in target environments to investigating the interaction between neural learning and evolution in a variety of robot tasks. After reviewing the historical roots of the subject, the article provides a general introduction to biologically inspired robotics, with special emphasis on the ideas that the robot is situated in the world and that many complex behaviors are emergent properties of the collective effects of linking a variety of simple behaviors. **REACTIVE ROBOTIC SYSTEMS** provides a conceptual framework for robotics that is rooted in “Schema Theory” (q.v.) rather than sym-

bolic AI. Here, robot behavior is controlled by the activation of a collection of low-level primitive behaviors (schemas), and complex behavior emerges through the interaction of these schemas and the complexities of the environment in which the robot finds itself. This work was inspired in part by studies of animal behavior (see, e.g., “Neuroethology, Computational” and related articles discussed in the road maps on **Motor Pattern Generators** and **Neuroethology and Evolution**). However, the article not only shows the power of reactive robots in many applications, it also notes the utility of hybrid systems capable of using deliberative reasoning as well as reactive execution (which fits with an evolutionary view of the human brain in which reactive systems handle many functions but can be overruled or orchestrated by, e.g., the deliberative activities of prefrontal cortex).

ROBOT NAVIGATION examines how to get a mobile robot to move to its destination efficiently (e.g., along short trajectories) and safely (i.e., without colliding). If a target location is either visible or identified by a landmark (or sequence of landmarks), a simple stimulus-response strategy can be adopted. However, if targets are not visible, the robot needs a model (or map) of the environment encoding the spatial relationships between its present and desired locations. Sensor uncertainty, together with the inaccuracy of the robot’s actuators and the unpredictability of real environments, makes the design of mobile robot controllers a difficult task. It has thus proved desirable to endow robots with learning capabilities in order to acquire autonomously their control system and to adapt their behavior to never experienced situations. The article thus reviews neural approaches to localization, map building, and navigation. More specifically, **POTENTIAL FIELDS AND NEURAL NETWORKS** examines biological findings on the use of potential fields (which represent, e.g., the force field that drives the motor output of an animal or part of an animal, such as a limb) to characterize the control and learning of motor primitives. The notion of potential fields has also been used to model externally induced constraints as well as internally constructed sensorimotor maps for robot motion control. A robot can reach a stable configuration in its environment by following the negative gradient of its potential field. In this case, the configurations reached will be locally stable but may not be optimal with respect to some behavioral criterion. This deficit can be overcome either by incorporating a global motion planner or by using a harmonic function that does not contain any local minima. The article further indicates how potential field-based motion control can benefit from the use of ANN-based learning. There are links here to the more biological concerns of the articles “Cognitive Maps,” “Hippocampus: Spatial Models,” and “Motor Primitives.”

Motor Pattern Generators

CHAINS OF OSCILLATORS IN MOTOR AND SENSORY SYSTEMS
 COMMAND NEURONS AND COMMAND SYSTEMS
 CRUSTACEAN STOMATOGASTRIC SYSTEM
 GAIT TRANSITIONS
 HALF-CENTER OSCILLATORS UNDERLYING RHYTHMIC MOVEMENTS
 LOCOMOTION, INVERTEBRATE
 LOCOMOTION, VERTEBRATE
 LOCUST FLIGHT: COMPONENTS AND MECHANISMS IN THE MOTOR
 MOTOR PATTERN GENERATION
 MOTOR PRIMITIVES
 RESPIRATORY RHYTHM GENERATION
 SCRATCH REFLEX
 SENSORIMOTOR INTERACTIONS AND CENTRAL PATTERN GENERATORS
 SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS

MOTOR PATTERN GENERATION provides an overview of the basic building blocks of behavior (see “Motor Control, Biological and Theoretical” for more general background) to be expanded upon in many of the following articles. The emphasis is on rhythmic behaviors (such as flight or locomotion), but a variety of “one-off” motor patterns (as typified in a frog snapping at its prey) are also studied. The crucial notion is that a central pattern generator (CPG), an autonomous neural circuit, can yield a good “sketch” of a movement, but that the full motor pattern generator (MPG) augments the CPG with sensory input which can adjust the motor pattern to changing circumstances (e.g., the pattern of locomotion varies when going uphill rather than on level terrain, or when the animal carries a heavy load). **SENSORIMOTOR INTERACTIONS AND CENTRAL PATTERN GENERATORS** discusses both the impact of sensory information on CPGs and the influence of motor systems on sensory activity. It stresses that interaction between motor and sensory systems is pervasive, from the first steps of sensory detection to the highest levels of processing, emphasizing that descending motor commands are only acted upon by spinal circuits when these circuits integrate their intrinsic activity with all incoming information.

COMMAND NEURONS AND COMMAND SYSTEMS analyzes the extent to which an MPG may be activated alone or in concert with others through perceptual stimuli mediated by a single “command neuron” or by more diffuse “command systems.” Command functions provide the sensorimotor interface between sensory pattern recognition and localization, on the one side, and motor pattern generation on the other. For example, if a certain interneuron is stimulated electrically in the brain of a marine slug, the animal then displays a species-specific escape swimming behavior, although no predator is present. If in a toad a certain brain area of the optic tectum is stimulated in this manner, snapping behavior is triggered, although no prey is present. In both cases, a stimulus produces a rapid ballistic response.

MOTOR PRIMITIVES and SCRATCH REFLEX look at two behaviors (the former studied in frogs, the latter primarily in turtles) elicited by an irritant applied to the animal’s skin. In each case, the position at which the limb is aimed varies with the position of the irritant; there is somatotopic (i.e., based on place on the body) control of the reflex. In both frog and turtle, and thus more generally, spinal cord neural networks can by themselves generate complex sensorimotor transformations even when disconnected from supraspinal structures. Moreover, each reflex has different “modes.” To understand this, just think of scratching your lower back. As the scratch site moves higher, the positioning of the limb changes continuously with the position of the irritant until the irritant moves up so much that you make a discontinuous switch to the “over-the-shoulder” mode of back-scratching. The mode changes in these two articles may be compared to the **GAIT TRANSITIONS** (q.v.), discussed below. In any case, we see here two important issues: how is an appropriate pattern of action chosen, and how is the chosen pattern parameterized on the basis of sensory input? **MOTOR PRIMITIVES** advances the idea that CPGs construct spinal motor acts by recruiting a few motor primitives from a set encoded in the spinal cord. The best evidence comes from examination of wiping movements and microstimulation of frog spinal cord, where movements are constructed as a sequencing and combination of a collection of force-field motor primitives or fundamental elements. “Visuomotor Coordination in Frog and Toad” discusses how the frog’s motor acts may be assembled on the basis of visual input.

With this, we switch to articles in which the emphasis is on rhythmic behavior, with rather little concern for the spatial structure of the movement (for example, the discussion of locomotion will focus on coordinating the rhythms of the legs when the animal progresses straight ahead, rather than on how these rhythms are modified when the animal traverses uneven terrain or turns to avoid

an obstacle). **CRUSTACEAN STOMATOGASTRIC SYSTEM** analyzes specific circuits of identified neurons controlling the chewing (by teeth inside the stomach) behavior of crustaceans. Of particular interest is the finding that neuropeptides (see “Neuromodulation in Invertebrate Nervous Systems”) can change the properties of cells and the strengths of connections so that, e.g., a cell can become a pacemaker or a previously ineffective connection can come to exert a strong influence, and with this a network can dramatically change its overall behavior. Thus, the change of “mode” may be under the control of an explicit chemical “switch” of underlying cellular properties. Of course, in some systems, different input patterns of excitation and inhibition may enable a given circuit to act in one of several modes; while in other cases the change of mode may involve the transfer of control from one neural circuit to another. **LOCOMOTION, INVERTEBRATE** focuses on invertebrate locomotion systems for which quantitative modeling has been done, reviewing computer models of swimming, flying, crawling, and walking, paying special attention to the interaction of neural networks with the biomechanical systems they control. The article also reviews the use of biologically inspired locomotion controllers in robotics, stressing their distributed nature, their robustness, and their computational efficiency. Conversely, robots can serve as an important new modeling methodology for testing biological hypotheses. **LOCUST FLIGHT: COMPONENTS AND MECHANISMS IN THE MOTOR** narrows the focus to one specific invertebrate motor system. The article emphasizes the interactions of the intrinsic properties of flight neurons, the operation of complex circuits, and phase-specific proprioceptive input, all subject to the concentrations of circulating neuromodulators. Locust flight can adapt to the demands of a constantly changing sensory environment, and the flight system is flexible and able to operate despite severe ablations and then to recover from these lesions.

HALF-CENTER OSCILLATORS UNDERLYING RHYTHMIC MOVEMENTS looks at a set of minimal circuits for generating rhythmic behavior, starting with the half-center oscillator model first proposed to account for the observation that spinal cats (i.e., cats in which connections between brain and spinal cord had been severed) could produce stepping movements even when all sensory feedback from the animal’s motion was eliminated. The article shows the utility of models of this type in analyzing rhythms in invertebrates as well as vertebrates—the pelagic mollusk *Clione*, tadpoles, and lampreys—in terms of the intrinsic membrane properties of the component neurons interacting with reciprocal inhibition to initiate and sustain oscillation in these networks. **SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS** marks an important transition: from seeing how one network can oscillate to seeing how the oscillation of a series of networks can be coordinated. Experiments show that neural circuitry in isolated pieces of the spinal cord of lamprey (a jawless, primitive type of fish) can exhibit oscillations, and when these pieces constitute an intact spinal cord, they all oscillate with the same frequency but form a “traveling wave” with a phase relationship that in the complete fish would yield a wave of bending progressing down the fish from head to tail to yield the coordinated “wiggling” that yields swimming. The article reviews the interaction between experimentation and modeling stimulated by such findings. **RESPIRATORY RHYTHM GENERATION** presents several alternative models of breathing and evaluates them against mammalian data. These data point to the importance both of endogenous bursting neurons and of network interactions in generating the basic rhythm. In most models, rhythmogenesis is either pacemaker or network driven. The article reviews the data and these models, and then points the way to future models that clarify the integration of endogenous bursting with network interactions. **LOCOMOTION, VERTEBRATE** shows how neural networks in the spinal cord generate the basic rhythmic patterns necessary for vertebrate locomotion, while higher control centers

interact with the spinal circuits for posture control and accurate limb movements, and by sending higher-level commands such as stop and go signals, speed, and heading of motion. In mammals, evolution of the CPGs has been accompanied by important modifications of the descending pathways under the requirements of complex posture control and accurate limb movements, although the extent of the respective changes remains unknown. Computer models that combine neural models with biomechanical models are seen as having an important role to play in studying these issues. One example uses “genetic algorithms” to model the transition from a lamprey-like spinal cord that supports traveling waves to a salamander-like spinal cord that supports both traveling waves for swimming and “standing waves” for terrestrial locomotion, and shows how vision may modulate spinal activity to yield locomotion toward a goal (see also “Visuomotor Coordination in Salamander”).

CHAINS OF OSCILLATORS IN MOTOR AND SENSORY SYSTEMS abstracts from the specific circuitry to show how oscillators and their coupling can be characterized in a way that allows the proof of mathematical theorems about patterns of coordination. CPGs are discussed not only for the spinal cord of lamprey, but also for the crayfish swimmeret system and the leech network of swimming. In the context of locomotion, each oscillator is likely to be a local subnetwork of neurons that produces rhythmic patterns of membrane potentials. Since the details of the oscillators often are not known and are difficult to obtain, the object of the mathematics is to find the consequences of what is known, and to generate sharper questions to motivate further experimentation. **GAIT TRANSITIONS** also studies its topic (e.g., the transition from walking to running) from the abstract perspective of dynamical systems.

Mammalian Motor Control

ACTION MONITORING AND FORWARD CONTROL OF MOVEMENTS
ARM AND HAND MOVEMENT CONTROL

BASAL GANGLIA

CEREBELLUM AND MOTOR CONTROL

COLLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE
CONTROL

EQUILIBRIUM POINT HYPOTHESIS

EYE-HAND COORDINATION IN REACHING MOVEMENTS

GEOMETRICAL PRINCIPLES IN MOTOR CONTROL

GRASPING MOVEMENTS: VISUOMOTOR TRANSFORMATIONS

HIPPOCAMPUS: SPATIAL MODELS

IMAGING THE MOTOR BRAIN

LIMB GEOMETRY, NEURAL CONTROL

MOTOR CONTROL, BIOLOGICAL AND THEORETICAL

MOTOR CORTEX: CODING AND DECODING OF DIRECTIONAL
OPERATIONS

MOTONEURON RECRUITMENT

MUSCLE MODELS

OPTIMIZATION PRINCIPLES IN MOTOR CONTROL

PROSTHETICS, MOTOR CONTROL

PURSUIT EYE MOVEMENTS

REACHING MOVEMENTS: IMPLICATIONS FOR COMPUTATIONAL
MODELS

REINFORCEMENT LEARNING IN MOTOR CONTROL

RODENT HEAD DIRECTION SYSTEM

SENSORIMOTOR LEARNING

VESTIBULO-OCULAR REFLEX

Muscle transduces chemical energy into force and motion, thereby providing power to move the skeleton. Because of the intricacies of muscle microstructure and architecture, no comprehensive models are yet able to predict muscle performance completely. **MUSCLE MODELS** reviews three classes of models each fulfilling a more

narrowly defined objective, ranging from attempts to understand the molecular level (cross-bridge models) through lumped parameter mechanical models to input-output models of whole muscle behavior that can be used as part of a broader study of basic musculoskeletal biomechanics or issues of neural control. A motor neuron together with the muscle fibers that it innervates constitutes a motor unit, and each muscle is a composite structure whose force-generating components, the motor units, are typically heterogeneous. Such aggregates can produce much larger forces than a single motor unit. **MOTONEURON RECRUITMENT** shows how the motor units can be recruited in the service of reflexes, voluntary movement, and posture. The article considers mechanisms that compensate for muscle fatigue and yielding, models the possible role of Renshaw cells in linearization or equalization of motor neuron pool responses, and considers the possible role of cerebellum in control of motor neuron gain, as well as the roles of motor cortex in motor neuron recruitment.

PROSTHETICS, MOTOR CONTROL deals with the use of electrical stimulation to alter the function of motor systems, either directly or indirectly. The article presents three clinical applications. Therapeutic electrical stimulation is electrically produced exercise in which the beneficial effect occurs primarily off-line as a result of trophic effects on muscles and perhaps the CNS. Neuromodulatory stimulation is preprogrammed stimulation that directly triggers or modulates a function without ongoing control or feedback from the patient, and functional electrical stimulation (FES) provides precisely controlled muscle contractions that produce specific movements required by the patient to perform a task. The article also describes subsystems for muscle stimulation, sensory feedback, sensorimotor regulation, control systems, and command signals, most of which are under development to improve on-line control of FES.

MOTOR CONTROL, BIOLOGICAL AND THEORETICAL sets forth the basic cybernetic concepts. A motor control system acts by sending motor commands to a controlled object, often referred to as "the plant," which in turn acts on the local environment. The plant or the environment has one or more variables that the controller attempts to regulate. If the controller bases its actions on signals that are not affected by the plant output, it is said to be a feedforward controller. The full understanding of movement must rest on a full analysis of the integration of neural networks with the biomechanics of the skeletomuscular system. Nonetheless, much has been learned about limb control from a more abstract viewpoint, as the next four articles show. Optimization theory has become an important aid to discovering organizing principles that guide the generation of goal-directed motor behavior, specifying the results of the underlying neural computations without requiring specific details of the way those computations are carried out. **OPTIMIZATION PRINCIPLES IN MOTOR CONTROL** concedes that not all motor behaviors are necessarily optimal but argues that attempts to identify optimization principles can yield a useful taxonomy of motor behavior. The hypothesis is that in performing a motor task, the brain produces coordinated actions that minimize some measure of performance (such as effort, smoothness, etc.). The article reviews several studies in which such ideas were examined in the context of planar upper limb movements, comparing the purely kinematic minimum jerk model with the more dynamics-based minimum torque change model. But how does one go from a kinematic description of the movement of the hand to the pattern of muscle control that yields it? There are still many competing hypotheses. One approach seeks to find control systems that yield optimal trajectories in the absence of disturbances. Another starts from the observation that a muscle is like a controlled-length spring: set its length, and it will naturally return to the equilibrium length that was set. The **EQUILIBRIUM POINT HYPOTHESIS** builds on this a systems-level description of how the nervous system controls the

muscles so that a stable posture is maintained or a movement is produced. In this framework, the controller is composed of muscles and the spinal-based reflexes, and the plant is the skeletal system. The controller defines a force field that is meant to capture the mechanical behavior of the muscles and the effect of spinal reflexes. The equilibrium point hypothesis views motion as a gradual postural transition, and it is suggested that for the case of multijoint arm movements, one can predict the hand's motion if the supraspinal system smoothly shifts the equilibrium point from the start point to a target location. **GEOMETRICAL PRINCIPLES IN MOTOR CONTROL** considers a different transition, that from the spatial representation of a motor goal to a set of appropriate neuromuscular commands, which is in many respects similar to a coordinate transformation. (A word of caution: The matter is subtle because the brain rarely has neurons whose firing encodes a single coordinate. Consider, for example, retinotopic coding as distinct from the specific use of (x, y) or (r, θ) coordinates. Thus the issue is whether the activity in certain networks is better described as encoding one representation than another, such as those related to the eye rather than those related to the shoulder.) The article describes three types of coordinate system—end-point coordinates, generalized coordinates, and actuator coordinates—each representing a particular "point of view" on motor behavior, then examines the geometrical rules that govern the transformations between these classes of coordinates. It shows how a proper representation of dynamics may greatly simplify the transformation of motor planning into action. **LIMB GEOMETRY, NEURAL CONTROL** offers another perspective, starting from a discussion of the role of extrinsic and intrinsic coordinates when a human makes a movement. Multijointed coordination complicates the problem of motor control. Consider the case of arm movements. The activation of an elbow flexor will always contribute a flexor torque at the elbow, but the resulting elbow movement can be flexion, extension, or no motion at all, depending on the actively produced torque at the shoulder. Although in principle a coordinated motor action could be planned muscle by muscle, a more parsimonious solution is to plan more global goals at higher levels of organization and let the lower-level controllers specify the implementation details. The article reviews issues related to the kinematic aspects of limb geometry control for arm movements and for posture and gait.

Fast, coordinated movements depend on the nervous system being able to use copies of motor control signals (the corollary discharge) to compute expectations of how the body will move, rather than always waiting for sensory feedback to signal the current state of the body. **ACTION MONITORING AND FORWARD CONTROL OF MOVEMENTS** spells out three functions of corollary discharge. The stability of visual perception during eye movements was one of the first physiological applications proposed for an internal comparison between a movement and its sensory outcome. Second, goal-directed behavior implies that the action should continue until the goal has been satisfied, so that motor representations must involve not only forward mechanisms for steering the action but also mechanisms for monitoring its course and checking its completion. Third, similar processes have been postulated for actions aimed at complex and relatively long-term goals, for comparing the representation of the intended action to the actual action and compensating for possible mismatch between the two. Clearly, the effective use of corollary discharge rests on the brain having learned the relation between current state, motor command, and the movement that ensues. **SENSORIMOTOR LEARNING** explains how neural nets can acquire forward and inverse "models" of some desired sensorimotor transformation. The managing of multiple models, each with its own range of applicability in given tasks, is given special attention. The relevance of such models to the role of cerebellum (**CEREBELLUM AND MOTOR CONTROL**) is briefly noted, as is the idea that these models may act by controlling lower-level "Motor

Primitives” (q.v.). **REINFORCEMENT LEARNING IN MOTOR CONTROL**, which presents general learning strategies based on adaptive neural networks, is treated further in the road map **Robotics and Control Theory**.

With this background, we turn to articles primarily concerned with visually controlled behaviors for which neurophysiological data are available from the mammalian (and in many cases the monkey) brain, as well as behavioral and, in some cases, imaging data for humans. The road map takes us from basic unconscious behaviors to those involving skilled action. The vestibulo-ocular reflex (VOR) serves to stabilize the retinal image by producing eye rotations that counterbalance head rotations. Vestibular nuclei neurons are much more than a simple relay; their functions include multimodality integration, temporal signal processing, and adaptive plasticity. **VESTIBULO-OCULAR REFLEX** reviews the empirical data, as well as control-theoretic and neural network models for the neural circuits that mediate the VOR. These perform diverse computations that include oculomotor command integration, temporal signal processing, temporal pattern generation, and experience-dependent plasticity.

COLLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL analyzes the role of superior colliculus in the control of the rapid movement, called a saccade, of the eyes toward a target. The article touches on afferent and efferent mapping, target selection, visuomotor transformations in motor error maps, remapping models, and coding of dynamic motor error. The theme of remapping is pursued in **DYNAMIC REMAPPING**, which distinguishes “one-shot” remapping (updating the internal representation in one operation to compensate for an entire movement) from a continuous remapping process based on the integration of a velocity signal or the relaxation of a recurrent network. In both cases, the problem amounts to moving a hill of activity in neuronal maps. The article uses data on arm movements as well as saccades. Models can be constrained by considering deficits that accompany localized lesions in humans. These data not only provide valuable insights into the nature of remappings but they might also help bridge the gap between behavior and single-cell responses. **PURSUIT EYE MOVEMENTS** takes us from saccadic “jumps” to those smooth eye movements involved in following a moving target. Current models of pursuit vary in their organization and in the features of pursuit that they are designed to reproduce. Three main types of model are “image motion” models, “target velocity” models, and models that address the role of prediction in pursuit. However, these models make no explicit reference to the neural structures that might be responsible. The article thus analyzes the neural pathways for pursuit, stressing the importance of both visual areas of the cerebral cortex and oculomotor regions of the cerebellum, to set goals for future modeling.

IMAGING THE MOTOR BRAIN shows that the behavioral form and context of a movement are important determinants of functional activity within cortical motor areas and the cerebellum, stressing that functional imaging of the human motor system requires one to study the interaction of neurological and cognitive processes with the biomechanical characteristics of the limb. Neuroimaging shows that multiple neural systems and their functional interactions are needed to successfully perform motor tasks, encode relevant information for motor learning, and update behavioral performance in real time. The article discusses how evidence from functional imaging studies provides insight into motor automaticity as well as the role of internal models in movement.

Two articles explore the way in which the rat charts the spatial structure of its environment, using both “landmark cues” and a sense of its head orientation with respect to some key aspects of its environment. **HIPPOCAMPUS: SPATIAL MODELS** starts with the finding that single-unit recordings in freely moving rats have revealed “place cells” in fields CA3 and CA1 of the hippocampus,

so called because their firing is restricted to small portions of the rat’s environment (the corresponding place fields), but the firing properties of place cells change when the rat is placed in a new environment. The article focuses on data and models for the role of place cell firing in the rat’s navigation (see “Cognitive Maps” for a less neurophysiological approach to the same general issues). **RODENT HEAD DIRECTION SYSTEM** focuses on head direction cells in a number of brain areas that fire maximally when the rat’s head is pointed in a specific preferred direction, with a gradual falloff in firing as the heading departs from that direction. Head direction is not a simple reflection of sensory stimuli since, for example, the neural coding can be updated when the animal turns in the dark. The authors analyze such phenomena using attractor networks.

The next six articles are concerned with reaching and grasping. **MOTOR CORTEX: CODING AND DECODING OF DIRECTIONAL OPERATIONS** spells out the relation between the direction of reaching and changes in neuronal activity that have been established for several brain areas, including the motor cortex. The cells involved each have a broad tuning function, the peak of which denotes the “preferred” direction of the cell. A movement in a particular direction will engage a whole population of cells. It is found that the weighted vector sum of these neuronal preferences is a “population vector” that points in (close to) the direction of the movement for discrete movements in 2D and 3D space. Further observations link this population encoding to speed of movement as well as to preparation for movement. The present article addresses the question of how movement variables are encoded in the motor cortex and how this information could be used to drive a simulated actuator that mimics the primate arm. **ARM AND HAND MOVEMENT CONTROL** discusses some of the most prominent regularities of arm and hand control, and examines computational and neural network models designed to explain them. The analysis reveals the controversies engendered by competition between explanations sought on different levels—neural, biomechanical, perceptual, or computational. Although some topics, such as internal model control, have gained solid grounding, the importance of the dynamic properties of the musculoskeletal system in facilitating motor control, the role of real-time perceptual modulation of motor control, and the balance between dynamical systems models versus optimal control-based models are still seen as offering many open questions. **REACHING MOVEMENTS: IMPLICATIONS FOR COMPUTATIONAL MODELS** reviews a number of issues that are emerging from neurophysiological studies of motor control and stresses their implications for development of future models. Data on movement planning, trajectory generation, temporal features of cortical activity, and overlapping polymodal gradients are used to set challenges for computational models that will meet the demands of both functional competence and biological plausibility.

EYE-HAND COORDINATION IN REACHING MOVEMENTS focuses on possible mechanisms responsible for visually guiding the hand toward a point within the prehension space. Reaching at a visual target requires transformation of visual information about target position into a frame of reference suitable for the planning of hand movement. Accurate encoding of target location requires concomitant foveal and extraretinal signals. The most popular hypothesis to explain how trajectories are planned is that the trajectory is specified as a vector in the arm’s joint space, with joint angle variations controlled in a synergic way (temporal coupling). The motor command initially sent to the arm is based on an extrafoveal visual signal; at the end of the ocular saccade, the updated visual signal is used to adjust the ongoing trajectory. Because of consistent delays in sensorimotor loops, the rapid path corrections observed during reaching movements cannot be attributed to sensory information only but must rely on a “forward model” of arm dynamics. In any case, where this article focuses on how the hand is brought to a target, **GRASPING MOVEMENTS: VISUOMOTOR TRANSFORMA-**

TIONS emphasizes the neural mechanisms that control the shaping of the hand itself to grasp an object, noting the crucial preshaping of the hand during reaching prior to grasping the object. The analysis emphasizes the cooperative computation of visual mechanisms in parietal cortex with motor mechanisms in premotor cortex to integrate sensing and corollary discharge throughout the movement.

CEREBELLUM AND MOTOR CONTROL reviews a number of models of the role of the cerebellum in building “internal models” to improve motor skills. The article asserts that motor control and learning in the brain employ a modular approach in which multiple controllers coexist, with each controller suitable for one or a small set of contexts. The basic idea is that, to select the appropriate controller or controllers at each moment, each of the multiple in-

verse models is augmented with a forward model that determines the responsibility each controller should assume during movement. This view is exemplified in the MOSAIC (MODular Selection And Identification Control) model. Recent human brain imaging studies have started to accumulate evidence supporting multiple internal models of tools in the cerebellum. (One caveat: The article stresses the idea that the cerebellum provides complete motor controllers; other authors emphasize the idea that the cerebellum provides a corrective side path that learns how best to augment controllers located elsewhere in the brain.) Finally, BASAL GANGLIA reviews the structure of this system in terms of multiple loops, with special emphasis on those involved in skeletomotor and oculomotor functions. It also reviews the role of dopamine in motor learning and the mechanisms underlying Parkinson’s disease.

II.9. Applications, Implementations, and Analysis

Applications

BRAIN-COMPUTER INTERFACES
DECISION SUPPORT SYSTEMS AND EXPERT SYSTEMS
FILTERING, ADAPTIVE
FORECASTING
KALMAN FILTERING: NEURAL IMPLICATIONS
PROSTHETICS, MOTOR CONTROL
PROSTHETICS, NEURAL
PROSTHETICS, SENSORY SYSTEMS

The road map **Robotics and Control Theory** presents a number of applications of neural networks. Here we offer a representative (but by no means exhaustive) set of other applications, a list that can be augmented by the study of many other road maps. Examples include a variety of topics in vision and speech processing (see the road maps **Vision** and **Linguistics and Speech Processing**, respectively). As noted in the Preface, the discussion of applications of ANNs in areas from astronomy to steel making was a feature of the first edition of the *Handbook* that is not reproduced in the second edition.

Several articles review the various contributions of adaptive neural networks to signal processing. FILTERING, ADAPTIVE notes that adaptive filtering has found widespread use in noise canceling and noise reduction, channel equalization, cochannel signal separation, system identification, pattern recognition, fetal heart monitoring, and array processing. The parameters of an adaptive filter are adjusted to “learn” or track signal and system variations according to a task-specific performance criterion. The field of adaptive filtering was derived from work on neural networks and adaptive pattern recognition. An adaptive filter can be viewed as a signal combiner consisting of a set of adjustable weights (or coefficients represented by a polynomial) and an algorithm (learning rule) that updates these weights using the filter input and output, as well as other available signals. The filter may include internal signal feedback, whereby delayed versions of the output are used to generate the current output, and it may contain some nonlinear components. The single-layer perceptron is a well-known type of adaptive filter that has a binary output nonlinearity (see “Perceptrons, Adalines, and Back-propagation”). The article focuses on the most widely used adaptive filter architecture and describes in some detail two representative adaptive algorithms: the least-mean-square algorithm and the constant modulus algorithm. KALMAN FILTERING: NEURAL IMPLICATIONS then introduces Kalman filtering, a powerful idea rooted in modern control theory and adaptive signal processing. Under linear

and Gaussian conditions, the Kalman filter produces a recursive estimate of the hidden state of a dynamic system, i.e., one that is updated with each subsequent (noisy) measurement of the observed system, with the estimate being optimum in the mean-square-error sense. The Kalman filter provides an indispensable tool for the design of automatic tracking and guidance systems, and an enabling technology for the design of recurrent multilayer perceptrons that can simulate any finite-state machine. In the context of neurobiology, Kalman filtering provides insights into visual recognition and motor control. Related applications are discussed in FORECASTING. Neural nets, mostly of the standard backpropagation type, have been used with great success in many forecasting applications. This article looks at the use of neural nets for forecasting with particular attention to understanding when they perform better or worse than other technologies, showing how the success of neural networks in forecasting depends significantly on the characteristics of the process being forecast.

A decision support system is an information system that helps humans make a decision on a given problem, under given circumstances and constraints. Expert systems are information systems that contain expert knowledge for a particular problem area and perform inferences when new data are entered that may be partial or inexact. They provide a solution that is expected to be similar to the solution provided by experts in the field. DECISION SUPPORT SYSTEMS AND EXPERT SYSTEMS uses the collective term *decision system* to refer to either a decision support system or an expert system. The article discusses how neural networks can be employed in a decision system. Such systems help humans in their decision process and so should be comprehensible by humans. The article reviews results of connectionist-based decision systems. In particular, trainable knowledge-based neural networks can be used to accumulate both knowledge (rules) and data, building adaptive decision systems with incremental, on-line learning. (For further developments related to the construction of expert systems, see “Bayesian Networks” and the three articles on “Graphical Models.”)

BRAIN-COMPUTER INTERFACES discusses the use of on-line analysis of brainwaves to derive information about a subject’s mental state as a basis for driving some external action, such as selecting a letter from a virtual keyboard or moving a robotics device, providing an alternative communication and control channel that does not depend on the brain’s normal output pathway of peripheral nerves and muscles, which may be nonfunctional in some patients. The brainwave signals may be evoked potentials generated in response to external stimuli or components associated with sponta-

neous mental activity. Targets for current research include the extraction of local components of brain activity with fast dynamics that subjects can consciously control. The article reviews the challenge of developing classifiers that work while the subject operates a brain-actuated application, with ANNs providing robust approaches to on-line learning. These studies are complemented by a range of articles on prosthetics. PROSTHETICS, NEURAL provides an overview of the physical components that tend to be common to all neural prosthetic systems. It emphasizes the biophysical factors that constrain the sophistication of those interfaces. Electro-neural interfaces for both stimulation of and recording from neural tissue are analyzed in terms of biophysics and electrochemistry. It is also shown how the design of practical neural prostheses must address the systems hardware issues of power and data management and packaging. PROSTHETICS, SENSORY SYSTEMS focuses on sensory prostheses, in which information is collected by electronic sensors and delivered directly to the nervous system by electrical stimulation of pathways in or leading to the parts of the brain that normally process a given sensory modality. After assessing the amenability of all sensory modalities (hearing, vision, touch, proprioception, balance, smell, and taste) the article focuses on auditory and visual prostheses. The great success story has been with cochlear implants. Here the article reviews improved temporospatial representations of speech sounds, combined electrical and acoustic stimulation in patients with residual hearing, and psychophysical correlates of performance variability. Visual prostheses are still in their early days, with no general agreement on the most promising site to apply electrical stimulation to the visual pathways. The article reviews the cortical approach and the retinal approach. Finally, it is noted that since a prosthesis does not necessarily match natural neural encoding of a stimulus, the success of the prosthesis depends in part on the plasticity of the human brain as it remaps to accommodate this new class of signals. PROSTHETICS, MOTOR CONTROL deals with the subset of neural prosthetic interfaces that employ electrical stimulation to alter the function of motor systems, either directly or indirectly. The article presents three clinical applications. Therapeutic electrical stimulation is electrically produced exercise in which the beneficial effect occurs primarily off-line as a result of trophic effects on muscles and perhaps the CNS; neuromodulatory stimulation is preprogrammed stimulation that directly triggers or modulates a function without ongoing control or feedback from the patient; and functional electrical stimulation (FES) provides precisely controlled muscle contractions that produce specific movements required by the patient to perform a task. The article describes subsystems for muscle stimulation, sensory feedback, sensorimotor regulation, control systems, and command signals, most of which are under development to improve on-line control of FES. Electrical stimulation of the nervous system is also being used to treat other disorders, including spinal cord stimulation to control pain and basal ganglia stimulation to control parkinsonian dyskinesias.

To close this road map, we note the importance of using special-purpose VLSI chips to gain the full efficiency of artificial neural network in various applications. Such chips are among the methods for implementation of neural networks discussed in the next road map, **Implementation and Analysis**.

Implementation and Analysis

ANALOG VLSI IMPLEMENTATIONS OF NEURAL NETWORKS
BIOPHYSICAL MECHANISMS IN NEURONAL MODELING
BRAIN SIGNAL ANALYSIS
DATABASES FOR NEUROSCIENCE
DIGITAL VLSI FOR NEURAL NETWORKS
GENESIS SIMULATION SYSTEM
NEUROINFORMATICS

NEUROMORPHIC VLSI CIRCUITS AND SYSTEMS
NEURON SIMULATION ENVIRONMENT
NEUROSIMULATION: TOOLS AND RESOURCES
NSL NEURAL SIMULATION LANGUAGE
PHOTONIC IMPLEMENTATIONS OF NEUROBIOLOGICALLY INSPIRED NETWORKS
PROGRAMMABLE NEUROCOMPUTING SYSTEMS
SILICON NEURONS
STATISTICAL PARAMETRIC MAPPING OF CORTICAL ACTIVITY PATTERNS

Briefly, a neural network (whether an artificial neural network for technological application or a simulation of a biological neural network in computational neuroscience) can be implemented in three main ways: by programming a general-purpose electronic computer, by programming an electronic computer designed for neural net implementation, or by building a special-purpose device to emulate a particular network or parametric family of networks. We discuss these three approaches in turn, and then review a number of articles describing tools and methods for the analysis of brain signals and related activity.

NEUROSIMULATION: TOOLS AND RESOURCES reviews neurosimulators, i.e., programs designed to reduce the time and effort required to build models of neurons and neural networks. A neurosimulator requires, at the very least, a highly developed interface, a scalable design (e.g., through parallel hardware), and extensibility with new neural network paradigms. The review includes programs for modeling networks of biological neurons as well as programs for kinetic modeling of intracellular signaling cascades and regulatory genetic networks but does not cover connectionist simulators. It provides a general picture of the capabilities of several neurosimulators, highlighting some of the best features of the various programs, and also describes ongoing efforts to increase compatibility among the various programs. Compatibility allows models built with one neurosimulator to be independently evaluated and extended by investigators using different programs, thereby reducing duplication of effort, and also allows models describing different levels of complexity (molecular, cellular, network) to be related to one another. The next three articles present some of the methods necessary for efficient simulation of detailed models of single neurons (see, e.g., the articles “Axonal Modeling” and “Dendritic Processing” in the **Biological Neurons and Synapses** road map). BIOPHYSICAL MECHANISMS IN NEURONAL MODELING is a primer on biophysically detailed compartmental models of single neurons (see the road map **Biological Neurons and Synapses** for a fuller précis), but contributes to the topic of neurosimulators by illustrating examples of model definitions using the Surf-Hippo Neuron Simulation System, providing a minimal syntax that facilitates model documentation and analysis. GENESIS SIMULATION SYSTEM describes GENESIS (GENERAL NEURAL SIMULATION SYSTEM), which was developed to support “structurally realistic” simulations, computer-based implementations of models designed to capture the anatomical structure and physiological characteristics of the neural system of interest. GENESIS has been widely used for single-cell “compartmental” modeling but is also used for large network models, using libraries of ion channels and complete cell models, respectively. NEURON is a neurosimulator that was first developed for simulating empirically based models of biological neurons with extended geometry and biophysical mechanisms that are spatially nonuniform and kinetically complex. This functionality has been enhanced to include extracellular fields, linear circuits to emulate the effects of nonideal instrumentation, models of artificial (integrate-and-fire) neurons, and networks that can involve any combination of artificial and biological neuron models. NEURON SIMULATION ENVIRONMENT shows how these capabilities have been implemented so as to achieve computational efficiency

while maintaining conceptual clarity, i.e., the knowledge that what has been instantiated in the computer model is an accurate representation of the user's conceptual model. Where NEURON has been primarily used for detailed modeling of single neurons, NSL NEURAL SIMULATION LANGUAGE provides methods for simulating very large networks of relatively simple (artificial or biological simulation) neurons. NSL (pronounced "Nissl") models focus on modularity, a well-known software development strategy in dealing with large and complex systems. Full understanding of a system is gained both by simulating modules in isolation and by designing computer experiments that follow the dynamics of the interactions between the various modules. An NSL model can be described either by direct programming in NSLM, the NSL (compiled) Modeling language, or by using the Schematic Capture System (SCS), a visual programming interface to NSLM supporting the description of module assemblages. "Phase-Plane Analysis of Neural Nets" introduces the qualitative theory of differential equations in the plane for analyzing neural networks. Computational methods are a very powerful adjunct to this type of analysis. The article concludes with comments on numerical methods and software. Between them, the articles reviewed in this paragraph make clear the challenge of providing multilevel neurosimulation environments in which one can move effortlessly between the levels of schemas (functional decomposition of an overall behavior), large neural networks, detailed models of single neurons, and neurochemical models of synaptic plasticity. To be fully effective, such an environment will also need visualization tools, and the ability to access a database to provide experimental results for comparison with model-based predictions.

The next two articles address the *digital*, parallel implementation of neural networks. DIGITAL VLSI FOR NEURAL NETWORKS starts by looking at the differences between digital and analog design techniques, with a focus on analyzing cost-performance trade-offs in flexibility (Amdahl's Law), and then considers the use of standard VLSI processors in parallel configurations for ANN emulation. The Adaptive Solutions CNAPS custom digital ANN processor is then discussed to convey a sense of some of the issues involved in designing digital structures for ANN emulation. Although this chip is no longer produced, it is still being used and provides a good vehicle for understanding the trade-offs inherent in emulating neural structures digitally. Finally, the article looks at field programmable gate array (FPGA) technology as a promising vehicle for digital implementation of ANNs. PROGRAMMABLE NEUROCOMPUTING SYSTEMS emphasizes that the design of specialized digital neurocomputers has exploited three items common to many neural (ANN) algorithms to improve cost/performance: the limited numeric precision required; the inherently high data parallelism, where the same operation is performed across large arrays of data; and communication patterns restricted enough to allow broadcast buses or unidirectional rings to support parallel execution of many common neural network algorithms. However, in the future, the work of commercial design teams to incorporate multimedia-style kernels into the workloads they consider during the design of new microprocessors will have as a by-product the ability to dramatically improve performance for ANN algorithms. This suggests that in the future there will be greatly reduced interest in special-purpose neurocomputers but much attention to software strategies to optimize ANN performance on commercially available microprocessors.

However, the above three assumptions are not so useful in the implementation of detailed "compartmental" models of neurons. Here, attention has been paid to the design of highly special-purpose *analog* VLSI circuits. Digital VLSI assigns a different circuit to each bit of information that is to be stored and processed. Each circuit is driven to the limit so that it settles into a 0-state or a 1-state, passing through a linear voltage-current regime to get

from one saturation state to the other. Thus, if a synaptic weight is to be stored with eight-bit precision in digital VLSI, it requires eight such circuits. By contrast, the linear regime of a single circuit element on a VLSI chip can store data with about three bits of precision with far less "real estate" on the chip, and with far less power loss. The price, of course, is that precision cannot be guaranteed on the same scale as for digital circuits, but in many neural net applications, analog precision is more than adequate. ANALOG VLSI IMPLEMENTATIONS OF NEURAL NETWORKS provides an overview of the implementation of circuitry in analog VLSI, and then summarizes a number of technological implementations of such analog chips for ANNs. The article introduces the difference between the constraints imposed by the biological and silicon media and emphasizes that letting the silicon medium constrain the design of a system results in more efficient methods of computation. Special emphasis is given to five properties of a silicon synapse that are essential for building large-scale adaptive analog VLSI synaptic arrays. This article focuses on building neural network integrated circuits (IC), and especially on building connectionist neural network models. SILICON NEURONS takes the same implementation methodology into the realm of computational neuroscience. Biological neural networks are difficult to model because they are composed of large numbers of nonlinear elements and have a wide range of time constants. Simulation on a general-purpose digital computer slows dramatically as the number and coupling of elements increase. By contrast, silicon neurons operate in real time, and the speed of the network is independent of the number of neurons or their coupling. On the other hand, high connectivity still poses problems in 2D chip layouts, and the design of special-purpose hardware is a significant investment, particularly if it is analog hardware, since analog VLSI still lacks a general set of easy-to-use design tools. In any case, NEUROMORPHIC VLSI CIRCUITS AND SYSTEMS charts the virtues of using analog VLSI to build "neuromorphic" chips, i.e., chips whose design is based on the structure of actual biological neural networks. Biological systems excel at sensory perception, motor control, and sensorimotor coordination by sustaining high computational throughput with minimal energy consumption. Neuromorphic VLSI systems employ distributed and parallel representations and computation akin to those found in their biological counterparts. The high levels of system integration offered in VLSI technology make it attractive for the implementation of highly complex artificial neuronal systems, even though the physics of the liquid-crystalline state of biological structures is different from the physics of the solid-state silicon technologies. The article provides a basic foundation in device physics and presents a set of specific circuits that implement certain essential functions that exemplify the breadth possible within this design paradigm. However, VLSI-based neural networks have difficulty in scaling up or interconnecting multiple neural chips to incorporate large numbers of neuron units in highly interconnected architectures without significantly increasing the computational time. This motivates the use of optical interconnections. The success of optic fibers as media for telecommunications has been complemented by the use of holograms and spatial light modulators as mechanisms for storing and processing information via patterns of light (photonics) rather than patterns of electrons (electronics). The current state of photonic approaches to neural network implementation is charted in PHOTONIC IMPLEMENTATIONS OF NEUROBIOLOGICALLY INSPIRED NETWORKS, which provides a perspective on the use of holography as a technique for building adaptive connection matrices for ANNs, as well as earlier discussions of holography as a metaphor for the working of associative memory in actual brains. In photonic implementation of neurobiologically inspired networks, optical (free-space or through-substrate) techniques enable an increase in the number of neuron units and the interconnection complexity by using the

off-chip (third) dimension. This merging of optical and photonic devices with electronic circuitry provides additional features such as parallel weight implementation, adaptation, and modular scalability.

The remaining articles provide a number of perspectives on the analysis of data on the brain.

BRAIN SIGNAL ANALYSIS reviews applications of ANNs to brain signal analysis, including analysis of the EEG and MEG, the electromyogram (EMG), and computed tomographic (CT) images and magnetic resonance (MR) brain images, and to series of functional MR brain images (fMRI). Since most medical signals usually are not produced by variations in a single variable or factor, many medical problems, particularly those involving decision making, must involve a multifactorial decision process. In these cases, changing one variable at a time to find the best solution may never reach the desired objective, whereas multifactorial ANN approaches may be more successful. The review is organized according to the nature of brain signals to be analyzed and the role that ANNs play in the applications.

STATISTICAL PARAMETRIC MAPPING OF CORTICAL ACTIVITY PATTERNS describes the construction of statistical maps to test hypotheses about regionally specific effects like “activations” during brain imaging studies. Statistical parametric maps (SPMs) are image processes with voxel values that are, under the null hypothesis, distributed according to a known probability density function (usually Student’s T or F distributions), analyzing each and every voxel using any standard (univariate) statistical test. The resulting statistical parameters are assembled into an image, the SPM. $\text{SPM}\{T\}$ refers to an SPM comprising T statistics; similarly, $\text{SPM}\{F\}$ denotes an SPM of F statistics. SPMs are interpreted as spatially extended statistical processes by referring to the probabilistic behavior of stationary Gaussian fields. Unlikely excursions of the SPM are interpreted as regionally specific effects, attributable to the sensorimotor or cognitive process that has been manipulated experimentally.

NEUROINFORMATICS presents an integrated view of neuroinformatics that combines tools for the storage and analysis of neuroscience data with the use of computational models in structuring

masses of such data. In Europe, *neuroinformatics* is a term used to encompass the full range of computational approaches to brain theory and neural networks. In the United States, some people use the term neuroinformatics solely to refer to databases in neuroscience. Taking the perspective of the *Handbook*, this article sees the key challenge for neuroinformatics to be to integrate insights from synthetic data obtained from running a model with data obtained empirically from studying the animal or human brain. The problem is that the data, and thus the models, of neuroscience are so diverse. Neuroscience integrates anatomy, behavior, physiology, and chemistry, and studies levels from molecules to compartments and neurons up to biological neural networks and on to the behavior of organisms. The article thus presents an architecture for a federation of databases of empirical neuroscientific data in which results from diverse laboratories can be integrated. It further advocates a cumulative approach to modeling in neuroscience that facilitates the reusability (with appropriate changes) of modules within current neural models, with the pattern of re-use fully documented and tightly constrained by the linkage with this federation of databases. **DATABASES FOR NEUROSCIENCE** then focuses on the issues in constructing such databases. In order to be able to integrate such diverse sources, the various communities within the neurosciences must begin to develop standards for their community’s data. Neuroscientists use many different and incompatible data formats that do not allow for the free exchange of data, and the article stresses the need for standards for the description of the actual data (i.e., a formalized description of the metadata), possibly using extensible markup language (XML) technologies. (On a related theme, **NEUROSIMULATION: TOOLS AND RESOURCES** examines two of the enabling neurosimulation technologies that will allow modelers to compare and modify models, verify one another’s simulations, and extend models with their own tools.) One possible solution to integrating data from sources with heterogeneous data and representation is to extend the conventional wrapper-mediator architecture with domain-specific knowledge. The article concludes with analysis of a specific database of brain images (the fMRI Data Center) and a comprehensive table of neuroscience databases constructed to date.

Part III: Articles

Action Monitoring and Forward Control of Movements

Marc Jeannerod

Introduction

Monitoring its own output is thought to be a basic principle of functioning of the nervous system. This idea, inherited from the cybernetic era, and still operational now, is based on the notion of a comparison of the actual output of the system with the expected, or desired, output. In the domain of motor control, for example, it is assumed that each time the motor centers generate an outflow signal for producing a movement, a "copy" of this command (the "efference copy") is retained. The reafferent inflow signals generated by the movement (e.g., visual, proprioceptive) are compared with the copy. If a mismatch between the two types of signals is recorded, new commands are generated until the actual outcome of the movement corresponds to the desired movement. This comparison cannot be made, however, until the reafferent signals and the efference copy have been rendered compatible with one another. Proprioceptive signals, in principle, should be directly compatible with motor output (they arise from the same muscles and joints that are activated during the movement). Visual signals, by contrast, are generated in a set of coordinates quite different from those of motor output. Thus, a common set of coordinates must be computed to make the comparison useful.

The efference copy can only measure the performance error when the action comes to execution. In order to give this mechanism a predictive role in anticipating the effects of an action, one must assume the existence of a more complex "internal model" of that action. Such a model should be able to simulate the action generation process without waiting for the sensory reafference, or even without performing it. According to Wolpert, Ghahramani, and Jordan (1995), a combination of two processes is required: "The first process uses the current state estimate and motor command to predict the next state by simulating the movement dynamics with a forward model. The second process uses a model of the sensory output process to predict the sensory feedback from the current state estimate. The sensory error—the difference between actual and predicted sensory feedback—is used to correct the state estimate resulting from the forward model" (p. 1881). Several possible applications for this mechanism are reviewed in the following discussion.

Stability of Visual Perception and Target Localization

The stability of visual perception during eye movements was one of the first physiological applications proposed for an internal comparison between a movement and its sensory outcome. When one moves one's eyes across the visual field, objects tend to appear stationary in spite of their displacement on the retina; if, however, the same displacement is produced by an external agent (e.g., by gently pressing against the eye at the external canthus), objects no longer appear stationary. To account for this phenomenon, it has been conjectured that the command signals to the eye muscles are effective in remapping the visual scene and canceling out the visual displacement. In the absence of this signal, the visual displacement becomes visible. Sperry (1950) coined the term of "corollary discharge" (CD) for the centrally arising discharge that reaches the visual centers as a corollary of any command generated by the motor centers. In this way, the visual centers can distinguish the retinal displacement related to a self-generated movement from that produced by a moving scene. Visual changes produced by a movement of the eye are normally "canceled" by a CD of a corresponding size and direction. If, however, the CD is absent or does not

correspond to the visual changes (e.g., when the eye is pressed), these changes are not canceled and are read by the visual system as having their origin in the external world. The combination of the retinal signals and the extraretinal command signals (CD) thus produces a perceived stability of the visual world (Jeannerod, Kennedy, and Magnin, 1979). Signals arising from eye muscle proprioceptors also contribute to visual stability during eye movements (see Gauthier, Nommay, and Vercher, 1990). A CD type of regulation should in principle be more advantageous, however, because of its timing: a discharge propagating directly from the motor to the visual centers should be available to the visual system earlier than discharges arising from the periphery.

The same logic used for perceptual visual stability can also apply to egocentric localization of visual targets. The retinal position of a target cannot in itself be sufficient for its localization in space because, as the eyes move in the head, and the head moves on the trunk, several different retinal positions correspond to the same spatial locus. The spatial location of the target must therefore be reconstructed by combining eye/head position signals with the position of the target on the retina. The relationships between the retinal error signal (the position of the target on the retina) and the eye position signal were first formalized by Robinson (1975). In this influential model, the efference copy from eye position is derived from the output of a neural integrator that maintains the eye at a given position during fixation. It is this signal of actual eye position that is combined with retinal error to provide other motor systems (e.g., the arm) with the target location information. Eye movements are not generated on the basis of retinal error. Instead, the driving signal for the eye to reach the desired eye position relative to the head is the eye motor error signal. This signal is obtained by "subtracting" the actual change in eye position in orbit from the desired position. The movement stops when the motor error equals zero.

Guittton (1992) was able to directly demonstrate the dynamic nature of this process, by showing that output neurons from the superior colliculus—the tectoreticular (TR) neurons—code the change in eye motor error during the movement. Before the movement takes place, a TR neuron with a preferred vector corresponding to the desired eye position will be activated and will drive the eye movement generator. As the movement progresses and motor error is reduced, other TR neurons coding for smaller vectors will be activated until the error is zero. At this point, a TR neuron coding for a zero vector will be activated and fixation will be maintained. Guittton postulates that an internal representation of change in gaze position is generated and compared with the desired gaze position to yield instantaneous gaze motor error. If one assumes that this error is the parameter represented topographically on the collicular map, one can conceive how this signal will activate the proper sequence of TR neurons. Hence Guittton's hypothesis of a moving "hill" of activity shifting across the collicular map, from the caudal part where large vectors are encoded to the rostral part where fixation vectors are encoded. There are some difficulties with this model, however, notably with the timing of discharges in the superior colliculus which, in order to be suitable for coding motor error, should precede those of the eye movement generator.

Representation of Goals of Movements

Goal-directed behavior implies that the action should continue until the goal has been satisfied. Motor representations must therefore involve not only forward mechanisms for steering and directing the

action, but also mechanisms for monitoring its course and for checking its completion. This error correction mechanism implies a short-term storage of outflow information processed at each level of action generation. Because reafferent signals during execution of a movement are normally delayed with respect to the command signal, the comparison mechanism must look ahead in time and produce an estimate of the movement velocity corresponding to the command. The image of this estimated velocity is used for computing the actual position of the limb with respect to the target (Hoff and Arbib, 1992). It is only because the current state of the action is monitored on-line (rather than after the movement terminates), that corrections can be applied without delay as soon as the deviation of the current trajectory from the desired trajectory is detected. A subtle mechanism postulated by Miles and Evarts (1979) for the regulation of movements could be useful here. They pointed out that the discharge of muscle spindles in the agonist muscle during a movement (due to the co-activation of the gamma motoneurons) exactly fulfills the criterion for an efference copy that propagates "upward" and is an exact copy of the motor input sent to the alpha motoneurons. This signal could well be used for on-line comparison with incoming signals resulting from the limb movement.

It has been proposed that the information stored in the comparison process should encode, not joint rotations or kinematic parameters, but final configurations (of the body, of the moving segments, etc.) as they should arise at the end of the action. In other words, the goal of the action, rather than the action itself, would be represented in the internal model of the action. This hypothetical mechanism is supported by experimental arguments. Desmurget et al. (1995) recorded reach and grasp movements directed at a handle that had to be grasped with a power grip. When the orientation of the handle was suddenly changed at the onset of a movement, the arm smoothly shifted from the optimal configuration initially planned to reach the object to another optimal configuration corresponding to the object in its new orientation. The shift was achieved by simultaneous changes at several joints (shoulder abduction, wrist rotation), so that the final grasp was effected in the correct position. In this case the comparison between the desired and the actual arm position could be effected dynamically through a process similar to that which has been proposed to solve the problem of coordinate transformation during goal-directed movements. The position of an object in space is initially coded in extrinsic (e.g., visual) coordinates. In order to be matched by the moving limb, however, this position must be transferred into an intrinsic coordinate frame. If the position of the object in extrinsic coordinates and the position of the extremity of the limb in intrinsic coordinates coincide (that is, if these positions correspond to the same point in the two systems of coordinates), the action should logically be considered as terminated (see Carrozzo and Lacquaniti, 1994).

In addition to matching the movement trajectory to the representation of the intended movement, this mechanism has also other potential functions for the control of movements. The comparison between corollary and incoming signals might be used to produce a correspondence between the motor command and the amount of muscular contraction, even if the muscular plant is not linear. Other nonlinearities may also arise from interaction of the moving limb with external forces, especially if it is loaded (for a review, see Weiss and Jeannerod, 1998). This mapping problem, which is a critical factor for producing accurate limb movements, is less important for eye movements, where interactions with the external force field are minimal and where the load of the moving segment is constant. In this case, the pattern of command issued by the eye movement generator should unequivocally reflect the final desired position of the eye, that is, the position where the retinal error is zero.

Action Monitoring

At a still higher level, that of actions aimed at complex and relatively long-term goals, similar processes have been postulated for comparing the representation of the intended action to the actual action and compensating for possible mismatch between the two. Several studies, using brain imaging techniques, have focused on identifying neural structures that would fulfill the requirements for a comparison mechanism or an error detecting device. Carter et al. (1998) studied the activity of the anterior cingulate gyrus, a region lying on the medial cortical surface of the frontal lobe, in a letter detection task designed to increase error rates and manipulate response competition. Activity was found to increase during erroneous responses, but also during correct responses in conditions of high levels of response competition. They concluded that the anterior cingulate gyrus detects conditions under which errors are likely to occur, rather than errors themselves. This result suggests that action-monitoring mechanisms anticipate the occurrence of errors, by using internal models of the effects of the action on the world. In other words, the sensory consequences of an action are evaluated before they occur, even in conditions in which the action may not be executed. This mechanism can also become a powerful means of determining whether a sensory event is produced by our own action or by an external agent (and ultimately, if an action is self-produced or not). Blakemore, Rees, and Frith (1998) compared brain activity during the processing of externally produced tones and the processing of tones resulting from self-produced movements. They found an increase in the right inferior temporal lobe activity when the tones were externally produced, suggesting that this area would be inhibited by the volitional system in the self-produced condition. This result raises interesting questions about the possible consequences of a dysfunction of such a system. Increased activity in the primary auditory areas in the temporal lobe has been observed during auditory hallucinations in psychotic patients (Dierks et al., 1999). Hence, it is possible that a defective self-monitoring system would produce false attribution of one's own speech to an external source.

Road Map: Mammalian Motor Control

Related Reading: Collicular Visuomotor Transformations for Gaze Control; Consciousness, Neural Models of; Eye-Hand Coordination in Reaching Movements; Schema Theory; Sensorimotor Learning

References

- Blakemore, S. J., Rees, G., and Frith, C. D., 1998, How do we predict the consequences of our actions? A functional imaging study, *Neuropsychologia*, 36:521–529. ♦
- Carrozzo, M., and Lacquaniti, F., 1994, A hybrid frame of reference for visuomanual coordination, *Neuroreport*, 5:453–456.
- Carter, C. S., Braver, T. S., Barch, D. M., Botwinick, M. M., Noll, D., and Cohen, J. D., 1998, Anterior cingulate cortex, error detection and the online monitoring of performance, *Science*, 280:747–749. ♦
- Desmurget, M., Prablanc, C., Rossetti, Y., Arzi, M., Paulignan, Y., Urquizar, C., and Mignot, J. C., 1995, Postural and synergic control for three-dimensional movements of reaching and grasping, *J. Neurophysiol.*, 74:905–910.
- Dierks, T., Linden, D. E. J., Jandl, M., Formisano, E., Goebel, R., Lanferman, H., and Singer, W., 1999, Activation of Heschl's gyrus during auditory hallucinations, *Neuron*, 22:615–621.
- Gauthier, G. M., Nommay, D., and Vercher, J. L., 1990, The role of ocular muscle proprioception in visual localization of targets, *Science*, 249:58–61.
- Guittion, D., 1992, Control of eye-head coordination during orienting gaze shifts, *Trends Neurosci.*, 15:174–179.
- Hoff, B., and Arbib, M. A., 1992, A model of the effects of speed, accuracy and perturbation on visually guided reaching, in *Control of Arm Move-*

ment in Space (R. Caminiti, P. B. Johnson, and Y. Burnod, Eds.), *Experimental Brain Research*, Series 22, pp. 285–306. ♦
 Jeannerod, M., Kennedy, H., and Magnin, M., 1979, Corollary discharge: Its possible implications in visual and oculomotor interactions, *Neuropsychologia*, 17:241–258.
 Miles, F., and Evarts, E. V., 1979, Concepts of motor organization, *Annu. Rev. Psychol.*, 30:327–362.
 Robinson, D. A., 1975, Oculomotor control signals, in *Basic Mechanisms*

of Ocular Motility and Their Clinical Implications (G. Lennerstrand and P. Bach-y-Rita, Eds.), Oxford, UK: Pergamon, pp. 337–374.
 Sperry, R. W., 1950, Neural basis of the spontaneous optokinetic response produced by visual inversion, *J. Comp. Physiol. Psychol.*, 43:482–489.
 Weiss, P., and Jeannerod, M., 1998, Getting a grasp on coordination, *News in Physiological Science*, 13:70–75.
 Wolpert, D. M., Ghahramani, Z., and Jordan, M. I., 1995, An internal model for sensorimotor integration, *Science*, 269:1880–1882. ♦

Activity-Dependent Regulation of Neuronal Conductances

Larry F. Abbott and Eve Marder

Introduction

An enormous amount of both theoretical and experimental work has focused on the implications of activity-dependent synaptic plasticity for development, learning, and memory. Less attention has been paid to the fact that the intrinsic characteristics of individual neurons change during development (Spitzer and Ribera, 1998) and can be modified by activity (Franklin, Fickbohm, and Willard, 1992; Desai, Rutherford, and Turrigiano, 1999), yet these too play a vital role in shaping network function.

The electrical characteristics of a neuron depend on the numbers of channels of various types that are active within the cell membrane and on how these channels are distributed over the surface of the cell. The conventional approach to developing a conductance-based model is to attempt to measure all the ionic currents expressed by a neuron, describe them with Hodgkin-Huxley equations, and finally assemble the neuron model (Koch and Segev, 1998). This approach is based on the assumptions that individual neurons of a given class have the same ionic currents expressed at the same levels and that a neuron expresses the same currents whenever it is sampled under a specified set of experimental conditions. However, these assumptions appear to be contradicted by experimental evidence (see, e.g., Liu et al., 1998). In addition, it is rarely possible to make all of the measurements needed to construct a conductance-based model in this manner. Most neurons have many types of ion channels with complex spatial distributions. It is unlikely that all these currents and their spatial distributions can be measured. As a result, conventional conductance-based models depend on considerable hand-tuning of parameters. Each attempt to make neurons more biologically realistic is accompanied by a worsening of this problem. Hand-tuning a detailed, multicompartment, conductance-based model can be extremely time consuming and frustrating.

Neurons accomplish the feat of expressing appropriate numbers of ion channels in the relevant locations without running months of computer simulations. This suggests that a set of parameter adjustment mechanisms may allow neurons to self-tune their conductance densities to produce specific electrophysiological properties. A number of attempts have been made at incorporating such mechanisms into conductance-based neuron models (Bell, 1992; LeMasson, Marder, and Abbott, 1993; Siegel, Marder, and Abbott, 1994; Liu et al., 1998; Golowasch et al., 1999; Stemmler and Koch, 1999). Self-tuning activity-dependent models provide an alternative approach to modeling neurons and networks. This class of models does not assume that neurons necessarily have the same conductance densities over time, nor that individual neurons of a well-defined class are identical. Rather, they depend on simple negative feedback mechanisms to develop and maintain sets of con-

ductances that produce particular firing patterns and response characteristics. In these models, a second-messenger system, which may involve Ca^{2+} influx and a variety of Ca^{2+} sensors, guides the expression of the membrane conductances in a self-regulating manner.

Results

A Model Neuron with Self-Regulating Conductances

Self-regulating models are constructed by specifying a set of activity sensors and the rules by which they modify conductance densities. Experimental work indicates that the intracellular Ca^{2+} concentration is a good indicator of neuronal activity. Intracellular Ca^{2+} concentrations become elevated in response to activity and fall during inactive periods (Ross, 1989). Many Ca^{2+} -dependent cellular processes, controlled by a variety of Ca^{2+} sensors that monitor Ca^{2+} entry through different ion channels, can affect channel densities (Bito, Deisseroth, and Tsien, 1997; Barish, 1998; Finkbeiner and Greenberg, 1998). The intracellular Ca^{2+} concentration itself, or sensors of inward Ca^{2+} currents, can thus be used as feedback elements that monitor activity and change conductances. In general, when the neuron's activity is high, the activity-dependent rules decrease excitability, and when activity is low, they increase excitability by modifying appropriate conductances.

A model neuron constructed on these principles can start with almost any initial conductance densities and self-assemble a set of maximal conductances that produce particular intrinsic characteristics and patterns of activity. An example of a model spontaneously developing a set of maximal conductances that produces bursting behavior, starting from an initially silent state, is shown in Figure 1 (Liu et al., 1998). This illustrates but one example of the many ways that the model can generate bursting activity. Virtually any initial state leads ultimately to bursting, but, interestingly, the sets of conductances constructed by the model differ from trial to trial, although the final pattern of bursting is always similar to that shown in Figure 1. Thus, there is a non-unique map between maximal conductances and activity. The final set of conductances depends on initial conditions and is variable, even though the pattern of activity produced by the model is not.

A Model Network with Self-Regulating Conductances

Circuits of self-regulating neurons can self-assemble into functional circuits. This can be illustrated using a simplified model of the pyloric circuit of the crab stomatogastric ganglion (STG). The pyloric rhythm of the STG consists of alternating bursts of activity in several neurons, including the lateral pyloric (LP) and pyloric

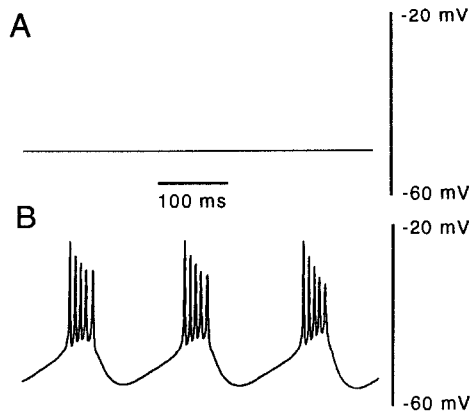


Figure 1. Self-assembly of a bursting model neuron (*B*) starting from different initial conditions (*A*). *A* and *B* represent the voltage traces at the beginning and end of the self-tuning process. (Adapted from Liu et al., 1998).

(PY) neurons, and the anterior burst (AB)/pyloric dilator (PD) pacemaker unit. The model shown in Figure 2 is a three-neuron circuit with individual neurons and synaptic connections similar to those of the LP and PY neurons and the AB/PD unit of the STG. Each model neuron consists of two compartments and has maximal conductances that are regulated by activity as described in the previous section.

When isolated from each other, the individual AB/PD, LP, and PY neurons of the model, like the neuron model shown in Figure 1, self-assemble their conductances. A novel feature of the circuit model is apparent when a realistic pattern of fixed synaptic connections is established between the model cells. In this case, the

entire network self-assembles to generate a pattern of activity similar to the triphasic rhythm recorded in the intact STG, and it can do so from any initial configuration of the maximal conductances of the three neurons (Golowasch et al., 1999). Interestingly, the intrinsic maximal conductances and responses properties of the individual model neurons are different if they self-assemble as a coupled circuit rather than in isolation. When assembled in a circuit, each of the model neurons ends up similar to its biological counterpart when acutely isolated, and the entire network generates realistic rhythmic activity. Thus, a cell-autonomous, activity-dependent regulatory rule is sufficient to self-assemble an entire circuit, at least in this example. It is not necessary to use any sensor that monitors the output of the whole circuit. Rather, each neuron takes care of its own activity, and the resultant circuit is tuned as a consequence of each cell's independent self-adjustment.

Comparison with Data from STG Organ Culture

Generation of the pyloric rhythm normally requires the presence of neuromodulatory substances released from axon terminals of the stomatogastric nerve (*stn*). If the *stn* is cut or blocked, rhythmic activity slows considerably or ceases. However, if the preparation is maintained over a period of days without *stn* modulatory input, rhythmic activity eventually resumes (Golowasch et al., 1999). Thus, it appears that prolonged removal of modulatory input alters the configuration of the pyloric circuit, allowing it to operate independently of the modulators that it normally requires.

The right side of Figure 2 shows the basic experimental result. Before blockade of the *stn*, the preparation shown on the right side of Figure 2A displayed a robust pyloric rhythm. Immediately following blockade of action potential conduction along the *stn*, the rhythm completely terminated (Figure 2B, right). However, when the block was maintained for approximately 24 hours, rhythmic pyloric activity resumed (Figure 2C, right). This recovery may be due, at least in part, to changes in the intrinsic properties of the neurons of the STG induced by the shift in activity following *stn* blockade, which allows the pyloric network to operate in the absence of neuromodulatory input.

This hypothesis can be studied, using the model discussed in the previous section, by including a proctolin conductance to simulate the effects of neuromodulators released by *stn* axons. The peptide proctolin is only one of many substances released from axon terminals of the *stn*, but it is a particularly potent modulator of the pyloric network. Figure 2 compares the behavior of the model with experimental results. Initially, the activity of the model network with the proctolin current included (Figure 2A, left) was similar to the pyloric activity of the experimental preparation with the *stn* intact (Figure 2A, right). To simulate the effects of blocking the *stn*, the proctolin conductance in the LP and AB/PD neurons was set to zero (Figure 2B, left). This immediately terminated the rhythmic activity of the model network, duplicating the effect of blocking the *stn* in the real preparation (Figure 2B, right). The suppression of the rhythm following the elimination of the proctolin conductance caused the activity-dependent conductance regulation mechanisms to modify the maximal conductances of the model neurons. This resulted in restoration of the pyloric rhythm in the model network after elimination of the proctolin conductance (Figure 2C, left), matching the natural resumption of the rhythm (Figure 2C, right).

It is important to stress that although the pyloric rhythms in Figures 2A and 2C look similar, they are produced by quite different cellular mechanisms. In Figure 2A, the existence of the rhythm depends on the presence of the modulatory proctolin current, while in Figure 2C the rhythms are produced in the absence of the modulator.

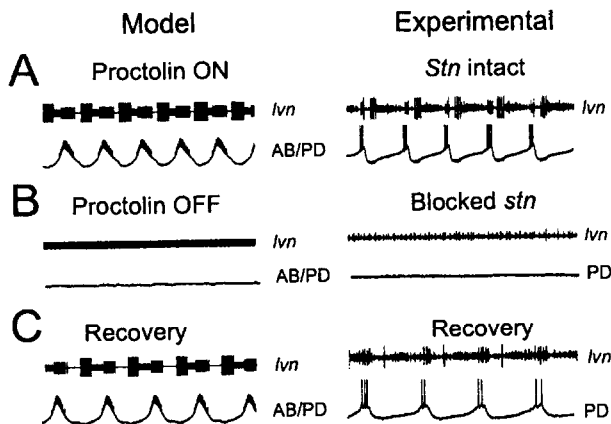


Figure 2. Comparison of a network model and experimental data. *A*, Control conditions in the model (left) and experiment (right). There is a triphasic motor pattern, revealed in the extracellular lvn recording. Intracellular PD recordings are also shown. In the model, the AB/PD and LP neurons contain a proctolin current. In the experiment, the modulatory inputs in the *stn* were left intact. *B*, Immediately after the modulatory inputs are removed, rhythmic activity is lost. Modulatory inputs are removed in the model by setting the proctolin current to zero. In the experiment, impulse activity in the *stn* was blocked to prevent the release of the neuromodulators. *C*, Activity eventually resumes. In the model, the activity-dependent sensors in each cell respond to the change in activity seen in *B*, and slowly modify the conductances of each of the model neurons, resulting in the recovery of rhythmic network activity, as occurred in the experimental case. (Adapted from Golowasch et al., 1999.)

Discussion

The homeostatic regulation of neuronal circuits is an essential element in their development and maintenance as functioning systems. This is often ignored in the construction of neural networks, because fixed parameters are adjusted and used to control network function. Biological systems do not have the luxury of using fixed constants, because of the continual recycling of the proteins from which they are built. As a consequence, biological networks are typically more robust than artificial networks, and they are self-assembling. The work described here is an attempt to incorporate these features into neural network models.

The model studies described have revealed several interesting consequences of activity-dependent regulation of conductances: (1) Conductance regulation stabilizes a model neuron against activity shifts caused by extracellular perturbations. (2) Intrinsic properties of model neurons are modified by sustained shifts in activity. (3) The regulation scheme described here, applied as a local regulator of channel density in a multicompartment model, can produce a realistic spatial distribution of conductances. (4) Regulation of the activity of individual neurons in a network may, in some case, be sufficient for the development and maintenance of a network pattern requiring coordination across neurons.

One of the most significant messages provided by models of conductance regulation is that the same mechanisms that develop and maintain membrane conductances are likely to modify these conductances in response to long-lasting changes in the activity of the neuron. Furthermore, different neurons, or the same neuron at different times, may exhibit similar characteristics and activity while expressing membrane conductances at quite different levels. These observations make it apparent that neuron models must be much more flexible and dynamic than has conventionally been the case.

Road Map: Biological Neurons and Synapses

Background: Ion Channels: Keys to Neuronal Specialization

Related Reading: Biophysical Mosaic of the Neuron

References

- Barish, M. E., 1998, Intracellular calcium regulation of channel and receptor expression in the plasmalemma: Potential sites of sensitivity along the pathways linking transcription, translation, and insertion, *J. Neurobiol.*, 37:146–157.
- Bell, A. J., 1992, Self-organization in real neurons: Anti-Hebb in “channel space,” in *Advances in Neural Information Processing Systems* (J. Moody, S. Hanson, and R. Lippmann), Eds., San Mateo, Morgan Kaufmann, pp. 59–66.
- Bito, H., Deisseroth, K., and Tsien, R. W., 1997, Ca^{2+} -dependent regulation in neuronal gene expression, *Curr. Opin. Neurobiol.*, 7:419–429.
- Desai, N. S., Rutherford, L. C., and Turrigiano, G. G., 1999, Plasticity in the intrinsic excitability of cortical pyramidal neurons, *Nature Neurosci.*, 2:515–520.
- Finkbeiner, S., and Greenberg, M. E., 1998, Ca^{2+} channel-regulated neuronal gene expression, *J. Neurobiol.*, 37:171–189. ♦
- Franklin, J. L., Fickbohm, D. J., and Willard, A. L., 1992, Long-term regulation of neuronal calcium currents by prolonged changes of membrane potential, *J. Neurosci.*, 12:1726–1735.
- Golowasch, J., Casey, M., Abbott, L. F., and Marder, E., 1999, Network stability from activity-dependent regulation of neuronal conductances, *Neural Computat.*, 11:1079–1096.
- Koch, C., and Segev, I., 1998, *Methods in Neuronal Modeling*, Cambridge, MA: MIT Press. ♦
- LeMasson, G., Marder, E., and Abbott, L. F., 1993, Activity-dependent regulation of conductances in model neurons, *Science*, 259:1915–1917.
- Liu, Z., Golowasch, J., Marder, E., and Abbott, L. F., 1998, A model neuron with activity-dependent conductances regulated by multiple calcium sensors, *J. Neurosci.*, 18:2309–2320.
- Ross, W. N., 1989, Changes in intracellular calcium during neuron activity, *Annu. Rev. Physiol.*, 51:491–506. ♦
- Siegel, M., Marder, E., and Abbott, L. F., 1994, Activity-dependent current distributions in model neurons, *Proc. Natl. Acad. Sci. USA*, 91:11308–11312.
- Spitzer, N. C., and Ribera, A. B., 1998, Development of electrical excitability in embryonic neurons: Mechanisms and roles, *J. Neurobiol.*, 37:190–197. ♦
- Stemmler, M., and Koch, C., 1999, How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate, *Nature Neurosci.*, 2:521–527.

Adaptive Resonance Theory

Gail A. Carpenter and Stephen Grossberg

Introduction

Principles derived from an analysis of experimental literatures in vision, speech, cortical development, and reinforcement learning, including attentional blocking and cognitive-emotional interactions, led to the introduction of adaptive resonance as a theory of human cognitive information processing (Grossberg, 1976a, 1976b). The theory has evolved as a series of real-time neural network models that perform unsupervised and supervised learning, pattern recognition, and prediction (Levine, 2000; Duda, Hart, and Stork, 2001). Models of unsupervised learning include ART 1 (Carpenter and Grossberg, 1987) for binary input patterns and fuzzy ART (Carpenter, Grossberg, and Rosen, 1991) for analog input patterns. ARTMAP models (Carpenter et al., 1992) combine two unsupervised modules to carry out supervised learning. Many variations of the basic supervised and unsupervised networks have since been adapted for technological applications and biological analyses.

Match-Based Learning, Error-Based Learning, and Stable Fast Learning

A central feature of all ART systems is a pattern matching process that compares an external input with the internal memory of an active code. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the same or incorporate new information from matched portions of the current input. If the search ends at a new code, the memory representation learns the current input. This *match-based learning* process is the foundation of ART code stability. Match-based learning allows memories to change only when input from the external world is close enough to internal expectations, or when something completely new occurs. This feature makes ART systems well suited to problems that require on-line learning of large and evolving databases.

Match-based learning is complementary to *error-based learning*, which responds to a mismatch by changing memories so as to reduce the difference between a target output and an actual output, rather than by searching for a better match. Error-based learning is naturally suited to problems such as adaptive control and the learning of sensorimotor maps, which require ongoing adaptation to present statistics. Neural networks that employ error-based learning include backpropagation and other multilayer perceptrons (MLPs) (Duda et al., 2001; see BACKPROPAGATION: GENERAL PRINCIPLES).

Many ART applications use *fast learning*, whereby adaptive weights converge to equilibrium in response to each input pattern. Fast learning enables a system to adapt quickly to inputs that occur rarely but that may require immediate accurate recall. Remembering details of an exciting movie is a typical example of learning on one trial. Fast learning creates memories that depend on the order of input presentation. Many ART applications exploit this feature to improve accuracy by voting across several trained networks, with voters providing a measure of *confidence* in each prediction.

Coding, Matching, and Expectation

Figure 1 illustrates a typical ART search cycle. To begin, an input pattern I registers itself as a short-term memory activity pattern x across a field of nodes F_1 (Figure 1A). Converging and diverging pathways from F_1 to a coding field F_2 , each weighted by an adaptive long-term memory trace, transform x into a net signal vector T . Internal competitive dynamics at F_2 further transform T , generating a compressed code y , or *content-addressable memory*. With

strong competition, activation is concentrated at the F_2 node that receives the maximal $F_1 \rightarrow F_2$ signal; in this *winner-take-all* mode, only one code component remains positive (see WINNER-TAKE-ALL NETWORKS).

Before learning can change memories, ART treats the chosen code as a *hypothesis*, which it tests by matching the *top-down expectation* of y against the input that selected it (Figure 1B). Parallel specific and nonspecific feedback from F_2 implements matching as a real-time locally defined network computation. Nodes at F_1 receive both learned excitatory signals and unlearned inhibitory signals from F_2 . These complementary signals act to suppress those portions of the pattern I of bottom-up inputs that are not matched by the pattern V of top-down expectations. The residual activity x^* represents a pattern of *critical features* in the current input with respect to the chosen code y . If y has never been active before, $x^* = x = I$, and F_1 registers a perfect match.

Attention, Search, Resonance, and Learning

If the matched pattern x^* is close enough to the input I , then the memory trace of the active F_2 code converges toward x^* . The property of encoding an *attentional focus* of critical features is key to code stability. This learning strategy differentiates ART networks from MLPs, which typically encode the current input rather than a matched pattern, and hence employ slow learning across many input trials to avoid catastrophic forgetting.

ART memory search begins when the network determines that the bottom-up input I is too novel or unexpected with respect to the active code to satisfy a matching criterion. The search process resets the F_2 code y before an erroneous association to x^* can form (Figure 1C). After reset, medium-term memory within the $F_1 \rightarrow F_2$ pathways (Carpenter and Grossberg, 1990) biases the network against the previously chosen node, so that a new code y^* may be chosen and tested (Figure 1D).

The ART matching criterion is determined by a parameter ρ called *vigilance*, which specifies the minimum fraction of the input that must remain in the matched pattern in order for resonance to occur. Low vigilance allows broad generalization, coarse categories, and abstract memories. High vigilance leads to narrow generalization, fine categories, and detailed memories. At maximal vigilance, category learning reduces to exemplar learning. While vigilance is a free parameter in unsupervised ART networks, in supervised networks vigilance becomes an internally controlled variable that triggers a search after rising in response to a predictive error. Because vigilance then varies across learning trials, the memories of a single ARTMAP system typically exhibit a range of degrees of refinement. By varying vigilance, a single system can recognize both abstract categories, such as faces and dogs, and individual examples of these categories.

Supervised Learning and Prediction

An ARTMAP system includes a pair of ART modules, ART_a and ART_b (Figure 2). During supervised learning, ART_a receives a stream of patterns $\{a^{(n)}\}$ and ART_b receives a stream of patterns $\{b^{(n)}\}$, where $b^{(n)}$ is the correct prediction given $a^{(n)}$. An associative learning network and a vigilance controller link these modules to make the ARTMAP system operate in real time, creating the minimal number of ART_a recognition categories, or *hidden units*, needed to meet accuracy criteria. A minimax learning rule enables ARTMAP to learn quickly, efficiently, and accurately as it conjointly minimizes predictive error and maximizes code compression in an on-line setting. A *baseline vigilance* parameter $\bar{\rho}_a$ sets the minimum matching criterion, with smaller $\bar{\rho}_a$ allowing broader categories to form. At the start of a training trial, $\rho_a = \bar{\rho}_a$. A predictive failure at ART_b increases ρ_a just enough to trigger a search, through a feedback control mechanism called *match tracking*. A

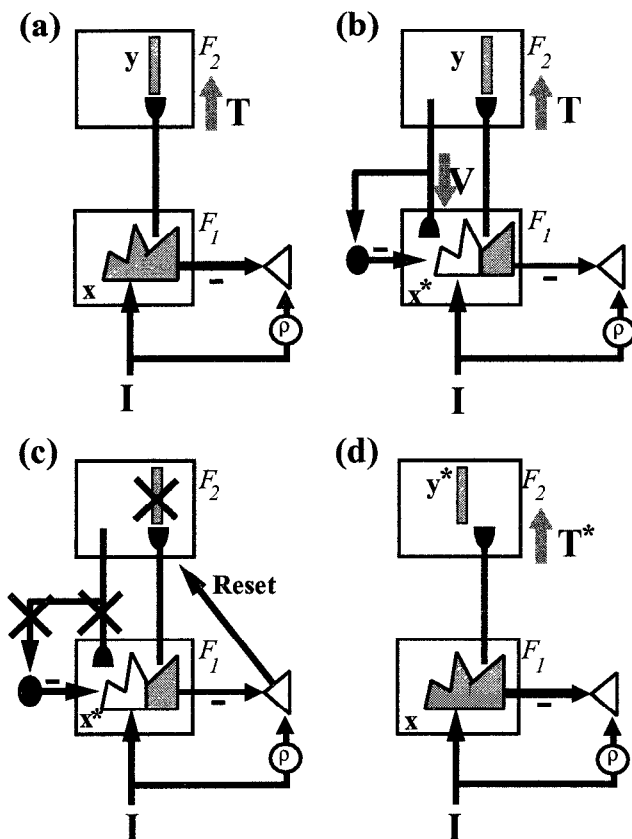


Figure 1. An ART search cycle imposes a matching criterion, defined by a dimensionless vigilance parameter ρ , on the degree of match between a bottom-up input I and the top-down expectation V previously learned by the F_2 code y chosen by I . See text for discussion of A through D sequence.

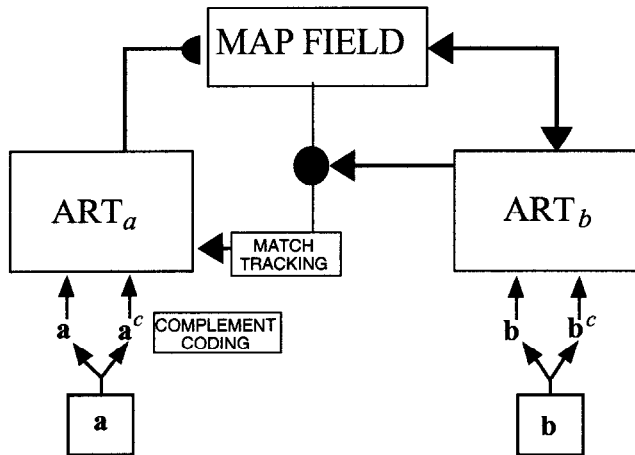


Figure 2. The general ARTMAP network for supervised learning includes two ART modules. For classification tasks, the ART_b module may be simplified.

newly active code focuses attention on a different cluster of input features, and checks whether these features are better able to predict the correct outcome. Match tracking allows ARTMAP to learn a prediction for a rare event embedded in a cloud of similar frequent events that make a different prediction.

ARTMAP employs a preprocessing step, called *complement coding*, which, by normalizing input patterns, solves a potential category proliferation problem (Carpenter et al., 1991). Complement coding doubles the number of input components, presenting to the network both the original feature vector and its complement. In neurobiological terms, complement coding uses both on-cells and off-cells to represent an input pattern. The corresponding on-cell portion of a weight vector encodes features that are consistently present in category exemplars, while the off-cell portion encodes features that are consistently absent. Small weights in complementary portions of a category representation encode as uninformative those features that are sometimes present and sometimes absent.

Distributed Coding

Winner-take-all activation in ART networks supports stable coding but causes category proliferation when noisy inputs are trained with fast learning. In contrast, distributed McCulloch-Pitts activation in MLPs promotes noise tolerance but causes catastrophic forgetting with fast learning (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS). *Distributed ART* (dART) models are designed to bridge these two worlds: distributed activation enhances noise tolerance, while new system dynamics retain the stable learning capabilities of winner-take-all ART systems (Carpenter, 1997). These networks automatically apportion learned changes according to the degree of activation of each coding node, which permits fast as well as slow distributed learning without catastrophic forgetting.

New learning laws and rules of synaptic transmission in the re-configured dART network (Figure 3) sidestep computational problems that occur when distributed coding is imposed on the architecture of a traditional ART network (Figure 1). The critical design element that allows dART to solve the catastrophic forgetting problem of fast distributed learning is the *dynamic weight*. This quantity equals the rectified difference between coding node activation and an *adaptive threshold*, thereby combining short-term and long-term memory in the network's fundamental computational unit.

Thresholds τ_{ij} in paths projecting directly from an input field F_0 to a coding field F_2 obey a *distributed instar* (dInstar) learning law,

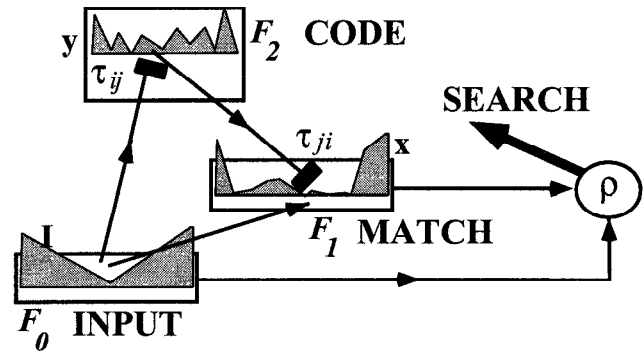


Figure 3. A distributed ART (dART) architecture retains the stability of WTA ART networks but allows the F_2 code to be distributed across arbitrarily many nodes.

which reduces to an instar law when coding is winner-take-all. Rather than adaptive gain, learning in the $F_0 \rightarrow F_2$ paths resembles the *redistribution of synaptic efficacy* (RSE) observed by Markram and Tsodyks (1996) at neocortical synapses. In these experiments, pairing enhances the strength, or efficacy, of synaptic transmission for low-frequency test inputs, but fails to enhance, and can even depress, synaptic efficacy for high-frequency test inputs. In the dART learning system, RSE is precisely the computational dynamic needed to support real-time stable distributed coding.

Thresholds τ_{ji} in paths projecting from the coding field F_2 to a matching field F_1 obey a distributed outstar (dOutstar) law, which realizes a principle of atrophy due to disuse to learn the network's expectations with respect to the distributed coding field activation pattern. As in winner-take-all ART systems, dART compares top-down expectation with the bottom-up input at the matching field, and quickly searches for a new code if the match fails to meet the vigilance criterion.

Discussion: Applications, Rules, and Biological Substrates

ART and dART systems are part of a growing family of self-organizing network models that feature attentional feedback and stable code learning. Areas of technological application include industrial design and manufacturing, the control of mobile robots, face recognition, remote sensing land cover classification, target recognition, medical diagnosis, electrocardiogram analysis, signature verification, tool failure monitoring, chemical analysis, circuit design, protein/DNA analysis, three-dimensional visual object recognition, musical analysis, and seismic, sonar, and radar recognition (e.g., Caudell et al., 1994; Griffith and Todd, 1999; Fay et al., 2001). A book by Serrano-Gotarredona, Linares-Barranco, and Andreou (1998) discusses the implementation of ART systems as VLSI microchips. Applications exploit the ability of ART systems to learn to classify large databases in a stable fashion, to calibrate confidence in a classification, and to focus attention on those featural groupings that the system deems to be important based on experience. ART memories also translate to a transparent set of IF-THEN rules that characterize the decision-making process and may be used for feature selection.

ART principles have further helped explain parametric behavioral and brain data in the areas of visual perception, object recognition, auditory source identification, variable-rate speech and word recognition, and adaptive sensorimotor control (e.g., Levine, 2000; Page, 2000). One area of recent progress concerns how the neocortex is organized into layers, clarifying how ART design prin-

ciples are found in neocortical circuits (see LAMINAR CORTICAL ARCHITECTURE IN VISUAL PERCEPTION).

Pollen (1999) resolves various past and current views of cortical function by placing them in a framework he calls *adaptive resonance theories*. This unifying perspective postulates resonant feedback loops as the substrate of phenomenal experience. Adaptive resonance offers a core module for the representation of hypothesized processes underlying learning, attention, search, recognition, and prediction. At the model's field of coding neurons, the continuous stream of information pauses for a moment, holding a fixed activation pattern long enough for memories to change. Intrafield competitive loops fixing the moment are broken by active reset, which flexibly segments the flow of experience according to the demands of perception and environmental feedback. As Pollen (1999, pp. 15–16) suggests, “[I]t may be the consensus of neuronal activity across ascending and descending pathways linking multiple cortical areas that in anatomical sequence subserves phenomenal visual experience and object recognition and that may underlie the normal unity of conscious experience.”

Road Maps: Learning in Artificial Networks; Vision

Related Reading: Competitive Learning; Helmholtz Machines and Sleep-Wake Learning; Laminar Cortical Architecture in Visual Perception

References

- Carpenter, G. A., 1997, Distributed learning, recognition, and prediction by ART and ARTMAP neural networks, *Neural Netw.*, 10:1473–1494.
- Carpenter, G. A., and Grossberg, S., 1987, A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision Graphics Image Process.*, 37:54–115.
- Carpenter, G. A., and Grossberg, S., 1990, ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures, *Neural Networks*, 3:129–152.
- Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B., 1992, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps, *IEEE Trans. Neural Netw.*, 3:698–713.
- Carpenter, G. A., Grossberg, S., and Rosen, D. B., 1991, Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Netw.*, 4:759–771.
- Caudell, T. P., Smith, S. D. G., Escobedo, R., and Anderson, M., 1994, NIRS: Large scale ART-1 neural architectures for engineering design retrieval, *Neural Netw.*, 7:1339–1350.
- Duda, R. O., Hart, P. E., and Stork, D. G., 2001, *Pattern Classification*, 2nd ed., New York: Wiley, section 10.11.2. ♦
- Fay, D. A., Verly, J. G., Braun, M. I., Frost, C., Racamato, J. P., and Waxman, A. M., 2001, Fusion of multi-sensor passive and active 3D imagery, in *Proc. SPIE Enhanced Synthet. Vision*, vol. 4363.
- Griffith, N., and Todd, P. M., Ed., 1999, *Musical Networks: Parallel Distributed Perception and Performance*, Cambridge, MA: MIT Press.
- Grossberg, S., 1976a, Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors, *Biol. Cybern.*, 23:121–134.
- Grossberg, S., 1976b, Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, and illusions, *Biol. Cybern.*, 23:187–202.
- Levine, D. S., 2000, *Introduction to Neural and Cognitive Modeling*, Mahwah, New Jersey: Erlbaum, chap 6. ♦
- Markram, H., and Tsodyks, M., 1996, Redistribution of synaptic efficacy between neocortical pyramidal neurons, *Nature*, 382:807–810.
- Page, M., 2000, Connectionist modelling in psychology: A localist manifesto, *Behav. Brain Sci.*, 23:443–512.
- Pollen, D. A., 1999, On the neural correlates of visual perception, *Cereb. Cortex*, 9:4–19.
- Serrano-Gotarredona, T., Linares-Barranco, B., and Andreou, A. G., 1998, *Adaptive Resonance Theory Microchips: Circuit Design Techniques*, Boston: Kluwer Academic.

Adaptive Spike Coding

Adrienne Fairhall and William Bialek

Introduction

The meaning of any signal that we receive from our environment is modulated by the context within which it appears. Our interpretation of color, a spoken phoneme, or a patch of luminance depends critically on its context. Although “context” may be a rather abstract notion, it is often reasonable to understand the term as meaning the statistical ensemble in which the signal is embedded. Interpreting a message requires both registering the signal itself and knowing something about this statistical ensemble. The relevant temporal or spatial ensemble depends on the task. The context may be highly local; we interpret appropriately gradations of light and dark in a scene where local brightness typically varies over orders of magnitude (see FEATURE ANALYSIS). For tasks such as decision making, the relevant statistics may reflect complex descriptions of the world accumulated over long periods.

Neural representations at every level of information processing should be similarly modulated by context. Information theoretically, this has measurable advantages: representations that appropriately take into account the statistical properties of the incoming signal are more efficient. Since the 1950s it has been suggested that efficiency is a design principle of the nervous system, allowing neurons to transmit more useful information with their limited dynamic range (see OPTIMAL SENSORY ENCODING). Thus, one expects that learning the context and implementing this knowledge

through coding strategy is inherent in the formation of representations.

Such adjustments occur over a wide range of time scales. Through the genetic code, species adapt to environmental changes over many generations. In a single individual, learning, implemented through neural plasticity, continues throughout life in response to experience of the world; perceptual learning is stored even at low levels of neural information processing (see SOMATOTOPY: PLASTICITY OF SENSORY MAPS). In the article, we discuss even more rapid changes: neural adaptation, which we take to mean reversible change in the response properties of neurons on short time scales.

Since Adrian's first observations of adaptation in spiking neurons, it has been suggested that adaptation serves a useful function for information processing, preventing a neuron from continuing to transmit redundant information and increasing its responsiveness to new stimuli. Within the simplified picture of a neuron as a combination of linear filtering followed by a threshold, or a decision rule for spiking, either or both of the two components—the filter and the threshold function—may be adaptive functions of the input, and both may implement the goal of increasing information transmission. We will discuss both of these possibilities.

Neurons in every sensory modality have been shown to have adaptive properties, and the mechanisms governing various types of adaptation have been at least partially explored (Torre et al.,

1995). Here we will discuss adaptation as the simplest form of learning and memory. We describe recent experiments that explicitly aim to link the phenomenology of adaptive spike coding to its functional relevance, in particular to improved information transmission. A common feature of adaptation is the existence of multiple time scales. In examining mechanisms, we concentrate on recent work suggesting that the long time scales retaining short-term memory can be generated through single-cell properties.

Adaptive Coding

Adaptation of neural firing rate to stationary stimuli has been seen in all modalities of the primary sensory system. In the visual system, photoreceptors adapt to light level, and retinal ganglion cells show rapid contrast gain control. The trade-offs and information processing gains due to adaptation in insect eyes, relevant also for the vertebrate retina, are discussed in Laughlin (1989). Mechanoreceptors in the somatosensory system have been classified into four main types of cells, three of which are distinguished by the time scales of their adaptation (rapidly and slowly adapting), and these time scales in part determine the cells' function: slowly adapting cells are implicated in the perception of spatial form and texture, while the experience of flutter and of motion is mediated by rapidly adapting cells (Johnson, 2001). Thus, the dynamics of adaptation can determine a neuron's functional role.

Adaptation is not limited to primary receptors. In visual cortex, V1 neurons show contrast adaptation, which is thought to occur entirely at the level of cortex. The motion aftereffect, a familiar phenomenon whereby following exposure to motion in one direction, the visual field appears to move in the opposite direction, is thought to be due to adaptation of direction-sensitive neurons in visual cortex.

Adaptation to a Distribution

Understanding the significance of adaptation for information processing requires going beyond fixed stimuli. Recently, studies have focused on adaptation to the stimulus *distribution*. This approach is necessary to characterize coding information theoretically: the evaluation of a coding strategy requires considering the entire ensemble of inputs and outputs. In Smirnakis et al. (1997), retinal ganglion cells were stimulated with dynamic movies of flickering light intensity where the mean light level was fixed but the variance was switched periodically from one value to another. The spike rate of the neurons showed typical adaptive behavior (Figure 1): following an increase in variance, the firing rate increased initially, but gradually returned to a considerably lower level; a decrease in variance led to a sudden dip in firing rate, with eventual recovery.

The experiments of Smirnakis et al. (1997) consider only firing rate. However, the timing of single spikes can convey a great deal of information about the stimulus. In the visual system of the fly, in particular the motion-sensitive identified neuron H1 in the fly's lobula plate, much is understood about single-spike coding, providing an excellent opportunity to study the effects of adaptation in detail.

H1 responds to a simple stimulus, wide-field horizontal motion. The neuron is characterized by its input/output relation $P(\text{spikels})$, or the probability of a spike given the projection s of the dynamic stimulus onto a relevant feature, determined by reverse correlation.

When the system has reached steady state through exposure to a zero-mean, white noise velocity stimulus with a given variance σ^2 , its input/output relation is measured. The resulting curves, measured for a range of values of the variance, are shown in Figure 2. Clearly, the input/output relation is not a fixed property of the system but adapts to the distribution of inputs. Indeed, it does so in such a way that the stimulus appears to be measured in units of its

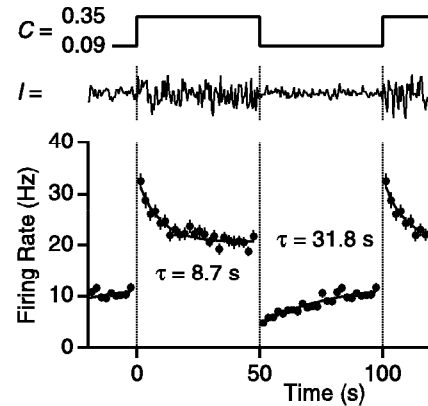


Figure 1. Firing rate of rabbit retinal ganglion cells in response to a flicker stimulus where the variance of the light intensity I switches periodically in time. (From Smirnakis S. M., et al., 1997, Adaptation of retinal processing to image contrast and spatial scale, *Nature*, 386: 69–73. Copyright 1997, Macmillan Publishers Ltd.; reprinted with permission.)

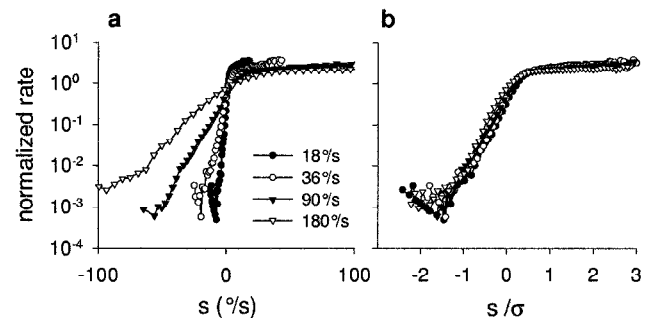


Figure 2. (a) A set of input/output relations relating the probability of spiking to the velocity stimulus, measured for stationary white noise stimuli with different variances. (b) The curves differ only by a scale factor as is shown by normalizing the stimulus by its standard deviation. In this case the curves coincide. (From Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. R., 2001, Adaptive rescaling maximizes information transmission, *Neuron*, 26:695–702. Copyright 2000, Elsevier Science, reprinted with permission.)

standard deviation; when the curves are replotted with the stimulus normalized by its RMS value, they superimpose. Thus, a scale factor λ multiplying the stimulus, and thereby matching the dynamic range of the response to the distribution of the inputs, is a degree of freedom for the system. The value of λ chosen by the system achieves a maximum of information transmission (Brenner, Bialek, and de Ruyter van Steveninck, 2000).

This is a simple form of learning: the system gauges the standard deviation of the signal and modifies its response properties to adjust its dynamic range to the range of inputs. The adjustment must take some time, as the new distribution must be sampled from examples. This sets fundamental physical and statistical limits for the system's estimate of the current variance. We can examine the time scale for learning (Fairhall et al., 2001) by, as in the retina experiments described earlier, switching periodically between two distributions. The firing rate shows the same pattern of adaptation as was seen in the experiments of Smirnakis et al. (1997), but this pattern need not correspond to the time scale for adjustment of the input/output relation. Indeed, it was found that the scale factor of the input/output relations, measured dynamically, adjusts much

more rapidly than the relaxation time of the rate—on the order of 100 ms, compared with several seconds. This short time scale is consistent with the limits imposed by estimates of noise from the photoreceptors. One can verify that the dynamic adaptation of the input/output relation maintains information transmission through the system by computing how much information one can extract from the spikes about the stimulus (see SENSORY CODING AND INFORMATION TRANSMISSION and Fairhall et al., 2001). The information rate recovers on comparably short time scales.

For the decoder, a potential drawback of adaptive coding is *ambiguity*: it is necessary to know the context in order to interpret the signal correctly. Thus, information about the context must be conveyed independently. Although this information might be carried by other neurons in the network, here the information about the ensemble is carried simultaneously by the same spike train: it can be read off, either through the rate (taking into account the delays due to the slow relaxation) or, more accurately, through the variance dependence of the statistics of *spike time differences* (Fairhall et al., 2001). Thus, for the code of H1, spikes carry multiple meanings: in absolute timing, as precise markers of single stimulus events, and in relative timing, as indicators of the stimulus ensemble.

Multiple Time Scales

The slow relaxation of the rate appears to be related to a commonly observed property of adapting primary sensory neurons: a power law decay of the firing rate, $r \sim t^{-\alpha}$. More generally, in the case just presented, the rate is close to the *fractional derivative* of the logarithm of the stimulus variance. For each frequency ω , fractional differentiation shifts the frequency component by a constant phase, and scales each component by ω^α , where α is a power less than 1. Some of the properties of a fractional differentiator are illustrated in Figure 3. Several examples of a power law decay of the rate following a step change in stimulus amplitude were collected by Thorson and Biederman-Thorson (1974; Figure 4) and more have since been observed; examples include various invertebrate mechanoreceptors and photoreceptors, mammalian carotid sinus baroreceptors, and cat retinal ganglion cells.

We have noted a separation of time scales in the adaptation of the input/output relation compared with the rate. This type of adaptation on its own signals the existence of many time scales. Power-law scaling implies the lack of a typical time scale or the presence of multiple time scales. Fractional differentiation is non-local; the response at time t_0 is affected by times $t \ll t_0$. This is a linear “memory” mechanism.

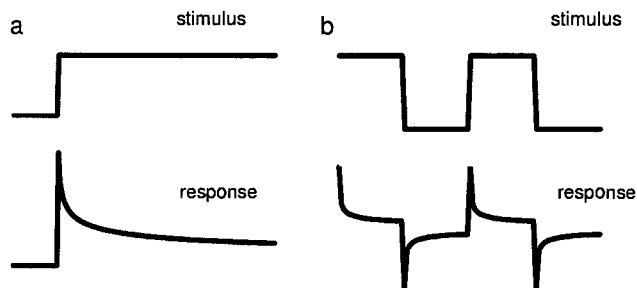


Figure 3. Illustration of some properties of a fractional differentiator with exponent $\alpha = 0.3$. (a) A step function stimulus leads to a power law decaying rate. In a log-log plot the curve would appear as a straight line with slope $-\alpha$. (b) A square wave leads to a similar adaptation curve as shown in Figure 1.

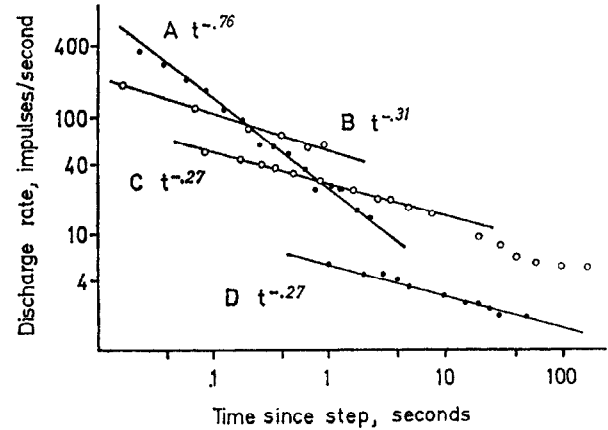


Figure 4. Four curves showing power law adaptation in response to a step increase in stimulus in four different receptors: cockroach leg mechanoreceptor, in response to distortion of the tactile spine on the femur (curve A); slit sensillum on the leg of the hunting spider in response to 1,200 Hz sound (curve B); slowly adapting stretch receptor of the crayfish (curve C); and increase of response over light-adapted level of *Limulus* lateral-eye eccentric cell to an increase in light intensity (curve D). (Examples from Thorson, J., and Biedermann-Thorson, M., 1974, Distributed relaxation processes in a sensory adaptation, *Science*, 183:161–172. Copyright 1974, American Association for the Advancement of Science; reprinted with permission.)

Such adaptation is particularly interesting both because it is so prevalent and because it may have an important role in optimizing information transmission. Fractional-differentiation-like behavior is observed in fly photoreceptors, and in that case, the exponent of the fractional differentiator appears to be matched to the spectrum of natural stimuli (van Hateren and Snippe, 2001). Thus the effect of the transformation is to whiten the spectrum of natural signals. Because many natural stimuli have power-law characteristics, it is intriguing to speculate that fractional differentiation at the sensory periphery may be a general neural mechanism for whitening input statistics.

Mechanisms

Adaptation requires retaining memory of activity over extended time scales. These long time scales can arise from a number of sources. Intracellular calcium concentration has been identified as playing an important role in information processing, acting as a slowly changing “integrator” of activity. Other forms of adaptation, particularly the power-law-like behavior discussed in the previous section, are also likely to be a property of single cells rather than of the network. Recent biophysical studies show that membrane dynamics can have long time scales that retain memory of the history of stimulation/activity over hundreds of seconds (Marom, 1998). This could be brought about either by the modification of intrinsic properties or by intrinsic properties that have built-in long time scales through *state-dependent inactivation* (Turrigiano, Marder, and Abbott, 1996; Marom, 1998).

Calcium as an Integrator of Activity

Each spike introduces a roughly constant amount of calcium into the cell through voltage-dependent Ca^{2+} channels. The Ca^{2+} concentration then decays slowly. Thus, $[\text{Ca}^{2+}]$ can be modeled as a leaky integrator of activity, with a decay time scale of ca. 100 ms. This calcium signal can allow activity-dependent regulation of sub-

sequent neural activity through the modification of conductances (see ACTIVITY-DEPENDENT REGULATION OF NEURONAL CONDUCTANCES).

Recent evidence indicates that single-cell properties may contribute to contrast adaptation in cortex (Sanchez-Vives, Nowak, and McCormick, 2000). Previous work has shown that contrast adaptation is associated with hyperpolarization of the membrane potential in cat area 17 neurons. By stimulating the neurons directly with injected current, effects similar to contrast adaptation are seen (though less dramatically than to real visual input). This suggests that these effects can be induced through the modulation of intrinsic cell properties; the activation of Ca^{2+} - and Na^{+} -dependent potassium conductances is indicated.

State-Dependent Channel Dynamics

In some cases the relevant dynamics may be due to the complex behavior of the channels themselves. Recently it has become clear that the dynamics of inactivation provide the membrane with the possibility for extended history dependence (Marom, 1998).

A simplified picture of the gating of voltage-gated ion channels is a three-state scheme:



where channels can be either closed (C), open (O), or inactivated (I). Generally, the transition between closed and open is voltage-dependent and rapid, on the order of the duration of an action potential. The transition between open and inactivated, on the other hand, is voltage-independent and can have very long time scale dynamics. Intriguingly, studies in vitro show that some sodium channel types have inactivation rates that scale with the duration of the input (Marom, 1998), providing time scales of up to several minutes. The precise mechanism underlying this large variety of time scales is not yet well understood; it is hypothesized that the system cascades through a multiplicity of inactivation states. Earlier theoretical work has shown that the coupling of many states leads to a scaling relation between the duration of activity and the rate of recovery from inactivation.

In a step closer to a realistic preparation, the dynamic clamp was applied to cultured stomatogastric ganglion neurons to add an effective slowly inactivating potassium current (Turrigiano et al., 1996). As had been observed previously, this produced long delays to firing during depolarization, and an increase in excitability with a time scale much longer than the duration of the input. Further, the slow channel dynamics produced a long-lasting effect on the firing properties of the neuron.

In vivo, the contribution of slowly inactivating sodium channels to power-law-like adaptation has been suggested. Mechanosensory neurons in the cockroach femoral tactile spine have been shown to display power-law adaptation. From intracellular measurements, Basarsky and French (1991) found that the spike rate adaptation is due to cumulative slowing of the recovery of the membrane potential between spikes. Previous work had demonstrated that calcium channel blockers or blockers of Ca^{2+} -activated K^{+} channels did not reduce adaptation, while modifying sodium channel inactivation did.

These mechanisms might be seen as primitives for short-term "learning and memory."

Modeling

Historically, attempts to model adaptation have considered the process to involve a dynamic threshold. More recently, modeling approaches have taken a functional perspective on the outcome of adaptation and have proposed algorithms whereby the conductances may adjust to provide the cell with desirable properties, such

as approximately constant activity (see ACTIVITY-DEPENDENT REGULATION OF NEURONAL CONDUCTANCES). Closer to our earlier discussion, Stemmler and Koch (1999) derive a learning rule for conductances that maximizes the mutual information between input and output, where the output is taken to be the neuron's firing rate. The learning rule adjusts conductances at every new presentation of the stimulus, subject to biologically plausible constraints. Under this learning rule, a realistic conductance-based model neuron was indeed able to learn a changing distribution and adjust its firing statistics accordingly. The time scales treated were orders of magnitude longer than those observed experimentally in Fairhall et al. (2001) and predicted theoretically from statistical considerations. Experimental evidence is still required to determine whether such a model is realistic.

As noted, many adaptation processes in sensory receptors follow a power-law relaxation. Assuming that most elementary processes involve a single time scale, with exponential dynamics, Thorson and Biederman-Thorson (1974) proposed that power laws may arise from a superposition of many elementary processes with a wide range of time scales. From the definition of the gamma function,

$$t^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} dr r^{\alpha-1} e^{-rt} \quad (2)$$

a power law may be generated by a weighted sum of exponentials with a range of time scales. This distribution was considered to be generated through geometric factors, such as the inhomogeneous distribution of elements within the receptor.

This model has met with some skepticism because of the requirements both for a continuous distribution of time scales and for these to be present in the appropriate proportions. It has been noted that power-law-like behavior results from much less stringent conditions: the superposition of only a few exponentials can produce a power law over the decade or two normally available to experiment. However, recent experimental advances, outlined in the previous section, may provide a better underpinning for the derivation of power-law adaptation from membrane mechanisms.

Adaptation of Receptive Fields

As noted in the Introduction, a neuron can be modeled as a combination of feature extraction (linear filtering) and a nonlinear decision function (or threshold). Although we have discussed the effects of adaptation on the nonlinear decision function, adaptation can also affect the feature that causes the neuron to spike: the receptive field can depend on the ensemble of inputs. Although this result had been frequently observed in work on invertebrate vision, recent experiments demonstrate analogous results for cortical receptive fields. Sceniak et al. (1999) show that the extent of spatial summation implemented by neurons in V1 depends adaptively on contrast; this has parallels in the adaptation of filters in retina (Laughlin, 1989). Theunissen, Sen, and Doupe (2000) found that the spatiotemporal receptive fields of neurons in auditory cortex showed a strong dependence on the stimulus ensemble. This is a natural consequence of neural nonlinearity, but such a dependence is also necessary for optimal information processing.

Discussion

The ubiquity of adaptation throughout the nervous system should be proof of its fundamental importance. Although the phenomenology of adaptation, particularly to constant stimuli, has been extensively explored, recent experimental and theoretical approaches have made contact with the principles of information theory in order to evaluate adaptive coding. For fly motion-sensitive neurons,

it was found that the coding strategy of the system adapts rapidly and continuously to track dynamic changes in the statistics of the stimulus.

We have discussed a variety of mechanisms that may implement adaptive coding at the level of single cells. Although it is likely that systems will implement such important behavior at many levels, it is appealing that the simplest elements of neural computation have the power to carry out dynamic aspects of information processing.

Road Map: Neural Coding

Related Reading: Population Codes; Sensory Coding and Information Transmission

References

- Basarsky, T., and French, A., 1991, Intracellular measurements from a rapidly adapting sensory neuron, *J. Neurophysiol.*, 65:49–56.
- Brenner, N., Bialek, W., and de Ruyter van Steveninck, R., 2000, Adaptive rescaling maximizes information transmission, *Neuron*, 26:695–702.
- Fairhall, A. L., Lewen, G., Bialek, W., and de Ruyter van Steveninck, R. R., 2001, Efficiency and ambiguity in an adaptive neural code, *Nature*, 412:787–792.
- Johnson, K. O., 2001, The roles and functions of cutaneous mechanoreceptors, *Curr. Opin. Neurobiol.*, 11:455–61. ♦
- Laughlin, S. B., 1989, The role of sensory adaptation in the retina, *J. Exp. Biol.*, 146:39–62. ♦
- Marom, S., 1998, Slow changes in the availability of voltage-gated ion channels: Effects on the dynamics of excitable membranes, *J. Membr. Biol.*, 161:105–113. ♦
- Sanchez-Vives, M., Nowak, L., and McCormick, D., 2000, Membrane mechanisms underlying contrast adaptation in cat area 17 in vivo, *J. Neurosci.*, 20:4267–4285.
- Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R., 1999, Contrast's effect on spatial summation by macaque V1 neurons, *Nature Neurosci.*, 2:733–739.
- Sminakakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., and Meister, M., 1997, Adaptation of retinal processing to image contrast and spatial scale, *Nature*, 386:69–73.
- Stemmler, M., and Koch, C., 1999, How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate, *Nature Neurosci.*, 2:521–527.
- Theunissen, F., Sen, K., and Doupe, A., 2000, Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds, *J. Neurosci.*, 20:2315–2331.
- Thorson, J., and Biederman-Thorson, M., 1974, Distributed relaxation processes in a sensory adaptation, *Science*, 183: 161–172. ♦
- Torre, V., Ashmore, J. F., Lamb, T. D., and Menini, A., 1995, Transduction and adaptation in sensory receptor cells, *J. Neurosci.*, 15:7757–7763. ♦
- Turrigiano, G., Marder, E., and Abbott, L., 1996, Cellular short-term memory from a slow potassium conductance, *J. Neurophysiol.*, 75:963–968.
- van Hateren, J. H., and Snippe, H. P., 2001, Information theoretical evaluation of parametric models of gain control in blowfly photoreceptor cells, *Vision Res.*, 41:1851–1865.

Amplification, Attenuation, and Integration

H. Sebastian Seung

Introduction

Differential equations such as

$$\tau \dot{x}_i + x_i = f\left(\sum_j W_{ij}x_j + b_i\right) \quad (1)$$

have long been used to model networks of interacting neurons (Ermentrout, 1998; PHASE-PLANE ANALYSIS OF NEURAL NETS). The activity of neuron i is represented by a single dynamical variable x_i , and its input-output characteristics by a single transfer function f . There are more biophysically realistic descriptions of neural networks that include many dynamical variables per neuron, in order to explicitly model dendritic integration, action potential generation, and synaptic transmission. Nevertheless, simplified models like that in Equation 1 have been useful for understanding how the computational properties of neural networks are related to their synaptic organization.

The parameter W_{ij} in Equation 1 represents the strength of the synapse from neuron j to i . These synapses are termed *recurrent*, as they connect to other neurons in the same network. Feedforward synaptic input from outside the network is implicit in the bias b_i . The feedforward synapses could be made explicit by writing $b_i = b_i^0 + \sum_a V_{ia}z_a$, where z_a are input neuron activities, V_{ia} the strengths of the feedforward synapses, and b_i^0 any intrinsic tendency of neuron i to be active. But the feedforward connections will be left implicit in the following, so as to focus on the computational role of the recurrent connections.

Accordingly, the biases b_i in Equation 1 will be regarded as the inputs to the network, while the activities x_i are the outputs. If there were no recurrent synapses ($W_{ij} = 0$ for all i and j), then each neuron i would respond by low-pass filtering the signal $f(b_i)$ with time constant τ . When there are recurrent synapses, a general char-

acterization of the response properties of a network is difficult, but the situation is greatly simplified when nonlinearity is neglected. Putting the transfer function $f(u) = u$ in Equation 1 yields the linear network

$$\tau \dot{x}_i + x_i = \sum_j W_{ij}x_j + b_i \quad (2)$$

which can be completely analyzed using the tools of linear systems theory. The modest goal of this article is to describe some properties of linear networks and give examples of their application to neural modeling.

In particular, the focus is on the role of recurrent synaptic connectivity. Provided that they do not lead to instability, the recurrent connections alter both the gain and speed of response to feedforward input. Either they amplify and slow down responses to feedforward input, or they attenuate and speed up responses. Both effects can occur simultaneously in the same network, as can be seen by mathematically transforming the network of interacting neurons into a set of noninteracting eigenmodes. The effect of the recurrent synapses generally varies from mode to mode.

Besides amplification and attenuation, a linear network can also carry out the operation of temporal integration, in the sense of Newtonian calculus. This happens when the strength of feedback is precisely tuned for an eigenmode, so that its gain and time constant diverge to infinity.

Admittedly, the neglect of nonlinearity is a step away from biological realism. Nevertheless, linear models are important because they give insight into the local behavior of nonlinear networks, which can often be linearly approximated in the vicinity of fixed points. And the linear computations of amplification, attenuation, and integration have been ascribed to a number of brain areas.

Autapse

The simplest example of a recurrent synapse is a single neuron with a synapse onto itself, or *autapse*, in the terminology of neurophysiology. For this case, the dynamics (Equation 2) takes the form

$$\tau \dot{x} + x = Wx + b \quad (3)$$

The autapse has strength W and is said to be excitatory if $W > 0$ and inhibitory if $W < 0$. The example is not meant to be a realistic model of a biological autapse; it is only a simple illustration of some of the effects of recurrent synaptic connections. The parameter W will also be called the strength of *feedback*, in the terminology of engineering. Without feedback ($W = 0$), the neuron acts as a low-pass filter of input b with time constant τ . When the effect of feedback is considered, the first distinction that has to be made is between the unstable $W > 1$ and the stable $W < 1$ cases. (Discussion of the borderline $W = 1$ case is postponed until later.)

If $W > 1$, the autapse is unstable, as can be seen by solving Equation 3 for input b that is constant in time. The solution diverges exponentially to infinity, because the feedback is so strong that it leads to runaway instability. Note that in a more realistic model, the growth of this runaway instability would eventually be limited by nonlinearity, but in the idealized linear model (Equation 3), divergence to infinity is possible.

If $W < 1$, the autapse is stable, and the dynamics (Equation 3) can be rewritten in the form

$$\frac{\tau}{1 - W} \dot{x} + x = \frac{b}{1 - W} \quad (4)$$

From this formula can be read two numbers that characterize the autapse: the steady-state gain, and the time constant of response. The gain is operationally defined by holding the input constant and allowing the output to relax to the steady-state value $x_\infty = b/(1 - W)$. Then the steady-state gain, defined as the ratio of output x_∞ to input b , is $1/(1 - W)$. By this definition, the gain is exactly unity in the case of no feedback ($W = 0$). Positive ($W > 0$) and negative ($W < 0$) feedback have different effects. Positive feedback amplifies, boosting the gain to a value greater than 1. Negative feedback attenuates, making the gain less than 1.

Positive and negative feedback also have opposite effects on the speed of response. The time constant of the exponential relaxation to the steady state is $\tau/(1 - W)$. In the case of no feedback, this is equal to the fundamental time constant τ . But positive feedback lengthens the time constant, while negative feedback shortens it. This means that there is a trade-off between amplification and speed, sometimes known as the gain-bandwidth trade-off. Intuitively speaking, the trade-off arises because feedback amplification requires that the signal circulate in the feedback loop, so that more amplification requires more time.

In summary, a feedback loop containing a perfectly linear element behaves in a simple way. Positive feedback ($W > 0$) amplifies and slows down response, assuming that it doesn't lead to instability. Negative feedback ($W < 0$) attenuates and speeds up response.

The idea of amplification by positive feedback has been prominent in a number of models of primary visual cortex (Douglas et al., 1995). Neurons in layer 4 receive both feedforward drive from the thalamus and recurrent input from other cortical neurons. It has been proposed that the recurrent interactions amplify the responses to feedforward input. To test this idea, Ferster and colleagues recorded from layer 4 neurons. They inactivated corticocortical inputs both by cooling (Ferster, Chung, and Wheat, 1996) and electrical stimulation (Chung and Ferster, 1998). In both cases, they measured a two- or threefold reduction in the amplitude of cortical

responses to visual stimulation, which was interpreted as a loss of amplification by positive feedback.

The above discussion omitted the special case of $W = 1$, which is the borderline between stability and instability. For $W \neq 1$, there was exactly one steady state, which was either stable or unstable, depending on whether W was less than or greater than 1. In contrast, if $W = 1$, there is not a unique steady state. The number of steady states depends on b . There are infinitely many if $b = 0$, and none at all if $b \neq 0$. To understand the case of non-zero b , it is helpful to return to Equation 3, which reduces to $\tau \dot{x} = b$. In other words, the response x is the time integral of b . Therefore, a linear autapse can act as an integrator, if the strength of feedback is precisely tuned (Seung et al., 2000). Variants of this idea have been used to model neural integrators, brain areas that integrate their inputs in the sense of Newtonian calculus (Robinson, 1989).

Mutually Inhibitory Pair

While the autapse illustrates the gain-bandwidth trade-off in feedback amplification, it involves only a single neuron, and cannot capture genuine population behaviors. A more interesting example consists of two linear neurons with mutual inhibition:

$$\tau \dot{x}_1 + x_1 = -\beta x_2 + b_1 \quad (5)$$

$$\tau \dot{x}_2 + x_2 = -\beta x_1 + b_2 \quad (6)$$

The parameter β is assumed to be positive, so that the interaction is inhibitory. This dynamics is more complex than Equation 3 because it involves two differential equations that are coupled to each other. Luckily, it turns out that the equations can be decoupled by adding and subtracting them.

Adding the two equations yields an equation for the common mode $x_c = x_1 + x_2$,

$$\tau \frac{d}{dt} (x_1 + x_2) + (x_1 + x_2) = -\beta(x_1 + x_2) + (b_1 + b_2) \quad (7)$$

Comparison with Equation 3 reveals that the common mode behaves like an autapse with negative feedback. Therefore the common mode attenuates its input $b_1 + b_2$ with steady-state gain $1/(1 + \beta)$ and time constant $\tau/(1 + \beta)$.

Similarly, subtracting the two equations yields an equation for the differential mode $x_d = x_1 - x_2$,

$$\tau \frac{d}{dt} (x_1 - x_2) + (x_1 - x_2) = \beta(x_1 - x_2) + (b_1 - b_2) \quad (8)$$

The differential mode behaves like an autapse with positive feedback. If $\beta > 1$, the differential mode is unstable. If $\beta < 1$, then the differential mode amplifies its input $b_1 - b_2$ with steady-state gain $1/(1 - \beta)$ and time constant $\tau/(1 - \beta)$.

To recapitulate, transforming from (x_1, x_2) to (x_c, x_d) formally decoupled the mutually inhibitory pair of neurons into two "virtual" autapses. Note that the transformation is reversible, as x_1 and x_2 can be reconstructed from the common and differential modes, e.g., $x_1 = (x_c + x_d)/2$.

A striking aspect of this example is that mutual inhibition has completely opposite effects on the common and differential modes. For the common mode, inhibition mediates negative feedback, which leads to attenuation. But inhibition mediates positive feedback for the differential mode, which leads to amplification.

The general lesson to be drawn is that no direct correspondence exists between the sign of synaptic connections and the sign of feedback. This is because a synapse is local, belonging to just two neurons. In contrast, feedback strength is global, belonging to a distributed mode of the network. As will be described below, the feedback strength is given in general by the eigenvalues of the synaptic weight matrix W . The autapse is a special exception for

which the sign of the synaptic connection directly corresponds to the sign of feedback, but this does not hold true in general.

The idea that inhibition between neurons can amplify differences has been used to explain the fact that visual systems are more sensitive to relative luminance, or contrast, than to absolute luminance. For example, the *Limulus* retina consists of visual receptors that are topographically organized in a two-dimensional network and interact via lateral inhibition. Measurements of retinal output reveal enhancement of luminance differences, a fact that has been successfully explained using network models that are generalizations of the mutually inhibitory pair considered here (Hartline and Ratliff, 1972).

The special case $\beta = 1$ is also of interest. It is the borderline of stability for the differential mode. If $b_1 - b_2$ is zero, then the differential mode $x_1 - x_2$ is constant in time, according to Equation 8, while the common mode $x_1 + x_2$ converges exponentially to the value $(b_1 + b_2)/2$. This is a simple example of a *line attractor*, a line of fixed points to which all trajectories are attracted (Seung, 1996). More complex nonlinear network models with approximate line attractors have been used to model the phenomenon of persistent neural activity (Seung, 1996; Zhang, 1996).

Note that having a continuous set of fixed points is an unusual situation, requiring the precise tuning of the inhibitory strength β and the differential input $b_1 - b_2$. When $b_1 - b_2$ is non-zero, then it is integrated by the differential mode. In this case, inhibitory interactions yield an integrator, in contrast to the autapse, which requires excitatory feedback to integrate. Robinson et al. proposed that lateral inhibition is the mechanism of the oculomotor neural integrator, which converts vestibular and other velocity-coded inputs into eye position outputs (Cannon, Robinson, and Shamma, 1983).

General Network

For a general network of N neurons, the effects of feedback can be understood via eigensystem analysis. It is convenient to rewrite the dynamics in Equation 2 in matrix-vector form as

$$\tau \frac{d}{dt} x + x = Wx + b \quad (9)$$

where x and b are vectors and W is the synaptic weight matrix.

Suppose that the weight matrix can be factorized as $W = S\Lambda S^{-1}$, where Λ is a real diagonal matrix and S is a real invertible matrix. A sufficient condition for a real diagonalization is that the weight matrix W be symmetric, but this is not a necessary condition. The diagonal entries of Λ are the eigenvalues of W . The columns of S are the right eigenvectors of W , and the rows of S^{-1} are the left eigenvectors.

Recall that transforming to the common and differential modes simplified the dynamics of the mutually inhibitory pair. The analogue here is to change from x and b to

$$\tilde{x} = S^{-1}x, \quad \tilde{b} = S^{-1}b$$

These vectors can be used to express x and b as linear combinations of the right eigenvectors, $x = S\tilde{x}$ and $b = S\tilde{b}$.

The transformation of Equation 9 is effected by multiplying with S^{-1} ,

$$\tau \frac{d}{dt} \tilde{x} + \tilde{x} = S^{-1}Wx + \tilde{b} \quad (10)$$

$$= S^{-1}WS\tilde{x} + \tilde{b} \quad (11)$$

$$= \Lambda\tilde{x} + \tilde{b} \quad (12)$$

Writing out the last expression component by component yields

$$\tau \frac{d}{dt} \tilde{x}_a + \tilde{x}_a = \lambda_a \tilde{x}_a + \tilde{b}_a$$

where λ_a is the a th diagonal element of Λ , or equivalently the a th eigenvalue of W . This is a great simplification: the network (Equation 9) of N interacting neurons has been transformed into N non-interacting “virtual” autapses. Each autapse has feedback with strength given by the eigenvalues λ_a . Assuming that the eigenvalues are less than or equal to 1, each autapse can perform the operations of amplification, attenuation, or integration.

Discussion

In this article, some effects of recurrent synaptic connectivity on linear networks were characterized. The autapse example demonstrated that there is a gain-bandwidth trade-off in amplification and attenuation by feedback, and the possibility of integration when feedback is precisely tuned. The mutually inhibitory pair illustrated the decoupling of an interacting network into “virtual” autapses, and also illustrated that the sign of feedback is not directly related to the sign of synaptic connections. Such a decoupling is generally possible for any synaptic weight matrix W that is diagonalizable with all real eigenvalues.

More generally, the eigenvalues (and eigenvectors) are complex numbers. When an eigenvalue of W has a non-zero imaginary part, the corresponding eigenmode exhibits oscillatory behavior. Accordingly, linear analyses have been used to explain the existence of oscillations in some neural network models (Li and Hopfield, 1989).

It is natural to ask whether the concepts introduced above have any relevance for *nonlinear* neural networks. A simple way of modeling nonlinearity is to introduce a threshold for activation by choosing $f(x) = \max\{x, 0\}$ for the transfer function in Equation 1. Because the resulting dynamics are piecewise linear, eigenvalues and eigenvectors are still essential for mathematical analysis (Haddeler and Kuhn, 1987; Hahnloser et al., 2000), but the threshold nonlinearity leads to a richer variety of dynamical behaviors. A full discussion of threshold linear networks is beyond the scope of this article, but let us briefly reconsider the example of a mutually inhibitory pair of neurons presented with inputs that are constant in time. For linear neurons, the mutual inhibition caused differences in input to be amplified in the steady-state response. If the neurons are instead threshold linear, *winner-take-all* behavior can result for some choices of model parameters. Then only a single neuron is active at steady state, no matter how small the difference in inputs may be (Amari and Arbib, 1977). As in the purely linear case, the difference in steady-state outputs is greater than the difference in inputs. However, this behavior cannot be explained in terms of a simple linear amplification. For a more detailed explanation, see WINNER-TAKE-ALL NETWORKS.

Road Map: Dynamic Systems

Background: I.3. Dynamics and Adaptation in Neural Networks

Related Reading: Pattern Formation, Neural; Winner-Take-All Networks

References

- Amari, S., and Arbib, M. A., 1977, Competition and cooperation in neural nets, in *Systems Neuroscience* (J. Metzler, Ed.), New York: Academic Press, pp. 119–165. ♦
- Cannon, S. C., Robinson, D. A., and Shamma, S. 1983, A proposed neural network for the integrator of the oculomotor system, *Biol. Cybern.*, 49:127–136.
- Chung, S., and Ferster, D., 1998, Strength and orientation tuning of the thalamic input to simple cells revealed by electrically evoked cortical suppression, *Neuron*, 20:1177–1189.

- Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., and Suarez, H. H., 1995, Recurrent excitation in neocortical circuits, *Science*, 269:981–985.
- Ermentrout, B., 1998, Neural networks as spatio-temporal pattern-forming systems, *Rep. Prog. Phys.*, 61:353–430. ♦
- Ferster, D., Chung, S., and Wheat, H., 1996, Orientation selectivity of thalamic input to simple cells of cat visual cortex, *Nature*, 380(6571):249–252.
- Hadel, K. P., and Kuhn, D., 1987, Stationary states of the Hartline-Ratcliff model, *Biol. Cybern.*, 56:411–417.
- Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S., 2000, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, *Nature*, 405(6789):947–951.
- Hartline, H. K., and Ratcliff, F., 1972, Inhibitory interaction in the retina of *Limulus*, in *Handbook of Sensory Physiology: Physiology of Photore-*

- ceptor Organs* (M. G. F. Fuortes, Ed.), Berlin: Springer-Verlag, pp. 382–447.
- Li, Z., and Hopfield, J. J., 1989, Modeling the olfactory bulb and its neural oscillatory processes, *Biol. Cybern.*, 61:379–392.
- Robinson, D. A., 1989, Integrating with neurons, *Annu. Rev. Neurosci.*, 12:33–45.
- Seung, H. S., 1996, How the brain keeps the eyes still, *Proc. Natl. Acad. Sci. USA*, 93:13339–13344.
- Seung, H. S., Lee, D. D., Reis, B. Y., and Tank, D. W., 2000, The autapse: A simple illustration of short-term analog memory storage by tuned synaptic feedback, *J. Comput. Neurosci.*, 9:171–185. ♦
- Zhang, K., 1996, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory, *J. Neurosci.*, 16:2112–2126.

Analog Neural Networks, Computational Power

Bhaskar DasGupta and Georg Schnitger

Introduction

The last two decades have seen a surge in theoretical techniques to design and analyze the performance of neural nets as well as novel applications of neural nets to various applied areas. Theoretical studies on the computational capabilities of neural nets have provided valuable insights into the mechanisms of these models.

In subsequent discussion, we distinguish between feedforward neural nets and recurrent neural nets. The architecture of a feedforward net \mathcal{N} is described by an interconnection graph and the activation functions of \mathcal{N} . A node (processor or neuron) v of \mathcal{N} computes a function

$$\gamma_v \left(\sum_{i=1}^k a_{v_i,v} u_{v_i} + b_v \right) \quad (1)$$

of its inputs u_{v_1}, \dots, u_{v_k} . These inputs are either external (i.e., representing the input data) or internal (i.e., representing the outputs of the immediate predecessors of v). The coefficients $a_{v_i,v}$ (respectively b_v) in Equation 1 are the *weights* (respectively *threshold*) of node v , and the function γ_v is the *activation function* of v . No cycles are allowed in the interconnection graph, and the output of designated nodes provides the output of the network. A recurrent neural net, on the other hand, allows cycles, thereby providing potentially higher computational capabilities. The *state* u_v of node v in a recurrent net is updated over time according to

$$u_v(t+1) = \gamma_v \left(\sum_{i=1}^k a_{v_i,v} u_{v_i}(t) + b_v \right) \quad (2)$$

In this article, we emphasize the exact and approximate representational power of feedforward and recurrent neural nets. This line of research can be traced back to Kolmogorov (1957), who essentially proved the first existential result on the (exact) representation capabilities of neural nets (cf. UNIVERSAL APPROXIMATORS). The need to work with “well-behaved” activation functions, however, enforces approximative representations of target functions, and the question of the approximation power (with limited resources) becomes fundamental.

The representation power of neural nets has immediate consequences for learning, since we cannot learn (approximately) what we cannot represent (approximately). On the other hand, the complexity of learning increases with increasing representational power of the underlying neural model, and care must be exercised to strike a balance between representational power, on the one hand, and

learning complexity, on the other. The emphasis of this article is on representational power, i.e., what can be represented with networks using a given set of activation functions, rather than on learning complexity.

In this article, we discuss only a small subset of the literature on this topic. After introducing the basic notation, we discuss the representational power of feedforward and recurrent neural nets. There follows a brief discussion of networks of spiking neurons and their relation to sigmoidal nets, with a summary statement.

Models and Basic Definitions

In this section we present the notation and basic definitions used in subsequent sections. For real-valued functions we measure the approximation quality of function f by function g (over a domain $D \subseteq \mathbb{R}^n$) by the Chebychev norm,

$$\|f - g\|_D = \sup\{|f(x) - g(x)| : x \in D\}$$

(the subscript D will be omitted when clear from the context). To emphasize the selection of activation functions, we introduce the concept of Γ -nets for a class Γ of real-valued activation functions. A Γ -net \mathcal{N} assigns only functions in Γ to nodes. We assume that each function in Γ is defined on some subset of \mathbb{R} , and require that Γ contain the identity function by default (thus allowing weighted additions as node outputs). Finally, we restrict our attention to Γ -nets with a single output node.

The *depth* of a feedforward net \mathcal{N} is the length of the longest path of the (acyclic) interconnection graph of \mathcal{N} , and the *size* of \mathcal{N} is the number of nodes. The *hidden nodes* are all nodes excluding all input nodes and the output node.

The class of important activation functions is rather large and includes, among others, the binary threshold function

$$\mathcal{H}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0, \end{cases}$$

the cosine squasher, the Gaussian, the standard sigmoid $\sigma(x) = 1/(1 + e^{-x})$, the hyperbolic tangent, (generalized) radial basis functions, polynomials and trigonometric polynomials, splines, and rational functions.

Care must be exercised when using a neural net with continuous activation functions to compute a Boolean-valued function, since

in general, the output node computes a real number. A standard output convention in this case is as follows (see Maass, 1994):

Definition 1. A Γ -net \mathcal{N} computes a Boolean function $F: \mathbb{R}^n \rightarrow \{0, 1\}$ with separation $\varepsilon > 0$ if there is some $t \in \mathbb{R}$ such that for any input $x \in \mathbb{R}^n$, the output node of \mathcal{N} computes a value that is at least $t + \varepsilon$ if $F(x) = 1$, and at most $t - \varepsilon$ otherwise.

Recurrent neural nets, unlike their feedforward counterparts, allow loops in their interconnection graph. Certainly *asynchronous* recurrent nets are an important neural model, but we assume in Equation 2 that all nodes update *synchronously* at each time step. Typically, besides internal and external data lines, some of the inputs and outputs are validation lines, indicating if there is any input or output present at the time.

Computational Power of Feedforward Nets

The simple perceptron as a feedforward neural net with one layer has only limited computational abilities. For instance, if we restrict ourselves to one-node simple perceptrons and assume monotone, but otherwise arbitrary, activation functions, then the XOR function $\text{XOR}(x_1, x_2) = x_1 \oplus x_2$ cannot be computed.

On the other hand, if we choose the binary threshold function \mathcal{H} as activation function, then the learning problem for simple perceptrons is efficiently solvable by linear programming. This positive result is also extendable to a large class of activation functions, including the standard sigmoid. But simple perceptrons should not be underestimated, since the problem of approximately minimizing the misclassification ratio (when the target function is not representable as a simple perceptron) has been shown to be (probably) intractable (Arora et al., 1997).

However, the power of feedforward nets increases significantly when networks of more layers are considered. In fact, a result of Kolmogorov (refuting Hilbert's 13th problem for continuous functions), when translated into neural net terminology, shows that any continuous function can be computed *exactly* by a feedforward net of depth 3.

Theorem 1 (Kolmogorov, 1957). Let n be a natural number. Then there are continuous functions $h_1, \dots, h_{2n+1}: [0, 1] \rightarrow \mathbb{R}$ such that any continuous function $f: [0, 1]^n \rightarrow \mathbb{R}$ can be represented as

$$f(x) = \sum_{j=1}^{2n+1} g\left(\sum_{i=1}^n \alpha_i h_j(x_i)\right)$$

where the function g as well as the weights $\alpha_1, \dots, \alpha_n$ depend on f .

But, unfortunately, the function g depends on the function to be represented. Moreover, the functions h_j are nondifferentiable and hence cannot be used by current learning algorithms. For further discussion, we refer the reader to Poggio and Girosi (1989).

However, if we only allow everywhere differentiable activation functions (such as the standard sigmoid), then we can only represent everywhere differentiable target functions. Thus, one has to relax the requirement of exact representation, and demand only that the approximation error (in an appropriate norm) is small. Applying the Stone-Weierstrass theorem one obtains, for instance, (trigonometric) polynomials as universal approximators, and hence we get neural nets with one hidden layer as universal approximators.

Cybenko (1989) considers activation functions from the class of continuous *discriminatory* functions. This class contains, for instance, *sigmoidal* functions, i.e., continuous functions σ satisfying

$$\sigma(t) \rightarrow \begin{cases} 1 & \text{as } t \rightarrow +\infty \\ 0 & \text{as } t \rightarrow -\infty \end{cases}$$

Theorem 2. Let σ be a continuous discriminatory function and let $f: [0, 1]^n \rightarrow \mathbb{R}$ be a continuous target function. Then, for every $\varepsilon > 0$ and for sufficiently large N (where N depends on the target function f and ε), there exist weights α_{ij} , w_j and thresholds β_j , such that $\|f - g\| < \varepsilon$, where $g = \sum_{j=1}^N w_j \cdot \sigma(\sum_{i=1}^n \alpha_{ij} \cdot x_i + \beta_j)$.

In particular, one hidden layer suffices to approximate any continuous function by sigmoidal nets within arbitrarily small error. Further results along this line are shown by Hornik; Stinchcombe and White; Funahashi, Moore, and Poggio; and Poggio and Girosi, to mention just a few. Whereas most arguments in the above-mentioned results are nonconstructive, Carroll and Dickinson describe a method using Radon transforms to approximate a given L_2 function to within a given mean square error.

Barron (1993) discusses the approximation quality achievable by sigmoidal nets of small size. In particular, let $B_r^n(0)$ denote the n -dimensional ball with radius r around 0 and let $f: B_r^n(0) \rightarrow \mathbb{R}$ be the target function. Assume that F is the magnitude distribution of the Fourier transform of f .

Theorem 3 (Barron, 1993). Let σ be any sigmoidal function. Then for every probability measure μ and for every N there exist weights α_{ij} , w_j and thresholds β_j , such that

$$\int_{B_r^n(0)} \left(f(x) - \sum_{j=1}^N w_j \cdot \sigma\left(\sum_{i=1}^n \alpha_{ij} \cdot x_i + \beta_j\right) \right)^2 \mu(dx) \leq \frac{(2fr \cdot |w| \cdot F(dw))^2}{N}$$

Set $C_f = \int r \cdot |w| \cdot F(dw)$, and the approximation error achievable by sigmoidal nets of size N is bounded by $(2 \cdot C_f)^2/N$. However, C_f may depend superpolynomially on n , and the curse of dimensionality may strike. As an aside, the best achievable squared error for linear combinations of N *basis functions* will be at least $\Omega(C_f/n \cdot N^{1/n})$ for certain functions f (Barron, 1993), and hence neural networks are superior to conventional approximation methods from this point of view.

The results just enumerated show that depth-2 feedforward nets are universal approximators. This dramatically increased computing power, however, has a rather negative consequence. Kharitonov (1993) showed that under certain cryptographic assumptions, no efficient learning algorithm will be able to predict the input-output behavior of binary threshold nets with a fixed number of layers. This result holds even when experimentation is allowed, that is, when the learning algorithm is allowed to submit inputs for classification.

In the next section, we compare important activation functions in terms of their approximation power, when resources such as depth and size are limited. The following section discusses networks of spiking neurons. Lower size bounds for sigmoidal nets are mentioned when we compare networks of spiking neurons and sigmoidal nets.

Efficient Approximation by Feedforward Nets

Our discussion will be informal, and we refer the reader to DasGupta and Schnitger (1993) for details. Our goal is to compare activation functions in terms of the size and depth required to obtain tight approximations. Another resource of interest is the *Lipschitz bound* of the net, which is a measure of the numerical stability of the circuit. Informally speaking, for a net \mathcal{N} to have Lipschitz-bound L , we first demand that all weights and thresholds of \mathcal{N} be bounded in absolute value by L . Moreover, we require that each activation function of \mathcal{N} have (the conventional) Lipschitz-bound L on the inputs it receives. Finally, the actually received inputs

have to be bounded away from regions with higher Lipschitz bounds.

We formalize the notion of having *essentially the same approximation power*.

Definition 2. Let Γ_1 and Γ_2 be classes of activation functions.

- (a) We say that Γ_2 simulates Γ_1 (denoted by $\Gamma_1 \leq \Gamma_2$) if and only if there is a constant $k \geq 1$ such that for all Γ_1 -nets C_2 of size at most s , depth at most d , and Lipschitz bound 2^s , there is a Γ_2 -circuit C_1 of size at most $(s+1)^k$, depth at most $k \cdot (d+1)$, and Lipschitz bound $2^{(s+1)^k}$, such that

$$\|C_1 - C_2\|_{[-1,1]^n} \leq 2^{-s}$$

- (b) We say that Γ_1 and Γ_2 are equivalent if and only if $\Gamma_1 \leq \Gamma_2$ and $\Gamma_2 \leq \Gamma_1$.

In other words, when simulating classes of gate functions, we allow depth to increase by a constant factor size and the logarithm of the Lipschitz bound to increase polynomially. The relatively large Lipschitz bounds should not come as a surprise, since the negative exponential error 2^{-s} requires correspondingly large weights in the simulating circuit.

Splines (i.e., piecewise polynomial functions) have turned out to be powerful approximators, and they are our benchmark class of activation functions; in particular, we assume that a spline net of size s has as its activation functions splines of degree at most s with at most one knot. Which properties does a class Γ of activation functions need to reach the approximation power of splines? The activation functions should be able to approximate polynomials as well as the binary threshold \mathcal{H} with few layers and relatively few nodes.

Tightly approximating polynomials is not difficult as long as there is at least one “sufficiently smooth” nontrivial function $\gamma \in \Gamma$. The crucial problem is to obtain a tight approximation of \mathcal{H} . It turns out that γ -nets achieve tight approximations of \mathcal{H} whenever

$$|\gamma(x) - \gamma(x + \varepsilon)| = O(\varepsilon/x^2) \quad \text{for } x \geq 1, \quad \varepsilon \geq 0 \quad \text{and} \quad \left| \int_1^\infty \gamma(u^2) du \right| \neq 0$$

Let us call a function with these two properties *strongly sigmoidal*. (We are actually demanding too much, since it suffices to tightly approximate a strongly sigmoidal function γ by small Γ -nets with few layers.) Let us call a class Γ *powerful* if there is at least one “sufficiently smooth” nontrivial function in Γ and if a strongly sigmoidal function can be approximated as demanded above.

Examples of powerful singleton classes include, for instance, $1/x$ as a prime example, and more generally any rational function that is not a polynomial, $\exp(x)$ (since $\exp(-x)$ is strongly sigmoidal) and $\ln(x)$ (since $\ln(x+1) - \ln(x)$ is strongly sigmoidal), any power x^α provided α is not a natural number, and the standard sigmoid as well as the Gaussian $\exp(-x^2)$.

Theorem 4.

- (a) Assume that Γ is powerful. Then splines $\leq \Gamma$.
 (b) The following classes of activation function have equivalent approximation power: splines (of degree s for nets of size s), any rational function that is not a polynomial, any power x^α (provided α is not a natural number), the logarithm (for any base), $\exp(x)$, and the Gaussian $\exp(-x^2)$.

Notably missing from the list of equivalent activation functions are polynomials, trigonometric polynomials, and the binary threshold function \mathcal{H} (or, more generally, low-degree splines). Low-

degree splines turn out to be properly weaker. The same applies to polynomials, even if we allow any polynomial of degree s an activation function for nets of size s . Finally, sine nets cannot be simulated (as defined in Definition 2), for instance by nets of standard sigmoids, and we conjecture that the reverse is also true, namely, that nets of standard sigmoids cannot be simulated efficiently by sine nets.

What happens if we relax the required approximation quality from 2^{-s} to s^{-d} , when simulating nets of depth d and size s ? Linear splines and the standard sigmoid are still not equivalent, but the situation changes completely if we *count* the number of inputs when determining size and if we restrict the Lipschitz bound of the target function to be at most s^{-d} . With this modification an even larger class of important functions, including linear splines, polynomials, and the sine function, turn out to be equivalent with the standard sigmoid.

The situation for Boolean input and output is somewhat comparable. Maass, Schnitger, and Sontag, and subsequently DasGupta and Schnitger constructed Boolean functions that are computed by sigmoidal nets of constant size (i.e., independent of the number of input bits), whereas \mathcal{H} -nets of constant size do not suffice. (See Maass, 1994, for a more detailed discussion.) However, Maass (1993) showed that spline nets of constant degree, constant depth, and polynomial size (in the number of input bits) can be simulated by \mathcal{H} -nets of constant depth and polynomial size. This simulation holds without any restriction on the weights used by the spline net.

Thus, analog computation does help for discrete problems, but apparently by not too much. For a thorough discussion of discrete neural computation, see Siu, Roychowdhury, and Kailath (1994).

Sigmoidal Nets and Nets of Spiking Neurons

A formal model of networks of spiking neurons is defined in Maass (1997); see SPIKING NEURONS, COMPUTATION WITH. The architecture is described by a directed graph $G = (V, E)$, with V as the set of nodes and E as the set of edges. We interpret nodes as neurons and edges as synapses, and assign to each neuron v a threshold function $\Theta_v: \mathbb{R}^+ \rightarrow \mathbb{R}^+$. (\mathbb{R}^+ denotes the set of nonnegative reals.) The value of $\Theta_v(t - t')$ measures the “reluctance” (or the threshold to be exceeded) of neuron v to fire at time t ($t > t'$), assuming that v has fired at time t' . This reluctance can be overcome only if the potential $P_v(t)$ of neuron v at time t is at least correspondingly as large.

The potential of v at time t depends on the recent firing history of the presynaptic neurons (or the immediate predecessors) u of v . In particular, if the synapse between neurons u and v has the efficacy (or weight) w_{uv} , if $\varepsilon_{uv}(t - s)$ is the response to the firing of neuron u at time s ($s < t$) and if the presynaptic neuron u has fired previously for the times in the set Fire_u^t , then the potential at time t is defined as

$$P_v(t) = \sum_{(u,v) \in E} \sum_{s \in \text{Fire}_u^t} w_{uv} \cdot \varepsilon_{uv}(t - s) \quad (3)$$

Two models, namely deterministic (respectively noisy) nets of spiking neurons, are distinguished. The deterministic version assumes that neuron v fires whenever its potentials $P_v(t)$ reach $\Theta_v(t - t')$ (with most recent firing t'), whereas the more realistic noisy version assumes that the firing probability increases with increasing difference $P_v(t) - \Theta_v(t - t)$.

Thus we can complete the definition of the formal model, assuming that a response function $\varepsilon_{uv}: \mathbb{R}^+ \rightarrow \mathbb{R}$ as well as the weight w_{uv} is assigned to the synapse between u and v . The model computes by transforming a spike train of inputs into a spike train of outputs. For instance assuming temporal coding with constants T and c , the output of a designated neuron with firing times $T + c \cdot t_1, \dots, T + c \cdot \sum_{i=1}^k t_i, \dots$ is defined as $t_1, \dots, \sum_{i=1}^k t_i, \dots$.

The power of spiking neurons shows for the example of the element distinctness function ED_n with real inputs x_1, \dots, x_n , where

$$ED_n(x) = \begin{cases} 1 & \text{if } x_i = x_j \text{ for some } i \neq j, \\ 0 & \text{if } |x_i - x_j| \geq 1 \text{ for all } i \neq j, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

We assume that the inputs x_1, \dots, x_n are represented by n input trains of single spikes. Now it is easy to choose a simple threshold function as well as simple (and indeed identical) response functions such that even a single spiking neuron with unit weights is capable of computing ED_n . On the other hand, any sigmoidal net computing ED_n requires at least $(n - 4)/2 - 1$ hidden units (Maass, 1997). This result is also the strongest lower size bound for sigmoidal nets computing a specific function; the argument builds on techniques from Sontag (1997).

Certainly this one-neuron computation requires time, because of the temporal input coding, but the same applies to sigmoidal networks, since, from the point of neurobiology, the x_i 's will be obtained after sampling the firing rate of their input neurons.

Nets of spiking neurons are capable of simulating \mathcal{H} -nets with at most the same size, and hence are properly stronger than \mathcal{H} -nets and at least in some cases stronger than sigmoidal nets. Thus, careful timing is an advantage that synchronized models cannot overcome.

Computational Power of Recurrent Nets

The computational power of recurrent nets is investigated in Siegelmann and Sontag (1994, 1995). (See also Siegelmann, 1998, for a thorough discussion of recurrent nets and analog computation in general.) Recurrent nets include feedforward nets, and thus the results for feedforward nets apply to recurrent nets as well. But recurrent nets gain considerably more computational power with increasing computation time. In the following discussion, for the sake of concreteness, we assume that the piecewise linear function

$$\pi(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

is chosen as the activation function. We concentrate on binary input and assume that the input is provided one bit at a time.

First of all, if weights and thresholds are integers, then each node computes a bit. Recurrent nets with integer weights thus turn out to be equivalent to finite automata, and they recognize exactly the class of regular language over the binary alphabet $\{0, 1\}$.

The computational power increases considerably for rational weights and thresholds. For instance, a "rational" recurrent net is, up to a polynomial time computation, equivalent to a Turing machine. In particular, a network that simulates a universal Turing machine does exist, and one could refer to such a network as "universal" in the Turing sense. It is important to note that the number of nodes in the simulating recurrent net is fixed (i.e., *does not grow* with increasing input length).

Irrational weights provide a further boost in computation power. If the net is allowed exponential computation time, then arbitrary Boolean functions (including noncomputable functions) are recognizable. However, if only polynomial computation time is al-

lowed, then nets have less power and recognize exactly the languages computable by polynomial-size Boolean circuits.

Discussion

We have discussed the computing power of neural nets, including universal approximation results for feedforward and recurrent neural networks as well as efficient approximation by feedforward nets with various activation functions. We emphasize that this survey is far from complete. For instance, we omitted important topics such as the VAPNIK-CHEVONENKIS DIMENSION OF NEURAL NETWORKS (q.v.) and the complexity of discrete neural computation.

Important open questions relate to proving better upper and lower bounds for sigmoidal nets computing (or approximating) specific functions, and achieving a better understanding of size and depth trade-offs for important function classes. Other neural models, such as networks of spiking neurons, significantly change the computing power, and the questions we have identified apply to these models as well.

Road Map: Computability and Complexity

Background: I.3. Dynamics and Learning in Neural Networks

Related Reading: Neural Automata and Analog Computational Complexity; PAC Learning and Neural Networks; Universal Approximators

References

- Arora, S., Babai, L., Stern, J., and Sweedyk, Z., 1997, The hardness of approximate optima in lattices, codes and systems of linear equations, *J. Comput. Syst. Sci.*, 54:317–331.
- Barron, A. R., 1993, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory*, 39:930–945.
- Cybenko, G., 1989, Approximation by superposition of a sigmoidal function, *Math. Control Signals Syst.*, 2:303–314.
- DasGupta B., and Schnitger, G., 1993, *The Power of Approximating: A Comparison of Activation Functions*, NIPS 5, 615–622. Available: <http://www.cs.uic.edu/~dasgupta/resume/publ/papers/approx.ps.Z>
- Kharitonov, M., 1993, Cryptographic hardness of distribution specific learning, in *Proceedings of the 25th ACM Symposium on the Theory of Computing*, pp. 372–381.
- Kolmogorov, A. N., 1957, On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk USSR*, 114:953–956.
- Maass, W., 1993, Bounds for the computational power and learning complexity of analog neural nets, in *Proceedings of the 25th Annual ACM Symposium on the Theory of Computing*, pp. 335–344.
- Maass, W., 1994, Sigmoids and Boolean threshold circuits, in *Theoretical Advances in Neural Computation and Learning* (V. P. Roychowdhury, K. Y. Siu, and A. Orłitsky, Eds.), Boston: Kluwer, pp. 127–151.
- Maass, W., 1997, Networks of spiking neurons: The third generation of neural network models, *Neural Netw.*, 10:1659–1671.
- Poggio, T., and Girosi, F., 1989, A theory of networks for approximation and learning, *Artif. Intell. Memorandum*, No. 1140.
- Siegelmann, H. T., 1998, *Neural Networks and Analog Computation: Beyond the Turing Limit*, Boston: Birkhäuser. ♦
- Siegelmann, H. T., and Sontag, E. D., 1994, Analog computation, neural networks, and circuits, *Theoret. Comput. Sci.*, 131:331–360.
- Siegelmann, H. T., and Sontag, E. D., 1995, On the computational power of neural nets, *J. Comput.*, 50:132–150.
- Siu, K.-Y., Roychowdhury, V. P., and Kailath, T., 1994, *Discrete Neural Computation: A Theoretical Foundation*, Englewood Cliffs, NJ: Prentice Hall. ♦
- Sontag, E. D., 1997, Shattering all sets of k points in general position requires $(k - 1)/2$ parameters, *Neural Computat.*, 9:337–348.

Analog VLSI Implementations of Neural Networks

Paul Hasler and Jeff Dugger

Introduction

The primary goal of analog implementations of neural networks is to incorporate some level of realistic biological modeling of adaptive systems into engineering systems built in silicon. We cannot simply duplicate biological models in silicon media because the constraints imposed by the biological media and the silicon media are not identical. Approaches that have been successful begin with the constraints that the silicon medium imposes on the learning system. Therefore, letting the silicon medium constrain the design of a system results in more efficient methods of computation.

We will focus our attention on issues concerning building neural network integrated circuits (ICs), and in particular on building connectionist neural network models. Connectionist neural networks, loosely based on biological computation and learning, can be useful for biological modeling if the limitations are understood. These neural systems are typically built as mappings of mathematical models into analog silicon hardware either by using standard building blocks (i.e., Gilbert multipliers: Mead, 1989) or by taking advantage of device physics (Hasler et al., 1995). This approach, related to investigations of adaptation and learning in neurobiological systems, provides the minimum necessary model of synaptic interaction between neurons, even for biological models. Neuromorphic (Mead, 1989, see also NEUROMORPHIC VLSI CIRCUITS AND SYSTEMS) and connectionist approaches develop adaptive systems from different levels of neural inspiration, and therefore lead to different levels of model complexity. Adding dendritic interactions and precise models of biological learning (e.g., LTP) to the connectionist model yields more biological realism. Implementations of fuzzy systems typically follow a similar approach to implementations of neural networks. The related field of cellular neural networks (CNNs), started by Chua, is particularly concerned with the circuit techniques used to build locally connected two-dimensional (2D) meshes of neuron processors (Chua, 1998), but the architecture design is fundamentally different and imposes different constraints on implementation.

Neural Network Basics Focused on Implementation Issues

Figure 1 shows the basic feedforward structure typically used in neural network implementations. Most approaches focus on feedforward structures, since feedback systems and networks with time dynamics (e.g., time delays) are straightforward extensions for silicon implementation, although the algorithm design is considerably more difficult. In this model, we encode a neuron's activity as an analog quantity based on the mean spiking rate in a given time window. One can build linear or nonlinear filters at the input to the sigmoid function. Typically, a low-pass filter is built or modeled, since that will naturally occur for a given implementation or will set a desired convergence to an attractor (i.e., recurrent networks). This model is excellent for describing neurobiology if only mean-firing-rate behavior with minimal dendritic interactions is considered.

A basic model synapse (either digital or analog) must be able to store a weight, multiply its input with the stored weight, and adapt that weight based on a function of the input and a feedback error signal. We model feedforward computation mathematically as

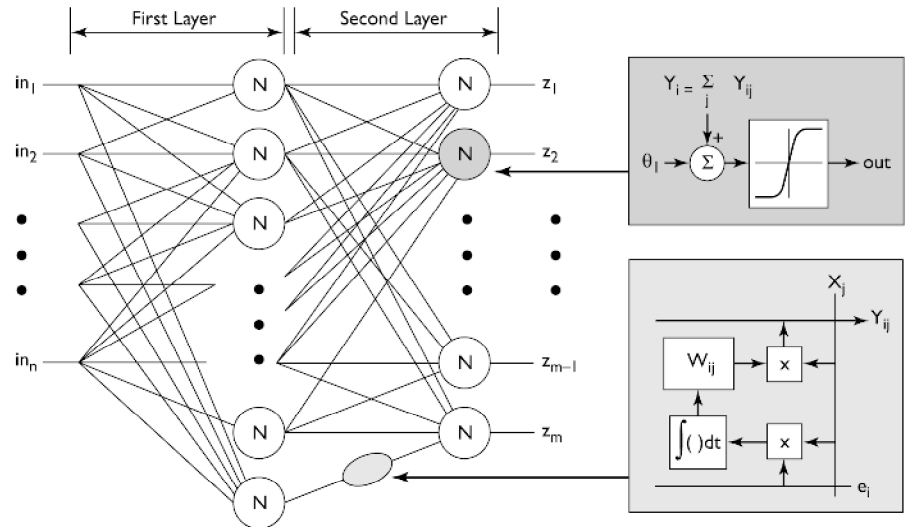
$$y_i = w_{ij}x_j \rightarrow \mathbf{y} = \mathbf{W}\mathbf{x} \quad (1)$$

where x_j is the j th input (\mathbf{x} is a vector of inputs), y_i is the i th output (\mathbf{y} is a vector of outputs), and w_{ij} is the stored weight at position (i,j) (\mathbf{W} is a matrix of weights). The result of this output is passed through a nonlinear function

$$z_i = \tanh(a(y_i - \theta_i)) \quad (2)$$

where we designate z_i as the result of the computation, a is a gain factor, and θ_i is a variable threshold value. Other nonlinear functions, like radial basis functions (see RADIAL BASIS FUNCTION NETWORKS), are also often used, which would typically modify the

Figure 1. Classic picture of a two-layer neural network from the perspective of implementing these networks in hardware. The neural networks are layers of simple processors, called neurons, that are interconnected through weighting elements, called synapses. The neurons aggregate the incoming inputs (including a threshold or offset) and are applied through a $\tanh(\cdot)$ nonlinearity. The synapse elements, which in general are far more numerous than the neuron elements, must multiply the incoming signal by an internally stored value, called the weight, and must adapt this weight according to a particular learning rule. Learning rules implemented in silicon are typically functions of correlations of signals passing through each synapse processor.



Wx computation. We model the weight adaptation mathematically as

$$\tau \frac{d\mathbf{W}}{dt} = f(\mathbf{W}, \mathbf{x}\mathbf{e}^T) \quad (3)$$

where \mathbf{e} is a vector of error signals that is fed back along various rows. We call this an outer-product learning rule, or a *local* learning rule, because of the $\mathbf{x}\mathbf{e}^T$ computation. The outer-product learning rule is dependent on the choice of $f(\mathbf{W}, \mathbf{x}\mathbf{e}^T)$ and the choice of the error signal.

Several learning algorithms have been proposed that conform to this functional form; representative examples can be found elsewhere in the *Handbook*. Learning algorithms usually divide into two categories, supervised and unsupervised. *Supervised algorithms* adapt the weights based on the input signals and a supervisory signal to train the network to produce an appropriate response. In many supervised algorithms (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION) this weight change is a time average of the product of the input and some fed-back error signal ($\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the target signal). *Unsupervised algorithms* adapt the weights based only on the input and output signals, and in general the weights are a function of the input statistics. Although these learning algorithms result in very different results, both weight-update rules are similar from an implementation viewpoint. Most unsupervised algorithms are based on Hebbian learning algorithms, which correspond to neurobiological evidence of learning (see HEBBIAN SYNAPTIC PLASTICITY). For a Hebbian synapse, the weight change is a time average of the product of the input and output activity ($\mathbf{e} = \mathbf{y}$).

Neural Network Implementations: Architecture Issues

Before considering circuit implementations of neurons and synapses, we first frame the overall architecture issues involved in implementing neural networks. In most implementations, a single layer of synapses is built as mesh architectures connected to a column of neuron processors (Figure 2). Because silicon ICs are 2D, mesh architectures work optimally with 2D routing constraints.

Feedforward Computation

Figure 2A shows the typical mesh implementation of feedforward computation for a single-layer architecture. A mesh of processors is an optimal communication architecture for interconnect limited systems, which is the case for small synapse elements. Currents are preferred for outputs, because the summation typically required for most connectionist models is easily performed on a single wire, and voltages are preferred for inputs because they are easy to broadcast. Local processing is defined as interaction between *physically close* elements, voltage broadcast along global lines (inputs), or current/charge summation along a wire (outputs). As a result, each synapse has only to compute the local computation: $\mathbf{W}_{ij}\mathbf{x}_j$. Because the synapses store a weight value, the picture in Figure 2A resembles an analog memory that allows a full matrix-vector multiplication in the equivalent of one memory column access. This approach, called analog computing arrays, is defined and its implication for signal processing is described elsewhere (Kucic et al., 2001). Figure 2B shows how to modify a mesh architecture when considering *m*-nearest-neighbor connections. Other sparse

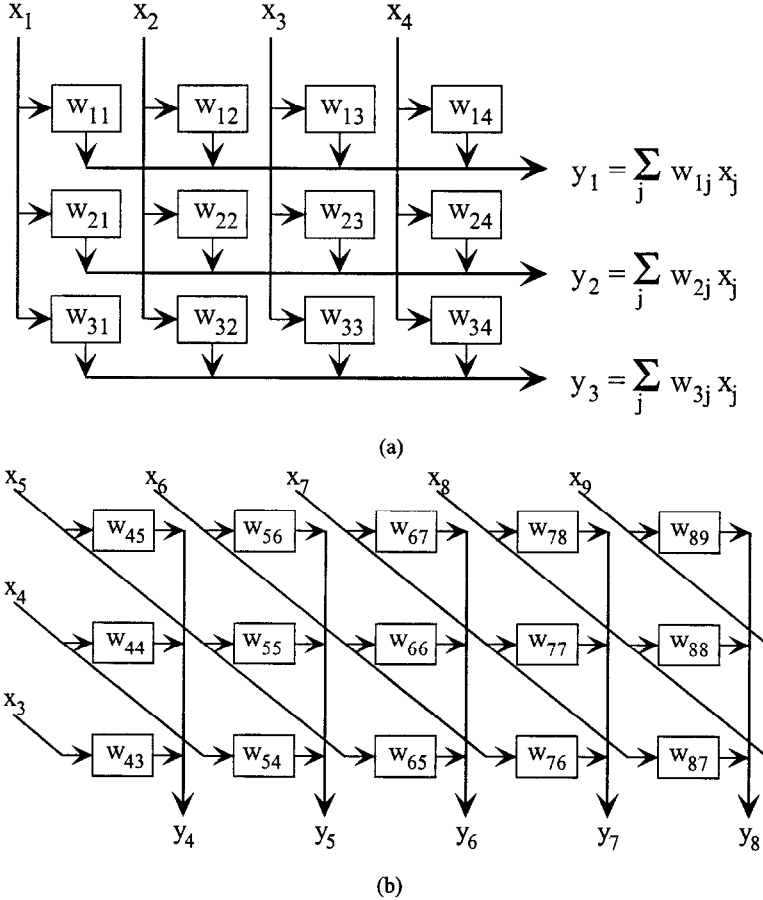


Figure 2. Typical architectures for neural network implementations. Although the routing looks complicated in Figure 1, it can easily be implemented in a mesh architecture. A, Diagram of the classic mesh architecture, typically used for fully connected architectures. B, Diagram of a mesh processor architecture optimized for nearest-neighbor computations.

encodings require digital communication and processing to handle the addressing schemes (i.e., address translation tables) and additional complexity (i.e., multiplexing scheme to access the inputs of each synapse).

To implement a neuron, we need a function that can compute a $\tanh(\cdot)$ function. Fortunately, this function occurs in many IC circuits using either BJT or MOSFET (subthreshold or above-threshold) devices, such as the differential transistor pair (Mead, 1989). Since we only need a column of neuron circuits, they do not have the same area constraints that are imposed on synapse elements. Dynamics (e.g., low-pass filtering) are usually achieved by adding additional capacitance. Often one needs a current to perform voltage conversion between the summed synapse outputs and $\tanh(\cdot)$ output, as well as at the output of a differential transistor pair. This conversion often can be nonlinear, or it may have to be nonlinear to interface with later processing stages.

Adaptive Neural Network Architectures

Synapses require both feedforward and adaptation computations; therefore, architectural constraints imposed by the learning algorithm are an essential consideration for any neural network. Only learning algorithms that scale to large numbers of inputs and outputs are practical. A single-layer architecture with a local supervised or unsupervised rule of the form of Equation 3 only requires communicating the *error* signal along each row (Figure 3). The complexity of the synapse computation will depend on the particular learning rule. Many complicated algorithms, such as the generalized Hebbian algorithm (GHA) (Hasler and Akers, 1992) and INDEPENDENT COMPONENT ANALYSIS (ICA) (q.v.), require additional matrix-vector multiplications, but can be developed into a mesh architecture. Algorithms requiring matrix-matrix multiplications are difficult in standard IC technologies.

For multilayer algorithms, the architecture gets more complicated, particularly for supervised algorithms such as multilayer backpropagation. To extend the basic silicon synapse to a back-propagating synapse, we need an additional function: we need an output current that is the product of the fed-back error signal (drain voltage) and stored weight. We show this architecture in Figure 4A. This additional function results in two issues, one concerning the signal-to-noise ratio of the resulting error signal and the other concerning the overall synapse size. The effect of these small error signals, even without the resolution issues, is a slow learning rate.

The neural network literature is replete with possible alternative approaches, but we will base our proposed research on the Helm-

holtz machine concept (see HELMHOLTZ MACHINES AND SLEEP-WAKE LEARNING). Our primary reason for using this approach rests on our desire to use single-layer networks as primitives for building larger networks, as well as the fact that this reciprocal adaptive single-layer network architecture is seen in various models of sensory neurosystems, such as the pathways from retina to LGN to V1 or some of the pathways between the cochlea and auditory cortex (A1). Figure 4B considers a two-layer network implementation of a backpropagation-like learning rule using this Helmholtz block. In this case, we double the number of layers, and therefore double the effective synapse size; for a backpropagation algorithm, we require the same number of floating-gate multipliers, but with significant additional implementation costs that greatly increase the synapse complexity. This approach seems more IC-friendly for the development of adaptive multilayer algorithms than backpropagation approaches, although its digital implementation is nominally equivalent to backpropagation approaches. This approach directly expands to multiple layers and could be used in limited reconfigurable networks because we are building networks with single adaptive layers. Starting with the single-layer network as the basic building block simplifies the abstraction toward system development.

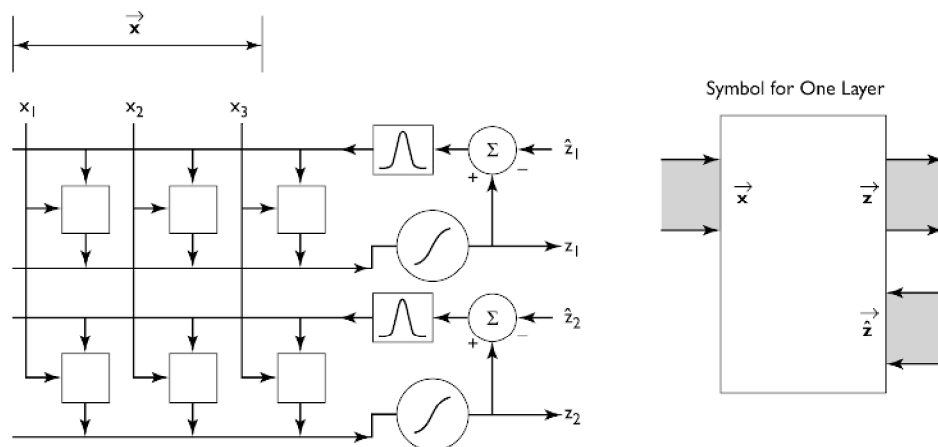
Resulting Synapse Design Criteria

Because the synapse is the critical element of any neural network implementation, we state five properties of a silicon synapse that are essential for building large-scale adaptive analog VLSI synaptic arrays (Hasler et al., 1995):

1. The synapse must store a weight permanently in the absence of learning.
2. The synapse must compute an output current as the product of its input signal and its synaptic weight.
3. The synapse must modify its weight at least using outer-product learning rules.
4. The synapse must consume minimal silicon area, thereby maximizing the number of synapses in a given area.
5. The synapse must dissipate a minimal amount of power; therefore, the synaptic array is not power constrained.

Achieving all five requirements requires a detailed discussion of the circuits used to implement a synapse, which is the subject of the next section.

Figure 3. Learning in a single layer. We can build either supervised algorithms (LMS is explicitly shown) or unsupervised one-layer networks in this architecture. For a one-layer supervised case, $\hat{\mathbf{z}}$ is the desired or target output signal vector, where $e_j = z_j - \hat{z}_j$. Further, one might apply a nonlinear function to the resulting error signal; in LMS, one applies a nonlinear function to counteract the effect of the sigmoid in the feedforward path. Many unsupervised rules, like Hebbian or Oja rules, can be formulated as $\hat{\mathbf{z}} = \mathbf{f}(\mathbf{z})$. One can schematically represent this network from its terminals, \mathbf{x} , \mathbf{z} , and $\hat{\mathbf{z}}$, as shown from its block diagram. Finally, the nonlinear (sigmoid) elements typically convert current to voltage.



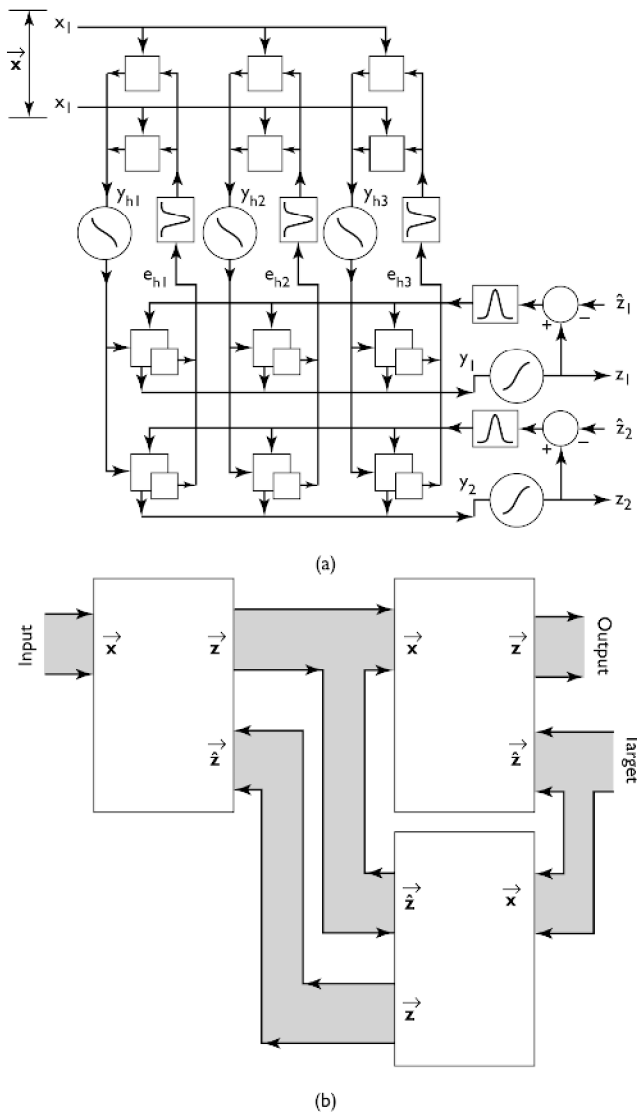


Figure 4. Possible architectures for adaptive multilayer neural networks. A, Implementation for backpropagation networks. There are many forms and modifications, but from an implementation viewpoint, these approaches can be modified toward this architecture. This approach significantly increases synapse size, because one typically requires the complexity of two synapses for weight feedback. Further, this approach limits some circuit approaches to building dense synapses. The output from the hidden layer, or layer 1, is y_h and the error signal given to the hidden layer is e_h . The synapses in the second layer must also output a current proportional to the product of error * stored weight; the sum of these currents along a column is the error for the next layer. As a result, the synapses on the second layer must be more complex. B, Implementation using Helmholtz machine concepts. This approach requires twice as many synapses for all but the first layer, which yields the same complexity as the backpropagation approaches. This approach will converge to the same steady states and requires only a modular tiling of single-layer networks; its reciprocal feedback has a similar feel to communication between layers of cortical neurons.

Neural Network Implementation: Synapse Circuits

Early Research in Synapse Design

Several neural networks have been built in analog silicon hardware. Several good recent implementation techniques can be found in

Cauwenberghs and Bayoumi (1999); here we present an overview. From the architecture discussions, we require a synapse block where an input voltage should modulate an output current, which is summed along a line; therefore, most implementations employ a variable resistance or transconductance element. As a result, a primary issue in synapse circuit designs is developing dense multiplier circuits, because multiplication of an input by a weight is fundamental to every synapse. Earlier approaches for implementing the feedforward synapse computation included fixed resistances (which were among the earliest implementations), switched-capacitor implementations (Tsividis and Satyanarayana, 1987), Gilbert multiplier cells (Mead, 1989), and linearized conductance elements (Dupuis and Ismail, 1990; Cauwenberghs, Neugebauer, and Yariv, 1991; Hasler and Akers, 1992). Intel's ETANN chip was the first commercially available neural network IC that used floating gates for weight storage (Holler et al., 1989). One of the most successful implementations of a large-scale adaptive neural system was the Heuralt-Juetten algorithm, but it required a great deal of circuit complexity (Cohen and Andreou, 1992). Other researchers have implemented unsupervised learning and backpropagation algorithms, with mixed success (Furman, White, and Abidi, 1988; Hasler and Akers, 1992). Successful analog implementations of connectionist networks have included algorithmic modifications that facilitate implementation in silicon. The history of this field has shown that the success of an implementation is strongly correlated with the degree to which the algorithm is adapted to the silicon medium.

Synapses in previous silicon implementations have required large circuit complexity because they have typically been constructed using traditional circuit building blocks to realize memory, computation, and adaptation functions separately, rather than taking advantage of device physics to combine these functions in a compact circuit element. Not only does large circuit complexity consume tremendous circuit area and power, but the chance of a network operating correctly decreases exponentially with cell size.

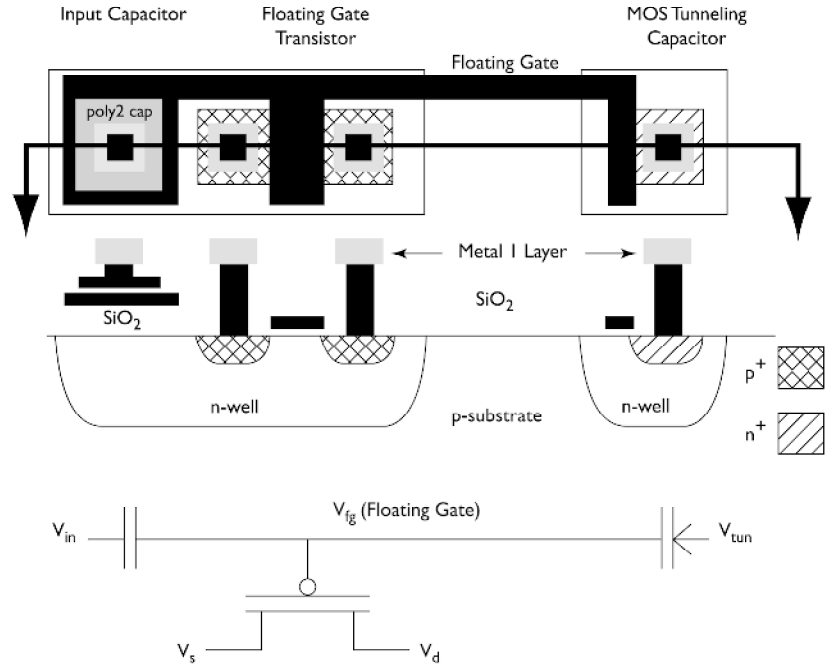
The most difficult problem to overcome when building efficient adaptive circuits is the effect of p-n junction leakage currents (Hasler et al., 1995; Hasler and Minch, 2002). First, since many implementations dynamically store their weight parameters on a capacitor, these junction leakage currents typically limit the hold time, on the order of seconds; therefore, weight storage often becomes a critical concern in many of these applications. Several on-chip refreshing schemes have been proposed and built (Hasler and Akers, 1992) and are currently finding applications in various ICs (Cauwenberghs and Bayoumi, 1999). Second, since real-time learning often requires time constants from 10 ms to days, junction leakage currents limit the use of capacitor storage techniques, unless prohibitively large capacitor areas are used. Weight update schemes based on weight perturbation methods, i.e., where the error signal is based on random known changes in the weights, can often work in these constraints if some form of dynamic refreshing scheme is used (Cauwenberghs and Bayoumi, 1999). Often, junction leakage is too large for many adaptive system problems.

Single-Transistor Learning Synapses

Current research into analog neural network ICs pursues two directions. The first direction is based on refreshable DRAM elements with adaptation using weight perturbation techniques (Cauwenberghs and Bayoumi, 1999). The second direction is based on a wide range of techniques using floating-gate synapses. Floating gates have seen use in neural networks as storage elements (Holler et al., 1989), which eliminates the long-term weight storage issues but still results in fairly complex synapse circuits. Here, we briefly describe the potential of using floating-gate synapses.

The single-transistor learning synapse (STLS), or transistor synapse, makes use of device physics and constraints inherent to the

Figure 5. Layout, cross-section, and circuit diagram of the floating-gate pFET in a standard double-poly n -well MOSIS process. The cross-section corresponds to the horizontal line slicing through the layout view. The pFET transistor is the standard pFET transistor in the n -well process. The gate input capacitively couples to the floating gate by either a poly-poly capacitor, a diffused linear capacitor, or a MOS capacitor, as seen in the circuit diagram (not explicitly shown in the other figures). We add floating-gate charge by electron tunneling, and we remove floating-gate charge by hot-electron injection. The tunneling junctions used by the single-transistor synapses is a region of gate oxide between the polysilicon floating gate and n -well (a MOS capacitor). Between V_{tun} and the floating gate is our symbol for a tunneling junction, a capacitor with an added arrow designating the charge flow.



silicon medium to realize learning and adaptation functions, rather than direct implementation of learning rules using traditional circuit building blocks (Hasler et al., 1995). This technology is rooted in floating-gate circuits (Hasler and Lande, 2001; Hasler and Minch, 2002) in which multiple features of a floating-gate transistor are used, not just the nonvolatile storage (Figure 5). These elements utilize physical characteristics of the silicon medium, such as electron tunneling and hot-electron injection, which traditionally have posed problems for engineers. The starting point for this technology is a floating-gate transistor (Hasler et al., 1995; Kucic et al., 2001) operating with subthreshold currents and configured to simultaneously store permanently the weight charge, compute an output current that is the product of the input signal and the synaptic weight, and modify its weight charge based on many outer-product learning rules. This approach meets all five requirements for a silicon synapse. These weights can be automatically programmed, which enables setting fixed weights, setting initial bias conditions, and employing weight perturbation learning rules (Kucic et al., 2001). Further, by setting the appropriate boundary circuits for the synapse array, we can get a wide range of learning rules by continuously enabling our *programming mechanisms* during computation (Kucic et al., 2001). One form of the learning rules looks like

$$\tau \frac{dw_{ij}}{dt} = \eta E[x_i e_j] - w_{ij} \quad (4)$$

where τ is the adaptation time constant and η is the strength of the correlating term.

Road Map: Implementation and Analysis

Related Reading: Digital VLSI for Neural Networks; Photonic Implementations of Neurobiologically Inspired Networks; Silicon Neurons

References

- Chua, L. O., 1998, *A Paradigm for Complexity*, vol. 31, in World Scientific Series on Nonlinear Science, Series A, Singapore: World Scientific Publishing.
- Cauwenberghs, G., and Bayoumi, M. A., Eds., 1999, *Learning in Silicon*, Boston: Kluwer Academic.
- Cauwenberghs, G., Neugebauer, C., and Yariv, A., 1991, An adaptive CMOS matrix vector multiplier for large scale analog hardware neural network applications, in *Proceedings of the International Joint Conference on Neural Networks*, vol. 1, pp. 507–512.
- Cohen, M., and Andreou, A. G., 1992, Current-mode subthreshold MOS implementation of the Herault-Jutten autoadaptive network, *IEEE Trans. Solid State Circuits*, 27:714–727.
- Dupuis, S. T., and Ismail, M., 1990, High frequency CMOS transconductors, in *Analog IC Design: The Current-Mode Approach* (C. Toumazou, F. J. Lidgey, and D. G. Haigh, Eds.), London: Peter Peregrinus, pp. 181–238.
- Furman, B., White, J., and Abidi, A. A., 1988, CMOS analog IC implementing the backpropagation algorithm, in *Abstracts of the First Annual INNS Meeting*, vol. 1, p. 381.
- Hasler, P., and Akers, L., 1992, Circuit implementation of a trainable neural network using the generalized Hebbian algorithm with supervised techniques, in *Proceedings of the International Joint Conference on Neural Networks*, vol. 1, Baltimore, pp. 1565–1568. ♦
- Hasler, P., Diorio, C., Minch, B. A., and Mead, C., 1995, Single transistor learning synapses, in *Advances in Neural Information Processing Systems 7*, Cambridge, MA: MIT Press, pp. 817–824.
- Hasler, P., and Lande, T. S., Eds., 2001, *Floating-Gate Circuits* (special issue), *IEEE Trans. Circuits Syst II*, 48(1).
- Hasler, P., and Minch, B. A., 2002, *Floating-Gate Devices, Circuits, and Systems*, New York: IEEE Press.
- Holler, M., Tam, S., Castro, H., and Benson, R., 1989, An electrically trainable artificial neural network with 1024 “floating gate” synapses, in *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, Washington, D.C., pp. 191–196. ♦
- Kucic, M., Low, A.-C., Hasler, P., and Neff, J., 2001, A programmable continuous-time floating-gate Fourier process, *IEEE Trans. Circuits and Systems II*, 48:90–99.
- Mead, C., 1989, *Analog VLSI and Neural Systems*, Reading, MA: Addison-Wesley. ♦
- Tsividis, Y., and Satyanarayana, S., 1987, Analogue circuits for variable synapse electronic neural networks, *Electron. Lett.*, 24(2):1313–1314.

Analogy-Based Reasoning and Metaphor

Dedre Gentner and Arthur B. Markman

Introduction

Analogy and metaphor have been characterized as comparison processes that permit one domain to be seen in terms of another. They are important to connectionism for two reasons. First, there is an affinity at the descriptive level: many of the advantages suggested for connectionist models—representation completion, similarity-based generalization, graceful degradation, and learning—also apply to analogy (Barnden, 1994). Second, analogical processing poses significant challenges for connectionist models. Analogy involves the comparison of *systems of relations* between items in a domain. To model analogy requires representations that include internal relations. Many connectionist models have concentrated instead on statistical learning of correlational patterns over featural or dimensional representations.

Tenets of Analogy and Metaphor

Analogy derives from the perception of relational commonalities between domains that are dissimilar on the surface. These correspondences often suggest new inferences about the target domain. Analogy has been widely studied in humans. In the past decade, psychological research on analogy has converged on a set of benchmark phenomena against which models of analogy can be measured. These eight benchmarks, shown in Table 1, can be organized according to four processing principles. Analogy and metaphor involve (1) structured pattern matching; (2) structured pattern completion, (3) a focus on common relational structure rather than on common object descriptions, and (4) flexibility in that (a) the same domain may yield different interpretations in different comparisons, and (b) a single comparison may yield multiple distinct interpretations. Any model of analogy must account for these phenomena.

We begin by reviewing the principles and benchmarks, and then discuss current connectionist models of analogy and metaphor. Our discussion takes place at Marr's *computational and algorithmic*

levels, at which cognition is explained in terms of representations and associated processes. We will not evaluate the models in terms of brain function, partly because the neural basis is not yet understood, but also because we believe a computational model must first justify itself as a cognitive account. We will focus mainly on analogy, which has been well studied at the processing level. Much of what we know about analogy can be applied to metaphor as well. Later, we will explore ways in which analogy and metaphor may differ.

Structured Pattern Matching

The defining characteristic of analogy and many metaphors is the alignment of relational structure. Alignment involves finding *structurally consistent matches* (those observing parallel connectivity and one-to-one correspondence). *Parallel connectivity* requires that matching relations have matching arguments; *one-to-one correspondence* limits any element in one representation to at most one matching element in the other representation (Gentner and Markman, 1997; Holyoak and Thagard, 1995). For example, in the analogy "The atom is like the solar system," the nucleus in the atom (the *target*) corresponds to the sun in the solar system (the *base*) and the electrons to the planets, because they play similar roles in a common relational structure: e.g., **revolve** (sun, planets) and **revolve** (nucleus, electron). The sun is not matched to both the nucleus and the electron, as that violates one-to-one correspondence. Another characteristic of analogy is *relational focus*: objects correspond by virtue of playing like roles and need not be similar (e.g., the nucleus need not be hot).

There is considerable evidence that people can align two situations, preserving connected systems of commonalities and making the appropriate lower-order substitutions. For example, Clement and Gentner (1991) showed people analogous stories and asked them to state which of two assertions shared by base and target was most important to the match. Subjects chose the assertion connected to matching causal antecedents. More generally, people's correspondences are based both on the goodness of the local match and on its connection to a larger matching system (Markman and Gentner, 1993). This finding demonstrates the systematicity principle: Analogies seek *connected systems of matching relations* rather than isolated relational matches.

When making comparisons, it often occurs that nonidentical items are matched by virtue of playing a common role in the matching system. These corresponding but nonidentical elements give rise to *alignable differences*, and have been shown to be salient outputs of the comparison process (Gentner and Markman, 1997). In contrast, aspects of one situation that have no correspondence in the other, called *nonalignable differences*, are not salient outputs of comparison. For example, when comparing the atom to the solar system, the fact that atoms have electrons and solar systems have planets is an alignable difference. The fact that solar systems have asteroids, while atoms have nothing that corresponds to asteroids, is a nonalignable difference.

Structured Pattern Completion

Analogical reasoning also involves the mapping of inferences from one domain to another. Thus, a partial representation of the target is completed based on its structural similarity to the base. For example, Clement and Gentner (1991) extended the findings described earlier by deleting some key matching facts from the target

Table 1. Eight Benchmark Phenomena of Analogy

1. Relational Similarity	Analogies involve relational commonalities; object commonalities are optional.
2. Structured Pattern Matching	Analogical mapping involves one-to-one correspondence and parallel connectivity.
3. Systematicity	In interpreting analogy, connected systems of relations are preferred over sets of isolated relations.
4. Candidate Inferences	Analogical inferences are generated via structural completion.
5. Alignable Differences	Differences that are connected to the commonalities of a pair are rendered more salient by a comparison.
6. Flexibility (1): Interactive Interpretation	Analogy interpretation depends on both terms. The same term yields different interpretations in different comparisons.
7. Flexibility (2): Multiple Interpretation	Analogy allows multiple interpretations of a single comparison.
8. Cross-mapping	People typically perceive both interpretations of a cross-mapping and prefer the relational interpretation.

story and asking subjects to make a new prediction about the target based on the analogy with the base story. Consistent with the previous result, subjects mapped just those predicates that were causally connected to other matching predicates.

Flexibility: Interactive Interpretation

Analogy and metaphor are flexible in important ways. Indeed, Barnden (1994) suggests that analogy and metaphor may reconcile connectionism's flexibility with symbolic AI's structure-sensitivity. One way that analogy and metaphor are flexible is that the interpretations are interactions between the two terms. The same item can take part in many comparisons, with different aspects of the representation participating in each comparison.

For example, Spellman and Holyoak (1992) compared politicians' analogies for the Gulf War. Some likened it to World War II, implying that the United States was acting to stop a tyrant, whereas others likened it to Vietnam, implying that the United States had embroiled itself in a potentially endless conflict between two other opponents. Comparisons with different bases highlighted different features of the target. Flexibility is also evident when the same base term is combined with different targets. For example, the metaphor "A lake is a mirror" highlights that a lake has a flat reflective surface, whereas "Meditation is a mirror" highlights the self-examination aspect of meditation.

Flexibility: Multiple Interpretations of the Same Comparison

A more striking kind of flexibility is that a single base-target comparison can give rise to multiple distinct interpretations. For a comparison like "Cameras are like tape recorders," people can readily

provide an object-level interpretation ("Both are small mechanical devices") or a relational interpretation ("Both record events for later display"). Interestingly, children tend to prefer the former and adults the latter.

Despite this flexibility, people generally maintain structural consistency within an interpretation. In one study, Spellman and Holyoak (1992) asked subjects to map the Gulf War onto World War II (WWII). They asked "If Saddam Hussein corresponds to Hitler, who does George Bush correspond to?" Some subjects chose Franklin Delano Roosevelt, whereas others chose Winston Churchill. The key finding was that, when asked to make a further mapping for the United States in 1991, subjects chose structurally consistent correspondences. Those who mapped Bush to Roosevelt usually mapped the US-1991 to the US-during-WWII, and those who mapped Bush to Churchill mapped the US-1991 to Britain-during-WWII.

An extreme case of conflicting interpretations is *cross-mapping*, in which the object similarities suggest different correspondences than do the relational similarities. For example, in the comparison between "Spot bit Fido" and "Fido bit Rover," Fido is cross-mapped. When presented with cross-mapped comparisons, people can compute both alignments. Research suggests that adding higher-order relational commonalities increases people's preference for the relational alignment, whereas increasing the richness of the local object match increases people's preference for the object match. For example, people are more likely to select the relational correspondence in Figure 1B than in Figure 1A. This example also illustrates that the analogical processes we describe can apply to perceptual as well as conceptual materials. The ability to compute relational interpretations (even for the cross-mappings) is central to human analogizing across a wide range of domains.

This flexibility and the ability to process cross-mappings have significant implications for the comparison process, because they

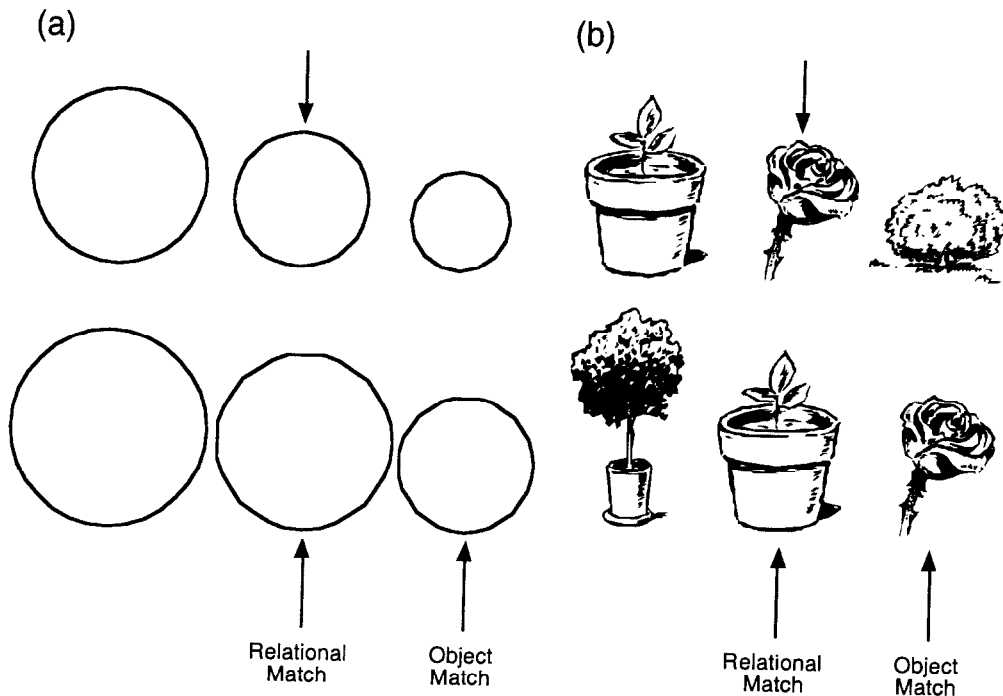


Figure 1. Sets of object triads containing a cross-mapping. A cross-mapping occurs when two similar objects play different roles in a matching relational system. In this case, the similar objects have different relative

sizes. (A) shows a sparse pair of objects that are likely to have few distinguishing attributes, whereas (B) shows a rich pair of objects that are likely to have many distinguishing attributes.

mean that simulations cannot simply be trained to generate a particular kind of interpretation. Rather, the comparison process must be able to determine both object matches and structural matches and to attend selectively to one or the other.

Connectionism and Analogical Mapping

As the preceding review makes clear, a central aspect of analogical reasoning and metaphor is alignment and mapping between structured representations. Symbolic models—e.g., Falkenhainer, Forbus, and Gentner's (1989) structure-mapping engine (SME)—have been able to pass the eight benchmarks in Table 1. Advances in connectionist models of analogy and metaphor have come with the development of techniques for representing structure (e.g., Hinton, 1991; STRUCTURED CONNECTIONIST MODELS). The best-developed models to date have been models of analogy rather than models of metaphor, and so we will focus on the analogy models here. We will discuss differences between analogy and metaphor in the following section.

An early connectionist model of analogy was ACME (Holyoak and Thagard, 1989). This model was a localist constraint-satisfaction network in which the nodes represented possible correspondences between elements in the base and target. Nodes were created using the constraint of semantic matching via a table that determined which predicates were seen as semantically similar. Nodes were connected in accordance with structural consistency, with nodes for consistent matches getting excitatory links and nodes for inconsistent matches getting inhibitory links. Finally, a pragmatic constraint was added by activating nodes related to goals and correspondences known in advance. After this activation was set up, the network was allowed to settle, and the most active nodes (above some threshold) determined the correspondences between base and target found by the system. The interpretation found by ACME need not maintain structural consistency, which can lead to problems in making inferences. Hummel, Burns, and Holyoak (1994) point out that the implementation of the pragmatic constraint often causes the important node(s) to map to everything in the other analog. Finally, because ACME settles on a single interpretation of an analogy, its solution to cross-mappings merges the object and relational interpretations.

A model of analogy has also been developed using *tensor product representations* (Smolensky, 1990). In a tensor product, two vectors \mathbf{X} and \mathbf{Y} are bound by taking the outer product of these vectors, \mathbf{YX}^T . The outer product normally forms a matrix, but a vector can be constructed from this matrix by concatenating its columns. Given \mathbf{X} and \mathbf{YX}^T , the vector \mathbf{Y} can be obtained as $\mathbf{YX}^T\mathbf{X}$ if \mathbf{X} is a unit vector. Variable bindings can thus be captured by using one vector to represent a predicate and the other to represent its argument.

Tensor products have been used in a distributed connectionist model—STAR—that performs $a:b::c:d$ analogies (Halford et al., 1994). STAR represents binary relations ($R(a, b)$) using tensor products of rank 3 (which are like the binary tensor products just described except that three vectors are bound together). In this model, long-term memory consists of a matrix of tensor products corresponding to various relations the system knows about. To process an analogy, the model takes the a and b terms and probes long-term memory to find a relation between them. It then takes this relation and the c term of the analogy and finds a fourth term that shares that relation with the c term. This model uses a distributed connectionist representation to perform a one-relation analogical reasoning task. Thus, STAR performs analogy through retrieval of known relations. STAR cannot generate multiple distinct interpretations of a comparison. If the system knows many different items that could be the answer to the analogy, the output vector is a

combination of them all. Finally, this model does not make use of higher-order relational structure to constrain its matches.

Perhaps the most complete connectionist model of analogy is Hummel and Holyoak's (1997) LISA, which operates over structured representations by using temporal synchrony in unit firing to encode relations. The connections between relations and their arguments are maintained by having individual units, which represent concepts, fire in phase with units that represent particular relational bindings (STRUCTURED CONNECTIONIST MODELS). For example, to represent **kiss** (John, Mary) nodes for **kiss**, John and *agent* fire in phase. Nodes for **kiss**, Mary and *patient* also fire in phase (but out of phase with those for John and *agent*). Furthermore, each node representing a concept is connected to a distributed representation designed to capture the meaning of that concept. The semantic similarity of any two concepts is just the dot product of the vectors in the distributed representations of those concepts. Finally, higher-order relations are represented in LISA by chunking relations that are arguments to higher-order relations into a single node.

Mapping takes place in LISA by selecting one domain (either the base or the target) as a driver. A role-argument binding is activated in the driver, and activation flows from the active nodes to the distributed semantic representation, and from the semantic nodes to localist concept nodes in the other domain. LISA has a limited-capacity working memory of 4–6 nodes. This working memory holds onto the correspondences from a small number of previous firings, thus allowing some influence of higher-order relational structure. If the role bindings for a higher-order relation are fired followed by the role bindings for the relational arguments of that higher order relation, then the correspondences suggested by the higher order relation can influence the mapping given to its argument. Trainable connections between nodes in the base and target are updated only after a certain number of firing cycles (depending on the size of working memory). LISA has been tested on a number of analogy problems. It tends to make relational mappings for analogies, and generally finds structurally consistent correspondences. The model selects either the relational mapping or the object mapping for a cross-mapping. On any given run, LISA arrives at only one interpretation; however if the order in which nodes in the driver are activated is varied, the system can find different interpretations on different runs. Finally, because the model can use complex representations, it can use different aspects of the representation of a domain in different comparisons involving that domain.

LISA is the only extant model of analogical mapping to include an explicit working memory constraint. At present, two major questions remain. First, the order in which statements are activated in the driver—a crucial determinant of the outcome of a match—is currently decided by the modeler. Second, the model has not been tested on large representations of the base and target. Thus, it is not clear how it will perform on these representations.

How Metaphor Differs from Analogy

The previous section focused on connectionist models of analogy. There has been little work on connectionist models of metaphor. To some degree, models of analogy could be extended to metaphor. In this section, we discuss some differences between analogy and metaphor that are relevant for developing a connectionist model of metaphor.

Metaphors are nonliteral assertions of likeness. They may be phrased as comparisons, in *simile* form (“A cloud is like a sponge”) or as class inclusions, in *metaphor* form (“A cloud is a sponge”). When a novel metaphor is being processed, the two domains in the metaphor are compared using the same process that is applied to analogy. Unlike analogy, however, metaphors need not focus ex-

clusively on relations. For example, the example above could be interpreted as a cloud that is fluffy, which would focus on an attribute of sponges that clouds also possess. This metaphor can also be given a relational interpretation. For example, it might be interpreted to mean that both clouds and sponges soak up water and give it back later. Typically, adults (but not children) prefer relational interpretations of metaphors to attribute interpretations.

There are three key ways in which metaphors differ from analogies. First, whereas analogies tend to have explanatory-predictive functions, metaphors may have expressive purposes and may affect the mood of the piece in which they are embedded. Thus, the primary impact of a metaphor might come in the emotions that it brings out rather than on the information in the comparison that is promoted. Second, not all metaphors are necessarily processed as comparisons. Glucksberg and his colleagues (e.g., Glucksberg and Keysar, 1990) suggest that metaphors might be processed as class inclusion statements rather than as comparisons. While there is debate as to exactly when metaphors are processed as comparisons or as class inclusion statements, some evidence suggests that novel metaphors and similes (e.g., "Some cults are termites") are processed by alignment and mapping, whereas conventional metaphors (e.g., "Some people are sheep") may be processed by accessing a stored metaphorical word sense. Finally, there are often systems of related metaphors in a language (Lakoff and Johnson, 1980). For example, English has a system of metaphors in which anger is described as heated fluid in a container (e.g., "Mary was boiling mad. The pressure built up in her until she finally exploded with rage."). These metaphorical systems might reflect a large-scale mapping between a base and target domain.

One model of system metaphors has been developed by Narayanan (1999). This model uses a localist connectionist system to handle extended metaphors such as the anger as heated fluid example above. In this system, the connection between a base and target domain is assumed to be established by convention, so there is no mapping mechanism for constructing new correspondences. Instead, the model focuses on how understanding a physical base domain can aid the comprehension of an abstract target. The model has a detailed localist network representation of the base domain in which actions can be simulated as transitions through the network. After simulating a possible outcome in the physical domain, the established mapping to the target domain is used by passing activation from the base to corresponding nodes in a belief network representing the target. In this way, metaphorical inferences can be drawn from base to target. These inferences are confined to existing correspondences between the domains; there is no mechanism for establishing new correspondences. A variety of constraints on metaphor interpretation such as the intent of the speaker can be incorporated into the model by treating them as additional sources of activation.

It may be possible to extend connectionist models of analogy to metaphor. Connectionist models may be well suited to capturing emotional aspects of metaphor. Associations between emotions and words (and word sounds) are unlikely to be mediated by strictly symbolic and rule-based processes. Thus, the kinds of soft constraints that are easily implemented in connectionist models might be particularly well suited to understanding this aspect of metaphor comprehension.

Discussion

Analogical and metaphor processing rely heavily on structurally governed correspondences between the two domains. This leads to

the eight benchmarks summarized in Table 1 that pose a challenge for any model of analogy. Connectionist models that address these phenomena have focused on techniques for representing and processing structured representations. LISA, which uses structured representations and structure-sensitive processing, accounts for many of the phenomena in Table 1, although some additional specification and testing of the model is still required.

Some challenges for future research include (1) building analogical models that can preserve structural relations over incrementally extended analogies such as are used in reasoning, (2) developing models that can be used as components of a broader cognitive system such as one that would perform problem solving, and (3) developing models that can handle novel and conventional metaphors.

Road Map: Psychology

Related Reading: Associative Networks; Compositionality in Neural Systems; Concept Learning; Systematicity of Generalizations in Connectionist Networks

References

- Barnden, J. A., 1994, On the connectionist implementation of analogy and working memory matching, in *Advances in Connectionist and Neural Computation Theory, Vol 3: Analogy, Metaphor, and Reminding* (K. J. Holyoak and J. A. Barnden, Eds.), Norwood, NJ: Ablex Publishing Company, pp. 327–374. ♦
- Clement, C. A., and Gentner, D., 1991, Systematicity as a selection constraint in analogical mapping, *Cognitive Sci.*, 15:89–132.
- Falkenhainer, B., Forbus, K. D., and Gentner, D., 1989, The structure-mapping engine: An algorithm and examples, *Artificial Intelligence*, 41:1–63.
- Gentner, D., and Markman, A. B., 1997, Structural alignment in analogy and similarity, *Am. Psychol.*, 52(1):45–56. ♦
- Glucksberg, S., and Keysar, B., 1990, Understanding metaphorical comparisons: Beyond similarity, *Psychol. Rev.*, 97(1):3–18.
- Halford, G. S., Wilson, W. H., Guo, J., Wiles, J., and Stewart, J. E. M., 1994, Connectionist implications for processing capacity limitations in analogies, in *Advances in Connectionist and Neural Computation Theory, Vol. 2: Analogical Connections* (K. J. Holyoak and J. Barnden, Eds.), Norwood, NJ: Ablex, pp. 363–415.
- Hinton, G. E., Ed., 1991, *Connectionist Symbol Processing*, Cambridge, MA: MIT Press.
- Holyoak, K. J., and Thagard, P., 1989, Analogical mapping by constraint satisfaction, *Cognit. Sci.*, 13(3):295–355.
- Holyoak, K. J., and Thagard, P., 1995, *Mental Leaps: Analogy in Creative Thought*, Cambridge, MA: MIT Press. ♦
- Hummel, J. E., Burns, B., and Holyoak, K. J., 1994, Analogical mapping by dynamic binding: Preliminary investigations, in *Advances in Connectionist and Neural Computation Theory: Vol. 2: Analogical Connections* (K. J. Holyoak and J. A. Barnden, Eds.), Norwood, NJ: Ablex.
- Hummel, J. E., and Holyoak, K. J., 1997, Distributed representations of structure: A theory of analogical access and mapping, *Psychol. Rev.*, 104(3):427–466.
- Lakoff, G., and Johnson, M., 1980, *Metaphors We Live By*, Chicago, IL: The University of Chicago Press.
- Markman, A. B., and Gentner, D., 1993, Structural alignment during similarity comparisons, *Cognitive Psychology*, 25(4):431–467.
- Narayanan, S., 1999, Moving right along: A computational model of metaphorical reasoning about events, in *The Proceedings of AAAI-99*, Orlando, FL: AAAI.
- Smolensky, P., 1990, Tensor product variable binding and the representation of symbolic structures in connectionist systems, *Artificial Intelligence*, 48:159–216.
- Spellman, B. A., and Holyoak, K. J., 1992, If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles, *J. Personality Soc. Psychol.*, 62(6):913–933.

Arm and Hand Movement Control

Stefan Schaal

Introduction

The control of arm and hand movements in human and nonhuman primates has fascinated researchers in psychology, neuroscience, robotics, and numerous related areas. To the uninitiated observer, movement appears effortless. It is only when trying to duplicate such skills with artificial systems or when examining the underlying neural substrate that one discovers a surprising complexity that, so far, has prevented us from understanding the biological implementation, how to repair neural damage, and how to create human-like robots with a human level of movement skills.

Research directed toward understanding motor control can be approached on different levels of abstraction. For example, such research may entail examining the biochemical mechanisms of neuronal firing, the representational power of single neurons and populations of neurons, neuroanatomical pathways, the biomechanics of the musculoskeletal system, the computational principles of biological feedback control and learning, or the interaction of perception and action. No matter which level of inquiry is chosen, however, ultimately we need to solve the “reverse engineering” problem of how the properties of each level correlate with the characteristics of skillful behavior. Motor control of the arm and hand is an excellent example of the difficulties that arise in the reverse engineering problem. Behavioral research has discovered a variety of regularities in this movement domain, but it is hard to determine on which level they arise. Moreover, most of these regularities were examined in isolated arm or hand movement studies, whereas coordination of the arm and hand is a coupled process in which hand and arm movement influence each other. In this article, we discuss some of the most prominent regularities of arm and hand control and consider where these regularities come from, with a particular focus on computational and neural network models. It will become apparent that an interesting competition exists among explanations sought on the neural, biomechanical, perceptual, or computational level. These competing explanations have created a large amount of controversy in the research community over the years.

Behavioral Phenomena of Arm and Hand Control

Most movement skills can be achieved in an infinite number of ways. For instance, during reaching for an object, an arbitrary hand path can be taken between starting point and end point, and the path can be traversed at arbitrary speed profiles. Moreover, because of the excess of the number of degrees of freedom in primate movement systems, there is an infinite number of ways for realizing a chosen hand path through postural configurations (see *ROBOT ARM CONTROL*). On the biomechanical level there is even more redundancy, because there are many more muscles than degrees of freedom in the human body, and the redundancy increases on the neuronal level. Thus, it is extremely unlikely that two different individuals would use similar movement strategies to accomplish the same movement goal. Surprisingly, however, behavioral research did find a large number of regularities, not just across individuals of a given species but also across different species (see, e.g., Flash and Sejnowski, 2001). These regularities or invariants have become central to understanding perceptuomotor control, as they seem to indicate some fundamental organizational principles in the central nervous system (CNS).

Bell-Shaped Velocity Profiles and Curvature in Reaching Movements

About 20 years ago, Morasso (see *OPTIMIZATION PRINCIPLES IN MOTOR CONTROL*) discovered that in point-to-point reaching

movements in humans, the hand path in Cartesian (external) space was approximately straight and the tangential velocity trajectory along the path could be characterized by a symmetric bell shape, a result that was duplicated in monkeys. In contrast, velocity profiles in joint space and muscle space were much more complex. These findings gave rise to the hypothesis that point-to-point reaching movements are planned in external coordinates and not in internal ones. Later, more detailed examinations of reaching movements revealed that, although *approximately* straight, reaching movement showed a characteristic amount of curvature as a function of where in the workspace the starting point and end point of the movement were chosen. Also, the symmetry of the velocity profile varies systematically as a function of movement speed (e.g., Bullock and Grossberg, 1988). These behavioral phenomena gave rise to a variety of models to explain them.

Initial computational models of reaching focused on accounting for the bell-shaped velocity profile of hand movement, employing principles of optimal control based on a kinematic optimization criterion for movement planning that favors smooth acceleration profiles of the hand (see *OPTIMIZATION PRINCIPLES IN MOTOR CONTROL*). As this theory would produce perfectly straight-line movements in Cartesian space and perfectly symmetric bell-shaped velocity profiles, the observed violation of these features in behavioral expression was explained by assuming that these movement plans were executed imperfectly by an equilibrium point controller (see *EQUILIBRIUM POINT HYPOTHESIS*). Thus, the behavioral features of point-to-point movements were attributed to perfect motor planning and imperfect motor execution.

An alternative viewpoint was suggested by Kawato and co-workers (see *OPTIMIZATION PRINCIPLES IN MOTOR CONTROL* and *EQUILIBRIUM POINT HYPOTHESIS*). Their line of research emphasizes that the CNS takes the dynamical properties of the musculoskeletal system into account and plans trajectories that minimize “wear and tear” in the actuators, expressed as a minimum torque-change or minimum motor-command-change optimization criterion. According to this overall view, the behavioral features of arm and hand control are an intentional outcome of an underlying computational principle that employs models of the entire movement system and its environment.

Recently, Harris and Wolpert (see *OPTIMIZATION PRINCIPLES IN MOTOR CONTROL*) suggested that the features of arm and hand movement could also be due to the noise characteristics of neural firing, i.e., the decreasing signal-to-noise ratio of motor neurons with increasing firing frequency. Thus, the neuronal level together with the behavioral goal of accurate reaching was held responsible for behavioral characteristics.

Several other suggestions were made to account for features of arm and hand control. Perceptual distortion could potentially contribute to the curvature features in reaching, and dynamical properties of feedback loops in motor planning could generate asymmetries of bell-shaped velocity profiles (Bullock and Grossberg, 1988). Moreover, imperfection of motor learning (see *SENSORI-MOTOR LEARNING*) and delays in the control system could equally play into explaining behavior.

Movement Segmentation

For efficient motor learning, it seems mandatory that movement systems plan on a higher level of abstraction than individual motor commands, as otherwise the search space for exploration during learning would become too large to find appropriate actions for a

new movement task (see ROBOT LEARNING). Movement primitives (see MOTOR PRIMITIVES), also called units of action, basis behaviors, or gestures (see SPEECH PRODUCTION), could offer such an abstraction. Pattern generators in invertebrates and vertebrates (see MOTOR PATTERN GENERATION) and the few different behavioral modes of oculomotor control (e.g., VOR, OKR, smooth pursuit, saccades, vergence) can be seen as examples of such movement primitives. For arm and hand control, however, whether some form of units of actions exist is a topic of ongoing research (Sternad and Schaal, 1999). Finding behavioral evidence for movement segmentation could thus provide some insight into the existence of movement primitives.

Since the 1980s, kinematic features of hand trajectories have been used as one major indicator to investigate movement segmentation. From the number of modes of the tangential velocity profile of the hand in linear and curvilinear drawing movements, it was concluded that arm movements may generally be created based on discrete strokes between start points, via points, and end points, and that these strokes are piecewise planar in three-dimensional movement (for a review, see Sternad and Schaal, 1999). From these and subsequent studies, stroke-based movement generation and piecewise planarity of the hand movement in Cartesian space became one of the main hypotheses for movement segmentation (Flash and Sejnowski, 2001).

Recent work (Sternad and Schaal, 1999), however, reinterpreted these indicators of movement segmentation partially as an artifact, in particular for rhythmic movement, that, surprisingly, was also assumed to be segmented into planar strokes. Human and robot experiments demonstrated that features of apparent movement segmentation could also arise from principles of trajectory formation that use oscillatory movement primitives in joint space. When such oscillations are transformed by the nonlinear direct kinematics of an arm (see ROBOT ARM CONTROL) into hand movement, complex kinematic features of hand trajectories can arise that are not due to movement segmentation. Sternad and Schaal (1999) therefore suggested that movement primitives may be better sought in terms of dynamic systems theory, looking for dynamical regimes like point and limit cycle attractors and using perturbation experiments to find principles of segmenting movements into these basic regimes.

The 2/3 Power Law

Another related behavioral feature of primate hand movements trajectories, the 2/3 power law, was discovered by Lacquaniti et al. (in Flash and Sejnowski, 2001). In rhythmic drawing movements, the authors noted a power law relationship with proportionality constant k between the angular velocity $a(t)$ of the hand and the curvature of the trajectory path $c(t)$:

$$a(t) = kc(t)^{2/3} \quad (1)$$

There is no physical necessity for movement systems to satisfy this relation between kinematic (i.e., velocity) and geometric (i.e., curvature) properties of hand movements. Since the power law has been reproduced in numerous behavioral experiments (Viviani and Flash, 1995, in Flash and Sejnowski, 2001) and even in population code activity in motor cortices (Schwartz and Moran, 1999, in Flash and Sejnowski, 2001), it may reflect an important principle of movement generation in the CNS.

The origins of the power law, however, remain controversial. Schaal and Sternad (2001) reported strong violations of the power law in large-scale drawing patterns and, in accordance with other studies, interpreted it as an epiphenomenon of smooth movement generation (Flash and Sejnowski, 2001). Nevertheless, the power law remains an interesting descriptive feature of regularities of human motor control and has proved to be useful even in model-

ing the perception of movement (see MOTOR THEORIES OF PERCEPTION).

The Speed-Accuracy Trade-off

In rapid reaching for a target, the movement time MT of reaching the target was empirically found to depend on the distance A of the start point of movement from the target and the target width W —equivalent to the required accuracy of reaching—in a logarithmic relationship: $MT = a + b \log_2(2A/W)$, where a and b are proportionality constants in this so-called Fitts' law or speed-accuracy trade-off. Since Fitts' law is a robust phenomenon of human arm and hand movement, many computational models used it as a way to verify their validity. Unfortunately, Fitts' law has been modeled in many different ways, including models from dynamic system theory, noise properties of neuronal firing, and computational constraints in movement planning (for a review, see Mottet and Bootsma, 2001; Bullock and Grossberg, 1988). Thus, it seems that the constraints provided by Fitts' law are too nonspecific to give clear hints as to the organization of the nervous system. Nevertheless, the empirical phenomenon of Fitts' law remains a behavioral landmark.

Resolution of Redundancy

As mentioned earlier, during reaching for a target in external space, the excess number of degrees of freedom in the human body's kinematic structure usually allows an infinite number of postures for each hand position attained during the reaching trajectory. An active area of research in motor control is thus concerned with how redundancy is resolved, whether there is within- or across-subject consistency of the resolution of redundancy, and whether it is possible to deduce constraints on motor planning and execution from the resolution of redundancy.

Early studies by Cruse et al. (in Bullock, Grossberg, and Guenther, 1993) demonstrated that redundancy resolution was well described by a multiterm optimization criterion that primarily tries to keep joint angular position as far as possible away from the extreme positions of each joint and also minimizes some physiological cost. According to this explanation, when a reaching movement is initiated in a rather unnatural posture, the movement slowly converges to the optimal posture on the way to the goal, rather than achieving optimality immediately. This strategy resembles the method of resolved motion rate control in control theory, suggested as a neural network model of human motor planning by Bullock et al. (1993). Grea, Desmurget, and Prablanc (2000) observed similar behavior in reaching and grasping movements. Noting that the final posture at a grasp target was highly repeatable even if the target changed its position and orientation during the course of the reaching movement, the authors concluded that the CNS plans the final *joint space* position for reaching and grasping, not just the final hand position. However, the optimization methods proposed by Bullock et al. (1993) could result in similar behavior, without the CNS explicitly planning the final joint space posture. An elegant alternative view to optimization methods is suggested in GEOMETRICAL PRINCIPLES IN MOTOR CONTROL (q.v.), where motor control and planning based on force fields is emphasized. It is evident more work will be needed before a final conclusion can be reached on the issue of redundancy resolution.

Reaching and Grasping

The coordination of reaching and grasping offers at least three important windows onto motor control. First, reaching and grasping require a resolution of redundancy, as outlined in the previous section. However, small changes in target orientation can lead to the

need for drastic changes in arm and hand posture at the target, such that movement planning requires carefully chosen strategies for successful control. Second, reaching and grasping are two separate motor behaviors that may or not be executed independently of each other. This issue allows researchers to examine the superposition and sequencing of movement primitives. Third, grasping has a more interesting perceptual component than reaching, since appropriate grasp points, grasping strategies, and grasping forces need to be selected as a function of target shape, size, and weight. The principles of perceptuomotor coordination can thus be examined in well-controlled experiments, including the grasping of objects that induce visual illusions.

Among the key features of reaching and grasping are the following: (1) a fast initial arm movement to bring the hand close to the target, (2) a slow approach movement when the hand is near the target, and (3) a preshaping phase of the hand with initial progressive opening of the grip, followed by closure of the grip until the object size is matched and the object is finally grasped (Jeannerod et al., 1995; Arbib and Hoff, 1993, in Jeannerod et al., 1995). Although early models of reaching and grasping assumed independence of these different phases and simply executed them in a programmatic way, behavioral perturbation studies that changed the target size, orientation, or distance revealed a coupling between the phases (for a review, see Jeannerod et al., 1995), such that, e.g., the preshaping partially reversed when the target distance was suddenly increased. Using optimization principles, Hoff and Arbib (in Jeannerod et al., 1995) developed a model of these interactions by structuring the reach-and-grasp system in appropriate perceptual and motor schemas (see SCHEMA THEORY), including abstraction of the multifingered hand in terms of two or more virtual fingers to simultaneously model different grip types (e.g., precision grip, power grip) and their opposition spaces for contact selection. This model can also be mapped onto the known functional cortical anatomy in primates. Grip force selection and the anticipation of object properties has been studied by a number of authors (e.g., Flanagan and Beltzner, 2000), who generally agree that the CNS seems to use internal models to adjust grip force. From studies of the resolution of redundancy, it was concluded that the entire arm posture at the target seems to be planned in advance (Grea et al., 2000), but this result may need differentiation as outlined in the previous section. In general, there seems to be a consensus that behavioral features of reaching and grasping are carefully planned by the CNS and are not accidental.

Motor Learning

Because of continuous change in body size and biomechanical properties throughout development, the ability to learn motor control is of fundamental importance in biological movement systems. Moreover, when it comes to arm and hand control, primates show an unusual flexibility in devising new motor skills to solve novel tasks. Learning must therefore play a pivotal role in computational models of motor control.

One of the most visible research impacts of motor learning was the controversy between equilibrium point control (see EQUILIBRIUM POINT HYPOTHESIS) and internal model control (see SENSORIMOTOR LEARNING and CEREBELLUM AND MOTOR CONTROL). Proponents of equilibrium point control believed that the learning of internal models is too complicated to be plausible for biological information processing, while proponents of internal model control accumulated evidence that various, in particular fast, movement behaviors cannot be accounted for by equilibrium point control. At present, there seems to be an increasing consensus that internal model control is a viable concept for biological motor learning, and that the equilibrium point control strategy in its original and appealing simplicity is not tenable. Behavioral learning experiments

that were created in the wake of the equilibrium point control discussion sparked a new branch of research on motor learning (see SENSORIMOTOR LEARNING and GEOMETRICAL PRINCIPLES IN MOTOR CONTROL). Adaptation to virtual force fields, to altered perceptual environments, or to virtual objects are among the main behavioral paradigms to investigate motor learning, with the goal of better understanding the time course, representations, control circuits, retention, and functional anatomy of motor learning (see SENSORIMOTOR LEARNING).

Interlimb Coordination

In robotics, the control of two limbs can be accomplished as if the two systems were completely independent, thus reducing the control problem to that of controlling two robots instead of one. In biological motor control, such independence does not exist, and a rich area of behavioral investigation examines the computational principles and constraints that arise from the coordination of multiple limbs. In arm and hand control, the approach of dynamic pattern formation (e.g., Kelso, 1995) has been a prominent methodology to account for interlimb coordination. In this approach, motor control in general and interlimb coordination in particular are viewed as an assembly of the required degrees of freedom of the motor system into a task-oriented attractor landscape (Saltzman and Kelso, 1987, in Kelso, 1995). Interlimb coordination is thus conceived of as the result of coupling terms in nonlinear differential equations. An important question thus arises as to what kind of equations model the control of movement, and what kind of variables cause the coupling. A variety of models of movement generation with nonlinear dynamics approaches were suggested, based on differential equations, that either generate movement plans (Kelso, 1995; Sternad, Dean, and Schaal, 2000) or directly generate forces. The origin of coupling between limbs, however, remains an issue of controversy. Possible sources could be perceptual, proprioceptive, purely planning-based, interaction force-based, a preference for homologous muscle activation, or neural crosstalk. By demonstrating that the orientation of limbs in external space can explain a certain class of interlimb coordination, recent behavioral results (Mechsner et al., 2001) emphasized that perceptual coupling may be much more dominant than previously suspected. In general, however, there seems to be a strong need for detailed computational modeling to elucidate the computational and neuronal principles of interlimb coordination.

Intralimb Coordination

Investigations of intralimb coordination seek to uncover the specific principles by which individual segments of a limb move relative to one other. Models of arm and hand control that are based on optimal control (see OPTIMIZATION PRINCIPLES IN MOTOR CONTROL) or optimal resolution of redundancy automatically solve the intralimb coordination problem by means of their optimization framework; any kind of special behavioral features would be considered accidental. However, some research has considered whether some special rules of information processing by the CNS can be deduced from the regularities of intralimb coordination. For reaching movements, the simple mechanism of joint interpolation can account for a large set of behavioral features when the onset times of the movements in individual degrees of freedom are staggered, an older observation that has been confirmed in more recent work (Desmurget et al., 1995). For rhythmic movement, it is of interest to know how the oscillations in individual degrees of freedom remain phase-locked to each other, and whether there are preferred phase-locked modes (Schaal et al., 2000). As in interlimb coordination, models of nonlinear differential equations seem the

most suitable for capturing the effects of rhythmic intralimb dynamics.

Perception-Action Coupling

Most of the behavioral studies outlined in the previous sections were primarily concerned with specific aspects of *motor control* and less with issues of *perceptuomotor control*. However, the interaction of perception and action reveals many constraints on the nervous system. In the behavioral literature, there is a large body of research that examines particular perceptuomotor skills, such as the rhythmic coordination of arm movement during the juggling of objects or the interaction of external forces and limb dynamics to generate movement (e.g., Sternad, Duarte, et al., 2000). This interesting topic cannot be discussed in detail here.

Discussion

Behavioral phenomena of arm and hand movement have sparked a rich variety of computational models on various levels of abstraction. Although some topics, such as internal model control, have gained solid ground in recent years (Flash and Sejnowski, 2001), many other issues remain controversial and deserve more detailed and computational investigations. Perhaps the most interesting topics for future research are the importance of the dynamic properties of the musculoskeletal system in facilitating motor control, the role of real-time perceptual modulation of motor control, and dynamic systems models versus optimal control-based models.

Road Maps: Mammalian Motor Control; Robotics and Control Theory

Related Reading: Eye-Hand Coordination in Reaching Movements; Grasping Movements: Visuomotor Transformations; Limb Geometry, Neural Control; Robot Arm Control; Sensorimotor Learning

References

- Bullock, D., and Grossberg, S., 1988, Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation, *Psychol. Rev.*, 95:49–90. ♦
- Bullock, D., Grossberg, S., and Guenther, F. H., 1993, A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm, *J. Cogn. Neurosci.*, 5:408–435.
- Desmurget, M., Prablanc, C., Rossetti, Y., Arzi, M., Paulignan, Y., Urquizar, C., and Mignot, J. C., 1995, Postural and synergic control for three-dimensional movements of reaching and grasping, *J. Neurophysiol.*, 74:905–910.
- Flanagan, J. R., and Beltzner, M. A., 2000, Independence of perceptual and sensorimotor predictions in the size-weight illusion, *Nature Neurosci.*, 3:737–741.
- Flash, T., and Sejnowski, T., 2001, Computational approaches to motor control, *Curr. Opin. Neurobiol.*, 11:655–662. ♦
- Grea, H., Desmurget, M., and Prablanc, C., 2000, Postural invariance in three-dimensional reaching and grasping movements, *Exp. Brain Res.*, 134:155–162.
- Jeannerod, M., Arbib, M. A., Rizzolatti, G., and Sakata, H., 1995, Grasping objects: The cortical mechanisms of visuomotor transformation, *Trends Neurosci.*, 18:314–320. ♦
- Kelso, J. A. S., 1995, *Dynamic Patterns: The Self-Organization of Brain and Behavior*, Cambridge, MA: MIT Press.
- Mechsner, F., Kerzel, D., Knoblich, G., and Prinz, W., 2001, Perceptual basis of bimanual coordination, *Nature*, 414:69–73.
- Mottet, D., and Bootsma, R. J., 2001, The dynamics of rhythmical aiming in 2D task space: Relation between geometry and kinematics under examination, *Hum. Movement Sci.*, 20:213–241.
- Schaal, S., and Sternad, D., 2001, Origins and violations of the 2/3 power law in rhythmic 3D movements, *Exp. Brain Res.*, 136:60–72. ♦
- Schaal, S., Sternad, D., Dean, W., Kotoska, S., Osu, R., and Kawato, M., 2000, Reciprocal excitation between biological and robotic research, in *Sensor Fusion and Decentralized Control in Robotic Systems III, Proceedings of the SPIE*, Boston, MA: SPIE.
- Sternad, D., Dean, W. J., and Schaal, S., 2000, Interaction of rhythmic and discrete pattern generators in single joint movements, *Hum. Movement Sci.*, 19:627–665.
- Sternad, D., Duarte, M., Katsumata, H., and Schaal, S., 2000, Dynamics of a bouncing ball in human performance, *Phys. Rev. E*, 63:1–8.
- Sternad, D., and Schaal, D., 1999, Segmentation of endpoint trajectories does not imply segmented control, *Exp. Brain Res.*, 124:118–136.

Artificial Intelligence and Neural Networks

John A. Barnden and Marcin Chady

Introduction

This article surveys the distinctions between symbolic artificial intelligence (AI) systems and neural networks (NNs), their relative advantages, and ways of attempting to bridge the gap between the two.

For this review we can take AI to consist of the development, analysis, and simulation of computationally detailed, efficient systems for performing complex tasks, where the tasks are broadly defined, involve considerable flexibility and variety, and are typically similar to aspects of human cognition or perception. These broad tasks include natural language understanding and generation; expert problem solving; common-sense reasoning; visual scene analysis; action planning; and learning.

There is nothing in this description of AI that prevents the computational systems from being neural networks. Nevertheless, it is fair to say that the bulk of AI can be called “traditional” or “symbolic” AI, relying on computation over symbolic structures (e.g., logic formulae). The rest of the review will therefore discuss relationships between symbolic AI and NNs.

Relative Advantages

Advantages of Neural Networks

One of the main benefits claimed for NNs is graceful degradation, especially when they are of the distributed variety (LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS). A computational system is said to exhibit graceful degradation when it can tolerate significant corruption of its input or internal workings. The toleration consists of the system’s continuing to perform usefully, though not necessarily perfectly. In NNs, input imperfection is typically a matter of corruption of individual input activation vectors. Internal corruption usually takes the form of deletions of nodes or links or corruptions of the link weights.

Symbolic AI systems, on the other hand, tend not to degrade gracefully. Consider a simple rule-based system. A small corruption of an input data structure is likely to make it fail to match the precise form expected by the rules that would otherwise have applied, so that they totally fail to be enabled. Equally, other rules might erroneously be enabled. Similarly, even minor damage to a rule can have very large effects on how the system operates.

As a special case of graceful degradation, NNs sometimes exhibit *error correction*, whereby an erroneous or corrupted pattern on an input bank of units leads to a corrected version of the pattern appearing in the network, enabling the network to proceed as if the correct version had been provided. Related to this type of error correction is *pattern completion*, whereby an incomplete pattern is filled out to become a more complete pattern somewhere in the network.

Also related to graceful degradation is *automatic similarity-based generalization*, in which previously unseen inputs that are sufficiently similar to training inputs lead naturally to behavior that is usefully similar to (or captures central tendencies in) the behavior elicited by the training inputs. There is a sense in which similarity of representations induces similarity of processing more readily than it does in symbolic AI: there is, by and large, a higher degree of naturally achievable continuity in the mapping from inputs to outputs. In addition, previously unseen blends of different representations will naturally tend to lead to processing that is a blend of the processing that would have arisen from the different representations that have been blended together. Such behavior is possible in symbolic AI but specific system-design effort is needed to achieve it.

Importantly, NNs can *learn* generalizations or category prototypes by exposure to instances, through fairly straightforward, uniform weight modification procedures. These generalizations or prototypes come to be implicit in the adjusted weights. Although learning is intensively studied in symbolic AI, and some learning paradigms in symbolic AI involve adjustment of numerical parameters akin to NN weights, symbolic processing does not provide any specific support to these paradigms. The paradigms could therefore be said to arise less easily and naturally out of symbolic processing than out of NN activity.

The preceding properties of NNs have found their application in *content-based access* (or *associative access*) to long-term memory (see ASSOCIATIVE NETWORKS). This can take two different forms. First, let us assume, as usual, that a neural net's long-term memory is its set of weights. The manipulation of an input vector by the network can be thought of as the bringing to bear of particular content-relevant long-term memories on that vector. Second, in any NN that learns a map from particular inputs to particular outputs, an output can be thought of as a particular long-term memory recalled directly on the basis of the content of the input. Content-based access is not as easily provided in symbolic systems implemented in conventional computers, although it can be obtained to some useful degree by sophisticated indexing schemes (see Bonissone et al. in Barnden and Holyoak, 1994), associative computer memories, or hashing (see Touretzky in Hinton, 1991, for discussion).

NNs can have *emergent rule-like behavior*. Such behavior can be described, approximately at least, as the result of following symbolic rules, even though the system does not contain representations of explicit rules (see Elman, 1991). Emergent rule-like behavior is a central issue in the application of neural networks to high-level cognitive tasks.

More generally, NNs tend to be more sensitive to *subtle contextual effects* than symbolic AI systems are, because multiple sources of information can more easily be brought to bear in a gracefully interacting and parallel way. This property of NNs facilitates *soft constraint satisfaction*. That is, it is possible to arrange for some hypotheses to compete and cooperate with each other, gradually influencing each other's levels of confidence until a stable set of hypotheses is found. Each hypothesis is represented by a node or group of nodes in the neural network, and the constraints are encoded by links joining those nodes or groups. The constraint-satisfaction is soft because no individual constraint needs to be satisfied absolutely. By contrast, although many symbolic AI sys-

tems are designed to do constraint satisfaction, the symbolic framework provides no special support for it, particularly when the constraints are soft.

Finally, NNs are an inherently parallel model of computation whose parallelism is straightforwardly realizable in a physical substrate.

Advantages of Symbolic AI Systems

The symbolic framework is better than NNs at encoding and manipulating the *complex, dynamic structures* of information that appear to be needed in cognition. These structures can, for instance, be interpretations of natural language utterances, descriptions of complex scenes, complex plans of action, or conclusions drawn from other information. The encodings of such structures, whether these encodings are symbolic or otherwise, need to have the following important properties. (See also Shastri in Barnden and Pollock, 1991.)

1. The encodings must often be highly temporary—for instance, encodings of interpretations of natural language sentences and encodings of intermediate conclusions during reasoning need to be rapidly created, modified, and discarded. Although activation patterns in NNs are temporary, temporariness is challenging for NNs when it is combined with properties 2–6.
2. The encoding technique must allow the encoded structures to combine information items (e.g., word senses) that have never been combined before, or never been combined in quite the same way, in the experience of the system.
3. The encodings must allow the encoded information to have widely varying structural complexity. Natural language sentence interpretations provide illustrations of this point.
4. In particular, the encoded structures can be multiply nested. In the sentence “John believes that Peter’s angry behavior toward Mary caused her to write him a strongly worded letter,” the anger description is nested within a causation description that is nested within a belief report.
5. A given type of information can appear at different levels of nesting. A system might have to represent a sitting *room* that has a wall that bears a picture that itself depicts a dining *room*. As another illustration, a *belief* might be about a hope that is about a *belief*.
6. A given type of information may also have to be multiply instantiated in other ways, as when, for instance, there are three love relationships that need to be simultaneously represented.

Turning to manipulations, cognitive systems must exhibit strong properties of *systematicity* of processing—each information structure *J* that one cares to mention has an extremely large class of variants that must be able to be subjected to the same sort of processing as *J* is; and the class of variants is far too large to imagine that each variant is processed by a separate piece of neural network or a separate symbolic module. So, we must have symbolic AI systems and NNs capable of very flexible and general processing. (See also SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS.)

The *variable-binding* problem for neural networks is one manifestation of the need for systematicity of processing. Suppose one wishes a neural network to make inferences that obey the following rule: *X* is jealous of *Z* whenever *X* loves *Y*, *Y* loves *Z*, and *X*, *Y*, and *Z* are distinct people. In this statement of the rule, the variables *X*, *Y*, and *Z* can be replaced by any suitable people-descriptions, such as “Joe Bloggs’ father’s boss.” The systematicity issue in this example is that of avoiding having replication of machinery for all the different possible combinations of values for these variables. (Each such combination is a *J* in the terms of the previous para-

graph.) This issue can also be thought of as the variable binding problem for the example, even if a neural network dealing with it does not have any explicit representation of the rule or the variables in it. In the special case in which a neural network implements the rule as a subnetwork and has particular units, subnetworks, or activation patterns that play the role of the three variables, the variable binding problem, in a narrower sense now, is the problem of how the network is to be able to “bind” such a unit, subnetwork, or activation pattern to a particular value at any given moment, and how the binding is to be used in processing.

Cognitive systems must also exhibit a high degree of *structure-sensitivity* in their processing. Pieces of information that have complex structure must be processed in ways that are heavily dependent on their structure as such, not (just) on the nature of constituents taken individually. For example, consider the operation of inferring from “not both *A* and *B*” that “not *A* or not *B*.” The operation is independent of what *A* and *B* are—it is only the “not both . . . and . . .” structure that is important.

These features of information structure encodings and manipulations combine to distinguish the types of information that neural networks for reasoning, natural language understanding, etc. must deal with from the types of information that typical neural networks cater for. Traditional NN techniques were originally developed largely for specific “low level” applications, such as restricted forms of pattern recognition, or for limited forms of pattern association. Because of the resulting continuing limitations in most applications of NNs, it has been sufficient for NNs to adhere, by and large, to the following restrictions (although almost every restriction is violated by some NN subparadigm). These restrictions cause difficulty in trying to apply NNs to natural language understanding, common-sense reasoning, and the like.

1. There is typically no dynamic, rapid creation and destruction of nodes and links. Therefore, temporary information cannot be encoded in temporary network topology changes. (Some of this effect can, however, be obtained by techniques such as dynamic links described in DYNAMIC LINK ARCHITECTURE (q.v.), or by higher-order units, whose activation is sensitive to weighted sums of products of input values, rather than just to weighted sums of input values.)
2. Links in NNs are not differentiated by labeling, unlike the links in symbolic structures such as SEMANTIC NETWORKS (q.v.). Therefore, in an NN, information that could otherwise be put into link labels has to be encoded somehow in activation values, weights, extra links, or other features of network topology, adding significantly to the cumbersomeness of the net and its processing. (See Barnden and Srinivas, 1991, for more discussion.)
3. The resolution of NN activation values is generally not fine enough to allow them individually to encode complex symbolic structures. Most typically, activation values merely encode confidence levels of some sort.
4. Pointers are usually not allowed. That is, activation values or patterns are not allowed to act as names or addresses of parts of the network itself.
5. Stored programs are not allowed. That is, activation values or patterns cannot act as instructions (names of internal computational actions).

Further Comparative Remarks

The advantages claimed here for NNs are not clearcut. For instance, there are types of AI systems that readily exhibit forms of graceful degradation, pattern completion, and similarity-based generalization. In particular, as Barnden in Barnden and Holyoak (1994) argues in detail and other researchers have noted, these benefits are natural properties of suitably designed symbolic analogy-based rea-

soning systems (see ANALOGY-BASED REASONING AND METAPHOR and also MEMORY-BASED REASONING).

Just as NNs support some types of learning more readily than symbolic AI systems do, the converse holds as well. Symbolic AI systems, by virtue of their ability to handle complex temporary information structures, are in a better position to perform various types of rapid learning, proceeding in large steps rather than lengthy, gradual weight modification. For instance, a symbolic AI system is in a good position to reason about why some plan of action failed, and thus quickly and greatly amend relevant parts of its knowledge base or planning strategies. Also, the ability of neural networks to learn generalizations is often hindered by elaborate, extensive training regimes. It is true that in some learning regimes, such as some forms of Hebbian learning, final weights are calculated in a direct way from single presentations of training items. But more typically, the number of training-item presentations one needs to make to the network runs to tens or hundreds of thousands before useful results can be obtained.

NNs are good at allowing hypotheses to be held with varying degrees of confidence, the degrees being realized as activation levels. However, it is commonplace also in symbolic AI to have numerical degrees of confidence. These appear in DECISION SUPPORT SYSTEMS AND EXPERT SYSTEMS (q.v.) and elsewhere. However, the normal properties of activation spread and activation combination in NNs support confidence levels in a natural way. In symbolic AI systems the computations have to be specially and explicitly designed.

The contrasts between neural networks and symbolic AI that were presented earlier are clouded by the fact that NNs can be *implementational*. Implementational NNs are exact implementations of symbol processing schemes of the sort used in traditional symbolic AI systems. That is, network-unit activations (and/or link weights, possibly) can be regarded as exactly encoding symbolic representations as used in traditional AI systems—such as logic formulas, frames, schemas, or pieces of semantic network—and changes in network state can be regarded as exactly encoding traditional symbolic manipulation steps—such as traversal, concatenation, and rearrangement of structures—that are used in traditional AI for directly effecting reasoning, planning, natural language understanding, etc. See, for example, Barnden’s and Shastri’s chapters in Barnden and Pollack (1991) and Lange and Wharton’s chapter in Barnden and Holyoak (1994).

The *nonimplementational* style includes NNs that can be usefully viewed as *approximately* manipulating traditional symbolic objects in traditional ways. However, the nearer an NN is to being implementational, the more it runs the danger of inheriting the disadvantages of symbolic AI, such as the tendency to lack graceful degradation.

Bridging the Gap

The discrepancy in the relative advantages of (nonimplementational) NNs and symbolic AI systems has been the focus of much attention during the last decade or so (see, e.g., Barnden and Pollack, 1991; Browne and Sun, 2001; Hinton, 1991; Jagota et al., 1999; McGarry, Wermter, and McIntyre, 1999). We shall review here some representative attempts to tackle the problem.

A common approach to extending conventional types of NN processing to handle complex dynamic structures is to use *reduced representations*, also known as *compressed encodings*. See, e.g., Hinton, Pollack, and St. John’s chapter and McClelland’s chapter in Hinton (1991), as well as Elman (1991). A reduced representation is a single activation vector that is created from the several activation vectors that encode the constituents of the structure in such a way that the resulting vector is of roughly the same size as each of the constituents’ vectors. For example, the constituents

could be words, and a sequence of word encodings could represent a sentence. The reduced representation is then a roughly word-sized vector for the whole sentence.

One architecture used to produce reduced representations of sequences of items is a Simple Recurrent Network (SRN). An SRN is typically a three-layer network in which the input to the middle layer consists of an item in the sequence together with the previous activation pattern in the middle layer itself. As a result of back-propagation training, the encodings produced in the middle layer are compressed vectors representing the current input *in the context* of the history of items presented to the network so far. SRNs have been successfully used, for instance, for predicting the category of the next word in a sentence being inputted (Elman, 1991).

A more general-purpose architecture for producing recurrent compressed encodings is Pollack's Recursive Auto-Associative Memory (RAAM) (see Pollack in Hinton, 1991). The input and output layers are divided up into segments that hold constituent encodings. The net is trained to map sequences of constituent encodings to themselves. The activation pattern that appears on the hidden layer of the trained network in response to a particular sequence of constituent encodings on the input layer is the compressed encoding for the sequence. (And a compressed encoding can be decoded by placing it in the hidden layer: a close approximation, hopefully, to the sequence of constituent encodings appears on the output layer.) Also, during training, a hidden layer pattern can be copied into any of the segments in the input and output layer, leading to the ability of the network to handle recursive structures some of whose constituents are themselves sequences of constituents. An example of such a structure is the sentence "John knows that Sally is clever," thought of as having the sentence "Sally is clever" as a constituent.

There are some indications that compressed encodings can support *holistic* structure-sensitive processing by means of conventional NN techniques such as feedforward association networks (see, e.g., Chalmers, 1990; Pollack in Hinton, 1991). The processing is holistic in that the encodings are not uncompressed into the individual activation vectors that encode their notional constituents. For example, Chalmers successfully trained a three-layer backpropagation network to transform compressed encodings of active English sentences into compressed encodings of their passive counterparts. The hidden layer had the same size as the input and output layers (the size of a compressed encoding) and the net operated in one pass of activation, so that it cannot have been working by first decoding the input compressed encodings into the corresponding sequence of constituent encodings.

However, in-depth analysis of RAAM-like systems (Kolen, 1994) reveals that the computation depends on very fine tuning of synaptic weights and highly precise activation levels. The implication is that holistic computation based on RAAM-generated encodings is sensitive to noise, which is a significant drawback given that graceful degradation is a major argument for using neural networks. An additional complication associated with all of the aforementioned techniques of generating reduced representations is the lengthy process of weight training.

These problems are avoided in a related approach of which a central example is the Holographic Reduced Representation (HRR) technique of Plate (1995). See also Rachkovskij and Kussul (2001). HRRs use predefined combination operations (circular convolutions) to produce compressed encodings. No training is required and both encoding as well as decoding are performed in a single step. What is more important, though, is that these transformations offer a more comprehensive account of systematicity. Given a sufficiently large size of code vectors, not only can HRRs recursively bind any number of elements, but also, using simple vector addition, multiple bindings can be combined further to form collections of predicates. Such collections retain superficial and structural sim-

ilarity of structures, so that decoding is not necessary to estimate an item's relevance for certain types of computation. If necessary, any one predicate can be readily extracted using inverse operations, although not without some loss of information due to the nature of circular convolution and vector addition. The operations introduce some degree of noise, which is further amplified by the decoding transformation. However, it can be rectified using an auto-associative memory to recover the original elements. Using this approach, Plate (1995) shows how HRRs can be used to represent sequences and more complex structures, and how to achieve chunking and variable binding.

In a different approach to bridging the gap, Barnden in Barnden and Holyoak (1994) capitalizes on the comment made previously that symbolic analogy-based reasoning possesses many of the main advantages of nonimplementational NNs. The claim is that an implementational NN that implements a symbolic analogy-based reasoning system inherits those advantages, as well as the symbolic AI advantage with respect to complex dynamic information structures.

The preceding approaches assume that it is worthwhile to develop gap-bridging systems that are neural networks in their entirety, rather than developing systems that are some combination of NN machinery with symbolic AI machinery (where the latter is given no NN realization). The latter, hybrid, strategy is a popular approach to bridging the gap (McGarry et al., 1999). The simpler types of hybridization occur in systems that have largely separate neural and symbolic modules (see, e.g., Hendler in Barnden and Pollack, 1991). But more intimate hybridizations have been developed, for instance, in networks where an individual node or link can act partially like those in neural networks and partially like those in symbolic networks.

Although the more implementational an NN is the more it risks inheriting disadvantages of symbolic AI, it may still be that some of the implementational NN techniques could be adapted for use in gap-bridging systems that escape those disadvantages. Therefore, we will now look at some of the techniques.

A crucial aspect of implementational NNs is the way in which they allow representational items to be rapidly and temporarily combined so as to form encodings of temporary complex information structures. One form of this *dynamic combination* (or *temporary association*) issue is the variable binding problem, and a closely related form is the role binding problem. The variable binding problem was described previously. The role binding problem is concerned with giving specific values to the roles (slots) in predicates, frames, schemas, and the like.

An immediately obvious, and somewhat natural, approach to dynamic combination is to combine network nodes or assemblies by adding new links or giving non-zero weights to existing zero-weight links. However, this method is highly cumbersome because network structure is not data that is directly manipulable by the network itself. Another rather similar approach is to facilitate existing (non-zero-weight) connection paths, between nodes/assemblies that are to be combined, by activating intermediate nodes on the paths. These nodes are called binding nodes. Since the dynamic combination structure is now encoded in the activation levels of binding nodes, the net can more easily analyze that structure. However, the processing is still cumbersome (Barnden and Srinivas, 1991).

A distinctly different approach is to deem nodes/assemblies to be bound together when they fire in synchrony (see STRUCTURED CONNECTIONIST MODELS, COMPOSITIONALITY IN NEURAL SYSTEMS, and DYNAMIC LINK ARCHITECTURE). See in particular Shastri in Barnden and Pollack (1991), and Henderson and Lane (1998). The method is an important special case of the more general notion of binding nodes together by giving them similar spatiotemporal activation patterns. This is the pattern-similarity association tech-

nique: see Barnden and Srinivas (1991) and Barnden in Barnden and Pollack (1991).

Distinctly different again is the use of positional encodings of dynamic combinations. In the more developed forms of this idea (see Barnden in Barnden and Pollack, 1991), activation patterns are dynamically combined by being put into suitable relative positions with respect to each other, much as bit-strings in computer memory can be put into contiguous memory locations to form records.

A somewhat pointer-like technique has been implemented: see Lange and Wharton in Barnden and Holyoak (1994). Different parts of the network are capable of emitting activation patterns that are thought of as their "signatures." Other parts can then temporarily hold signatures and thereby point, in a sense, to the parts that possess the signatures.

One noteworthy way of achieving temporary association is the use of auto-associative memories with rapid Hebbian learning, as in van der Velde (1995). Van der Velde demonstrates how multiple elements can be stored in a single network while preserving their ordering. Each element in the sequence refers to the next and previous ones by unique pointers that constitute part of the memory trace in an auto-associative module. Using this approach, van der Velde builds a conventional stack-based generator of center-embedded sentences. Despite its implementational architecture, this model manages to retain the graceful-degradation property of nonimplementational NNs. Hebbian association was also used by Hadley et al. (2001) to achieve a strong form of systematicity.

Finally, Smolensky in Hinton (1991) proposed an abstract but influential binding and structure-representation approach based on tensors. Some realizations of this approach involve binding nodes, but the approach can be seen to subsume other concrete techniques as well.

Discussion

One theme of this review has been that the relative advantages of symbolic AI and NNs are less clear-cut than is usually implied. In particular, although NNs have been successful for some purposes and can have advantages such as graceful degradation, most NN research has not addressed the complex information processing issues routinely tackled in symbolic AI research. The latter field has contributed much more, for instance, to the study of how natural language discourse can be understood and common-sense reasoning performed. Nevertheless, pursuing nonsymbolic approaches to the problems is beneficial for as long as the symbolic approaches fail to provide all the answers.

Some of the open questions in the area of this review are: Is it actually necessary to go beyond symbolic AI in order to account for complex cognition? If it is, should symbolic AI be dispensed with entirely, or is some amount of complex symbol-processing unavoidable? How can reasoning, natural language understanding,

etc. be effected by neural networks without just implementing conventional symbol processing? How can different styles of system, e.g., implementational and nonimplementational neural networks, or neural networks and non-neural systems, be gracefully combined into hybrid systems?

Road Map: Artificial Intelligence

Related Reading: Compositionality in Neural Systems; Connectionist and Symbolic Representations; Hybrid Connectionist/Symbolic Systems; Multiagent Systems; Philosophical Issues in Brain Theory and Connectionism; Structured Connectionist Models; Systematicity of Generalizations in Connectionist Networks

References

- Barnden, J. A., and Holyoak, K. J. (Eds.), 1994, *Advances in Connectionist and Neural Computation Theory, Vol. 3: Analogy, Metaphor and Reminding*, Norwood, NJ: Ablex Publishing Corp.
- Barnden, J. A., and Pollack, J. B. (Eds.), 1991, *Advances in Connectionist and Neural Computation Theory, Vol. 1: High Level Connectionist Models*, Norwood, NJ: Ablex Publishing Corp. ♦
- Barnden, J. A., and Srinivas, K., 1991, Encoding techniques for complex information structures in connectionist systems, *Connection Science*, 3:263–309.
- Browne, A., and Sun, R., 2001, Connectionist inference models, *Neural Networks*, 14:1331–1355. ♦
- Chalmers, D. J., 1990, Syntactic transformations on distributed representations, *Connection Science*, 2:53–62.
- Elman, J. L., 1991, Distributed representations, simple recurrent networks, and grammatical structure, *Machine Learning*, 7:195–225.
- Hadley, R. F., Rotaru-Varga, A., Arnold, D. V., and Cardei, V. C., 2001, Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network, *Connection Science*, 13:73–94. ♦
- Henderson, J., and Lane, P., 1998, A connectionist architecture for learning to parse, in *Proceedings of COLING-ACL* (Montreal, Canada, 1998), pp. 531–537.
- Hinton, G. E. (Ed.), 1991, *Connectionist Symbol Processing*, Cambridge, MA: MIT Press.
- Jagota, A., Plate, T., Shastri, L., and Sun, R. (Eds.), 1999, Connectionist symbol processing: Dead or alive?, *Neural Computing Surveys*, 2:1–40. ♦
- Kolen, J. F., 1994, Fool's gold: Extracting finite state machines from recurrent network dynamics, in *Advances in Neural Information Processing Systems, Vol. 6* (J. D. Cowan, G. Tesauro, and J. Alspector, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 501–508.
- McGarry, K., Wermter, S., and MacIntyre, J., 1999, Hybrid neural systems: From simple coupling to fully integrated neural networks, *Neural Computing Surveys*, 2:62–93.
- Plate, T. A., 1995, Holographic reduced representations, *IEEE Transactions on Neural Networks*, 6:623–641.
- Rachkovskij, D. A., and Kussul, E. M., 2001, Binding and normalization of binary sparse distributed representations by context-dependent thinning, *Neural Computation*, 13:411–452.
- van der Velde, F., 1995, Symbol manipulation with neural networks: Production of a context-free language using a modifiable working memory, *Connection Science*, 7:247–280.

Associative Networks

James A. Anderson

Introduction

The operation of *association* involves the linkage of information with other information. Although the basic idea is simple, association gives rise to a particular form of computation, powerful and idiosyncratic. The mechanisms and implications of association have a long history in psychology and philosophy. Association is

also the most natural form of neural network computation. This article will discuss association as realized in neural networks as well as association in the more traditional senses.

Neural networks are often justified as abstractions of the architecture of the nervous system. They are composed of a number of computing units, roughly modeled on neurons, joined together by

connections that are roughly modeled on the synapses connecting real neurons together. The basic computational entity in a neural network is related to the pattern of activity shown by the units in a group of many units.

Because of the use of activity patterns—mathematized as state vectors—as computational primitives, the most common neural network architectures are pattern transformers which take an input pattern and transform it into an output pattern by way of system dynamics and a set of connections with appropriate weights. In a very general sense, therefore, neural networks are frequently designed as *pattern associators*, which link an input pattern with the “correct” output pattern. Learning rules are designed to construct accurate linkages. The most common feedforward neural network architectures realize this linkage by way of connections between layers of units (Figure 1). There may be a single set of modifiable connections between input and output (Figure 1B), or multiple layers of connections (Figure 1C). Another common architecture is realized by a single layer of units where the units in the layer are recurrently interconnected (Figure 1A).

One common design goal of a feedforward associator (Figures 1B and 1C) is to realize what Kohonen (1977) has labeled *hetero-association*, that is, to link input and output patterns that need have no relation to each other. Another possibility is to realize what Kohonen has called *autoassociation*, where the input and output patterns are identical. Recurrent networks (Figure 1A) are well suited to autoassociation.

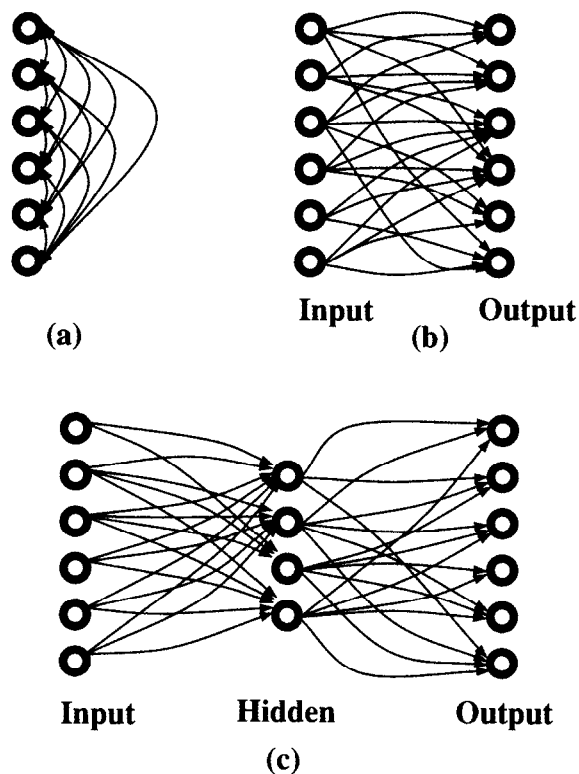


Figure 1. Three common basic neural network architectures. A, A set of units connects recurrently to itself by way of modifiable connections. (The connections are drawn as reciprocal.) B, A feedforward network in which an input pattern is transformed to an output pattern by way of a layer of modifiable connections. C, A more general feedforward network. An input layer projects to an intermediate layer of units. The intermediate layer is often called a *hidden layer* because it may not be accessible from outside the network. The hidden layer then projects to the output units.

Because the input and output patterns must correspond to information about the real world, the *data representation* is of critical importance at all levels of network operation. For example, simple pattern recognizers are often realized by neural networks as a special form of pattern associator by assuming a particular output representation, one where a single active output unit corresponds to the category of the input. Different categories correspond to different active output units. This highly localized representation is sometimes called a *grandmother cell* representation, because it implies that only when one particular unit is active is “grandmother” being represented. The alternative representation is called a *distributed representation*, where representation of a concept like “grandmother” may contain many active units. Choice of representation makes a major difference in how networks are used and how well they work, and is usually more important than the exact choice of network architecture and learning rule. A common situation in engineering applications of neural networks is to have a distributed representation at the input of the network and a grandmother cell representation at the output. In the vertebrate nervous system there is little evidence for this output representation; essentially all normal motor acts involve the coordinated discharge of large groups of neurons. Distributed activity patterns are associated with distributed patterns from one end of most biological networks to the other in vertebrates (but see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS), though there are some examples of extreme selectivity in invertebrates. The degree of distribution is a matter for experimental investigation.

Neural Network Associators

Let us give an example of how easily neural network learning rules and architectures give rise to associative behavior. Consider the two-layer network diagrammed in Figure 1B. Consider a situation where a pattern of activity, a state vector f , is present at the input set of units and another pattern, state vector g , is shown by the output set of units.

We want to link two patterns so that when f is presented to the input of the network, g will be generated at the output. In this two-layer network (two layers of units, one layer of connections), we will assume that the connections initially are zero and we want to change them to make the association between patterns f and g . We will also assume that all connection strengths are potentially changeable and the set of connection strengths forms a *connection matrix* (or *synaptic matrix*) which we will call W , for “weights.”

We have to propose a learning rule, but we also have to make some additional assumptions about the entire system. For example, virtually all artificial neural network learning assumes that the network is learning discrete pairs of patterns, that is, learning takes place only occasionally, when the time is ripe. One could speculate that learning in animals is a dangerous operation—after all, the nervous system is being rewired—and is kept under tight control. Primates are unusual in the degree of learned flexibility their nervous system allows. There is physiological evidence that amount of learning is controlled by diffuse biochemical processes. Dangerous and striking events, causing a biochemical upheaval, give rise to what have been called “flashbulb memories” where everything, including totally irrelevant detail, is learned. (“Where were you when John F. Kennedy was assassinated?” is practically guaranteed to involve a flashbulb memory in those old enough to remember it. September 11, 2001, provides a modern example.) Presumably this corresponds to an indiscriminating “learn” command. In terms of modeling, these observations mean that the decision to learn is decoupled from the act of learning.

Let us assume that we have an input pattern and an output pattern and we wish to associate them for good and sufficient reasons. We assume that we can impress pattern f on the input set of units and pattern g on the output set of units. By far the most common net-

work learning rule used is one or another variant of what is called the “Hebb synapse,” described in Hebb (1949). Perhaps the most quoted sentence in the neural network literature is from Hebb: “When an axon of cell *A* is near enough to excite a cell *B* and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells, such that *A*’s efficiency as one of the cells firing *B*, is increased.” (Hebb, 1949:62). The essence of the Hebb synapse is that there has to be a *conjunction* of activity on the two sides of the connection.

There is good physiological evidence for the existence of some form of Hebb synapse in parts of the mammalian central nervous system (see HEBBIAN SYNAPTIC PLASTICITY). However, there are a number of “technical” problems involved in mathematically describing the resulting system. The original formulation by Hebb was concerned with coincident excitation. Nothing was said about coincident inhibition or about coincident excitation and inhibition. Also, the exact function determining strength of modification was not given, and, in fact, is not known. A common assumption in *artificial* network theory is to assume some version of what is called the *generalized Hebb rule* or the *outer product rule*. This states that the change in strength of a connection during learning is given by the *product* of activities on the two sides of the connection, that is, if W_{ij} is the strength of the connection, then the change in strength ΔW_{ij} is proportional to the product, $f_j g_i$, where f_j is the activity of the j th input unit and g_i is the activity of the i th output unit. This convenient expression may have only a weak relationship to physiological reality.

Given the generalized Hebb rule, if we have only a single pair of vectors to associate, the results can be written compactly as

$$W = \eta g f^T$$

where η is a learning constant and W is the connection (or weight) matrix.

By making an additional assumption about the properties of the individual neural elements, this rule leads almost immediately to a simple pattern associator called the *linear associator*. Suppose the elementary computing units are linear, so that the output is given by the inner product between input activity and connection strengths. Then the output pattern is given by the matrix product of an input pattern f and the connection matrix W ; that is, the output of the network is Wf . Because we know what W is—it was constructed by the generalized Hebb rule—we can compute the output pattern,

$$(\text{output pattern}) = Wf = \eta g f^T f = (\text{constant})g$$

since $f^T f$ is a constant, the squared length of f . The output pattern is a constant multiple of g and, except for length, we have reconstructed the learned associate of f , that is, g .

Suppose we have a whole set of associations $\{f^i \rightarrow g^i\}$ that we want to teach the network. (Superscripts stand for individual pattern vectors.) If we assume that the overall strength of a connection is the algebraic sum of its past history (an unsupported assumption), then we have the weight matrix W given by

$$W = \sum_i \eta g^i f^{iT}$$

Notice that in the special case where the input patterns $\{f^i\}$ are orthogonal, that is, $f^i f^j = 0$ if $i \neq j$,

$$Wf^i = (\text{constant})g^i$$

because the contributions to the output pattern from the other terms forming W are identically zero since they involve the inner product $[f, f_j] = f_j^T f_j$. This model, and in fact most simple network models, make the prediction that outer product associators will work best and most reliably with representations where different input associations are as orthogonal as possible. For this reason, some cortical

models in the neuroscience literature have explicitly discussed aspects of cortical processing in terms of orthogonalization. The most complete reference for the linear associator and related models is Kohonen (1977, 1984).

It is possible to change almost any assumption and still have an associator. *Hebb learning rules of virtually any kind give rise to associative systems*. As only one example, the nonlinear Hebbian associator proposed by Willshaw, Buneman, and Longuet-Higgins (1969) used binary connections—with strengths either one or zero—and the resulting system still worked nicely as a pattern associator.

Supervised Networks

The outer product associator is less accurate with nonorthogonal patterns. However, observed distortions and human performance are sometimes remarkably similar. (See Anderson, 1995, chap. 11, for a model of “concept formation” that emerges when correlated inputs are stored in the linear associator.)

Most designers of artificial networks prefer networks to produce accurate reproductions of learned associations rather than interesting distortions. (This seemingly natural assumption is not necessarily a good one.) *Supervised* network algorithms can perform more accurate association. Examples of such algorithms would include the Widrow-Hoff (LMS) algorithm, the perceptron, back-propagation, and many others. The basic mechanism employed is *error correction*. Suppose we have an initial training set of patterns to be learned. This means we know what the output patterns are for a number of input patterns. We take an input from the training set and let the network generate an output pattern. We then compare the desired output pattern and the actual output pattern in some way. This process generates an *error signal*. The network is then modified using a learning rule so as to *reduce* the error signal.

The most commonly used error signal is based on the distance between the actual and desired output; however, other error signals can be more desirable. For example, one could incorporate a term penalizing large numbers of connections or large values of connection strength. The network learning problem reduces to a minimization problem where the space formed by the connection strengths (*weight space*) is searched to find the point where error is reduced to as low a value as possible. This process requires the use of control structures that can be complex; for example, there is assumed to be an omniscient *supervisor* who compares desired and actual network output and computes the error term as well as implements the mechanisms to change connection strengths appropriately. The structure of these algorithms is designed to produce good pattern association whether or not this is the aim of the network architects. (See PERCEPTRONS, ADALINES, AND BACKPROPAGATION.)

Autoassociative Models

We have described association as pattern linkage. However, there are alternative descriptions in the neural network literature. For example, in the first sentence of the second chapter of their textbook, *Introduction to the Theory of Neural Computation*, Hertz, Krogh, and Palmer (1991) write, “Associative memory is the ‘fruit fly’ or ‘Bohr atom’ problem of the field” (p. 11). Their definition of association is: “Store a set of patterns ζ_1, \dots in such a way that when presented with a new pattern ζ_i , the network responds by producing whichever one of the stored patterns most closely resembles ζ_i ” (p. 11). This is not, however, a description of association but of a *content addressable memory* where input of partial or noisy information is used to retrieve the correct stored information. The source of this limited view of association lies in the ability of auto-

associative systems to reconstruct missing or noisy parts of learned patterns.

Consider the autoassociative version of the linear associator. Suppose we learn one pattern, f , of length 1, with learning constant $\eta = 1$. Then

$$W = ff^T \quad \text{and} \quad Wf = f$$

Suppose we take vector f , with n elements, and set to zero some of the elements, forming a new vector, f' . Let us make a second vector, f'' , from only the elements that were set to zero in f' . Then $f' + f'' = f$ and $f' f'^T = 0$. If f' is input to the autoassociator,

$$Wf' = (f' + f'')(f' + f'')^T f' = (\text{constant})f$$

where the constant is related to the length of f' . In operation, by putting a part of f , f' , into the network, we retrieve all of f , bar a constant. This behavior is often referred to as the *reconstructive* or *holographic* property of neural networks. Of course, more subtle problems arise when W stores multiple vectors. Anyway, this type of memory is associative because if, for example, the state vector was meaningfully partitioned, then f' is associatively linked to f'' and vice versa in the sense that input of one pattern will produce the other. This kind of associator produces intrinsically bidirectional links (i.e., $f' \rightarrow f''$ and $f'' \rightarrow f'$), unlike feedforward heteroassociators ($f \rightarrow g$).

Some nonlinear “attractor” neural networks with dynamics that minimize energy functions develop their associative abilities largely from their autoassociative architecture. The best-known examples of this kind of associator are Hopfield networks and parallel feedback networks such as the BSB (Brain State in a Box) model (Anderson, 1995, chap. 15). For a general review of attractor networks, see Amit (1989) and *COMPUTING WITH ATTRACTORS*. Multilayer autoassociators are also possible. The multilayer *encoder networks*, which require the output pattern to be as accurate a reconstruction as possible of the input pattern, also have this form. Many autoassociative networks have close ties to known statistical techniques such as PRINCIPAL COMPONENT ANALYSIS.

A related associative attractor model, called a *bidirectional associative memory*, or BAM (Kosko, 1988), is a nonlinear dynamical system with a reciprocal feedback structure. It assumes two layers of units, as well as pairs of associations to be learned, as in a heteroassociator. There are connections from both input to output and output to input. Given f and g patterns to be learned, assumed to be binary vectors, we can form both a forward and a backward connection matrix. If f is input, then g will be given as the output; g at the output will give rise to f at the input because of the backward connections. Suppose the input is not exactly what was learned. After a few passes back and forth through the system, it can be shown that the network will stabilize, in the noise-free case, to the learned f and g .

Psychological Association

We have shown how neural networks easily form associators of many different kinds. We will now discuss a little of the history of association in psychology to show how associators form a style of computation with considerable power as well as severe limitations.

The major outlines of one way to use an associative computer can be found clearly expressed in Aristotle in the fourth century B.C. Aristotle made two important claims about memory structure: First, the elementary unit of memory is a *sense image*, that is, a sensory-based set of information. Second, links between these elementary memories serve as the basis for higher-level cognition. An English translation by Richard Sorabji (1969) used the term *memory* for the elementary memory unit and *recollection* for reasoning by associations between elementary units. Aristotle dis-

cussed at length how one “computes” with memorized sense images. The word *recollection* was used in the translation to denote this process: “Acts of recollection happen because one change is of a nature to occur after another.” That is, Aristotle proposed a linkage mechanism between memories. He suggested several ways that linkage could occur: by temporal succession or by “something similar, or opposite, or neighboring.” This list of the mechanisms for the formation of associations is approximately what would be given today by psychologists.

Recollection in Aristotle’s sense was computation. It was a dynamic and flexible process: “[R]ecollecting is, as it were, a sort of reasoning.” Aristotle argued that properly directed recollection is capable of discovering new truths, using memorized sense images as the raw material and learning to traverse new paths through memory (Figure 2).

A practical problem with such an associative net is branching, that is, what to do if there is more than one link leaving an elementary memory. Aristotle was aware of this problem: “[I]t is possible to move to more than one point from the same starting point.” A general solution to the branching problem requires a nonlinear mechanism to select one or the other branch.

The most influential psychologists in the twentieth century were the behaviorists, in particular B. F. Skinner, the Harvard psychologist whose ideas about reinforcement learning unfortunately dominated much of the theoretical discussion in psychology for several decades. This school held that learning formed an associative link between a stimulus and a specific response. The link could be strengthened by positive reinforcement (to a first approximation, something useful or pleasant, or the cessation of something unpleasant) or weakened by negative reinforcement (either absence of something pleasant or something actively unpleasant) when the response followed the stimulus. A number of careful experiments showed that there were accurate quantitative “laws of learning” that were followed by animals in some simple situations.

It was debatable whether this view of association is useful in more complex situations. From the beginning, human behavior has seemed to humans to be far richer than stimulus-response ($S \rightarrow R$) association. In the 1950s Skinner wrote a book attempting to explain language behavior using associative rules. In a famous book review, Chomsky (1957) pointed out that simple $S \rightarrow R$ association cannot do some kinds of linguistic computation. The argument used

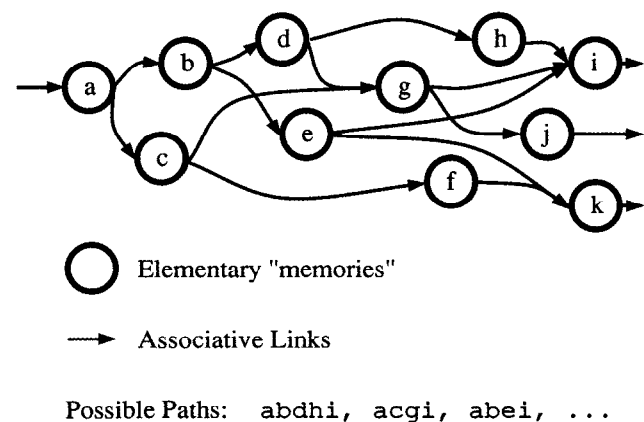


Figure 2. A simple model of associative computation. Elementary memories (“sense images,” according to Aristotle) are associatively linked (arrows) to other sense images. Branches are possible, and they present some difficulties. There are many possible paths through the network. Forming and traversing links between elementary memories is the basis of mental computation.

was that Skinner was proposing a well-defined computing machine with his associative model and that this computing machine was not powerful enough to do the computations we know language users perform. The simple $S \rightarrow R$ models of Skinner had about as much computing power as the simplest heteroassociative neural networks, which no one claimed were general-purpose computers. However, supervised network learning algorithms applied without insight may produce systems with only this degree of overall computational power.

“Connectionist” Models

Much modern work using association assumes that the entities linked, and the links themselves, can have complex internal structure. Flexible systems capable of complex reasoning can be produced by using labeled links: for example, a robin IS-A bird, an IS-A link, or “Fred is the father of Herb,” meaning that there is an associative link between Fred and Herb and that the link carries the relationship “Father-of.” Complex and sophisticated computational models, *semantic networks*, can be built from these pieces.

In the 1980s, many of those interested in semantic network models started working with neural networks. The term *connectionism* was often used to indicate the application of neural networks to high-level cognition. Recently there have been many attempts to apply networks to reasoning, to complex concept structures, and, in particular, to language understanding. A heated but illuminating debate arose from an early connectionist paper by Rumelhart and McClelland (1986) that used a neural network to simulate the way young children learn past tenses of verbs. Past tense learning had always been considered to be a good example of the application and misapplication of a specific rule, suggesting symbolic processing. Rumelhart and McClelland’s neural network acted as if it were using rules, but the rule-like behavior was the result of generalizing from examples and learning specific cases (see PAST TENSE LEARNING). Perhaps because this model was such a direct attack on the existence of rules in language, a vigorous counterattack developed. As one example, a long paper by Pinker and Prince (1988) finished its abstract with the sentence, “We conclude that connectionists’ claims about the dispensability of rules in explanations in the psychology of language must be rejected, and that, on the contrary, the linguistic and developmental facts provide good evidence for such rules” (p. 74). The vigor of the attack is perhaps due in part to the authors’ feeling that the connectionists had violated the “central dogma of modern cognitive science, namely that intelligence is the result of processing symbolic expressions” (Pinker and Prince, p. 74). Many other cognitive scientists feel that the “central dogma” is actually more like a central, and open, question.

Less well known outside psychology are several associative neural network models that were constructed to explain the fine structure of experimental data in more traditional areas of psychology such as verbal learning. An interesting example of such a model is the TODAM model of Murdock (see CLASSICAL LEARNING THEORY AND NEURAL NETWORKS in the first edition). TODAM and variants blur the distinction between the network and the representation. In the associative networks we have discussed, there are two formally distinct entities, state vectors and connection matrices. In the TODAM class of models, the association is stored with the items themselves and is therefore the same type of entity. TODAM makes a number of testable qualitative predictions about a wide range of data from the classical verbal learning literature. Recently, models assuming networks composed of large numbers of local networks (a “network of networks”) suggest that networks like TODAM might be realizable with neural networks.

Discussion and Open Questions

An often proclaimed virtue of neural networks is their ability to generalize effectively and to do computation based on similarity. Having learned example associations from a training set, the network can then generate correct answers to new examples. Many have pointed out the formal similarity of neural networks to approximation and interpolation as studied in numerical analysis. A properly designed neural network can act as a useful adaptive interpolator with good, even optimal, generalization around the region of the learned examples. However, it is not easy for neural networks to make good generalizations other than by approximation and interpolation. On this basis, Fodor and Pylyshyn (1988) made some telling arguments against the promiscuous application of connectionism to cognition (see SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS). The essential criticism they made is one that an engineer would be happy to make: Associative neural networks are such an inefficient way to compute that it would be foolish to build a cognitive system like that. Neural networks do not generalize well outside of a restricted definition based on mathematical interpolation, they cannot reason effectively, and they cannot extrapolate in any meaningful sense. These criticisms are part of a battle involving the limitations of association that has been going on for centuries. Fodor and Pylyshyn commented, “It’s an instructive paradox that the current attempt to be thoroughly modern and ‘take the brain seriously’ should lead to a psychology not readily distinguishable from the worst of Hume and Berkeley” (p. 64).

Fodor and Pylyshyn contrasted neural network associators with what they call the classical view of mental operation. In essence, this view postulates “a language of thought”; that is, “mental representations have a *combinatorial syntax and semantics*” (p. 12). The classical view is dominant in virtually all branches of traditional artificial intelligence and linguistics. The power of the digital computer arises in part from the fact that it is designed to be an extreme example of this organization: a programming language operating on data is the prototype of the classical view.

Suppose we have a sentence of the form *A and B* that we hold is true. An example Fodor and Pylyshyn used is *John went to the store and Mary went to the store*. The truth of this sentence logically entails the truth of *Mary went to the store*. This conclusion arises from the rules of logic and of grammar. It is not easy for an associative neural network to handle this problem. Such a network could easily learn that *John went to the store and Mary went to the store* is associated with *Mary went to the store*. But the power of the classical approach arises from the fact that every sentence of this form gives rise to the same result. Given the huge number of possible sentences, it *makes practical sense* to assume that some kind of logical syntax exists. It would be hard to figure out how language could function without some global rule-like operations, however implemented.

The ability to understand and answer sentences or phrases that are new to the listener is hard to explain purely with association. To give one example (see MENTAL ARITHMETIC USING NEURAL NETWORKS in the first edition), consider number comparisons such as “Is 7 bigger than 5?” There are nearly 100 such single-digit comparisons, nearly 10,000 two-digit comparisons, and so on. Children cannot possibly learn them as individual cases.

If there is a qualitative difference between human and animal cognition, it lies right here. There have been attempts to build neural networks that realize parts of the classical account, with indifferent success (see Hinton, 1991). Is it possible to build a neural network based largely on natural associators that can reproduce the kind of rule-governed behavior—even in limited domains—that does in fact seem to be part of human cognition? A neural network with this ability would allow for much more powerful and useful

generalization than current networks provide. It may not be easy to find this solution. There are many animals with complex nervous systems capable of associative learning, but only our own species, one out of millions of species, is really effective at using these powerful extensions to association.

[Reprinted from the First Edition]

Road Maps: Grounding Models of Networks; Learning in Artificial Networks

Background: I.3. Dynamics and Adaptation in Neural Networks

Related Reading: Artificial Intelligence and Neural Networks; Computing with Attractors

References

- Amit, D. J., 1989, *Modelling Brain Function: The World of Attractor Neural Networks*, Cambridge, Engl.: Cambridge University Press.
- Anderson, J. A., 1995, *Introduction to Neural Networks*, Cambridge, MA: MIT Press. ♦
- Anderson, J. R., 1983, *The Architecture of Cognition*, Cambridge, MA: Harvard University Press.
- Chomsky, N., 1957, A review of Skinner's *Verbal Behavior*, *Language*, 35:26–58.

- Fodor, J. A., and Pylyshyn, Z. W., 1988, Connectionism and cognitive architecture: A critical analysis, in *Connections and Symbols* (S. Pinker and J. Mehler, Eds.), Cambridge, MA: MIT Press.
- Hebb, D. O., 1949, *The Organization of Behavior*, New York: Wiley.
- Hertz, J., Krogh, A., and Palmer, R. G., 1991, *Introduction to the Theory of Neural Computation*, Redwood City, CA: Addison-Wesley. ♦
- Hinton, G. E., 1991, *Connectionist Symbol Processing*, Cambridge, MA: MIT Press. ♦
- Kohonen, T., 1977, *Associative Memory: A System Theoretic Approach*, Berlin: Springer-Verlag. ♦
- Kohonen, T., 1984, *Self-Organization and Associative Memory*, Berlin: Springer-Verlag. ♦
- Kosko, B., 1988, Bidirectional associative memories, *IEEE Trans. Sys., Man Cybern.*, 18:49–60.
- Pinker, S., and Prince, A., 1988, On language and connectionism: Analysis of a parallel distributed processing model of language acquisition, in *Connections and Symbols* (S. Pinker and J. Mehler, Eds.), Cambridge, MA: MIT Press.
- Rumelhart, D. E., and McClelland, J. L., 1986, On learning the past tenses of English verbs, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (D. E. Rumelhart, J. L. McClelland, and PDP Research Group, Eds.), vol. 2, *Psychological and Biological Models*, Cambridge, MA: MIT Press.
- Sorabji, R., 1969, *Aristotle on Memory*, Providence, RI: Brown University Press.
- Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C., 1969, Non-holographic associative memory, *Nature*, 222:960–962.

Auditory Cortex

Shihab A. Shamma

Introduction

The auditory cortex plays a critical role in the perception and localization of complex sounds. It is the last station in a long chain of processing centers that begins with the cochlea of the inner ear and passes through the cochlear nuclei (CN), the superior olivary complex (SOC), the lateral lemniscus, the inferior colliculus (IC), and the medial geniculate body (MGB) (Figure 1). Recent studies have expanded our knowledge of the neuroanatomical structure, the subdivisions, and the connectivities of all central auditory stages (Winer, 1992). However, apart from the midbrain cochlear

and binaural SOC nuclei, relatively little is known about the functional organization of the central auditory system, especially compared to the visual and motor systems. Consequently, modeling cortical auditory networks is complicated by uncertainty about exactly what the cortical machinery is trying to accomplish.

One exception to this state of affairs is the highly specialized echolocating bat, in which these uncertainties are much relieved by the existence of a stereotypical behavioral repertoire that is closely linked to the animal's acoustic environment (see ECHOLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS). This has made it possible to construct a functional map of the auditory cortex, which

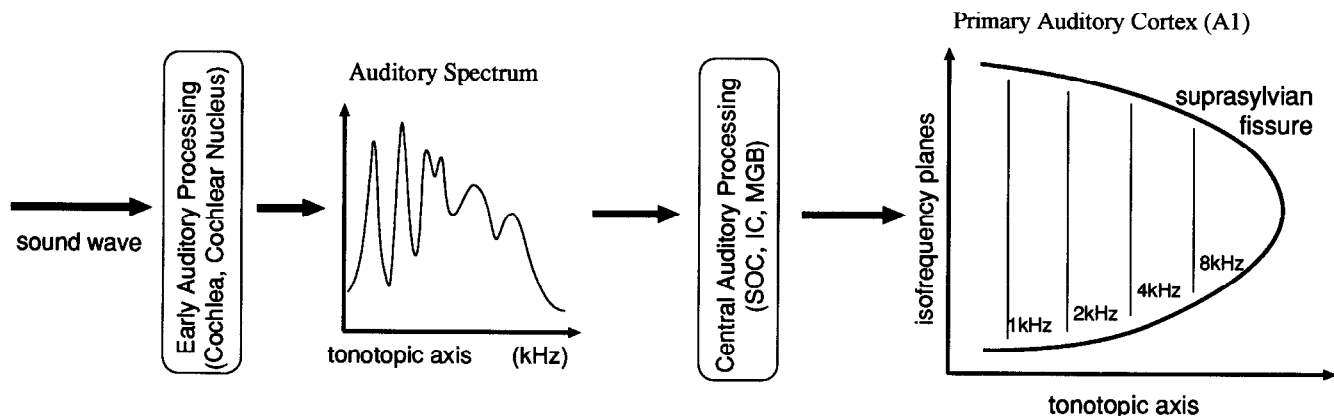


Figure 1. Schematic representation of the multiple stages of processing in the mammalian auditory pathway. Sound is analyzed in the cochlea, and an estimate of the acoustic spectrum (an auditory spectrum) is known to be extracted at the cochlear nucleus (Blackburn and Sachs, 1990). The tonotopic organization of the cochlea is preserved all the way up to the cortex,

where it has a two-dimensional layout. The isofrequency plane encodes perhaps other features of the stimulus.

revealed the specific acoustic features extracted and represented in the cortex. In turn, these cortical maps have acted as a guide to discovering the organization and nature of the transformations occurring in lower auditory centers such as the MGB, IC, and SOC. Thus, it has become meaningful in these species to investigate and model cortical and other central auditory neural networks.

In other mammals, it is more difficult to isolate an auditory behavior and its associated stimulus features with comparable specificity. Nevertheless, a few tasks have been broadly accepted as vital for all species, such as sound localization, timbre recognition, and pitch perception. For each, evidence of various functional and stimulus feature maps has been found or postulated, a significant number of them in the last few years. In this review, we elaborate on a few examples of such maps and relate them to the more intuitive and better understood case of the echolocating bats. In each example, our goal is to determine how and whether models of the underlying neural networks can further our understanding of the auditory cortex.

Parcellization and Neuroanatomy of the Auditory Cortex

The layout and neural structure of the auditory cortex is in many respects similar to that of other sensory cortices (Winer, 1992). For instance, based on cytoarchitectonic criteria and patterns of connectivity, it is subdivided into a primary auditory field (AI) and several other surrounding fields, e.g., the anterior auditory field (A) and the secondary auditory cortex (AII). The number and specific arrangement of surrounding fields vary among different species, reflecting presumably the complexity of the animal's acoustic environment. The AI, and possibly other fields, is further subdivided into smaller regions, serving perhaps different functional roles, such as echo delay and amplitude measurements in the bat (see ECHOLLOCATION: COCHLEOTOPIC AND COMPUTATIONAL MAPS).

The anatomical parcellization of the auditory cortex into different fields is mirrored by physiologically based divisions. Most important is the systematic frequency organization in different fields, or so-called tonotopic maps. For example, AI cells are spatially ordered based on the tone frequency to which they best respond, i.e., their best frequency (BF). They also respond vigorously to the onset of a tone and exhibit little evidence of adaptation to its repeated presentations. In other fields, cells may be less frequency selective, may respond more adaptively, or may be totally unresponsive to single tones, preferring more spectrally or temporally complex stimuli. A sudden change in these response patterns or in the gradual spatial order of the tonotopic map is usually taken to signify a border between different fields. In the cat, which has the most extensively mapped auditory cortex, four well-ordered tonotopic fields have been described, together with many other less precise secondary areas (Clarey, Barone, and Imig, 1992).

Timbre: Models for the Encoding of Spectral Profiles

Recognizing and classifying environmental sounds is critical for the survival and propagation of many animals. Although a multitude of cues are responsible, the single most important one is the shape of the so-called spectral envelope (or the spectral profile) of the sound. It is largely this cue that allows us to distinguish between speech vowels or between different instruments playing the same note. The spectral profile emerges early in the auditory system as the sound is analyzed into different frequency bands, in effect distributing its energy across the tonotopic axis (the auditory sensory epithelium) (Figure 1). As far as the central auditory system is concerned, the spectral profile is a one-dimensional (1D) pattern of activation analogous to the two-dimensional (2D) distribution of light intensity on the retina.

An important organizational feature of the central auditory system is the expansion of the 1D tonotopic axis of the cochlea into a 2D sheet, with each frequency represented by an entire sheet of cells (Figure 1). An immediate question thus arises as to the functional purpose of this expansion and the nature of the acoustic features that might be mapped along these isofrequency planes. For example, one might conjecture that the amplitude or the local shape of the spectrum is explicitly represented along this new dimension.

In general, there are two ways in which the spectral profile can be encoded in the central auditory system. The first is *absolute*, that is, the spectral profile is encoded in terms of the absolute intensity of sound at each frequency. Such an encoding would in effect combine both the shape information and the overall level. The second way is *relative*, in which the spectral profile shape is encoded separately from the overall loudness of the stimulus. Examples of each of these two hypotheses are discussed next.

The Best-Intensity Model

The first hypothesis is motivated primarily by the strongly non-monotonic responses as a function of stimulus intensity observed in many cortical and other central auditory cells (Clarey et al., 1992). In a sense, one can view such a cell's response as being selective to (or encoding) a particular intensity. Consequently, a population of such cells, tuned to different frequencies and intensities, can provide an explicit representation of the spectral profile by their spatial pattern of activity (Figure 2). This scheme is not a true transformation of the spectral features represented, but rather is strictly a change in the means of the representation. The most compelling example of such a representation is that in the DSCF area of AI in the mustache bat. However, an extension of this hypothesis to multicomponent stimuli (as depicted in Figure 2) has not been demonstrated in any species.

The Multiresolution Analysis Model

The second hypothesis, in which the relative shape of the spectrum is encoded, is supported by physiological experiments in cat and

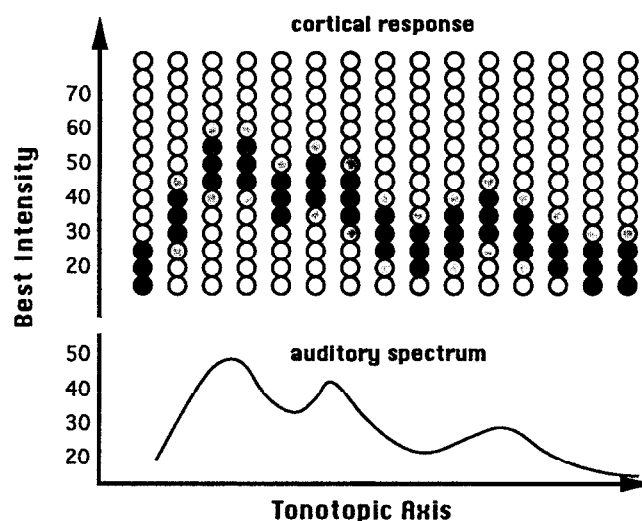


Figure 2. Schematic diagram of the way in which the spectral profile (lower plot) can be encoded by arrays of nonmonotonic cells (circles) tuned to different BFs (along the tonotopic axis) and best intensities (BIs). The black circles signify strongly activated cells, whereas the white circles indicate weakly activated cells. Thus, a peak in the input pattern located at a given BF and at an intensity of 40 dB would best activate cells with the same BF and BI.

ferret AI, coupled with psychoacoustical studies in human subjects. The data reveal a substantial transformation of the way the spectral profile is represented centrally. Specifically, besides the tonotopic axis, two features of the response areas of AI neurons (the analogue of the receptive fields in the visual system) are found to be topographically mapped across the isofrequency planes. They are the bandwidth and symmetry of the response areas, depicted schematically in Figure 3A as the *scale* and *symmetry* axes, respectively. In addition, auditory cortical units exhibit systematic response patterns to dynamic spectra that give rise to complex and varied spectrotemporal response areas, as depicted in Figure 3B. These response properties are discussed in greater detail below.

Changes in response area bandwidths. Cell response areas, i.e., the excitatory and inhibitory responses they exhibit to a tone of various frequencies and intensities, change their bandwidth in orderly fashion along the isofrequency planes (Mendelson and Schneiner, 1990). Near the center of AI, cells are narrowly tuned. Toward the edges, they become more broadly tuned. This orderly progression occurs at least twice, and it correlates with several other response parameters such as increasing response thresholds toward the edges.

An intuitively appealing implication of this finding is that response areas of different bandwidths are selective to spectral profiles of different widths. Thus, broad spectral profiles (e.g., broad peaks or gross trends, such as spectral tilts due to preemphasis) would best drive cells with wide response areas. Similarly, narrower spectral profiles (e.g., sharp peaks or edges, or fine details of the spectral profile) would best be represented in the responses of cells with more compact response areas. In effect, having a range of response areas at different widths allows us to encode the spectral profile at different scales or levels of detail (resolution). From a mathematical perspective, this is basically equivalent to analyzing the spectral profile into different scales or “bands,” much like performing a Fourier transform of the profile, hence representing it as a weighted sum of elementary sinusoidal spectra (usually known as *ripples*; Shamma, Versnel, and Kowalski, 1995). Coarser scales then correspond to the “low-frequency” ripples, while finer scales correspond to the “high-frequency” ripples.

Changes in response area asymmetry. Response areas exhibit systematic changes in the symmetry of their inhibitory response areas. For instance, cells in the center of AI have sharply tuned excitatory responses around a BF, flanked by symmetric inhibitory response areas. Toward the edges, the inhibitory response areas become significantly more asymmetric, with inhibition dominated by either higher or lower than BF frequencies. This trend is repeated at least twice across the length of the isofrequency plane.

It is intuitively clear that response areas with different symmetries would respond best to input profiles that match their symmetry. For instance, an odd-symmetric response area would respond best if the input profile had the same local odd-symmetry and worst if it had the opposite odd-symmetry. As such, one can state that a range of response areas of different symmetries (symmetry axis in Figure 3A) is capable of encoding the shape of a local region in the profile. From an opposite perspective, it can be shown mathematically that the local symmetry of a pattern can be changed by manipulating only the phase of its Fourier transform (Wang and Shamma, 1995). Therefore, the axis of response area asymmetries in effect is able to encode the phase of the profile transform, thus providing a complementary description to that of the magnitude along the scale axis described above.

Dynamics of cortical responses to spectral profile changes. Auditory cortical units also exhibit systematic and selective responses to dynamic spectra. Specifically, when stimulated by com-

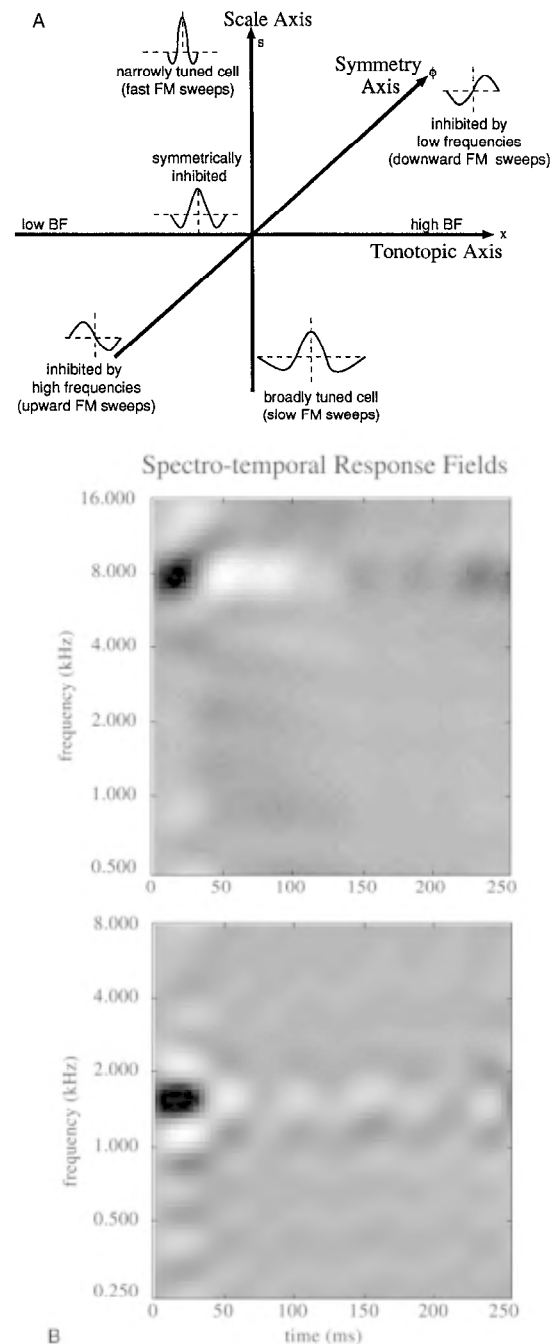


Figure 3. A, Schematic diagram of the three representational axes thought to exist in AI: the tonotopic (BF) axis, the scale (or bandwidth) axis, and the symmetry axis. B, Examples of spectrotemporal response fields measured from two auditory cortical units of the ferret. In each panel, the strength of the response is represented by the darkness of the display, with black indicating excitatory areas and white indicating regions of suppressed activity. Note that the excitatory central region defines the BF of the unit. Such STRFs exhibit a variety of bandwidths, asymmetry of inhibition relative to the BF, directional selectivity, and temporal dynamics. For instance, the unit in the top panel has significantly slower dynamics and much more asymmetric inhibition about the BF than the unit in the bottom panel. (From Simon, J. Z., Depireux, D. A., and Shamma, S. A., 1998, Representation of complex spectrain auditory cortex, in *Psychophysical and Physiological Advances in Hearing: Proceedings of the 11th International Symposium on Hearing* (A. R. Palmer, A. Ress, A. Q. Summerfield, and R. Meddis Eds.), London: Whurr, 1998, pp. 513–520. Reprinted with permission.)

plex sounds with rippled spectra like those described above, cortical units display preference not only to ripple density and phase, but also to the velocity at which the ripple is drifted past the BF of the cell. Unit selectivities span a wide range of best ripple velocities, from about 20 cycles/s (Hz), down to as low as 1–2 Hz (Kowalski, Depireux, and Shamma, 1996). In addition, auditory cortical units usually exhibit a range of directional sensitivities to upward- and downward-moving ripples.

This directional selectivity is probably directly linked to responses to frequency-modulated (FM) tones, a subject that has been the focus of extensive neural network modeling. These stimuli are important because they mimic the dynamic aspects of many natural vocalizations, as in speech consonant-vowel combinations or the trills of many birds and other animal sounds. The effects of manipulating two specific parameters of the FM sweep, its direction and rate, have been well studied. In several species and at almost all central auditory stages, cells can be found that are selectively sensitive to the FM direction and rate. Most studies have confirmed a qualitative theory in which directional selectivity arises from an asymmetric pattern of inhibition in the response area of the cell

(Wang and Shamma, 1995), whereas rate sensitivity is correlated to the bandwidth of the response area (Heil, Langner, and Scheich, 1992).

The full spectrotemporal response fields. All of above mentioned response area features are integrated into a unified spectrotemporal response area (or field) as illustrated in Figure 3B (deCharms, Blake, and Merzenich, 1998). The full spectrotemporal response field (STRF) summarizes all the response selectivities of a unit by the relative locations, widths, duration, and orientation of its excitatory and inhibitory fields. The overall picture that emerges from these findings is that AI decomposes the auditory spectrum into a multidimensional representation with multiple resolutions along both spectral and temporal dimensions, as illustrated in Figure 4. This spectrotemporal decomposition essentially segregates diverse perceptual features into different streams, e.g., fast, spectrally broad sounds (consonants) from the relatively slow, voiced vowels and the finely resolved harmonics (pitch cues) (Wang and Shamma, 1995). This kind of multiscale analysis is closely analogous to the well-studied organization of receptive fields in the primary visual

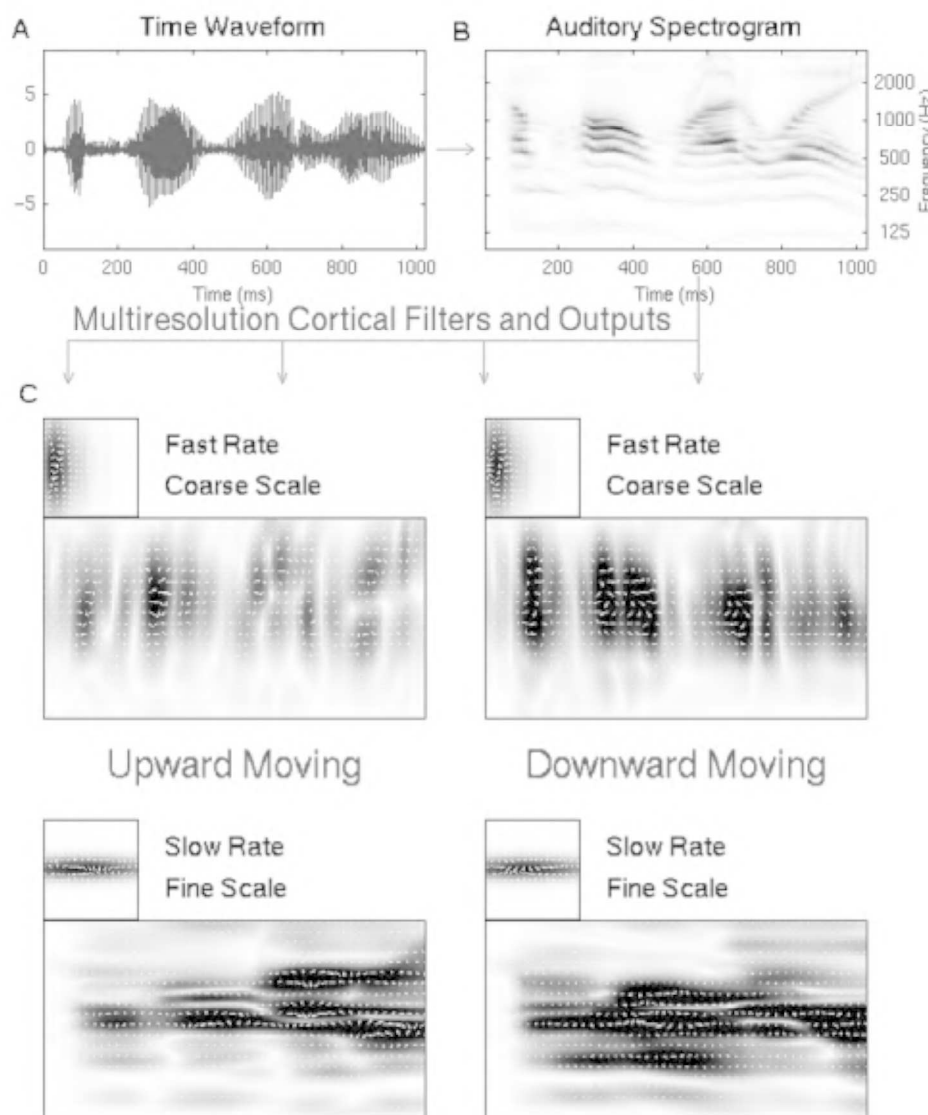


Figure 4. Schematic of the cortical representation of complex dynamic sound spectra. *A*, The time waveform of the acoustic signal /Come home right away/. *B*, The time-frequency representation of the signal (or the auditory spectrogram) generated in the early stages of the auditory system. The y-axis represents the logarithmic frequency axis of the cochlea (or the tonotopic axis, as depicted in Figure 1). *C*, Cortical multiscale analysis of the auditory spectrogram along the spectral and temporal dimensions. Each panel represents the activity of a population of cortical cells with the (idealized model) STRF shown in the inset above it. Arrow direction represent the phase of the response; the strength of the response is indicated by the darkness of the display, as in Figure 3B. The two top panels are for broadly tuned but relatively fast STRFs that are selective to motion in opposite directions. The bottom panels are for narrowly tuned and relatively slow units. Different features of the spectrogram are emphasized in different panels. For instance, harmonics (pitch cues) are seen in the lower (fine-scale) panels, whereas onsets due to different consonants are seen only in the upper (fast-rate) panels.

cortex (De-Valois and De-Valois, 1990), and may reflect a general principle of analysis of sensory patterns in all other sensoricortical areas.

Models of Pitch Representation in the Central Auditory System

A sound complex consisting of several harmonics is heard with a strong pitch at the fundamental frequency of the harmonic series, even if there is no energy at all at that frequency. This percept has been variously called the missing fundamental, virtual pitch, or residue pitch. A large number of psychoacoustical experiments have been carried out to elucidate the nature of this percept and its relationship to the physical parameters of the stimulus. Basically, all models fall into one of two camps. In the first camp, the pitch is extracted explicitly from the harmonic spectral pattern. This can be accomplished in a variety of ways, such as by finding the best match between the input pattern and various harmonic templates assumed to be stored in the brain (Goldstein, 1973). In the second camp, the pitch is extracted from the periodicities in the time-waveform of responses in the auditory pathway, which can be estimated, for example, by computing their autocorrelation functions. In this kind of model, some form of organized delay lines are assumed to exist so that the computations can be done, much like those that seem to exist in the FM-FM area of the mustached bat.

In all pitch models, however, the extracted pitch is assumed to be finally represented as a *spatial* map in higher auditory centers. This is because many studies have confirmed that neural synchrony to the repetitive features of a stimulus, be it the waveform of a tone or its AM modulations, becomes progressively worse toward the cortex (Langner, 1992). It is a remarkable aspect of pitch that, despite its fundamental and ubiquitous role in auditory perception, only a few reports exist of physiological evidence of spatial pitch maps, and none has been independently confirmed. One source is NMR scans of the primary auditory cortex in human subjects. The other source of evidence is multiunit mappings in various central auditory structures (Schreiner and Langner, 1988).

Of course, the difficulty of finding spatial pitch maps in the auditory cortex may be due to the fact that it does not exist. This possibility is counterintuitive, given the results of ablation studies showing that bilateral cortical lesions in the auditory cortex severely impair the perception of pitch of complex sounds but do not affect the fine discrimination of frequency and intensity of simple tones. Another possibility is that the maps sought are not at all as straightforward as we imagine. For example, harmonic complexes may evoke stereotypical patterns that are distributed over large areas in the auditory cortex, and not localized, as the simple notion of a pitch map implies (Wang and Shamma, 1995). Finally, it is also possible that AI simply functions as one stage that projects sufficient temporal or spectral cues for later cortical stages to extract the pitch explicitly.

Models of Sound Localization

It has been recognized for many years that the auditory cortex (and especially the AI) is involved in sound localization. Detailed physiological studies further confirmed that AI cells are rather sensitive to all kinds of manipulations of the binaural stimulus (Clarey et al., 1992). For instance, changing either of the two most important binaural cues, the interaural level difference (ILD) or interaural time difference (ITD), causes substantial changes in their firing rate patterns. This sensitivity to interaural cues has its origins early in the auditory pathway, at the SOC, where the first convergence of binaural inputs occurs. However, despite this diversity, two elements typical of a functional organization of AI have been lacking. The first missing element is a significant transformation of the single-unit responses. For example, if ILD-sensitive cells are to

encode the location of a sound source based on this cue, they ought to become uniformly more stable with overall sound intensity. This, however, does not seem to be the case (Semple and Kitzes, 1993). The second element lacking is a topographical distribution of the responses with respect to these cues or to a more complex combination of features (e.g., a map of acoustic space derived from ILD and ITD cues, as in the barn owl) (Sullivan and Konishi, 1986).

A map of auditory space has indeed been found in the superior colliculus of several mammals. No such map, however, has yet been detected in AI or other cortical fields despite intensive efforts (Clarey et al., 1992). What has been found, however, is a topographic order of certain binaural responses along the isofrequency planes of AI. Specifically, cells excited equally well by sounds from both ears (called EE cells) and others inhibited by ipsilateral sounds (called EI cells) are found clustered in alternating bands that parallel the tonotopic axis. One possible functional model that utilizes such maps assumes that EI cells are tuned to particular ILDs, and hence encode the location of a sound source based on this cue. EE cells, in contrast, would encode the absolute level of the sound. However, there is little evidence to support this hypothesis in the sense that neither EE nor EI cells are particularly stable encoders of specific ILD or absolute sound levels. An alternative hypothesis recently proposed is that these cells encode the absolute levels of the stimulus at each ear, rather than the difference and average binaural levels, as previously postulated (Semple and Kitzes, 1993). Finally, it has also been proposed that AI units encode the spatial location of a stimulus through unique patterns of temporal firing, ones that can be discerned using more elaborate pattern recognition neural networks (Middlebrooks et al., 1994).

Discussion

The study of central auditory function has reached a sufficiently advanced stage to allow meaningful quantitative and neuronal network models to be formulated. In most mammals, these models are still systemic in nature, with a primary focus on understanding the overall functional organization of the cortex and other central auditory structures. In the bat and other specialized animals, the models are somewhat more detailed, addressing specific neuronal mechanisms, such as the coincidences and the delay lines of the FM-FM areas. The auditory system, with its multitude of diverse functions and its combination of temporal and spatial processes, should thus prove to be a valuable window into the brain and an effective vehicle for understanding the brain's underlying mechanisms.

Road Maps: Mammalian Brain Regions; Other Sensory Systems

Related Reading: Auditory Periphery and Cochlear Nucleus; Auditory Scene Analysis; Echolocation: Cochleotopic and Computational Maps; Sound Localization and Binaural Processing

References

- Blackburn, C. C., and Sachs, M. B., 1990, The representations of the steady-state vowel sound phoneme *e* in the discharge patterns of cat anteroventral cochlear nucleus neurons, *J. Neurophysiol.*, 63(5):1191–1212.
- Clarey, J., Barone, P., and Imig, T., 1992, Physiology of thalamus and cortex, in *The Mammalian Auditory Pathway: Neurophysiology* (R. Fay, D. Webster, and A. Popper, Eds.), New York: Springer-Verlag, pp. 232–334.
- deCharms, R. C., Blake, D. T., and Merzenich, M. M., 1998, Optimizing sound features for cortical neurons, *Science*, 280:1439. ♦
- De-Valois, R., and De-Valois, K., 1990, *Spatial Vision*, New York: Oxford University Press.
- Goldstein, J., 1973, An optimum processor theory for the central formation of pitch of complex tones, *J. Acoust. Soc. Am.*, 54:1496–1516.
- Heil, P., Langner, G., and Scheich, H., 1992, Processing of FM stimuli in the chick auditory cortex analogue: Evidence of topographic representations and possible mechanisms of rate and directional sensitivity, *J. Comp. Physiol. A*, 171:583–600.
- Kowalski, N., Depireux, D., and Shamma, S., 1996, Analysis of dynamic

- spectra in ferret primary auditory cortex: Characteristics of single unit responses to moving ripple spectra, *J. Neurophysiol.*, 76:3503–3523.
- Langner, G., 1992, Periodicity coding in the auditory system, *Hearing Res.*, 6:115–142.
- Mendelson, J., and Schreiner, C., 1990, Functional topography of cat primary auditory cortex: Distribution of integrated excitation, *J. Neurophysiol.*, 64:1442–1459.
- Middlebrooks, J. C., Clock, A. E., Xu, L., and Green, D. M., 1994, A panoramic code for sound location by cortical neurons, *Science*, 264:842–844.
- Schreiner, C., and Langner, G., 1988, Periodicity coding in the inferior colliculus of the cat: 2. Topographical organization, *J. Neurophysiol.*, 60:1823–1840.
- Semple, M., and Kitzes, L., 1993, Binaural processing of sound pressure level in cat primary auditory cortex: Evidence for a representation based

- on absolute levels rather than level differences, *J. Neurophysiol.*, 69:449–461.
- Shamma, S., Versnel, H., and Kowalski, N., 1995, Ripple analysis in the ferret primary auditory cortex: 1. Response characteristics of single units to sinusoidally rippled spectra, *J. Aud. Neurosci.*, 1:233–254.
- Sullivan, W., and Konishi, M., 1986, Neural map of interaural phase difference in the owl's brainstem, *Proc Natl Acad. Sci. USA*, 83:8400–8404.
- Winer, J., 1992, The functional architecture of the medial geniculate body and primary auditory cortex, in *The Mammalian Auditory Pathway: Neuroanatomy* (D. Webster, A. Popper, and R. Fay, Eds.), New York: Springer-Verlag, pp. 232–334.
- Wang, K., and Shamma, S., 1995, Representation of spectral profiles in primary auditory cortex, *IEEE Trans. Speech Audio Process.*, 3:382–395.

Auditory Periphery and Cochlear Nucleus

David C. Mountain

Introduction

The auditory periphery transforms a very high information rate acoustic signal into a group of lower information rate neural signals. This process of parallelization is essential because the potential information rate in the acoustic stimulus is on the order of 0.5 megabits per second, and yet typical auditory nerve (AN) fibers have maximum sustained firing rates of 200 per second. The cochlear nucleus (CN) continues the process of parallelization by creating multiple representations of the original acoustic stimulus, with each representation emphasizing different acoustic features.

The major ascending auditory pathways are summarized in Figure 1. Sound is collected by the external ear (pinna) and passes through the ear canal to the eardrum (tympanic membrane), where it excites the middle ear. The middle ear couples the acoustic energy to the fluids of the cochlea, where transduction takes place. The sensory cells of the cochlea (hair cells) convert the mechanical signal to an electrical signal, which is then encoded by the fibers of the auditory nerve and transmitted to the CN in the brainstem. Within the CN, parallel information streams are created that feed other brainstem structures such as the superior olivary complex (SOC), the nuclei of the lateral lemniscus (NLL), and the inferior colliculus (IC). These parallel pathways are believed to be specialized for the processing of different auditory features that are used for sound source classification and localization. From the IC, auditory information is passed on to the medial geniculate body (MGB) in the thalamus, and from there to the auditory cortex.

External Ear

The head and pinna modify the magnitude and phase of the acoustic signal reaching the tympanic membrane in such a way as to provide important cues for sound source localization (Shaw in Gilkey and Anderson, 1997). The transfer function relating tympanic membrane pressure to pressure in the free field is called the head-related transfer function (HRTF) and changes with sound source elevation and azimuth.

Three major mechanisms contribute to the creation of the HRTF. The distance between the ears in most mammals is sufficient to create significant interaural time delays (ITDs) for sound sources off to the side of the head, and the head is large enough to create interaural level differences (ILDs) for frequencies where the wavelength is comparable or smaller than the head. For higher frequencies (above 5 kHz in humans), multiple resonant modes in the pinna add further complexity. These modes are preferentially excited by

sound waves from some directions but not others, resulting in an HRTF with peaks and valleys that change with sound source direction.

Middle Ear and Cochlear Mechanics

The middle ear consists of the tympanic membrane, the three middle-ear bones (ossicles), and the Eustachian tube. The primary function of the middle ear is to match the low acoustic impedance of air to the high acoustic input impedance of the cochlea. The middle-ear transfer function (ratio of intracochlear pressure to ear canal pressure) is high-pass in nature (Rosowski in Hawkins et al.,

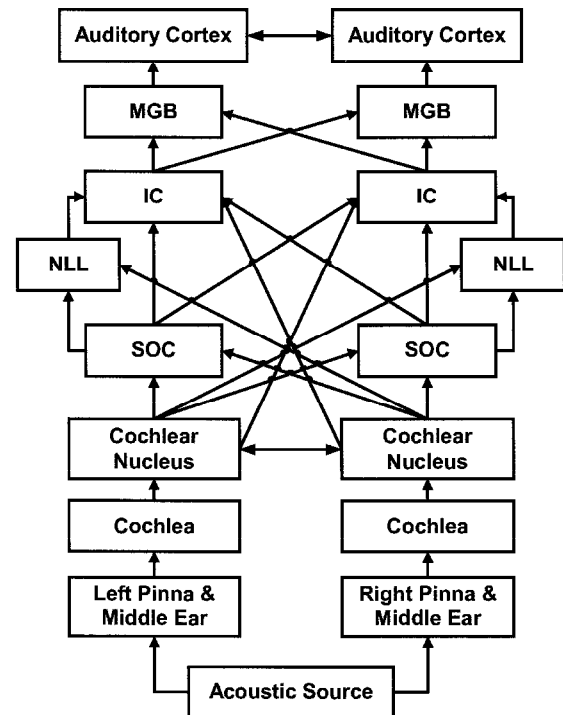


Figure 1. The major ascending auditory pathways. See text for explanation of abbreviations.

1996) and plays a major role in determining the audiogram for a given species.

The cochlea consists of a spiral-shaped, fluid-filled tube embedded in the temporal bone (Slepecky in Dallos, Popper, and Fay, 1996). It is separated into three longitudinal compartments by two membranes: the basilar membrane (BM) and Reissner's membrane. From a hydromechanical and physiological point of view, the BM is the more important of the two. It supports the organ of Corti, which contains the sensory hair cells. Pressure changes in the cochlear fluids produced by the middle ear excite a mechanical traveling wave that propagates along the BM. The traveling wave magnitude peaks at a location that depends on stimulus frequency: high frequencies peak near the base and low frequencies peak near the apex.

Direct measurements of BM motion demonstrate that, at low sound levels, the response can be highly tuned, with each cochlear location only responding to a narrow range of frequencies (Hubbard and Mountain in Hawkins et al., 1996). BM tuning decreases at high sound levels and appears to involve the presence of a group of sensory cells, the outer hair cells (OHCs). All hair cells respond to mechanical stimuli with voltage changes, but in the case of the OHCs, voltage changes result in cell length changes (Holley in Dallos et al., 1996). These voltage-dependent length changes appear to be mediated by voltage-sensitive transmembrane proteins. This novel form of electromotility is piezoelectric in nature, allowing the length changes to achieve very high velocities.

Many hydromechanical models have been proposed to explain these findings (Hubbard and Mountain in Hawkins et al., 1996; de Boer in Dallos et al., 1996), but these hydromechanical models are computationally intense. As a result, it is common practice to represent cochlear mechanics with a bank of digital bandpass filters that capture the salient features of the mechanical frequency response (Hubbard and Mountain in Hawkins et al., 1996). Filters of this type reproduce the magnitude of the cochlear frequency response reasonably well, but they cannot reproduce the changes in cochlear tuning that occur with changes in stimulus level. In order to replicate the nonlinear features of cochlear mechanics in filter-bank models, some authors have used filters with parameters that change with stimulus level (cf. Zhang et al., 2001).

Inner Hair Cells

The inner hair cells (IHCs) are the receptor cells that provide most of the input to the auditory nerve. Although much progress has been made in measuring basilar membrane motion, little direct data exist to explain how this motion gets coupled to the IHC hair bundle. Comparisons of IHC receptor potentials to inferred BM motion have led to the hypothesis that hair-bundle motion is a high-pass filtered version (cutoff frequency ~ 400 Hz) of BM motion. Alternatively, Mountain and Cody (1999) have proposed a model in which the OHCs, through their electromotility, displace the IHC hair bundles more directly, perhaps via movements of the tectorial membrane, rather than via enhanced BM motion.

The mechanical-to-electrical transduction process in hair cells is extremely sensitive, resulting in receptor potentials on the order of 1 mV for hair-bundle displacements of 1 nm. This transduction process is believed to be the result of tension-gated channels located in the hair bundle (Mountain and Hubbard in Hawkins et al., 1996). The relationship between stereocilia displacement x and the mechanically induced conductance change $G(x)$ is most commonly modeled using a first-order Boltzmann model (Mountain and Hubbard in Hawkins et al., 1996).

Although IHCs also contain voltage-dependent conductances (Kros in Dallos et al., 1996), most models include only the mechanically sensitive conductance coupled to a linear leakage resistance and a linear membrane capacitance. The RC nature of the

membrane acts as a low-pass filter with a cutoff frequency of around 1 kHz. The effect of this filter is to produce an IHC response that follows the fine structure of the stimulus waveform at low frequencies, while at high frequencies it follows the signal envelope (Mountain and Hubbard in Hawkins et al., 1996).

If a linear filter bank is used to represent cochlear mechanics, then it is often desirable to use a rectification function that includes considerable compression to accommodate the large dynamic range of many acoustic signals. Since the D.C. receptor potentials of IHCs measured using best-frequency tones appear to grow as a logarithmic function of sound pressure, a combination of a half-wave rectifier followed by a logarithmic compressor provides a reasonable model (Mountain and Hubbard in Hawkins et al., 1996).

Auditory Nerve

AN fibers, the cell bodies of which are located in the SG, are divided into two classes, depending on their morphology. Each IHC synapses with 10 to 30 type I AN (AN-I) fibers (Ryugo in Webster, Popper, and Fay, 1992). In most mammals, AN-I fibers synapse only with a single IHC. In contrast, type II fibers (AN-II), which innervate the OHCs, synapse with multiple hair cells. AN-I fibers exhibit spontaneous activity in the absence of sound, and they are often segregated into low (LSR), medium (MSR), and high (HSR) spontaneous rate categories. The pattern of this spontaneous activity is random and is usually modeled as a Poisson or dead-time modified Poisson process. Spontaneous rate tends to correlate with threshold, with HSR fibers being the most sensitive to sound stimuli.

The average firing rate of AN fibers in response to sustained tones is tuned, mimicking the responses at the BM. The peristimulus time histogram (PSTH) exhibits an initial rapid increase, followed by adaptation (Figure 2) to a lower steady-state rate (Ruggero in Popper and Fay, 1992). The steady-state response has only

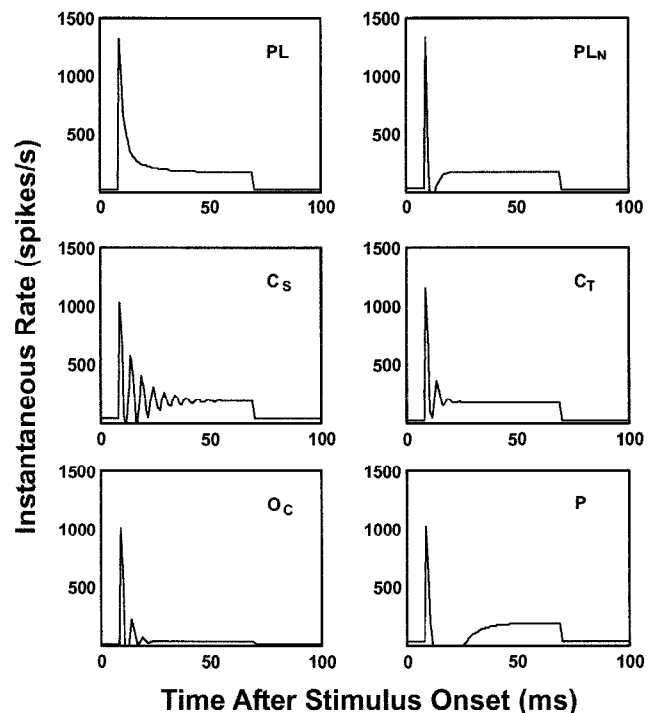


Figure 2. Typical auditory nerve and cochlear nucleus peristimulus time histograms. See text for explanation of abbreviations.

a limited dynamic range, typically saturating at sound levels of approximately 20 dB above the fiber's threshold. There are three components to the adaptation. The fastest component, rapid adaptation, has a time constant of a few milliseconds and creates an onset response with a large dynamic range. The second component, short-term adaptation, has a time constant of a few tens of milliseconds. It creates a slower component immediately after the onset response that has a smaller dynamic range, similar to that of the steady-state response. The third component of adaptation operates on a time scale of seconds and is not included in most auditory models.

On a finer time scale, the instantaneous firing rate (IFR) of AN fibers can be modulated on a cycle-by-cycle basis by the acoustic stimulus (phase locking) up to about 4 kHz (Ruggero in Popper and Fay, 1992). The fast dynamics of the AN IFR, coupled with only modest frequency resolution, suggests that we should think of the AN representation as that of a spectrogram that has been optimized more for temporal resolution than for spectral resolution. This excellent temporal resolution plays an important role in sound-source localization, which relies heavily on cues from interaural time delays.

Scant biophysical data are available for the IHC synapse, but since adaptation is not observed in the IHC receptor potentials, adaptation must be taking place in the IHC-AN synapse. The adaptation processes are most commonly assumed to be the result of synaptic vesicle depletion. Synaptic vesicles are typically divided into two or more pools. One of these pools represents vesicles that are docked at the active zones and is often referred to as the releasable pool or the immediate pool. Additional vesicles, which are located near the release sites but appear to be tethered to the cytoskeleton, are not available for immediate release (Mountain and Hubbard in Hawkins et al., 1996).

Cochlear Nucleus Anatomy

The two CN are the first and only brainstem structures to receive input from the AN. The CN can be anatomically subdivided into several subdivisions, each of which appears to perform a different physiological function. The major subdivisions are the ventral cochlear nucleus (VCN), which is further divided into anteroventral (AVCN) and posteroventral (PVCN) subdivisions, and the DC nucleus (DCN). Fibers of the AN travel through the core of the cochlear spiral and enter the AVCN, where they branch. The ascending branch innervates the AVCN and the descending branch travels through the PVCN and enters the DCN. The ventral regions are surrounded by the marginal shell, which is made up of the small-cell cap (SCC) and the granule-cell layer (GCL). Within a subdivision, the low-frequency fibers project to more ventral regions and the high-frequency fibers project to more dorsal regions. This orderly arrangement of characteristic frequencies is referred to as a *tonotopic projection*.

Cochlear Nucleus Response Types

The most commonly used physiological classification scheme in the CN is based on the PSTH. These histograms are derived by averaging the responses to short tone bursts presented at the cell's characteristic frequency (Rhode and Greenberg in Popper and Fay, 1992). Figure 2 illustrates six of the most common PSTH types found in the CN. The primary-like (PL) PSTHs are similar to the PSTHs recorded from AN fibers. The primary-like with notch (PL_N) response type is similar to the PL type but with better synchrony to the stimulus onset, followed by a transient dip in response due to refractory effects. The chopper-cell PSTHs exhibit periodically modulated activity at the beginning of the histogram, which is the result of the regular firing pattern of these cells becoming

synchronized to the stimulus onset. This chopping effect can either be sustained (C_S) or transient (C_T). The onset cell PSTHs all have large responses to the stimulus onset, followed by reduced or non-existent activity during the remainder of the stimulus. The onset chopper (O_C) PSTH exhibits a transient chopping response after the onset, whereas the O_L type (not shown) shows little or no response after the onset response. Other PSTH types include the pauser (P) and build-up (B) types. The P-type PSTH is characterized by an onset response followed by a period of no activity, which is then followed by a slow build-up of activity. The B-type (not shown) is similar to the P-type but lacks the initial onset response. The different PSTH types are believed to be the result, in part, of differences in intrinsic membrane properties and different degrees of AN fiber convergence and synaptic effectiveness.

Cochlear Nucleus Neural Circuits

Octopus Cells

Octopus cells are located in the PVCN and are characterized by long, thick primary dendrites that usually arise from one side of the cell body and give the cell the appearance of an octopus. The dendrites of octopus cells are oriented perpendicular to the path of incoming AN fibers (Oertel et al., 2000) which means that they receive input from a range of characteristic frequencies. The lower CF fibers synapse on the soma and the higher CF fibers synapse on the dendrites. Octopus cells generally exhibit onset (O_L) responses (Rhode and Greenberg in Popper and Fay, 1992) and appear to be more sensitive to broadband stimuli than AN fibers (Oertel et al., 2000). Functionally, octopus cells appear to act as coincidence detectors that detect synchronous events across AN fibers and may form part of networks involving other subthalamic nuclei devoted to processing temporal features such as duration, periodicity, and echo delay. The octopus cells project to contralateral VNLL terminating in calyx endings (Schwartz in Webster et al., 1992). The secure nature of these terminals reinforces the notion that octopus cells play a role in temporal processing. VNLL is primarily a monaural nucleus (Irvine in Popper and Fay, 1992) that provides inhibitory input to the IC (Schwartz in Webster et al., 1992). The octopus cells also provide diffuse innervation to periolivary areas of the SOC (Schwartz in Webster et al., 1992).

Stellate Cells

The stellate cells (SCs) of the VCN are hypothesized to be part of a system that uses a rate code to represent the acoustic spectrum (Rhode and Greenberg in Popper and Fay, 1992). Stellate cells have dendrites that extend away from the soma in all directions and often divide to form secondary and tertiary dendrites. The SCs can be divided into two classes, based on the path taken by their axons. The T-stellate cells project out of the CN by way of the trapezoid body (hence the name T), and the D-stellate cells are interneurons with axons that follow a descending path (hence the name D) on their way to the DCN and contralateral CN. The dendrites of T-stellate cells (also called planar cells) end in tufts and are generally aligned with the isofrequency plane created by the path of AN fibers, whereas those of D-stellate cells (also called radiate cells) extend radially across the isofrequency planes and branch sparingly. Both T- and D-stellate cells have terminal collaterals in the multipolar cell region of the PVCN and in the DCN (Oertel et al., 1990; Doucet and Ryugo, 1997).

D-stellate cells exhibit O_C responses (Figure 2), have a large dynamic range (80 dB or more), and, as would be expected from their dendritic morphology, are more broadly tuned than AN fibers. In contrast, T-stellate cells exhibit C_T responses (Figure 2) and have frequency tuning characteristics similar to those of AN fibers. Stel-

Figure 3. Examples of neural circuits in the cochlear nucleus. Excitatory connections are indicated by solid lines and inhibitory connections are indicated by broken lines. *A*, The stellate cell circuit. *B*, The spherical bushy cell circuit. *C*, The fusiform circuit. See text for explanation of abbreviations.

Fusiform and Giant Cells

The DCN is believed to be involved in processing spectral cues that are important for sound source location, especially source elevation. Cats with lesions to the DCN output pathways exhibit significant deficits in their ability to orient to sources at different locations (Young and Davis in Oertel et al., 2002). The DCN is usually subdivided into three layers, a superficial or molecular layer, an intermediate layer called the granular or fusiform cell layer, and a polymorphic or deep layer. The principal cells of the DCN are the fusiform cells (from which the fusiform cell layer gets its name) and the giant cells located in the deep layer. The apical dendrites of fusiform cells (also called pyramidal cells) are highly branched and extend up into the molecular layer, while the less highly branched basal dendrites extend down into the deep layer. The dendritic morphology of the giant cells is more diverse, ranging from elongate to radiate (Cant in Webster et al., 1992). Beneath the fusiform cells are a group of cells called vertical cells. These cells have their dendritic and axonal arbors confined to an isofrequency lamina. There are two groups of vertical cells. The more superficial group gives rise to only a local axon, the deeper group gives rise to axons that project to the VCN (Rhode, 1999).

The fusiform cell circuit is shown in Figure 3C. The basal dendrites of fusiform cells receive excitatory input from AN-I fibers with a limited range of CFs, a narrow-band inhibitory input from the vertical cells of the DCN, and a wide-band inhibitory input from the D-stellate cells of the VCN. The inhibitory input from the vertical cells is quite strong, and as a result, fusiform cells respond poorly or not at all to pure tones. They respond well to broadband stimuli except when there is a spectral notch at the characteristic frequency, in which case these cells are strongly inhibited (Rhode and Greenberg in Popper and Fay, 1992). As a result of these properties, fusiform cell models create a spectral representation that accentuates spectral notches (Hancock and Voigt, 1999), which are important features of the HRTF for determining sound source elevation. The fusiform cells also receive excitatory input on their apical dendrites from the granule cells, which in turn receive input from the dorsal column and spinal trigeminal nuclei of the somatosensory system. This somatosensory input appears to modify DCN response properties based on head and pinna position (Kanold and Young, 2001).

The giant cell circuit (not shown) is similar to the fusiform cell circuit except that the giant cells do not receive direct input from granule cells and AN input to the giant cells spans a large range of characteristic frequencies. Fusiform and giant cells project to the contralateral ICC and also project to the contralateral DNLL, which provides inhibitory input to both ICs (Oliver and Huerta in Webster et al., 1992).

Discussion

Significant progress has been made in understanding the anatomy and physiology of the subthalamic auditory pathways, but many questions remain. For example, the experimental data suggest that OHCs contribute to the tuned response of the BM and IHCs, but how OHCs perform their function is not well understood. Much of the basic circuitry of the CN has been worked out, but it is not clear how many subpopulations exist for each of the basic cell types

described in this article. And perhaps the greatest question of all is, how is information in the parallel pathways leaving the CN reintegrated into a unified percept by higher centers? To answer these questions, future research will need to take an integrated approach, with computational models being used to aid the design and interpretation of anatomical and physiological experiments. These models will need to incorporate the major features of individual cell types as well as the interactions between cell types at different levels of the auditory system. Efforts to create suitable large-scale models have begun (cf. Hawkins et al., 1996), but much remains to be done.

Road Maps: Mammalian Brain Regions; Other Sensory Systems

Related Reading: Auditory Cortex; Auditory Scene Analysis; Echolocation: Cochleotopic and Computational Maps; Sound Localization and Binaural Processing; Thalamus

References

- Dallos, P., Popper, A. N., and Fay, R.-R., 1996, *The Springer Handbook of Auditory Research*, vol. 8, *The Cochlea*, New York: Springer-Verlag. ♦
- Doucet, J. R., and Ryugo, D. K., 1997, Projections from the ventral cochlear nucleus to the dorsal cochlear nucleus in rats, *J. Comp. Neurol.*, 385:245–264.
- Ferragamo, M. J., Golding, N. L., and Oertel, D., 1998, Synaptic inputs to stellate cells in the ventral cochlear nucleus, *J. Neurophysiol.*, 79:51–63.
- Gilkey, R. H., and Anderson, T. R., 1997, *Binaural and Spatial Hearing in Real and Virtual Environments*, Mahwah, NJ: Erlbaum.
- Hancock, K. E., and Voigt, H. F., 1999, Wideband inhibition of dorsal cochlear nucleus type IV units in cat: A computational mode, *Ann. Biomed. Eng.*, 27:73–87.
- Hawkins, H. L., McMullen, T. A., Popper, A. N., and Fay, R.-R., 1996, *The Springer Handbook of Auditory Research*, vol. 6, *Auditory Computation*, New York: Springer-Verlag. ♦
- Kanold, P. O., and Young, E. D., 2001, Proprioceptive information from the pinna provides somatosensory input to cat dorsal cochlear nucleus, *J. Neurosci.*, 21:7848–7858.
- Mountain, D. C., and Cody, A. R., 1999, Multiple modes of inner hair cell stimulation, *Hear. Res.*, 132:1–14.
- Oertel, D., Bal, R., Gardner, S. M., Smith, P. H., and Joris, P. X., 2000, Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus, *Proc. Natl. Acad. Sci. USA*, 97:11773–11779.
- Oertel, D., Fay, R. R., and Popper, A. N., 2002, *The Springer Handbook of Auditory Research*, vol. 15, *Integrative Functions in the Mammalian Auditory Pathway*, New York: Springer-Verlag. ♦
- Oertel, D., Wu, S. H., Garb, M. W., and Dizack, C., 1990, Morphology and physiology of cells in slice preparations of the posteroventral cochlear nucleus of mice, *J. Comp. Neurol.*, 295:136–154.
- Popper, A. N., and Fay, R.-R., 1992, *The Springer Handbook of Auditory Research*, vol. 2, *The Mammalian Auditory Pathway: Neurophysiology*, New York: Springer-Verlag. ♦
- Rhode, W. S., 1999, Vertical cell responses to sound in cat dorsal cochlear nucleus, *J. Neurophys.*, 82:1019–1032.
- Webster, D. B., Popper, A. N., and Fay, R.-R., 1992, *The Springer Handbook of Auditory Research*, vol. 1, *The Mammalian Auditory Pathway: Neuroanatomy*, New York: Springer-Verlag.
- Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H., 2001, A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression, *J. Acoust. Soc. Am.*, 109:648–670.

Auditory Scene Analysis

Guy J. Brown

Introduction

We usually listen in environments that contain many simultaneously active sound sources. The auditory system must therefore parse the acoustic mixture that reaches our ears to segregate a target sound source from the background of other sounds. Bregman (1990) describes this process as *auditory scene analysis* (ASA) and suggests that it takes place in two conceptual stages. The first stage, *segmentation*, decomposes the acoustic mixture into its constituent components. In the second stage, acoustic components that are likely to have arisen from the same environmental event are *grouped*, forming a perceptual representation (*stream*) that describes a single sound source. Streams are subjected to higher-level processing, such as language understanding.

Bregman's account makes a distinction between *schema-driven* and *primitive* mechanisms of grouping. Schema-driven grouping applies learned knowledge of sound sources such as speech in a top-down manner (in this regard, the term "schema" refers to a recurring pattern in the acoustic environment). Primitive mechanisms operate on the acoustic signal in a bottom-up fashion and are well described by Gestalt heuristics such as proximity and common fate (see CONTOUR AND SURFACE PERCEPTION). Primitive organization is both *simultaneous* and *sequential*. Simultaneous grouping operates on concurrent sounds, using principles such as similarity of fundamental frequency. Sequential grouping combines acoustic events over time, according to heuristics such as temporal proximity and frequency proximity.

At the physiological level, segmentation corresponds (at least in part) to peripheral auditory processing, which performs a frequency analysis of the acoustic input. To a first approximation, this frequency analysis can be modeled by a bank of band-pass filters with overlapping passbands, in which each channel simulates the filtering characteristics of one location on the basilar membrane (see AUDITORY PERIPHERY AND COCHLEAR NUCLEUS). From the output of each filter, a simulation of the auditory nerve response can be obtained by rectification and compression or from a detailed model of inner hair cell function. In contrast, the physiological substrate of auditory grouping is much less well understood (Feng and Ratnam, 2000). As a result, models of ASA tend to be functional in approach. In the current review, we focus on models that are at least physiologically plausible; however, it should be noted that there is also a substantial literature that has addressed computational modeling of ASA from a more abstract information-processing perspective (e.g., Rosenthal and Okuno, 1998).

Models of Sequential Grouping

Sequential grouping can be demonstrated by playing listeners a repeating sequence of two alternating tones with different frequencies (ABAB . . .). When the sequence is played rapidly or when the frequency separation between the tones is large, the sequence is heard to split into separate streams (A-A- . . . and B-B- . . .). This phenomenon is known as *auditory streaming* (Bregman, 1990). Listeners are able to direct their attention to only one of the streams, which appears to be subjectively louder than the other. Auditory streaming may therefore be regarded as an example of figure-ground separation.

Auditory streaming may be viewed as a consequence of sequential grouping heuristics that allocate tones to streams depending on their proximity in time and frequency. Several modeling studies have demonstrated that such principles can be implemented by relatively low-level physiological mechanisms. For example, Beauvois

and Meddis (1996) describe a model of auditory streaming that has its basis in mechanisms of peripheral auditory function. The model utilizes two pathways: one in which auditory nerve activity is smoothed by temporal integration and an "excitation-level" path that adds a cumulative random element to the output of the temporal integration path. Firing activity is considered in three auditory filter channels: one at the frequency of each tone and one in between them. The channel with the highest activity in the excitation-level pathway is selected as the dominant "foreground" percept; the remaining channels are attenuated and become the "background." This simple model quantitatively matches auditory streaming phenomena, such as the effect of rate of presentation and frequency separation. Furthermore, the inclusion of a random element in the model (which is assumed to originate from the stochastic nature of auditory nerve firing patterns) explains how spontaneous shifts of attention can occur.

McCabe and Denham (1997) have extended the Beauvois and Meddis model by applying similar principles within a two-layer neural architecture. In their model, "foreground" and "background" neural arrays are connected by reciprocal inhibitory connections, which ensure that activity appearing in one array does not appear in the other (Figure 1). Their network is sensitive to frequency proximity because the strength of inhibitory feedback is related to the frequency difference between acoustic components. Additionally, each layer receives a recurrent inhibitory feedback related to the reciprocal of its own activity. As a result, previous activity in the network tends to suppress differences in subsequent stimuli. This mechanism may be viewed as a neural implementation of Bregman's (1990) "old plus new heuristic," which states that the auditory system prefers to interpret a current sound as a continuation of a previous sound unless there is strong evidence to the contrary. The inclusion of this heuristic within McCabe and Denham's model allows it to explain the effect of background organization on the perceptual foreground.

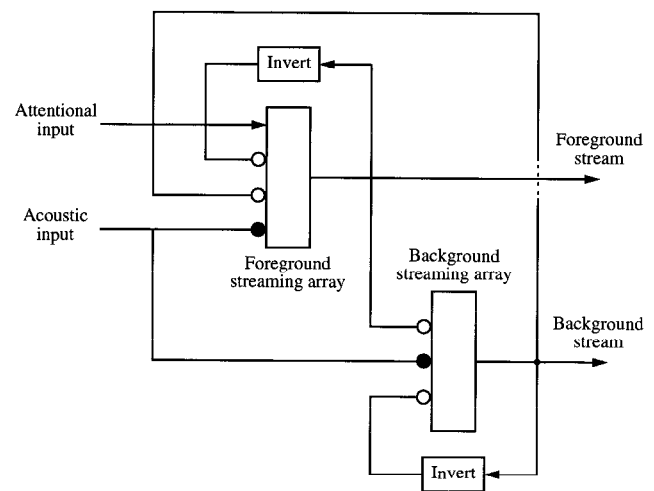


Figure 1. McCabe and Denham's model of auditory stream segregation. Foreground and background streams are encoded by separate neural arrays, which have reciprocal inhibitory connections. Each layer also receives recurrent inhibition. Solid circles indicate excitatory connections; open circles indicate inhibitory connections. (Modified from McCabe and Denham (1997).)

A more central explanation for auditory streaming is given by Todd (1996), who suggests that mechanisms of rhythm perception and stream segregation are underlain by cortical maps of periodicity-sensitive cells. In his model, periodicity detection leads to a spatial representation of the temporal pattern of the stimulus in terms of its amplitude modulation (AM) spectrum. Acoustic events whose AM spectra are highly correlated (as judged by a neural cross-correlation mechanism) are perceptually grouped, whereas events with uncorrelated AM spectra are segregated. Todd's model is able to qualitatively replicate the dependence of auditory streaming on tone frequency and temporal proximity.

The models of Todd (1996) and McCabe and Denham (1997) suggest that the auditory responses associated with different streams are encoded spatially in neural arrays. An alternative is that auditory streams are encoded temporally. For example, Wang (1996) suggests a principle of *oscillatory correlation*, which is a development of von der Malsburg's temporal correlation theory (von der Malsburg and Schneider, 1986). In Wang's scheme, neural oscillators alternate rapidly between relatively stable states of activity (the active phase) and inactivity (the silent phase). Oscillators that encode features of the same stream are synchronized (phase locked with zero phase lag) and are desynchronized from oscillators that represent different streams.

Wang has implemented the oscillatory correlation principle in a model of auditory streaming, in which oscillators are arranged within a two-dimensional time-frequency network. The time axis of the network is assumed to be constructed by a systematic arrangement of neural delay lines. Each oscillator is connected to others in its neighborhood with excitatory connections whose strength diminishes with increasing distance in time and frequency. In addition, every oscillator sends excitation to a global inhibitor, which feeds back inhibition to each oscillator in the network. Oscillators that are close in time and frequency tend to synchronize because the excitatory connections between them are strong. However, groups of oscillators that do not receive mutually supportive excitation tend to desynchronize because of the action of the global inhibitor. As the network dynamics evolve, the combined effects of local excitation and global inhibition cause streams of synchronized oscillators to form. The model qualitatively reproduces a number of auditory streaming phenomena. However, the oscillatory dynamics proceed rapidly, so Wang's network is not able to account for the gradual buildup of the auditory streaming effect over time.

Models of Simultaneous Grouping

Simultaneous grouping mechanisms exploit differences in the characteristics of concurrent sounds in order to perceptually segregate them. For example, the ability of listeners to identify two simultaneously presented vowels ("double vowels") can be improved by introducing a difference in fundamental frequency (F0) between the vowels (Bregman, 1990). Apparently, simultaneous grouping mechanisms are able to segregate the acoustic components related to each F0 and hence retrieve the spectra of the two vowels.

Meddis and Hewitt (1992) describe a model of double-vowel identification based on the *correlogram*, a model of auditory pitch analysis. A correlogram is formed by computing a running auto-correlation at the output of each auditory filter channel, giving a two-dimensional representation in which frequency and time lag are represented on orthogonal axes. Meddis and Hewitt suggest that the correlogram could be computed neurally by using a system of delay lines and coincidence detectors. The F0 of one of the vowels is identified from the correlogram, and channels whose response is dominated by that F0 are removed, thus allowing a clearer view of the second vowel. This mechanism fails to separate the vowels when they both have the same F0 and successfully predicts that

vowel identification performance improves when a difference in F0 is introduced.

The Meddis and Hewitt model is based on a strategy of "exclusive allocation" (Bregman, 1990); all of the energy in a single auditory filter channel is allocated to one vowel or the other. However, this need not be the case. De Cheveigné (1997) describes an approach that uses a "neural cancellation filter" to partition the energy in each channel between vowel percepts. In his scheme, a correlogram is computed, and the fundamental period of one of the vowels is identified. This period is canceled in each channel by a neural comb filter, which is implemented by a neuron with a delayed inhibitory input. This mechanism removes firing activity with a periodicity equal to the inhibitory delay. An advantage of de Cheveigné's approach is that it predicts an increase in listeners' performance with increasing difference in F0 for vowels that are weak in comparison to a harmonic background. The Meddis and Hewitt model is unable to reproduce this result because its exclusive allocation scheme tends to remove the evidence for a weak vowel when the stronger vowel is canceled.

Brown and Wang (1997) have described a neural oscillator model of vowel segregation, which is essentially an implementation of Meddis and Hewitt's (1992) scheme within an oscillatory correlation framework. In their model, each channel of the correlogram is associated with a neural oscillator. Oscillators corresponding to channels that are dominated by the same F0 become synchronized and are desynchronized from channels that are dominated by a different F0. When there is no difference in F0 between the two vowels, a single group of synchronized oscillators forms. However, when a difference in F0 is introduced, the two vowels are segregated according to their F0s, and the channels making up each vowel are encoded as separate groups of synchronized oscillators.

Von der Malsburg and Schneider (1986) describe a related model of simultaneous grouping based on temporal correlation of neural responses. Their scheme employs a neural architecture in which each member of a fully connected network of excitatory cells (E-cells) receives an input from one auditory filter channel. In addition, E-cells receive inhibition from a common inhibitory cell (H-cell). E-cells that receive simultaneous inputs tend to become synchronized by the excitatory links between them and tend to become desynchronized from other cells owing to the influence of inhibition from the H-cell. The network therefore displays a sensitivity to the common onset of acoustic components and may be regarded as implementing a Gestalt principle of common fate (Bregman, 1990).

The Role of Temporal Continuity

With the exception of that of Wang (1996), relatively few modeling studies have demonstrated the integration of simultaneous and sequential grouping principles within the same computational framework. However, several studies have shown how a complex time-frequency mixture can be organized using simultaneous grouping principles and temporal continuity constraints.

Grossberg (1999) describes a multistage model of ASA that implements grouping by common F0 and good continuation. The first stage of his model builds redundant spectral representations of the acoustic input in a "spectral stream" layer. Each stream is represented by a separate neural array. These representations are filtered by neural "harmonic sieves," which connect a node in a "pitch stream" layer with spectral regions near to the harmonics of the corresponding pitch value. Pitch representations compete across streams to select a winner, and the winning pitch node sends top-down signals via harmonic connections to the spectral stream layer. According to an adaptive resonance theory (ART) matching rule, frequency components in the spectral stream that are consistent with the top-down signal are selected, and others are suppressed (see ADAPTIVE RESONANCE THEORY). Selected components reac-

tivate their pitch node, and further top-down signals are produced. In this way, a resonance develops that binds together the frequency components constituting a sound source and its corresponding pitch.

Grossberg's model is able to account for simple simultaneous grouping phenomena, such as the perceptual fusion of components with the same F0. His model also reproduces the auditory continuity illusion (Bregman, 1990), in which a pure tone is heard to continue through a brief interrupting noise, even though the tone is not physically present during the noise burst. It is able to do so because a resonance develops for the tone that is maintained during the noise burst. The ART matching rule then selects the tone from the noise, and competitive interactions cause the tone and residual noise to be allocated to different streams.

Wang and Brown (1999) describe a neural oscillator model whose two-layer architecture echoes the two conceptual stages of ASA. The first (segmentation) layer consists of a two-dimensional time-frequency grid of oscillators with a global inhibitor (Figure 2). In this layer, excitatory connections are formed between auditory filter channels that have a similar temporal response. As a result, segments form in the time-frequency plane and thus correspond to harmonics and formants. The global inhibitor ensures that each segment desynchronizes from the others; the first layer therefore embodies the segmentation stage of ASA, in which the acoustic signal is split into its constituent elements. The second layer receives an input from the first layer. Also, segments in the second layer are connected by excitatory links if they represent time-frequency regions that are dominated by the same F0. As a result, synchronized groups of segments emerge in the second layer, each of which corresponds to a stream with harmonically related components.

Models of Schema-Driven Grouping

Liu, Yamaguchi, and Shimizu (1994) describe a neural oscillator model of vowel recognition that may be regarded as an implementation of schema-driven grouping. The model consists of an input layer and three layers of oscillators labeled A, B, and C, which are likened to regions of the auditory cortex. The A ("feature extraction") layer identifies local peaks in the acoustic spectrum and en-

codes them as separate groups of oscillations, which are assumed to correspond to vowel formants. The B ("feature linking") layer acts as a simple associative memory, in which hardwired connections encode the relationship between formant frequencies for different vowels. Associative interactions between the B layer, together with top-down and bottom-up interactions between the A and B layers, lead to the activation of a vowel in terms of a global pattern of synchronized oscillations. The C ("evaluation") layer assesses the synchronization in each formant region and outputs a vowel category. Top-down reinforcement from the B center confers robustness in noise; it is demonstrated that the model is able to recognize vowels robustly in the presence of multispeaker babble.

Discussion

The modeling studies reviewed here propose a neurobiological basis for the principles of auditory organization expounded in Bregman's account of ASA. Clearly, the models differ in their level of explanation, ranging from peripheral (Beauvois and Meddis, 1996) to cortical (Todd, 1996). Without exception, their approach is functional; currently, there is insufficient knowledge about the physiological mechanisms of ASA to attempt a detailed physiological model. It is likely that future research in this field will see a closer synergy between computational modeling studies and neurophysiological investigation.

Various strategies have been proposed for the neural encoding of auditory streams. Beauvois and Meddis (1996) stress that the perceptual separation of sounds need not imply a physical separation of their corresponding representations; in their model, auditory filter channels belonging to a nonattended stream are simply attenuated. This contrasts with the approaches described by Grossberg (1999) and McCabe and Denham (1997), in which different auditory streams are encoded by separate neural arrays. A further approach is to encode streams temporally in the responses of synchronized neural firing patterns (von der Malsburg and Schneider, 1986; Wang, 1996). Although all of these approaches are plausible, none are currently supported by strong neurophysiological evidence.

Many models of ASA require systematic time delays longer than those currently known to exist in the auditory system. Models of double-vowel separation based on the correlogram require delays of the order of 20 ms (Meddis and Hewitt, 1992; de Cheveigné, 1997). Similarly, it is questionable whether the system of delay lines employed in Wang's (1996) neural oscillator model is physiologically realizable. The temporal correlation architecture of von der Malsburg and Schneider (1986) does not suffer from this difficulty, since their network does not have an explicit time axis. However, the explanatory power of their model is weak in comparison to Wang's model, because temporal and frequency relationships between acoustic inputs are not preserved.

Generally speaking, the role of auditory attention has been neglected in computer models of ASA. In auditory streaming, a listener's attention can shift randomly between organizations or may be consciously directed to the high or low tones. The model of Beauvois and Meddis (1996) accounts for the former but not the latter. Similarly, McCabe and Denham's model includes an attentional input (Figure 1), but it is not utilized in their simulations. Wang (1996) suggests that in a neural oscillator framework, attention is paid to a stream when its constituent oscillators reach their active phases; attention therefore alternates quickly among the streams in turn. However, such a scheme does not explain how listeners are able to direct their attention to a particular stream over a sustained period of time.

Also, few models have attempted to model the interaction between top-down and bottom-up grouping mechanisms. In principle, the mechanism of recurrent neural connections described by Liu et al. (1994) could form the basis for such a model. Similarly, Gross-

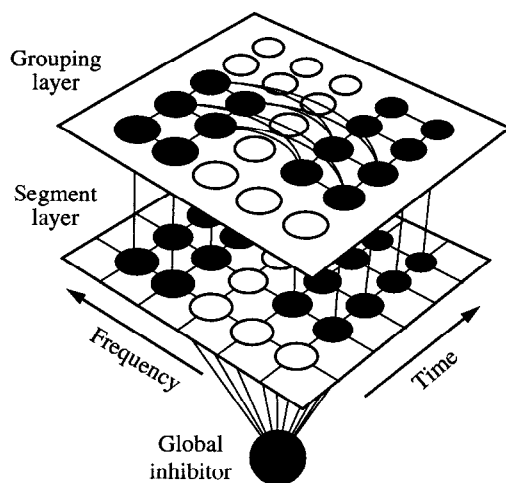


Figure 2. The two-layer neural oscillator model of Wang and Brown (1999). In the first layer, segments are formed that correspond to harmonics and formants. The second layer groups segments according to their fundamental frequency (F0); those with the same F0 form a stream in which all oscillators are synchronized and are desynchronized from other streams. (Modified from Wang and Brown (1999).)

berg's (1999) ART scheme could form the basis for a grouping mechanism in which bottom-up features interact with top-down information about the characteristics of sound sources.

The motivation for most of the studies reviewed here is to gain insight into the mechanisms of ASA through computational modeling. However, computer sound separation devices have many real-world applications, such as in hearing prostheses and as pre-processors for robust automatic speech recognition in noise. For example, Wang and Brown (1999) have applied their model to the separation of voiced speech from interfering sounds, with some success. Because they are founded on neurobiological principles, such approaches to sound separation may offer performance advantages over other techniques, such as blind statistical methods.

Road Map: Other Sensory Systems

Related Reading: Auditory Periphery and Cochlear Nucleus; Contour and Surface Perception; Dynamic Link Architecture; Echolocation: Cochleotopic and Computational Maps

References

- Beauvois, M. W., and Meddis, R., 1996, Computer simulation of auditory stream segregation in alternating-tone sequences, *J. Acoust. Soc. Am.*, 99:2270–2280.
- Bregman, A. S., 1990, *Auditory Scene Analysis*, Cambridge, MA: MIT Press. ♦
- Brown, G. J., and Wang, D., 1997, Modelling the perceptual segregation of double vowels with a network of neural oscillators, *Neural Networks*, 10:1547–1558.
- de Cheveigné, A., 1997, Concurrent vowel identification: III. A neural model of harmonic interference cancellation, *J. Acoust. Soc. Am.*, 101:2857–2865.
- Feng, A. S., and Ratnam, R., 2000, Neural basis of hearing in real-world situations, *Annu. Rev. Psychol.*, 51:699–725. ♦
- Grossberg, S., 1999, Pitch-based streaming in auditory perception, in *Musical Networks: Parallel Distributed Perception and Performance* (N. Griffith and P. Todd, Eds.), Cambridge, MA: MIT Press, pp. 117–140.
- Liu, F., Yamaguchi, Y., and Shimizu, H., 1994, Flexible vowel recognition by the generation of dynamic coherence in oscillator neural networks: Speaker-independent vowel recognition, *Biol. Cybern.*, 71:105–114.
- McCabe, S. L., and Denham, M. J., 1997, A model of auditory streaming, *J. Acoust. Soc. Am.*, 101:1611–1621.
- Meddis, R., and Hewitt, M. J., 1992, Modelling the identification of concurrent vowels with different fundamental frequencies, *J. Acoust. Soc. Am.*, 91:233–245.
- Rosenthal, D., and Okuno, H. G. (Eds.), 1998, *Computational Auditory Scene Analysis*, Mahwah, NJ: Lawrence Erlbaum Associates. ♦
- Todd, N., 1996, An auditory cortical theory of primitive auditory grouping, *Network: Comput. Neural Syst.*, 7:349–356.
- von der Malsburg, C., and Schneider, W., 1986, A neural cocktail-party processor, *Biol. Cybern.*, 54:29–40.
- Wang, D., 1996, Primitive auditory segregation based on oscillatory correlation, *Cogn. Sci.*, 20:409–456. ♦
- Wang, D., and Brown, G. J., 1999, Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Trans. Neural Networks*, 10:684–697.

Axonal Modeling

Christof Koch and Öjvind Bernander

Introduction

Axons are highly specialized “wires” that conduct the neuron's output signal to target cells—in the case of cortical pyramidal cells up to 10,000 other cortical neurons. As such they are highly specialized, with a relatively stereotypical behavior. Most authors agree that their role in signaling is largely limited to making sure that whatever pulse train is put into one end of the axon is rapidly and faithfully propagated to the other end. This is in contrast to the complexity of electrical events occurring at the cell body and in the dendritic tree, where the information from thousands of synapses is integrated.

Despite the uniformity of electrical behavior, there is great morphological variability that largely reflects a trade-off between propagation speed and packing density. Axonal size varies over four orders of magnitude: diameters range from 0.2- μ m fibers in the mammalian central nervous system to 1 mm in squid; lengths range from a few hundred microns to over a meter for motor neurons (Kandel, Schwartz, and Jessell, 2000). Some axons are bound only by the thin cellular membrane, while others are wrapped in multiple sheaths of myelin. An example of an axonal arbor is shown in Figure 1.

The majority of nerve cells encode their output as a series of brief voltage pulses. These pulses, also referred to as *action potentials* or *spikes*, originate at or close to the cell body of nerve cells and propagate down the axon at constant amplitude. Their shape is relatively constant across species and types of neurons. Common to all is the rapid upstroke (depolarization) of the membrane above 0 mV and the subsequent, somewhat slower, downstroke (repolarization) toward the resting potential and slightly beyond to more hyperpolarized potentials. At normal temperatures, the entire sequence occurs within less than 1 ms. A minority of cell types are axonless and appear to use graded voltage as output, such as cells

in the early part of the retina or interneurons in invertebrates (Roberts and Bush, 1981). Action potentials are such a dominant feature of the nervous system that for a considerable period of time it was widely held—and still is in parts of the neural network community—that all neuronal computations involve only these all-or-none events. This belief provided much of the impetus behind the neural network models originating in the late 1930s and early 1940s (see SINGLE-CELL MODELS).

The ionic mechanisms underlying the initiation and propagation of action potentials were elucidated in the squid giant axon by a number of workers, most notably Hodgkin and Huxley (1952). Today, with the widespread availability of cheap and almost unlimited computational power, it is very difficult to imagine the difficulty that Hodgkin and Huxley faced 50 years ago. Not only did they have to derive a proper mathematical formalism based on incomplete data, they also had to solve a nonlinear partial differential equation, the cable equation, using a very primitive hand calculator.

For this work they shared, together with Eccles, the 1963 Nobel prize in physiology and medicine (for a historical overview, see Hodgkin, 1976). Their model has played a paradigmatic role in biophysics; indeed, the vast majority of contemporary biophysical models use essentially the same mathematical formalism Hodgkin and Huxley introduced 50 years ago. This is all the more surprising because the kinetic description of the continuous, deterministic, and macroscopic membrane permeability changes within the framework of the Hodgkin-Huxley model was achieved without any knowledge of the underlying all-or-none, stochastic, and microscopic ionic channels.

Given its importance, we will describe the Hodgkin-Huxley model and its assumptions in some detail in the following section. We then introduce the two classes of axons found in most animals, myelinated and nonmyelinated, and describe their differences. Ax-

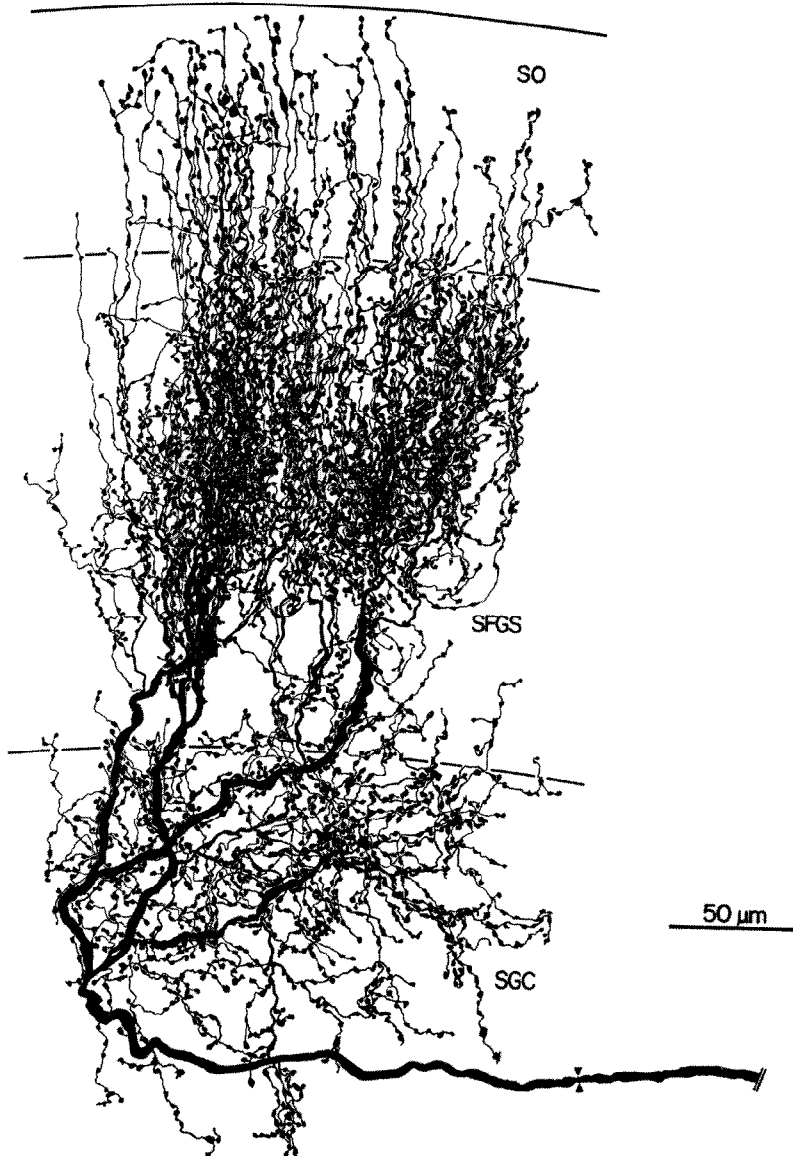


Figure 1. Axonal terminations. An axon from nucleus isthmi terminating in turtle tectum was labeled with horseradish peroxidase and reconstructed from a series of parallel sections. The thick parent trunk ($3\ \mu\text{m}$) is wrapped in myelin and shows a node of Ranvier (triangles). The thin ($<1\ \mu\text{m}$) branches are nonmyelinated and are home to approximately 3,600 synaptic boutons (bulbous thickenings), where contact is made onto other cells. The boutons vary greatly in size but average about $1.5\ \mu\text{m}$ in diameter. (From Sereno, M. I., and Ulinski, P. S., 1987, Caudal topographic nucleus isthmi and the rostral nontopographic nucleus isthmi in the turtle, *Pseudemys scripta*, *J. Comp. Neurol.*, 261:319–346. Copyright © 1987 by Wiley-Liss. Reprinted by permission of John Wiley & Sons, Inc.)

ons possess heavily branched axonal trees. We conclude the overview by briefly alluding to additional complications that arise when attempting to understand the role and function of axonal trees in information processing. For a useful book on the biophysics of dendrites and axons and their computational function, see Koch (1999). For a monograph on the axon in health and disease, see Waxman, Kocsis, and Stys (1995).

The Hodgkin-Huxley Model of Action Potential Generation

Electrical current in nerve cells is carried by the flow of ions through membrane proteins called *channels*. The concentration of sodium ions is high in the extracellular fluid and low in the intracellular axoplasm. This *concentration gradient* gives rise to a tendency for sodium ions to flow into the cell. At some membrane potential, termed the *reversal potential*, the effect of the concentration gradient will be canceled by the *electrical gradient*, and so the net flow of sodium ions will be zero at that point. The channel

transitions into its closed state by virtue of a conformational state in the underlying molecular structure. In the model, this is described by a change in the m variable. A similar situation holds for potassium ions flowing through separate potassium-selective channels, except that the concentration gradient is reversed.

In the squid giant axon, the membrane potential is determined by three conductances: a voltage-independent (passive) leak conductance, g_l , a voltage-dependent (active) sodium conductance, g_{Na} , and an active potassium conductance, g_{K} . The equivalent circuit used to model the membrane is shown in Figure 2. The conductances are in series with batteries, the values of which correspond to the respective reversal potentials of the ionic currents, E_l , E_{Na} , and E_{K} . The outside is connected to ground under the assumption that the resistivity of the external medium is negligible.

The time course of an action potential is illustrated in Figure 3. In this simulation of a membrane patch, a brief current pulse initiates an action potential. Before stimulation, the membrane voltage is at rest, $V_m = -65\ \text{mV}$. At this potential, g_{Na} and g_{K} are almost fully inactivated. g_{K} is still much larger than g_{Na} , and the membrane

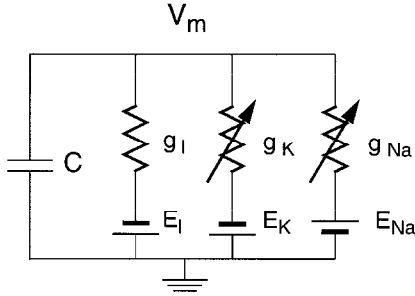


Figure 2. Schematic of ionic channel and neuronal membrane: Equivalent circuit of axonal membrane. The Hodgkin-Huxley model of squid axon incorporates a capacitance and three conductances. Two of the conductances are voltage dependent (active), g_{Na} and g_K , while the third is a passive “leak” conductance, g_l . The maximal conductances are 120, 36, and 0.3 ms/cm², respectively. Each conductance is in series with a battery that defines the *reversal potential* for each conductance type. The values are $E_{Na} = 50$, $E_K = -77$, and $E_l = -54.3$ mV. See text for the voltage dependences of g_{Na} and g_K . The top rail corresponds to the axoplasm (inside) of the axon, while the bottom rail, grounded, is the external medium. When a *membrane action potential* or *space clamp* is modeled, only one compartment is used, as shown, and the spatial structure of the membrane is ignored. When *propagating action potentials* are modeled, the specific resistivity of the axoplasm, $R_a = 34.5 \Omega\text{cm}$, cannot be ignored. R_a is then modeled as a series of resistors connecting identical compartments that correspond to different spatial locations along the axon. The membrane capacitance is $1 \mu\text{m F/cm}^2$.

is dominated by the leak current and the residual potassium current. The applied current slowly depolarizes the membrane by charging up the capacitance. As V_m approaches threshold ($V_t \sim -50$ mV), sodium channels begin to open up, allowing for the influx of Na^+ ions, which further depolarizes the membrane. About 1 ms later, two events occur to bring the voltage back toward and somewhat beyond the resting value: the sodium conductance inactivates, that is, the sodium channels slowly close again, and potassium channels open up, causing an outward current to flow. This outward current forces the membrane potential below the resting value of -65 mV (hyperpolarization), but the K^+ conductance too eventually deactivates, allowing g_l to pull V_m back to rest.

Mathematical Formulation

The equation describing the circuit in Figure 2 is:

$$C \frac{dV_m}{dt} = g_l(E_l - V_m) + g_{Na}(E_{Na} - V_m) + g_K(E_K - V_m) \quad (1)$$

While g_l is constant, g_{Na} and g_K are time and voltage dependent:

$$g_{Na} = \bar{G}_{Na} \cdot m(t)^3 h(t)$$

$$g_K = \bar{G}_K \cdot n(t)^4$$

where the constants \bar{G}_{Na} and \bar{G}_K are the maximal conductances and the time and voltage dependence reside in the so-called *gating variables*, described by the state variables m , h , and n . These fictitious variables follow first-order kinetics, relaxing exponentially toward a steady-state value x_∞ with a time constant τ_x :

$$\frac{dm}{dt} = \frac{m_\infty(V_m) - m}{\tau_m(V_m)}$$

$$\frac{dh}{dt} = \frac{h_\infty(V_m) - h}{\tau_h(V_m)}$$

$$\frac{dn}{dt} = \frac{n_\infty(V_m) - n}{\tau_n(V_m)}$$

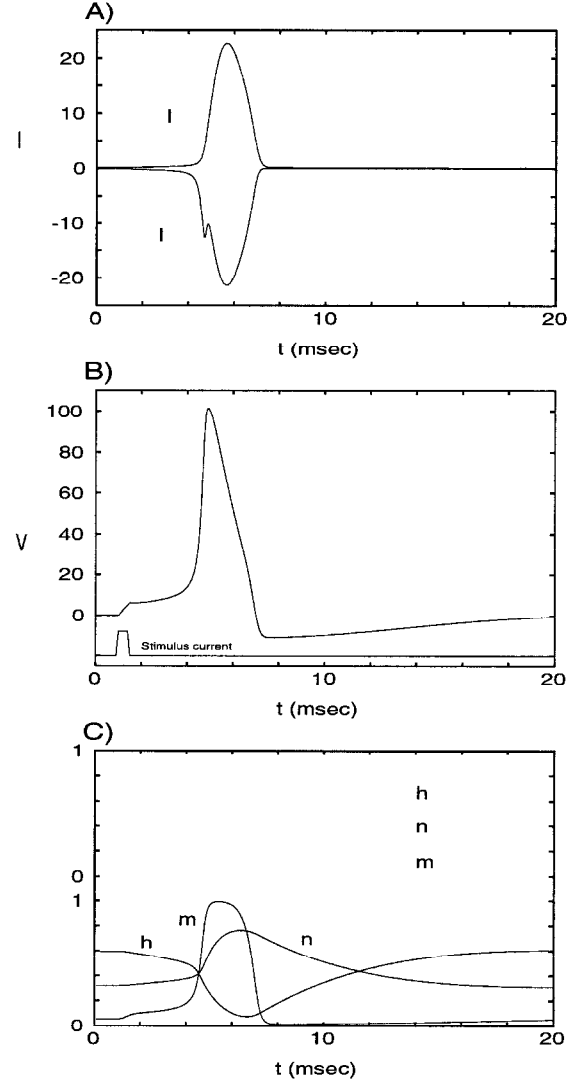


Figure 3. Action potential. Computed action potential in response to a 0.5-ms current pulse of 0.4-nA amplitude (solid lines) compared to a subthreshold response following a 0.35-nA current pulse (dashed lines). A, Time course of the two ionic currents. Note their large sizes compared to the stimulating current. B, Membrane potential in response to sub- and supra-threshold stimuli. The injected current charges up the membrane capacity (with an effective membrane time constant $\tau = 0.85$ ms), enabling sufficient I_{Na} to be recruited to outweigh the increase in I_K (due to the increase in driving potential). The smaller current pulse fails to trigger an action potential, but causes a depolarization followed by a small hyperpolarization due to activation of I_K . C, Dynamics of the gating variables. Sodium activation m changes much more rapidly than either h or n . The long time course of potassium activation n explains why the membrane potential takes 12 ms after the potential has first dipped below the resting potential to return to baseline level. (From Koch, C., 1999, *Biophysics of Computation*, Cambridge, MA: MIT Press, p. 150.)

The steady-state activations (m_∞ , h_∞ , and n_∞) have a sigmoidal dependence on voltage. The *activation* variables m and n have the asymptotes $\lim_{V_m \rightarrow -\infty} m_\infty, n_\infty = 0$, $\lim_{V_m \rightarrow \infty} m_\infty, n_\infty = 1$, while the reverse holds for the *inactivation* variable h . That is, for very negative voltages, the m and n variables shut off current flow through both channel types, while at very positive potentials, the h particle shuts off the sodium current. The time “constants” (τ_m , τ_h , and τ_n)

are not constant with respect to voltage but rather have a roughly bell-shaped dependence, with peaks in the -80 to -40 mV range. The x_∞ and τ_x values were the ones actually measured by Hodgkin and Huxley using a series of voltage clamp steps. Instead of fitting these curves directly with mathematical functions, which would be sufficient for simulation purposes, they chose to express x_∞ and τ_x in terms of the variables α_x and β_x :

$$x_\infty = \frac{\alpha_x}{\alpha_x + \beta_x}$$

$$\tau_x = \frac{1}{\alpha_x + \beta_x}$$

where α_x and β_x depend on V_m as follows:

$$\alpha_m = \frac{0.1(V_m - 40)}{e^{(V_m - 40)/10} - 1} \quad \beta_m = 4e^{(V_m - 65)/18}$$

$$\alpha_h = 0.07e^{(V_m - 65)/20} \quad \beta_h = \frac{1}{e^{(V_m - 35)/10} + 1}$$

$$\alpha_n = \frac{0.01(V_m - 55)}{e^{(V_m - 55)/10} - 1} \quad \beta_n = 0.125e^{(V_m - 65)/80}$$

Note that the dimensions of τ_x , α_x , and β_x are all in units of $1/s$, while n_x is a pure number.

These rate constants assume a temperature of 6.3°C . At higher temperatures, they should be multiplied by a factor of around 3 per 10°C . The functional forms were chosen by Hodgkin and Huxley for two reasons. First, they were among the simplest that fit the data, and second, they resemble the equations that govern the movement of a charged particle in a constant field.

There is no direct way to map this set of equations in a simple manner onto the known molecular correlates of ionic channels, except that many voltage-dependent ionic channels possess four identical subunits, close or identical to the exponent of the activation variable that determines the momentary conductance. How the molecular structure and physical chemistry of these membrane pores explain the high throughput (up to 10^8 ions per second) and selectivity (the potassium channel is at least 10,000 times more permeant to K^+ than to Na^+ ions) has been revealed in stunning detail for potassium channels by atomic-resolution pictures of them (Doyle et al., 1998).

Conceptually, $x_\infty(V_m)$ can be thought of as the probability that an x particle will be in the open state at potential V_m . Each particle follows a two-state Markov model, where α_x is the rate constant from the closed to the open state and β_x is the rate constant from the open to the closed state. The time courses of the three variables are graphed in Figure 3.

This mathematical formalism was laid down in 1952. Since then, most models of voltage-dependent conductances—not only in axons, but also in cell bodies, and dendrites—have used the same formalism, with only minor modifications (Koch and Segev, 1998).

The macroscopic Hodgkin-Huxley equations are both continuous and deterministic, yet the underlying microscopic ionic channels are binary and stochastic. That is, a correct biophysical formulation of the dynamics of the membrane potential needs to take into account the well-known probabilistic behavior of these ionic channels. However, given the large number of channels involved in axonal spike initiation and propagation, it is usually appropriate to approximate the system using the deterministic Hodgkin-Huxley equations. This is not to say, however, that for thin fibers with very high input impedances, a small number of channels, and close to the threshold, stochastic variation in channel behavior might not have large-scale effects on the timing of action potentials (Schneidman, Freedman, and Segev, 1998; Koch, 1999).

Action Potential Propagation

Equation 1 describes a patch of membrane with no spatial extent. This corresponds to the original experiments, in which the axon was “space-clamped”: a long electrode was inserted into the axon along its axis, removing any spatial dependence. In response to stimulation, the whole membrane would fire simultaneously as a single isopotential unit. More commonly, one end of the axon is stimulated and an action potential propagates to the other end. The equation that governs extended structures is the cable equation:

$$C \frac{\partial V_m}{\partial t} = \frac{d}{R_a} \frac{\partial^2 V_m}{\partial x^2} + g_l(E_l - V_m) + g_{\text{Na}}(E_{\text{Na}} - V_m) + g_{\text{K}}(E_{\text{K}} - V_m) \quad (2)$$

where d is the axon diameter, C is the membrane capacity, and R_a is the intracellular resistivity. The equation rests on the assumption of radial symmetry, i.e., radial current flow can be neglected, leaving only one spatial dimension, the distance x along the cable, in addition to t . If the last two (active) terms are dropped from the right-hand side, we are left with the classical cable equation for passive cables (see DENDRITIC PROCESSING). Associated with that equation is the *space constant* $\lambda = 1/\sqrt{g_l R_a}$, which is the distance across which the membrane potential decays a factor e in an infinite cable under steady-state conditions.

Figure 4 shows the result of a simulation of a 100-cm-long axon of diameter $d = 1$ mm. One end was stimulated with a brief current pulse and the voltage was graphed for five positions along the axon. The form of the action potential is very similar to that in Figure 3; furthermore, the action potential is self-similar as it propagates, showing no signs of dispersion.

The total delay from one end to the other is about 5 ms, giving an average velocity of about 20 m/s. By assuming a constant conduction velocity—that is, by postulating the existence of a wave, $V_m(x, t) = V_m(x - vt)$ —Equation 2 shows that the velocity is proportional to the square root of axon diameter: $v \propto \sqrt{d}$ (Rushton, 1951). Indeed, in a truly remarkable test of their model, Hodgkin and Huxley estimated the velocity to be 18.8 m/s, a value within 10% of the experimental value of 21.2 m/s. This is surprisingly

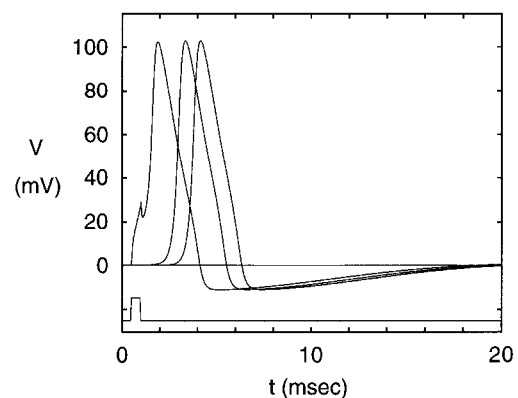


Figure 4. A propagating action potential. Solution to the complete Hodgkin-Huxley model for a long piece of squid axon for a brief suprathreshold current pulse. This pulse generates an action potential that travels down the cable and is shown here at the origin as well as 2 and 3 cm away from the stimulating electrode (solid lines). Note that the shape of the action potential remains invariant due to the nonlinear membrane. If the amplitude of the current pulse is halved, only a local depolarization is generated (dashed curve), which depolarizes the membrane 2 cm away by a mere 0.5 mV (not shown). This illustrates the dramatic difference between active and passive voltage propagation. (From Koch, C., 1999, *Biophysics of Computation*, Cambridge, MA: MIT Press, p. 162.)

accurate, considering that the model was derived from a space-clamped axon. This represents one of the rare instances in which a neurobiological model has made a successful quantitative prediction. The square root relationship had been discovered experimentally in the squid in the late 1930s.

Myelinated and Nonmyelinated Fibers

The principle of action potential generation and propagation appears to be very similar across neuronal types and species. One important evolutionary invention is that of myelination in the vertebrate phylum. Myelin sheaths are white fatty extensions of Schwann cells or neuroglial cells that are wrapped in many layers around axons. Myelin is a major component of the *white matter* of the brain, as opposed to the gray matter of neocortex, which has a high concentration of cell bodies and dendrites. The myelin sheaths extend for up to 1–2 mm along the axon (the *internodes*) and are separated by the *nodes of Ranvier*, which are only a few micrometers long. The internodal distance appears to be approximately linear in fiber diameter.

Myelin insulates the axon from the surrounding medium, increasing the membrane resistance and decreasing the capacitance. This reduces the electrotonic length of the axon for both DC and AC signals, making the cable electrically shorter, thereby significantly increasing the propagation speed. While a 1-mm nonmyelinated axon in the squid has an associated propagation speed of only about 20 m/s (Hodgkin and Huxley, 1952), a myelinated 20- μm vertebrate axon can reach over 100 m/s. For a nonmyelinated axon to reach that velocity, it would have to be an inch thick! This reduction in axon diameter allows for a much higher packing density while conserving speed.

It has been shown both experimentally and theoretically that the velocity of propagation is linear or slightly sublinear in the fiber diameter for myelinated axons. Figure 5 compares the spike propagation velocity for myelinated and nonmyelinated axons for small diameters. The myelinated axons overtake nonmyelinated ones already in the submicrometer range.

As opposed to their uniform distribution in nonmyelinated nerve, the voltage-gated channels in myelinated nerve are highly segregated between node and internode (Hille, 1992). The nodal membrane has a high concentration of fast sodium channels (between 700 and 2,000 per μm^2) and voltage-independent leak channels. The internodal membrane has a low concentration of potassium and leak channels and is virtually devoid of sodium channels. Here, the repolarization of the membrane following the initial phase of the spike is via the leak channels and sodium inactivation. This low density of channels in the internodal membrane, which makes up the more than 99% of the axonal membrane, reduces the average current density across the membrane, resulting in great savings in metabolic energy. Most of the activity occurs at the nodes of Ranvier, while the propagation along the internodes is chiefly passive.

In summary, myelin provides three advantages: propagation speed and packing density are both dramatically increased, while power consumption is decreased.

The Axonal Tree

Some axons branch profusely in the vicinity of the cell body. Others send off one or a few branches that course through the body for up to a meter before branching. Others extend for a few millimeters, giving rise to axonal arbors at regular intervals. The axon often arises at the “axon hillock,” a somatic bulge opposite from the trunk of the dendritic tree, though other arrangements are possible, such as the axon’s emanating from the dendrite rather than the soma.

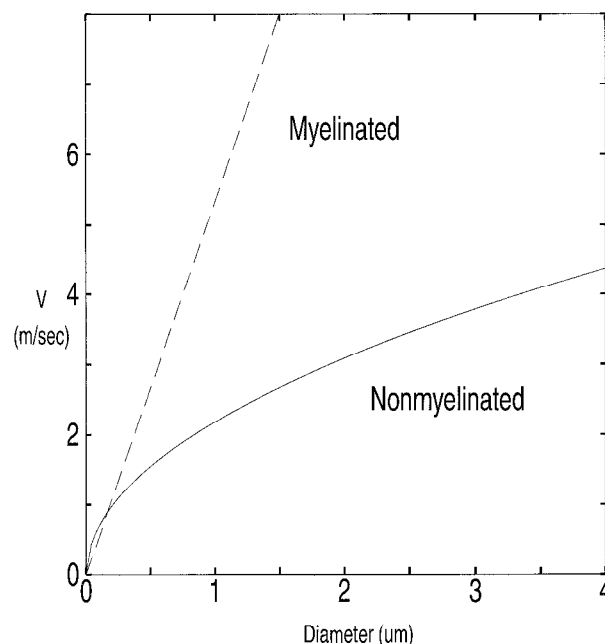


Figure 5. Spike propagation velocity and axonal diameter. The propagation velocity has a square root dependence on diameter for nonmyelinated axons. For myelinated axons, the dependence is linear or slightly sublinear. Myelination increases velocity for axons as thin as 0.2 μm , which are among the smallest found in the brain. (Adapted from Waxman and Bennett, 1972.)

Figure 1 shows an example of a terminal arbor in turtle tectum from a cell originating in nucleus isthmi. This particular axon has a 3- μm myelinated parent trunk and initial daughter branches. These give rise to hundreds of thin, highly varicosed daughter branches that lack myelin. The varicosities are usually the location of synaptic boutons, a local thickening where action potentials trigger the release of neurotransmitter, which in turn induces a conductance change in the postsynaptic target neuron. Boutons of some neurons may receive synaptic input that can inhibit this signal transmission, a process known as presynaptic inhibition. The 3,600 boutons on this arbor average 1.5 μm in diameter, though the size is highly variable, with a few boutons being as large as 7 μm .

The propagation speed along an unbranched axon depends on the diameter, as discussed earlier. In addition, a delay might be introduced at branch points, at varicosities at presynaptic terminals, and at locations where the diameter changes abruptly (Manor, Koch, and Segev, 1991). The delay may be negative (a speed-up), depending on the geometrical aspects, in particular the diameter of the parent branch in relation to that of the daughter branches. In a simulation of a 3.5-mm-long branched terminal axonal tree, Manor et al. found that the total axonal delay from the cell body to the synaptic terminals ranged from about 3 ms to 6 ms. Most of this delay (67%–78%) arose from the properties of unbranched, uniform cables; 16%–26% resulted from branch point delays, and 6%–7% from the presence of varicosities. In theory, the delay at a single branch point may be as large as 2 ms or more, if the temperature is low and the impedance mismatch is large. If the mismatch is too large, however, *branch point failure* may occur, a condition in which the action potential fails to propagate beyond the branch point. The concept of branch point filtering has been put forth by Chung, Raymond, and Lettvin (1970): the branch point may constitute a point of control where selective transmission occurs, allowing the axonal tree to distribute action potentials only to a sub-

set of nerve terminals. Experimentally, little is known concerning the amplitude of temporal dispersion of action potentials due to axonal branching.

While the axonal propagation delay may seem an unavoidable fact of life that slows down neural communication, it may also have important computational advantages. For instance, sound localization (see SOUND LOCALIZATION AND BINAURAL PROCESSING) in the barn owl depends on interaural time differences as small as a tenth of a millisecond and is apparently obtained by using the axon as a delay line (Konishi, 1992), and several models of brain function depend critically on the exact timing of inputs from different sources. Although delays may be imposed by the dendritic trees at the input end of the neuron, the axons are also important candidates for this function.

Debanne and colleagues (1997) discovered that action potentials can be selectively filtered at or beyond axonal branch points via a fast-inactivating A-type of potassium conductance. When a cell is hyperpolarized, a depolarizing step within 10–20 ms that would normally trigger an action potential fails to do so in hippocampal cell bodies. The reason for this selective block is a G_A -like K^+ conductance present somewhere along the axon. If de-inactivated by long-lasting hyperpolarization, it filters out single isolated spikes. It could thereby act to enhance the signal-to-noise ratio of neuronal firing. To what extent this is a general mechanism or an exception to the rule that axons faithfully transmit action potentials from their site of initiation close to the cell body to their postsynaptic target structures remains to be seen.

Over the past several decades, the formalism introduced by Hodgkin and Huxley in 1952—voltage- and time-dependent activation and inactivation variables that determine the current value of the various membrane conductances—has become the de facto standard for modeling an amazing variety of phenomena, including adaptation, calcium-dependent conductances, plateau potentials, first- and second-order inactivation, oscillatory discharges, and several varieties of bursting.

Road Maps: Biological Neurons and Synapses; Grounding Models of Neurons

Related Reading: Activity-Dependent Regulation of Neuronal Conductances; Ion Channels: Keys to Neuronal Specialization; Oscillatory and Bursting Properties of Neurons

Axonal Path Finding

Geoffrey J. Goodhill

Introduction

Many stages are involved in constructing a biological nervous system. Following the migration of neurons to their proper locations and their phenotypic specification, the initial pattern of connections forms between different regions (Sanes, Reh, and Harris, 2000). Making the right connections is crucial for proper function, and often requires axons to navigate over long distances with great precision (Tessier-Lavigne and Goodman, 1996). Until recently, relatively little was known about this process experimentally; however, the past decade has seen a dramatic increase in knowledge (at least qualitatively) concerning the molecules and mechanisms involved (Mueller, 1999). These insights are now being applied to understanding how axons can be made to regenerate to appropriate

References

- Chung, S. H., Raymond, S. A., and Lettvin, J. Y., 1970, Multiple meaning in single visual units, *Brain Behav. Evol.*, 3:72–101.
- Debanne, D., Guérineau, N. C., Gähwiler, B. H., and Thompson, S. M., 1997, Action-potential propagation gated by an axonal I_A -like K^+ conductance in hippocampus, *Nature*, 389:286–289.
- Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., and MacKinnon, R., 1998, The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity, *Science*, 280:69–77.
- Hille, B., 1992, *Ionic Channels of Excitable Membranes*, 2nd ed., Sunderland, MA: Sinauer. ♦
- Hodgkin, A. L., 1976, Chance and design in electrophysiology: An informal account of certain experiments on nerve carried out between 1934 and 1952, *J. Physiol.*, 263:1–21.
- Hodgkin, A. L., and Huxley, A. F., 1952, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.*, 117:500–544.
- Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds., 2000, *Principles of Neural Science*, 4th ed., New York: McGraw-Hill.
- Koch, C., 1999, *Biophysics of Computation*, Cambridge, MA: MIT Press. ♦
- Koch, C., and Segev, I., Eds., 1998, *Methods in Neuronal Modeling*, 2nd ed., Cambridge, MA: MIT Press.
- Konishi, M., 1992, The neural algorithm for sound localization in the owl, *Harvey Lect.*, 86:47–64.
- Manor, Y., Koch, C., and Segev, I., 1991, Effect of geometrical irregularities on propagation delay in axonal trees, *Biophys. J.*, 60:1424–1437.
- Roberts, A., and Bush, B. M. H., Eds., 1981, *Neurons without Impulses: Their Significance for Vertebrates*, Cambridge, Engl.: Cambridge University Press. ♦
- Rushton, W. A. H., 1951, A theory of the effects of fibre size in medullated nerve, *J. Physiol.*, 115:101–122.
- Schneidman, E., Freedman, B., and Segev, I., 1998, Ionic channel stochasticity may be critical in determining the reliability and precision of spike timing, *Neural Computat.*, 10:1679–1704.
- Sereno, M. I., and Ulinski, P. S., 1987, Caudal topographic nucleus isthmi and the rostral nontopographic nucleus isthmi in the turtle, *Pseudemys scripta*, *J. Comp. Neurol.*, 261:319–346.
- Waxman, S. G., and Bennett, M. V. L., 1972, Relative conduction velocities of small myelinated and non-myelinated fibers in the central nervous system, *Nature*, 238:217–219.
- Waxman, S. G., Kocsis, J. D., and Stys, P. K., Eds., 1995, *The Axon: Structure, Function and Pathophysiology*, New York: Oxford University Press. ♦

targets after injury to the adult nervous system, such as spinal cord injury.

Until now, the bulk of theoretical work in the neural network tradition has focused on changes in synaptic strengths within a fixed connectational architecture. Although local sprouting within the target has sometimes been considered (as has “sculpting,” based on the assumption that when synaptic strengths go to zero, the physical connection is lost), how axons chart their initial path toward the correct target structure has generally not been addressed. An example is the mapping from the eye to more central targets in the brain. Abundant theoretical models address how topographic maps form once axons reach the tectum or visual cortex (see DEVELOPMENT OF RETINOTECTAL MAPS; SELF-ORGANIZING FEATURE MAPS; and OCULAR DOMINANCE AND ORIENTATION COLUMNS),

but no theoretical work specifically addresses how retinal ganglion cell axons find the optic disc, how they then exit the retina, why they grow toward the optic chiasm, why some then cross at the midline while others do not, and so on. In recent years important insight into such issues has been gained through innovative experimental work, creating a body of knowledge that now has the potential to be framed and interpreted in terms of theoretical models. A crucial point is that, whereas work in neural networks has usually focused on processes such as synaptic plasticity that are dependent on neural activity, models for axon guidance must generally be phrased in terms of activity-independent mechanisms, particularly guidance by molecular gradients. In this article we first review some of the important experimental data regarding axon guidance, and then discuss some of the theoretical concepts that are relevant to this area.

Experimental Data

Several basic types of mechanisms have been identified to guide axons. (For more detailed discussions of the data summarized in this section, see Tessier-Lavigne and Goodman, 1996; Mueller, 1999; and Song and Poo, 2001.) First, axons can be channeled in particular directions by boundaries of permissive or inhibitory molecules. For instance, a "railroad track" of a permissive molecule may lock an axon into a particular trajectory, or a "wall" of an inhibitory molecule may keep it away from an undesired region. Second, axons can be pushed or pulled by "vector" signals in the form of molecular gradients. These gradients are often established by diffusion of a soluble molecule away from the target region. Third, the path to a distant target may be broken into several short segments, each involving a different type of cue, thus simplifying the problem of long-range guidance. Fourth, once one "pioneering" axon has reached the target, it is often the case that following axons simply fasciculate with (stick to) the pioneering axon. In each of these cases the molecules involved may be substrate-bound (expressed on cell membranes or bound to cells or to the extracellular space) or diffusible (diffusing through the extracellular space).

In the last few years the number of molecules specifically implicated in axon guidance has jumped from virtually none to around 100, most of them previously unknown. (Note that we distinguish between *guidance factors* and *growth factors*: the latter category, which includes the neurotrophins, are often essential for axons to extend, but so far have mostly not been shown to play an active role in axon guidance *in vivo*.) Guidance factors can be organized into several main families based on their molecular structure, including the netrins, semaphorins, slits, and ephrins. There is an astonishing amount of evolutionary conservation in these families. Homologous molecules perform analogous guidance functions in animals ranging from nematodes to flies to mammals, indicating that the basic molecular tools for wiring a nervous system were established hundreds of millions of years ago. Molecules involved in axon guidance are also often involved in the analogous chemotactic event of cell migration, and recent findings even suggest some commonality with the signal transduction mechanisms important for chemotaxis of leukocytes. Although it was originally thought that the different types of guidance mechanisms might be segregated between different families of molecules, it is now clear that this is not the case. For instance, the same molecule can be attractive in one context but repulsive in another, or it may normally be substrate-bound but have a soluble fragment that can diffuse.

Guidance signals for axons are detected and transduced by the growth cone, a dynamic and motile structure at the tip of the developing axon. This consists of a central region surrounded by web-like veils called lamellipodia, and long, finger-like protrusions called filopodia (Figure 1A). Receptors expressed on the surface of the growth cone can bind molecules of the families mentioned above. The resulting signals are then converted by complex internal transduction pathways into differential rates of actin polymerization in different parts of the growth cone so as to move it forward, left, or right. Dissection of the signaling networks responsible for converting a graded difference in receptor binding into directed movement is currently a very active area of research. One intriguing finding is that the concentration of cAMP within the growth cone

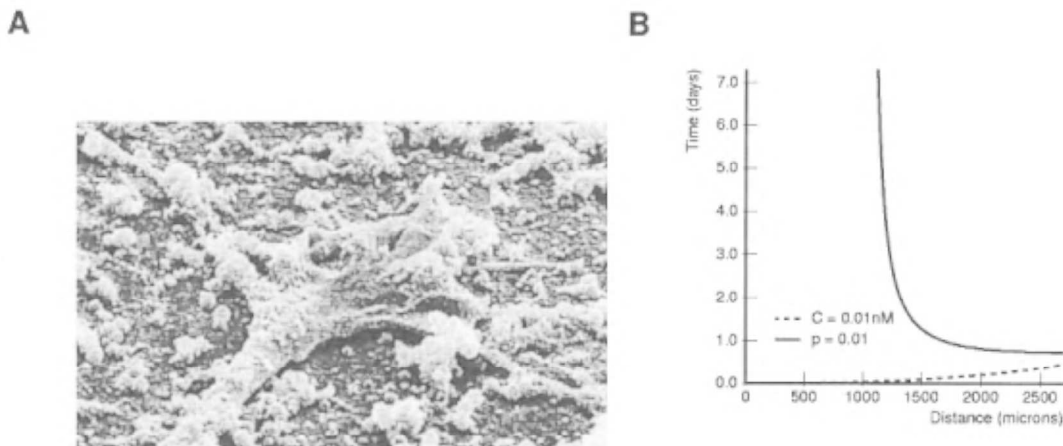


Figure 1. A, Electron micrograph of a growth cone at the end of an axon (here the growth cone is resting on a surface irregularly covered with coated beads). The long finger-like protrusions are filopodia. Growth cones are typically about 0.01 mm across. (From Rosentreter, S. M., Davenport, R. W., Lösinger, J. Huf, J., Jung, J., and Bonhoeffer, F., 1998, Response of retinal ganglion cell axons to striped linear gradients of repellent guidance molecules, *J. Neurobiol.*, 37:541–562. © 1998, John Wiley & Sons, Inc. Reprinted with permission.) B, Interaction of constraints for guidance by a target-derived diffusible factor. The graph shows, at each distance, the

time at which two constraints are satisfied: the low concentration limit, where not enough receptors are bound for a gradient signal to be detected (assumed to be $K_D/100$, with $K_D = 1$ nM), and the fractional change constraint (assumed to be $\Delta C/C = 1\%$). The region between the two curves in each graph is where guidance is possible. The guidance limit imposed by the fractional change constraint once the gradient has stabilized is 1 mm. However, guidance range is extended at earlier times, when the fractional change constraint has yet to take full effect.

helps determine how the growth cone responds to a gradient cue: if the cAMP level is above a certain threshold, the growth cone is attracted; if the cAMP level is below that threshold, the growth cone is repelled.

Theoretical Models

Gradient Detection

Several different areas of theoretical development are relevant to the emerging picture of axon guidance. The general topic of chemotaxis has inspired a large body of theoretical analysis. However, most work has focused on bacteria and leukocytes, and it remains to be established how relevant these models are to axon guidance. Perhaps most important in this category are theories describing the fundamental physical limits on the minimum steepness of gradients detectable by *any* small sensing device. The key hypothesis, first rigorously formulated by Berg and Purcell (1977), is that gradient detection is limited by inherent statistical fluctuations in receptor binding. They calculated the statistical noise in a concentration measurement ΔC_{noise} by a small sensing device that arises from inevitable stochastic variations in the number of receptors bound at any instant. The fractional root mean square error in the measurement of a concentration difference between two spatially or temporally separated points is then $\sqrt{2}\Delta C_{\text{noise}}/C$, where C is the average concentration at the sensing device. For a true gradient to be detected, it must be steep enough so that the actual concentration difference between the two points, ΔC_{grad} , is such that $\Delta C_{\text{grad}} > \sqrt{2}\Delta C_{\text{noise}}$. $\Delta C_{\text{noise}}/C$ can be calculated from first principles using various simplifying assumptions. This approach has been applied to growth cones by Goodhill and Urbach (1999), who derived estimates for the minimum gradient steepness detectable by an axon for a diffusible gradient of order 1% and for a substrate-bound gradient of order 10%. The difference arises because of the lower encounter rate between receptor and ligand molecules for a bound versus a diffusible gradient. Goodhill and Urbach (1999) also showed that the movement of filopodia does not significantly increase the encounter rate, which suggests that filopodia increase gradient sensitivity only by increasing the effective size of the growth cone. However, this approach assumes that the receptor-ligand reaction is diffusion-limited, which may not be the case for the molecules involved in axon guidance. Theoretical work following Berg and Purcell (1977), while still generally founded on the basic assumption that gradient detection is limited by the signal-to-noise ratio, has attempted to relax this and some other assumptions, but these models have not yet been specifically applied to growth cones.

Growth Cones

Theoretical models have been proposed to account for filopodial dynamics. Based on experimentally determined distributions for parameters such as rates of filopodial initiation, extension, and retraction, filopodial length, and angular orientation, Buettner and colleagues (e.g., Buettner, 1995) have developed simulation models describing filopodial structure as a function of time, and growth cone trajectories both during normal growth and when a target is encountered. Goodhill and Urbach (1999) presented a model of growth cone trajectories based on the assumption that each filopodium makes a noisy (in Berg and Purcell's sense) estimate of the concentration in the direction it is pointing, that more filopodia are generated in the direction of higher concentration, and that each filopodium exerts a pull on the growth cone. Other models have proposed hypotheses about how actin and microtubule dynamics lead to filopodia formation, though these models have yet to fully engage with what is known about these processes experimentally.

Another theoretically interesting aspect of growth cones is the signaling events that convert a small difference in receptor binding into a large directed movement. Meinhardt (1999) and others have proposed reaction-diffusion-type models in which a small inhomogeneity in an initially uniform system is amplified via the interaction of a short-range activator with a longer-range inhibitor. However, to return the system to a uniform state so that the directional preference of the growth cone can change with time, a second type of reaction with a longer time constant is invoked, and direct experimental evidence for such processes in growth cones is currently lacking. Tranquillo and Lauffenburger (1987), in the context of leukocyte chemotaxis, simulated and analyzed a model in which two pools of receptors (one on each side of the cell) communicated information about the degree of binding via a single intracellular messenger. This model was quite successful at accounting for various aspects of leukocyte movement, and subsequent versions by Tranquillo and colleagues have examined more complex assumptions regarding the internal signaling dynamics. Bacterial chemotaxis has been extensively studied from the perspective of signal transduction, and theoretical models have been effective at explaining the large amount now known experimentally about this system. A major focus of such models has been to explain the process by which bacteria adapt to background levels of ligand so that they can detect small changes in concentration over many orders of magnitude of absolute concentration. Although such analyses of signaling mechanisms in other chemotacting systems have the potential to be applied to growth cones, it is unclear how similar these systems really are. More generally there is increasing interest in mathematical modeling of the signal transduction pathways underlying cell behavior as a whole, although again, there is little application as yet of these theoretical ideas to axon guidance.

Diffusible Gradients and Optimal Gradients

An important class of gradients for guiding axons both in vivo and in vitro consists of gradients established by diffusion. Hentschel and van Ooyen (1999) investigated a possible role for diffusion in controlling axon fasciculation. They considered a population of axons being guided by a target-derived diffusible factor, and hypothesized that in addition, each axon releases a diffusible attractant that pulls it toward the other axons, hence leading to fasciculation as they grow together toward the target. To account for defasciculation at the target, they hypothesized that each axon also releases a repulsive factor for other axons at a rate dependent on the concentration of the target-derived factor. As the axons approach the target, this repulsive force overcomes the attractive force, leading to defasciculation.

Another approach is to analyze the gradient shapes expected from diffusion processes in particular situations and how these constrain the spatiotemporal domains in which guidance is possible (see Goodhill, 1998, for a review). Goodhill considered a source releasing a diffusible factor at a constant rate into an infinite, spatially uniform three-dimensional volume, a problem for which there is a closed-form solution. As long as the gradient is not too steep, the fractional change in concentration $\Delta C/C$ across the growth cone width a is $\Delta C/C = (\partial C/\partial r)(a/C)$, and can be straightforwardly calculated. It has the perhaps surprising characteristic that, for fixed r , $\Delta C/C$ decreases with t . That is, the largest fractional change at any distance occurs immediately after the source starts releasing factor. For large t , $\Delta C/C$ asymptotes at a/r . Thus: (1) At small times after the start of production the factor is very unevenly distributed. The concentration C falls quickly to almost zero moving away from the source, the gradient is steep, and the percentage change across the growth cone $\Delta C/C$ is everywhere large. (2) As time passes, the factor becomes more evenly distributed. C everywhere increases, but $\Delta C/C$ everywhere decreases. (3) For large times, C tends to an

inverse variation with the distance from the source r , while $\Delta C/C$ tends to a/r independent of all other parameters. The equation for $\Delta C/C$ can be compared with the size of the smallest gradient the growth cone can detect to yield the regions of parameter space found in which guidance is possible (Figure 1B). Based on data for leukocyte chemotaxis it was assumed that gradient detection occurs when $\Delta C/C \geq p$ and $C \geq C_{\min}$, where p is a threshold assumed independent of C . The positions and times for which the gradient calculated above satisfies these criteria were examined, given appropriate estimates for the relevant parameters. For large times (a few days) after the start of factor production, the maximum range is independent of the diffusion constant and is about 1 mm. This value fits well with both in vitro and in vivo observations. At earlier times, however, the factor is more unevenly distributed, being more concentrated around the source. This makes the fractional change larger than at later times, increasing the range over which guidance can occur. Depending on the parameters, the model predicts that guidance may be possible at distances of several millimeters before the distribution of factor equilibrates. It is conceivable that such a mechanism might be utilized in vivo to extend guidance range beyond the 1 mm limit imposed once the gradient has stabilized.

Similarly, one may inquire as to the optimal gradient shape, in the sense of the shape that guides an axon over the largest possible distance. Assuming the minimal fractional change is constant (not dependent on absolute concentration), the optimal shape is clearly exponential; the maximum guidance distance turns out to be about 1 cm (Goodhill, 1998). It is conceivable that substrate-bound gradients could achieve this shape, and in fact the size of the chick tectum at the time retinotectal maps are forming is about 1 cm. Assuming instead that the minimal fractional change varies with concentration, as predicted by Berg and Purcell, and also assuming a high concentration limit, the order of magnitude of the result remains at 1 cm (Goodhill and Urbach, 1999). More generally, this type of analysis raises the issue of overall scaling between different species. A guidance mechanism (e.g., target-derived diffusible gradient) that works for a small animal will not work in a large animal if the anatomy is simply scaled up. In general, the scale and structure of the anatomy of, say, the elephant nervous system at the time at which long-range navigation occurs are not known in sufficient detail to allow proper comparison with the same features in, say, the rat.

Retinotectal Maps

The most well-developed area of axon guidance modeling concerns the formation of topographic maps in the optic tectum (reviewed in Goodhill and Richards, 1999). The hypothesis of chemospecificity, that graded distributions of molecules are somehow matched to graded distributions of complementary molecules in the tectum so as to form a topographic map, was first proposed qualitatively by Sperry (1963). Although a great deal of experimental work ensued to investigate how such gradients may actually operate, matched gradients of receptors in the retina and ligands in the tectum were discovered only in the mid-1990s. These receptors/ligands are of the Eph/ephrin family, which currently are under intense experimental investigation. Theoretical modeling started in the 1970s (e.g., Willshaw and von der Malsburg, 1979), and early models were based directly on molecular gradients. A key finding was that some kind of normalization is essential to prevent all axons

from targeting the same part of tectum. Although modeling based on gradients has continued (e.g., Gierer, 1987), the focus of most modeling work changed to activity-dependent processes. Here only synaptic strength changes within a fixed architecture are generally considered, rather than the earlier stage of how axons traverse large expanses of the tectum. Data and models in this area are discussed in greater detail in DEVELOPMENT OF RETINOTECTAL MAPS (q.v.).

Discussion

Current experimental work in axon guidance is dominated by techniques and hypotheses at the molecular level. The data are also rapidly evolving, with new molecules and mechanisms important for guidance being discovered at a very fast rate. However, many fundamental questions remain unresolved, and theoretical models have the potential to make an important contribution to answering these questions. What is the minimum gradient steepness detectable by a growth cone, and how does this vary with the properties of the receptor-ligand interaction and the internal state of the growth cone? How is a graded difference in receptor binding internally converted into a signal for directed movement? How do axons integrate multiple cues? And, perhaps most relevant to human health, how can regenerating axons be encouraged to grow toward and reconnect with appropriate targets after injury?

Road Map: Neural Plasticity

Related Reading: Development of Retinotectal Maps

References

- Berg, H. C., and Purcell, E. M., 1977, Physics of chemoreception, *Biophys. J.*, 20:193–219.
- Buettner, H. M., 1995, Computer simulation of nerve growth cone filopodial dynamics for visualization and analysis. *Cell Motil. Cytoskelet.*, 32:187–204.
- Gierer, A., 1987, Directional cues for growing axons forming the retinotectal projection, *Development*, 101:479–489.
- Goodhill, G. J., 1998, Mathematical guidance for axons, *Trends Neurosci.*, 21:226–231. ♦
- Goodhill, G. J., and Richards, L. J., 1999, Retinotectal maps: Molecules, models, and misplaced data, *Trends Neurosci.*, 22:529–534.
- Goodhill, G. J., and Urbach, J. S., 1999, Theoretical analysis of gradient detection by growth cones, *J. Neurobiol.*, 41:230–241.
- Hentschel, H. G. E., and van Ooyen, A., 1999, Models of axon guidance and bundling during development, *Proc. R. Soc. Lond. B*, 266:2231–2238.
- Meinhardt, H., 1999, Orientation of chemotactic cells and growth cones: Models and mechanisms, *J. Cell Sci.*, 112:2867–2874.
- Mueller, B. K., 1999, Growth cone guidance: First steps towards a deeper understanding. *Annu. Rev. Neurosci.*, 22:351–388. ♦
- Sanes, D. H., Reh, T. A., and Harris, W. A., 2000, *Development of the Nervous System*, San Diego, CA: Academic Press.
- Song, H., and Poo, M.-M., 2001, The cell biology of neuronal navigation, *Nature Cell Biol.*, 3:E81–E88.
- Sperry, R. W., 1963, Chemoaffinity in the orderly growth of nerve fiber patterns and connections, *Proc. Natl. Acad. Sci. USA*, 50:703–710.
- Tessier-Lavigne, M., and Goodman, C. S., 1996, The molecular biology of axon guidance, *Science*, 274:1123–1133. ♦
- Tranquillo, R. T., and Lauffenburger, D. A., 1987, Stochastic model of leukocyte chemosensory movement, *J. Math. Biol.*, 25:229–262.
- Willshaw, D. J., and von der Malsburg, C., 1979, A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem, *Philos. Trans. R. Soc. B*, 287:203–243.

Backpropagation: General Principles

Michael A. Arbib

Introduction

Perceptrons are neural nets that use an *error-correction* rule to change the weights of each unit that makes erroneous responses to stimuli that are presented to the network. As already explained in PERCEPTRONS, ADALINES, AND BACKPROPAGATION (q.v.) and Section I.3: “Dynamics and Adaptation in Neural Networks,” *backpropagation* is a family of methods for training a *multilayer perceptron*, a loop-free network that has its units arranged in layers, with each unit providing input only to units in the next layer of the sequence. The first layer comprises input units; there may then be several layers of trainable “hidden units” carrying an internal representation, and finally there is the layer of output units, also with trainable synaptic weights.

Rumelhart, Hinton, and Williams (1986) is the most influential paper on the error backpropagation method, providing a formula (see the Proposition below) for propagating back the gradient of error evaluation from a unit to the units that provide its inputs. Since the formulas involve derivatives, the input and output of each unit must take continuous values in some range, here taken to be $[0, 1]$. The response is a sigmoidal function of the weighted sum. Werbos (1995) provides a historical perspective on precursors of their paper. As a specific example of such a precursor, LEARNING AND STATISTICAL INFERENCE (q.v.) presents a general stochastic descent on-line learning procedure (Amari, 1967), which, when applied to the multilayer perceptron, yields the error backpropagation method.

Proposition. Consider a layered loop-free net with error $E = \sum_k (t_k - o_k)^2$, where k ranges over designated output units, and let the weights w_{ij} be changed according to the gradient descent rule

$$\Delta w_{ij} = -\frac{\partial E}{\partial w_{ij}} = 2 \sum_k (t_k - o_k) \frac{\partial o_k}{\partial w_{ij}}$$

Then the weights may be changed inductively, working back from the output units, by the rule Δw_{ij} is proportional to $\delta_i o_j$, where

Basis Step: $\delta_i = (t_i - o_i)f'_i$ for an output unit.

Induction Step: If i is a hidden unit, and if δ_k is known for all units that receive unit i 's output, then $\delta_i = (\sum_k \delta_k w_{ki})f'_i$, where k runs over all units that receive unit i 's output. \square

Thus the “error signal” δ_i propagates back layer by layer from the output units. In $\sum_k \delta_k w_{ki}$, unit i receives error propagated back from a unit k to the extent to which i affects k .

The above proposition tells us how to compute Δw_{ij} for the *on-line* backpropagation algorithm that adjusts the weights in response to each single input pattern, using the “local error” of the network with its current weight settings for that input. It does not guarantee that the above step size is appropriate to reach the minimum, nor does it guarantee that the minimum, if reached, is global. The backpropagation rule defined by this proposition is, thus, a heuristic rule, not one guaranteed to find a global minimum. The *batch* version of the algorithm cycles through a complete training set of input-output pairs $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, with gradient descent applied to the cumulative error of each cycle, until no further changes are required.

As the index to this *Handbook* attests, backpropagation has been perhaps the most diversely used adaptive architecture, especially in technological applications. The purpose of this article is neither to introduce the basics of backpropagation (again, see PERCEPTRONS, ADALINES, AND BACKPROPAGATION and Section I.3: Dy-

namics and Adaptation in Neural Networks) nor to survey its applications (see SPEECH RECOGNITION TECHNOLOGY for one example of a careful analysis of the pros and cons of using multilayer perceptrons), but instead to place backpropagation in a broader context by providing a road map for a number of contributions elsewhere in the *Handbook* that enrich our basic understanding of this adaptive architecture. The article also assesses the biological plausibility of backpropagation.

Auto-Encoding

A basic application for backpropagation networks has been to find compressed representations. In this case, a network with one hidden layer is trained to become an *auto-encoder* or *auto-associator* by learning the identity function: making the desired states of the N output units identical to the states of the N input units for each input-output pair in the training sample. Data compression is achieved by making the number of hidden units $M < N$. Moreover, the features discovered by the hidden units may be useful for processing tasks, such as classification of the input patterns. However, as shown in UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS, it may not be possible to relate the activities of individual hidden units to specific features that may be found by other means to characterize complicated input patterns. One way to constrain the hidden unit representation is to add extra penalty terms to the error function. For example, a penalty term on hidden unit activations can be chosen that causes these units to represent high-dimensional data as localized bumps of activity in a lower-dimensional constraint surface. This encourages the hidden units to form a map-like representation that best characterizes the input. Other penalty terms lead to other encodings, such as sparse or combinatorial representations (see MINIMUM DESCRIPTION LENGTH ANALYSIS).

RAAM networks (Pollack, 1990) are three-layer backpropagation networks whose input and output layers are each divided into regions. The network is trained to “auto-associate,” i.e., to reproduce a given pattern of input on the output layer. The purpose of this training is to permit condensed, distributed encodings of K -tuples of information (i.e., the subpatterns presented to the K regions of the input layer) to be developed on the hidden layer. Once such a distributed encoding has been developed for a given K -tuple of information, that encoding may later be presented as input to a single region of the input layer, while the remaining input regions receive similarly derived distributed encodings, so that the network then develops codes for K -tuples of information. The network may then be trained to auto-associate on this more complex set of input information. Iterating the procedure yields codes for K -tuples of K -tuples, and so on, thus making possible condensed distributed encodings for entire tree structures in the RAAM's hidden layer. SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS (q.v.) discusses the implications of such techniques for the ability, or otherwise, of connectionist models to capture human abilities for symbol processing.

Stochasticity and Plateaus

STOCHASTIC APPROXIMATION AND EFFICIENT LEARNING (q.v.) notes that both on-line and batch backpropagation seek a weight vector w that minimizes the error function, but stresses the statistical notion that inputs must follow some probability distribution so that what we really seek to minimize is the error as averaged

over all examples. But should we average over the few examples available in the training set, or over all the complete probability distribution “given by Nature”? The first average is named *empirical risk* and measures only the training set performance. The second average is called the *expected risk* and measures the much more interesting generalization performance (Vapnik, 1998). The *Handbook* introduces a stochastic gradient descent algorithm in which each iteration consists of picking a single random example and updating the weight vector w accordingly. This stochastic gradient descent does not need to remember which examples were visited during the previous iterations, making this algorithm suitable for the on-line adaptation of deployed systems. In such a situation, the stochastic gradient descent directly optimizes the expected risk, since the examples are randomly drawn from the “ground truth” distribution. The stochastic gradient descent can also pick examples from a finite training set. This procedure optimizes the empirical risk. The number of iterations is usually larger than the size of the training set. The examples are therefore presented multiple times to the network.

LEARNING AND STATISTICAL INFERENCE (q.v.) offers a general method, called *Fisher efficiency*, of assessing the success of an estimator relating input and output patterns. It then notes that, although backpropagation learning has been used widely, it is not Fisher efficient. Moreover, the method may converge to one of the local minima of the error landscape, which might be different from the global minimum. Intriguingly, convergence may be drastically slow because of “plateaus.” The error decreases quickly at the beginning of learning, but its rate of decrease becomes extremely slow. After surprisingly many steps, the error again decreases rapidly. This is understood as showing that weights are trapped in a “plateau” that is not a local minimum but nonetheless provides a region of weight space that learning takes very long to escape from.

Saad and Solla (1995) used statistical mechanics to show that plateaus exist because of the “symmetry” in the hidden units: the output and hence the error measure is invariant under permutations of hidden units in the multilayer perceptron. Whereas this property leads to phase transitions in equilibrium batch training (see STATISTICAL MECHANICS OF GENERALIZATION), the effect in on-line training is that the system approaches a symmetric state from generic initial conditions. STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION (q.v.) discusses how the properties of such plateaus can be investigated in detail by linearizing the dynamics close to the fixed point (Biehl, Riegler, and Wöhlér, 1996). Figure 1 in STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION (q.v.) provides a simple example of the breaking of permutation symmetry during learning, showing a typical learning curve in which the learning process is dominated by a pronounced plateau state in which hardly any progress is made while the number of examples increases. Only after an extended period of time does the system leave the plateau and approach its asymptotic state exponentially fast. In the case displayed in the figure, the system is very close to a perfectly symmetric configuration.

There are various acceleration methods for the backpropagation learning rule, but they cannot eliminate plateaus. NEUROMANIFOLDS AND INFORMATION GEOMETRY (q.v.) shows that the natural gradient method (Amari, 1998), based on the Riemannian structure of a neuromanifold, not only eliminates plateaus but is Fisher efficient.

Recurrent Neural Networks

A feedforward network is just a static mapping of input vectors to output vectors, whereas our brain is a high-dimensional nonlinear dynamical system, replete with loops. This provides one motivation (another is technological) for the study of learning algorithms for recurrent neural networks, which have feedback connections and

time delays. In a recurrent network, the state of the system can be encoded in the activity pattern of the units, and a wide variety of dynamical behaviors can be encoded by the connection weights. Network dynamics that converge to a minimum of an “energy” function (see COMPUTING WITH ATTRACTORS) have proved important for associative memory tasks and optimization networks. However, steady-state solutions (fixed-point attractors) are only a limited portion of the capabilities of recurrent networks. A recurrent network can serve as a sequence recognition system or as a sequential pattern generator. RECURRENT NETWORKS: LEARNING ALGORITHMS (q.v.) reviews the learning algorithms for training recurrent networks, focusing on supervised learning algorithms for recurrent networks, with only a brief overview of reinforcement and unsupervised learning algorithms.

Recurrent neural networks use the additional degree of freedom provided by a priori unlimited processing time in order to map the information appropriately. For example, simple recurrent networks (SRNs; Elman, 1990) augments the three-layer backpropagation network with a supplementary context layer of the same size as the hidden layer. Reciprocal links between the hidden layer and the context layer create a loop enabling any activation pattern currently present on the hidden layer to be merged with the activation pattern currently present in the context layer, and vice versa. An extension of the backpropagation algorithm trains these connections as well. Essentially, the activity in the context and hidden layers may be seen as an internal state, so that training serves to update both the definition of a “next-state function” as well as the reading of the output from the internal state in such a way as to enable the system to better and better approximate a training set, which now consists of pairs of input and output sequences, rather than one-shot input vectors and output vectors. CONSTITUENCY AND RECURSION IN LANGUAGE (q.v.) exemplifies the use of SRNs in connectionist linguistics.

Other Perspectives

To create a neural network, a designer typically fixes a network topology and uses training data to tune its parameters, such as connection weights. The designer, however, often does not have enough knowledge to specify the ideal topology. In the case of a multilayer perceptron, the only free parameter in “topology space” is the number of hidden units. Too few hidden units and the current task is unlearnable; too many units and the network learns the noise as well as the task relationships. It is thus desirable to learn the topology from training data as well. LEARNING NETWORK TOPOLOGY (q.v.) looks at learning as a search in the space of topologies as well as in weight space. In particular, it provides a general measure of the “goodness” of a topology and some search strategies over the space of topologies to find the best one. This framework is applied to learning the topologies of both feedforward neural networks and Bayesian belief nets (see BAYESIAN NETWORKS).

A basic strategy to avoid false minima is Boltzmann learning (see SIMULATED ANNEALING AND BOLTZMANN MACHINES). Here the units respond in stochastic fashion to their inputs. The degree of “stochasticity” is controlled by a parameter T . As $T \rightarrow -\infty$, the unit becomes deterministic; as $T \rightarrow \infty$, the unit becomes very noisy. T is often referred to as “temperature,” as part of the comparison of large neural networks with the systems treated by statistical mechanics (see STATISTICAL MECHANICS OF NEURAL NETWORKS). Convergence to the global optimum is aided by starting at high T and gradually lowering it—this is the process of “simulated annealing”—with the intuition being that the initial high noise “bumps the system out of the high valleys” of the error landscape, while the eventual low noise allows it to settle in the “low valleys.”

MODULAR AND HIERARCHICAL LEARNING SYSTEMS (q.v.) replaces the training of a single network by the training of a set of networks that forms a “mixture of experts,” the idea being that each

network will become expert at processing inputs from a region of the input space, while a gating network will learn which experts to rely on for processing a given input. As an alternative to gradient methods, Jordan and Jacobs (1994) developed an Expectation-Maximization (EM) algorithm (McLachlan and Krishnan, 1997, give a general treatment of the EM algorithm) that is particularly useful for models in which the expert networks and gating networks have simple parametric forms. Each iteration of the algorithm consists of two phases: (1) a recursive propagation upward and downward in the tree of modules to compute posterior probabilities (the “E step”), and (2) solution of a set of local weighted maximum likelihood problems at the nonterminals and terminals of the tree (the “M step”). Jordan and Jacobs (1994) tested this algorithm on a nonlinear system identification problem (the forward dynamics of a 4-degrees-of-freedom robot arm) and reported that it converged nearly two orders of magnitude faster than backpropagation in a comparable multilayer perceptron network.

GRAPHICAL MODELS: PROBABILISTIC INFERENCE (q.v.) tells us that many neural network architectures are special cases of the general graphical model formalism that the article presents. Special cases of graphical models include essentially all of the models developed under the rubric of “unsupervised learning,” as well as Boltzmann machines, mixtures of experts, and radial basis function networks. It is argued that many other neural networks, including the classical multilayer perceptron of the present article, can be analyzed profitably from the point of view of graphical models.

Biological Considerations

REINFORCEMENT LEARNING IN MOTOR CONTROL (q.v.) notes the importance of supervised learning in motor control, but stresses that reinforcement learning (in which positive reinforcement signals success on a task and increasing negative reinforcement gives a measure of increasingly poor performance, but where no explicit error signal for the network’s output units is available) is more plausible in many situations involving motor learning; and, indeed, DOPAMINE, ROLES OF (q.v.) shows that certain reinforcement learning methods seem to fit well with the action of dopamine in the brain.

However, RECURRENT NETWORKS: NEUROPHYSIOLOGICAL MODELING (q.v.) argues for the utility of backpropagation as a tool for studying actual networks in the brain. The argument here is that backpropagation provides a means for the computational neuroscientist to adjust the parameters within a given neural network architecture to see whether there is indeed a parameter setting (whose robustness can then be studied) that yields a given type of behavior. The article presents backpropagation not as a model for biological learning, simply as an effective method of obtaining a solution. Biologically plausible learning algorithms will also find similar solutions, but usually take longer. For example, Mazzoni, Andersen, and Jordan (1991) argued that reinforcement learning gave a more biologically plausible learning rule than backpropagation in their study of a network model of cortical area 7a.

But what is the evidence that backpropagation is biologically implausible? HEBBIAN SYNAPTIC PLASTICITY (q.v.) makes a *partial* case for biological plausibility. While conceding that there is no evidence that the backpropagation formula represents actual brain mechanisms, it summarizes new evidence suggesting that activity in one neuron may affect presynaptic neurons, and even neurons presynaptic to those. One might call this *qualitative backpropagation* to stress that the evidence says nothing about the quantitative plausibility of the generalized delta rule. The ability to patch (make local electrode recordings and current injections) at different distances from the soma of a biological neuron has suggested that action potentials propagate back from the soma into the dendrites as well as in the “conventional” direction, from dendrites

to soma (see DENDRITIC PROCESSING). HEBBIAN SYNAPTIC PLASTICITY (q.v.) discusses three distinct mechanisms by which backpropagating spikes can be seen as the “binding signal” emitted by the soma to modify differentially synapses that are active within a precise temporal window. Moreover, the study of identified neurons and synapses in low-density hippocampal cultures has revealed extensive but selective spread of both long-term potentiation (LTP) and long-term depression (LTD) from the site of induction to other synapses in the network (see Bi and Poo, 2001, for a review). LTD induced at synapses between two glutamatergic neurons can spread to other synapses made by divergent outputs of the same presynaptic neuron, to synapses made by other convergent inputs on the same postsynaptic cell, and can even spread in a retrograde direction to depress synapses afferent to the presynaptic neuron (the evidence for qualitative backpropagation). In contrast, LTP can exhibit only lateral spread and backpropagation to the synapses associated with the presynaptic neuron.

Discussion

Backpropagation has provided an effective and widely used architecture for the training of artificial neural networks. We recalled the generalized delta rule for multilayer perceptrons, illustrated its utility with two examples of auto-encoder networks, and showed how the methodology could be extended to recurrent networks. However, statistical analysis showed that backpropagation has problems—a particular example being the likelihood of backpropagation training of multilayer perceptrons getting trapped in plateaus—as well as advantages. We thus provided pointers to stochastic descent methods that avoided these pitfalls, as well as noting extensions of, and alternatives to, backpropagation that can usefully be added to the repertoire of those who train artificial neural networks.

As for biology, we saw that backpropagation may serve as a computational tool to estimate the parameters of a particular biological network even though it does not model the actual learning processes within that network. On the other hand, evidence of “spike backpropagation” provides inspiration for a family of subtle new learning rules that allow the activity of a neuron to affect the neurons presynaptic to its input neurons, but this offers no direct support for the specific formulas of the generalized delta rule. Finally, it should be noted that the modeling and theory summarized in this article are based on neurons with sigmoid outputs. Such units are useful both in artificial neural networks and in connectionist modeling. They can also be considered biological models if their real-valued output is seen to represent a moving-window mean of spiking frequency of the biological neurons (see RATE CODING AND SIGNAL PROCESSING). However, there are cases in which it seems that a better fit to the biology can be obtained if the local temporal structure of spikes in the output of each neuron is taken into account (SPIKING NEURONS, COMPUTATION WITH). This suggests the importance of seeking to define learning rules that do take detailed spike placement, rather than local firing rates, into account. The data reviewed in HEBBIAN SYNAPTIC PLASTICITY (q.v.) may lead brain modelers in the right direction but, unfortunately, no efficient learning algorithm for networks of spiking neurons, whether biological or not, has yet gained wide acceptance.

Road Map: Learning in Artificial Networks

Background: I.3. Dynamics and Adaptation in Neural Networks; Perceptrons, Adalines, and Backpropagation

Related Reading: Computing with Attractors; Hebbian Synaptic Plasticity; Learning and Statistical Inference; Recurrent Networks: Learning Algorithms; Statistical Mechanics of On-Line Learning and Generalization; Stochastic Approximation and Efficient Learning

References

- Amari, S., 1967, Theory of adaptive pattern classifiers, *IEEE Trans. Elec. Comp.*, EC-16:299–307.
- Amari, S., 1998, Natural gradient works efficiently in learning, *Neural Computat.*, 10:251–276.
- Biehler, M., Riegler, P., and Wöhler, C., 1996, Transient dynamics of on-line learning in two-layered neural networks, *J. Phys. A*, 29:4769.
- Bi, G., and Poo, M., 2001, Synaptic modification by correlated activity: Hebb's postulate revisited, *Annu. Rev. Neurosci.*, 24:139–166. ♦
- Elman, J. L., 1990, Finding structure in time, *Cogn. Sci.*, 14:179–212. ♦
- Jordan, M. I., and Jacobs, R. A., 1994, Hierarchical mixtures of experts and the EM algorithm, *Neural Computat.*, 6:181–214.
- Mazzoni, P., Andersen, R. A., and Jordan, M. I., 1991, A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a, *Cereb. Cortex*, 1:293–307.
- McLachlan, G. J., and Krishnan, T., 1997, *The EM Algorithm and Extensions*, New York: Wiley-Interscience.
- Pollack, J. B., 1990, Recursive distributed representations, *Artif. Intell.*, 46:77–105.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986, Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*, (D. E. Rumelhart, J. L. McClelland, and PDP Research Group, Eds.), Cambridge, MA: MIT Press, pp. 318–362. ♦
- Saad, D., and Solla, S. A., 1995, On-line learning in soft committee machines, *Phys. Rev. E*, 52:4225–4243.
- Vapnik, V. N., 1998, *Statistical Learning Theory*, New York: Wiley.
- Werbos, P., 1995, Backpropagation: Basics and new developments, in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, Ed.), Cambridge, MA: MIT Press, pp. 134–139. ♦

Basal Ganglia

Tony J. Prescott, Kevin Gurney, and Peter Redgrave

Introduction

Lying on either side of the forebrain/midbrain boundary, at the hub of the mammalian brain, the basal ganglia are a group of highly interconnected brain structures with a critical influence over movement and cognition. The importance of these nuclei for a cluster of human brain disorders, including Parkinson's disease, Huntington's disease, and schizophrenia, has produced a century or more of strong clinical interest, and a prodigious volume of neurobiological research. Given the wealth of relevant data, and a pressing need for a better functional understanding of these structures, the basal ganglia provide one of the most exciting prospects for computational modeling of brain function.

This article will begin by summarizing aspects of the functional architecture of the mammalian basal ganglia and will then describe the computational approaches that have been developed over the course of the past decade (see also Houk, Davis, and Beiser, 1995; Wickens, 1997; Gillies and Arbuthnott, 2000). An important task for an appraisal of computational models is to provide a framework for comparing pieces of work that can differ radically in their breadth of focus, level of analysis, computational premises, and methodology, and whose relative merits can consequently be difficult to ascertain (I.2. Levels and Styles of Analysis). Here, we first distinguish between models that attempt to incorporate appropriate biological data (anatomical and/or physiological) and those that attempt an explanation of function using generic neural network architectures. This review will discuss only those models that incorporate known neurobiological constraints and will consider some of the implications for these models of recent biological data. The models can be divided in two main categories: (1) those that work at a comparatively low level of detail (membrane properties of individual neurons and micro-anatomical features) and that restrict themselves to a single component of the basal ganglia nucleus; and (2) those that deal at the "system level" with the basal ganglia as a whole and/or with their interactions with related structures (e.g., thalamus and cortex). In this article we will also seek to classify system level models in terms of the primary computational role that is being addressed by the neural substrate.

The neuromodulator dopamine is known to play a vital role in regulating basal ganglia processing and also in mediating learning within the basal ganglia. Although some of the likely regulatory functions of dopamine will be considered in this article, a fuller discussion of this topic, including hypotheses and models con-

cerned with the role of dopamine in learning from reinforcement, are the subject of a separate article (DOPAMINE, ROLES OF).

Key Architectural Features

There have been many excellent summaries of the functional anatomy of the basal ganglia (e.g., Mink, 1996; Smith et al., 1998), the following therefore focuses on those aspects most relevant to understanding the models discussed in this article.

The principle structures of the rodent basal ganglia (Figure 1) are the striatum (consisting of the caudate, the putamen, and the ventral striatum), the subthalamic nucleus (STN), the globus pallidus (GP), the substantia nigra (SN), and the entopeduncular nucleus (EP) (homologous to the globus pallidus internal segment in primates). These structures are massively interconnected and form a functional subsystem within the wider brain architecture. There is a growing consensus that the basal ganglia nuclei can be regionally subdivided on the basis of their topographically organized connectivity with each other and with cortical and thalamic regions. Current views of information processing within the basal ganglia are heavily influenced by this suggestion of multiple parallel loops or channels.

The principle input components of the basal ganglia are the striatum and the STN. Afferent connections to both of these structures originate from virtually the entire brain, including cerebral cortex, many parts of the brainstem (via the thalamus), and the limbic system. Input connections provide phasic (intermittent) excitatory input.

The main output nuclei of the basal ganglia are the substantia nigra pars reticulata (SNr) and the entopeduncular nucleus (EP). Output structures provide extensively branched efferents to the thalamus (which project back to the cerebral cortex), and to premotor areas of the midbrain and brainstem. Most output projections are normally (tonically) active and inhibitory.

To make sense of the intrinsic connectivity of the basal ganglia it is important to recognize that the main projection neurons from the striatum (medium spiny cells) form two widely distributed populations differentiated by their efferent connectivity and neurochemistry.

One population comprises neurons with mainly D1-type dopamine receptors and projects to the output nuclei (SNr and EP). In the prevailing informal model of the basal ganglia (Albin, Young, and Penney, 1989) this projection constitutes the so-called *direct*

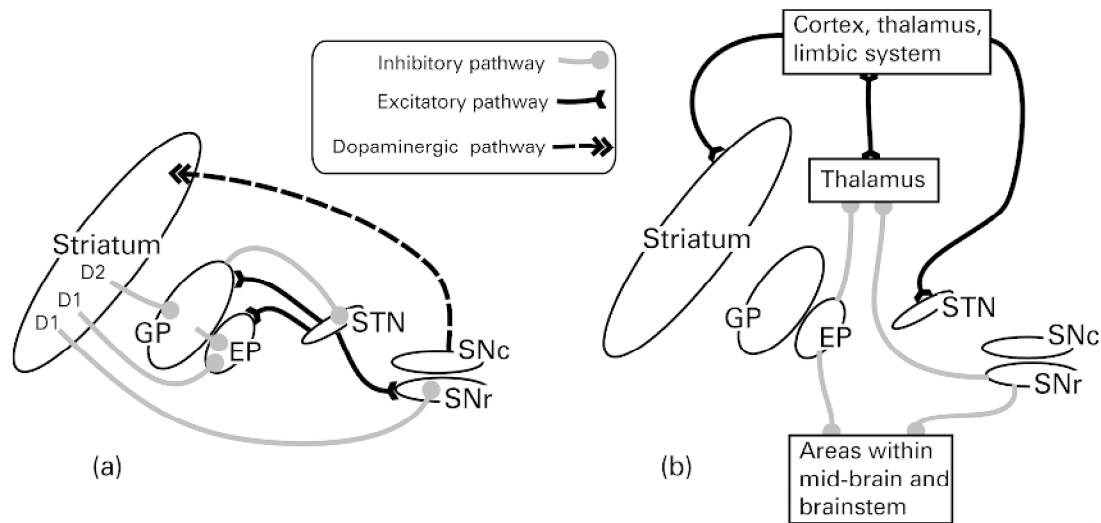


Figure 1. Basal ganglia anatomy of the rat: *A.* Internal pathways. *B.* External pathways. Excitatory and inhibitory pathways are denoted by solid and gray lines, respectively; not all connections are shown. See text for key to abbreviations.

pathway to the output nuclei (see Figure 2A). Efferent activity from these neurons suppresses the tonic inhibitory firing in the output structures, which in turn *disinhibits* targets in the thalamus and brainstem.

A second population of striatal output neurons has predominantly D2-type dopamine receptors. This group projects primarily to the globus pallidus (GP) whose tonic inhibitory outputs are directed both to the output nuclei (SNr and EP) and to the STN. The inhibitory projection from D2 striatal neurons constitutes the first leg of an *indirect pathway* to the output nuclei. Since this pathway has two inhibitory links (Striatum-GP, GP-STN), followed by an excitatory one (STN-EP/SNr), the net effect of striatal activity is

to activate output nuclei, which increases inhibitory control of the thalamus and brainstem.

The main source of dopamine innervation to the striatum is the substantia nigra pars compacta (SNc). Interestingly, the D1 and D2 striatal populations respond differently to dopaminergic transmission, activation of D1 receptors having a predominantly excitatory effect while D2 receptor activation appears to be mainly inhibitory. This arrangement seems to provide dopaminergic control of a “push/pull” mechanism subserved by the direct (inhibitory) and indirect (net excitatory) basal ganglia pathways. Importantly, a key input to the SNc is from striatal areas known as *striosomes* (areas that project to EP/SNr are known as *matri-*

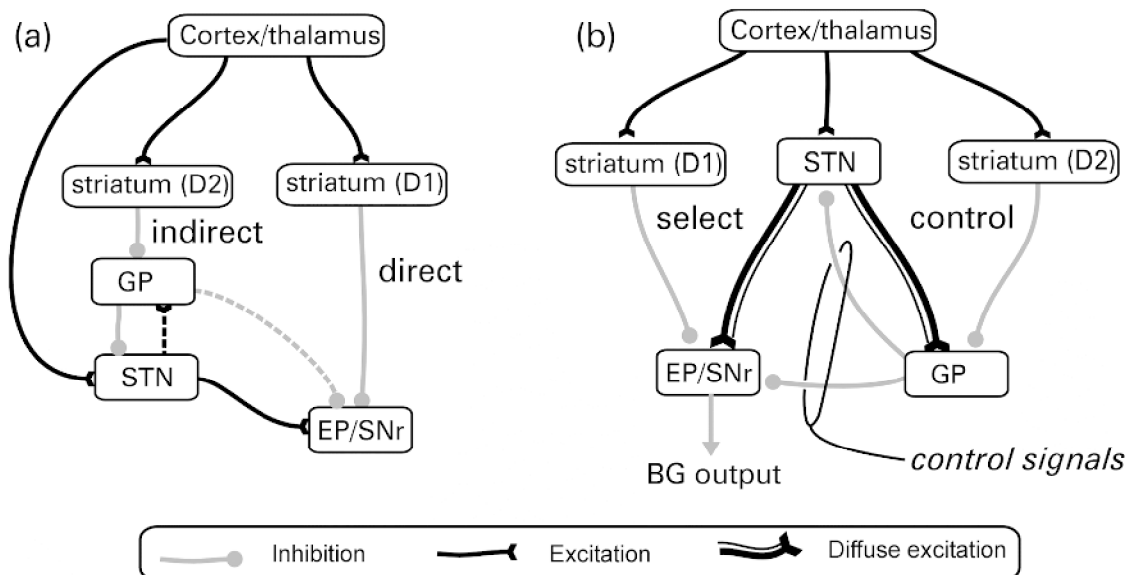


Figure 2. Functional interpretations of the basal ganglia: (a) *Informal* models stress the “direct” and “indirect” pathways and leave the functional consequences of their interactions ill-defined. Other pathways (indicated by dotted lines) have received less emphasis. (b) An alternative interpretation

arising from *computational* modeling by Gurney et al. (2001) specifies specific functional roles for the various intrinsic basal ganglia connections summarized by the concept of “selection” and “control” pathways. See text for further explanation.

somes), thus the striatum is a major player in modulating its own dopaminergic input.

Although the preceding description focuses on pathways originating from the striatum, the STN, though much smaller in size, is increasingly recognized as a second important input structure within the basal ganglia functional architecture (see Mink, 1996). STN's excitatory outputs project to both the output nuclei (SNr and EP) and to the intermediary structure GP.

Recent anatomical data by Wu, Richard, and Parent (2000) has suggested that the "direct pathway" is actually branched with a significant output going to GP. Other new data shows the existence of additional inhibitory projections from GP to EP/SNr implying a multiplicity of indirect pathways (Smith et al., 1998). The proliferation of intrinsic basal ganglia circuitry in recent literature has highlighted the need for: (1) a radical reinterpretation of basal ganglia functional anatomy; and (2) an increasingly important role for computational modeling in interpreting the functional properties of the multiple interconnections and loops within the basal ganglia.

Low-Level Models of Individual Basal Ganglia Nuclei

In contrast to the diverse nature of the brain regions projecting to the mammalian striatum, its internal organization appears surprisingly homogeneous. This finding offers hope that an understanding of striatal functioning in one local area could generalize across much of the entire structure. At any given moment the majority of striatal cells are in an inactive "down state," and can only be triggered into an active "up state" (in which they can fire action potentials) by a significant amount of coincident input. Since each neuron has a wide dendritic fan-in (with up to 30,000 synapses), but only a few synapses with any single source neuron, it must receive coincident signals from a large population of inputs to become active (see Wilson in Houk et al., 1995). This organization suggests that striatal spiny neurons may act as "context-specific filters," each one configured to match a specific pattern of activity distributed across multiple loci in one or more brain areas (Mink, 1996).

Recent studies have provided evidence for local inhibition within the striatum mediated either via local interneurons or by reciprocal inhibitory networks among the output cells themselves (see Oorschot et al. in Nicholson and Faull, 2002). Wickens and colleagues (see Wickens, 1997) have investigated the dynamics of such local neighborhoods of striatal neurons using network models. Under varying assumptions of topology and size, they concluded that reciprocal inhibition will usually lead to a network dynamic of competition, that is, the most active neurons will tend to suppress activity in their less active neighbors. This research also examined the effects of simulated dopamine inputs, showing that under circumstances of low dopamine, the dynamic of the network changes from competition to coactivation (where activity is uniformly distributed within the local population of neurons), a pattern that could provide a model for the muscular rigidity seen in dopamine-deficient Parkinson's patients. Using another variant of this model, Wickens explored the implications of dendritic asymmetries based on those observed in the early stages of Huntington's disease. Simulation of asymmetric interconnectivity generated slow traveling waves of activity (where normal symmetric configurations produce stationary activity patterns), suggesting that a similar abnormal network dynamic may underlie the sudden involuntary movements seen in Huntington's patients.

Apart from the striatum, relatively little attention has been given to modeling intrinsic processing within basal ganglia nuclei. One interesting exception is the work by Gillies and Willshaw (see Gillies and Arbuthnott, 2000) on a model of the STN. Having incorporated key physiological and anatomical properties, they showed that the widespread excitatory interconnectivity between STN neurons allows focused input to produce a widely distributed pulse of

excitation to SNr and EP. Given that the output of the basal ganglia is largely inhibitory, phasic STN activity could serve to break established patterns of activity in basal ganglia targets thereby acting as a form of interrupt or "reset" mechanism.

The preceding models have as their starting point a wealth of low-level biological constraints with the rationale that the resulting model behavior must approximate observed biological data. Nevertheless, since the phenomena discussed are related to the ability to resolve localized competitions (in striatum) and to interrupt ongoing behaviors (STN), we would argue that these models may be thought of as addressing components of the overall computational problem of action selection (see further discussion later in this article).

System Level Models of Basal Ganglia Circuits and External Functional Loops

Most of the effort so far directed at basal ganglia modeling has been concerned with simulating interactions between the various basal ganglia structures, and between the basal ganglia and other key brain regions such as cortex, thalamus, and brainstem. A comparatively high level of abstraction is usually adopted in this work, in which components of the basal ganglia are decomposed into functional units (e.g., multiple parallel channels). Most work to date has focused on a number of related computational hypotheses—that the basal ganglia function to (1) regulate the degree of action gating, (2) select between competing actions, (3) sustain working memory representations, and (4) store and enact sequences of behavior. These ideas will be the main focus of the remaining discussion.

Action Gating

A key function of the striatum is to provide intermittent, focused inhibition (via the "direct pathway") within output structures that otherwise maintain inhibitory control over motor/cognitive systems throughout the brain. This architecture strongly suggests that a core function of the basal ganglia is to gate the activity of these target systems via the mechanism of disinhibition. Many basal ganglia models employ selective gating, however, that of Contreras-Vidal and Stelmach (1995) is particularly interesting as it explores gating operations in both normal and dysfunctional model variants. These authors coupled a simulation of basal ganglia intrinsic circuitry to a neural network that computed arm movements. Excitatory striatal input resulted in a smoothly varying signal to thalamic targets that provided the "Go" signal for the motor command, and also set its overall velocity. The time taken to execute movements decreased with increasing basal ganglia input thereby matching the results of striatal microstimulation studies. A "dopamine depleted" version of the model exhibited akinesia and bradykinesia similar to that observed in Parkinson's disease.

Selecting Between Competing Actions

The proposal that the basal ganglia act to resolve action selection competitions is based on a growing consensus that a key function of these structures is to arbitrate between sensorimotor systems competing for access to the final common motor path. A computational hypothesis developed from this idea relies on the premise that afferent signals to the striatum encode the salience of "requests for access" to the motor system (Redgrave, Prescott, and Gurney, 1999). Multiple selection mechanisms embedded in the basal ganglia could resolve conflict between competitors and provide clean and rapid switching between winners. First, the up/down states of the striatal neurons may act as a first pass filter to exclude weakly supported "requests." Second, local inhibition within the striatum could selectively enhance the activity of the most salient channels.

Third, the combination of focused inhibition from striatum with diffuse (divergent) excitation from STN could operate as a feedforward, off-center/on-surround network across the basal ganglia as a whole (see Mink, 1996). Last, local reciprocal inhibition within the output nuclei could sharpen up the final selections.

Using the action selection hypothesis as an organizing principle, Gurney, Prescott, and Redgrave (2001) have proposed a reinterpretation of basal ganglia functional anatomy in which the direct/indirect classification is replaced by a new functional grouping based on *selection* and *control* circuits (Figure 2B). Specifically, the focused D1 inhibitory pathway from striatum to EP/SNr (originally the “direct pathway”), together with a *diffuse* excitatory pathway from STN to EP/SNr, form a primary feedforward *selection* circuit. A second group of intrinsic connections centered on the GP acts as a *control* circuit to regulate the performance of the main selection mechanism. Analytical and simulation studies of this model suggest two likely functional roles for this control circuit. First, the inhibition of STN by GP constitutes a negative feedback path that automatically scales the excitatory output of the STN with the number of channels. Second, GP inhibition of EP/SNr forms part of a mechanism that supports dopaminergic regulation of selection. Specifically, increased dopamine in these circuits promotes “promiscuous” selection in which channels are more easily disinhibited, while reduced dopamine results in a “stiffer” competition in which there are fewer winners and higher levels of general target inhibition. The adequacy of this model has been tested by embedding it in the control architecture of a mobile robot equipped with a small repertoire of animal-like behaviors (see Prescott, Gurney et al. in Nicholson and Faull, 2002). This work confirmed that the simulated basal ganglia can provide effective action selection in a real-world context requiring appropriate and timely behavioral switching. The robot model also provided an insight into the emergent consequences of abnormal dopamine modulation of action selection. For instance, and reminiscent of some motor symptoms of Parkinson’s disease, reduced dopamine was found to cause failures to select appropriate behavior or to complete behaviors once selected.

An earlier model of the basal ganglia proposed by Berns and Sejnowski (1995) shared the “action selection” premise of Gurney et al., but emphasized possible timing differences between the direct and indirect pathways in a model that included just the feedforward intrinsic basal ganglia connections. An interesting feature of this model is that it incorporated a version of the dopamine hypothesis for reinforcement learning (DOPAMINE, ROLES OF) as a means for adaptively tuning the selection mechanism.

Sustaining Working Memory Representations

The relationship between basal ganglia and cortex is characterized by relatively segregated parallel loops, in which cortical projections to the striatum are channeled through basal ganglia outputs to the thalamus and then back to their cortical areas of origin. The thalamic nuclei in this circuit also have reciprocal, net-excitatory, connections to their cortical targets. This architecture suggests a pattern of cortical-thalamic activity which, once initiated by disinhibitory signals from the basal ganglia, could be sustained indefinitely. Several authors have proposed that this circuit could act as a working memory store (see Houk et al., 1995). An example of this is provided by Arbib and Dominey’s model of basal ganglia control of the primate saccadic eye movement system (see article in Houk et al., 1995). These authors modeled an experimental task in which a monkey is required to make a saccade to a remembered target location. They simulated circuits in which cortical cells in the frontal eye fields were activated by the target, which, in turn, excited a population of striatal neurons specialized for delayed saccades. The basal ganglia loops involving these cells disinhibited their thalamic

targets so that the target location was maintained in the cortico-thalamic circuits until the saccade was made.

In our view, the selection and maintenance of specific working memory items can be viewed as an extension of action selection by the basal ganglia to the domain of cognition (selecting from a range of potential cognitive representations those which are to be sustained as working memory). It is interesting to speculate that deficits in this system may underlie the disorders of thought associated with schizophrenia, attention-deficit disorder, and obsessive-compulsive disorder.

Sequence Processing

A plausible use for the working memory mechanism outlined previously would be to link successful selections during the development of behavioral/cognitive sequences. This idea has therefore become a central theme in a number of basal ganglia models. According to Beiser and Houk (1998), sequence encoding can be viewed as the task of translating a temporal ordering into a spatial pattern of neural activity. They propose that the initial item in a sequence selects the basal ganglia loop whose striatal neurons are most attuned to that context. When this channel is disinhibited, the item then becomes encoded as a self-sustaining pattern of cortico-thalamic activity. Later sequence elements are recorded in an identical way, except that, as each new item is added, the cortical activity triggered by its predecessors becomes part of its context (thereby implicitly encoding its position in the temporal order). Rather than recording sequences as spatially distributed patterns, Fukai (1999) has suggested a form of cortical short-term memory for sequences that uses patterns of fast (gamma) and slow (theta) oscillatory activity. Reciprocal inhibition between striatal neurons would allow the basal ganglia to select the first item in such a sequence, while other striatal neurons (and their corresponding basal ganglia outputs) would be recruited to maintain the selection of that item for as long as required. Finally, an excitatory burst from STN terminates the movement and signals the transition to the next item in the sequence. Although differing considerably in detail, these two models share the premise that the basal ganglia is specialized to “unpack” a cortical representation of sequential behavior by selectively gating each of the component movements. Sequence learning is another important theme in basal ganglia modeling. For instance, Dominey, Arbib, and Joseph (1995) have extended their model of delayed saccade control (described previously) to include a mechanism for associative and SEQUENCE LEARNING (q.v.) based, again, on the hypothesis that dopamine provides a reinforcement learning signal.

Discussion

The preceding summary demonstrates that basal ganglia modeling is still at the stage of exploring the space of alternative hypotheses, seeking to operationalize theoretical proposals while trying to match known neurobiological constraints. As a result, there is now a candidate set of “global” basal ganglia functions whose computational requirements we are beginning to understand. It remains to be seen to what extent proposed functions are mutually exclusive and to what extent one may be subsumed within another (for instance, action gating can be viewed as an essential component of action selection). Similar considerations apply when appraising models directed at different levels of basal ganglia function. For example, lower-level models of the striatum or STN may, in the future, be imported as fully functional components into higher-level models. However, some system level models are clearly in direct competition with each other as they ascribe different functional roles to local pathways and nuclei. We anticipate that models based on correct computational assumptions will find it comparatively easy to incorporate new biological constraints, which in most

cases will improve their accuracy. In contrast, models making mistaken functional assignments will find it increasingly difficult to incorporate additional biological data while maintaining their functionality. Future work will therefore require ever-closer links between neurobiologists and modelers to refine the models, to formulate questions based on function, and to test the interesting and unforeseen predictions that can emerge from modeling studies.

Road Maps: Mammalian Brain Regions; Mammalian Motor Control

Related Reading: Action Monitoring and Forward Control of Movements; Arm and Hand Movement Control; Dopamine, Roles of; Motor Control, Biological and Theoretical; Reinforcement Learning in Motor Control

References

- Albin, R. L., Young, A. B., and Penney, J. B., 1989, The functional anatomy of basal ganglia disorders, *Trends Neurosci.*, 12(10):366–375.
- Beiser, D. G., and Houk, J. C., 1998, Model of cortical-basal ganglionic processing: Encoding the serial order of sensory events, *J. Neurophysiol.*, 79(6):3168–3188.
- Berns, G. S., and Sejnowski, T. J., 1995, How the basal ganglia make decisions, in *The Neurobiology of Decision Making* (A. Damasio, H. Damasio, and Y. Christen, Eds.), Berlin: Springer-Verlag, pp. 101–113.
- Contreras-Vidal, J. L., and Stelmach, G. E., 1995, A neural model of basal ganglia-thalamocortical relations in normal and Parkinsonian movement, *Biol. Cybernetics*, 73(5):467–476.
- Dominey, P., Arbib, M., and Joseph, J.-P., 1995, A model of corticostriatal plasticity for learning oculomotor associations and sequences, *J. Cognit. Neurosci.*, 7(3):311–336.
- Fukui, T., 1999, Sequence generation in arbitrary temporal patterns from theta-nested gamma oscillations: A model of the basal ganglia-thalamocortical loops, *Neural Networks*, 12(7–8):975–987.
- Gillies, A., and Arbuthnott, G., 2000, Computational models of the basal ganglia, *Movement Disorders*, 15(5):762–770. ♦
- Gurney, K., Prescott, T. J., and Redgrave, P., 2001, A computational model of action selection in the basal ganglia. I, II, *Biological Cybernetics*, 84(6):401–423.
- Houk, J. C., Davis, J. L., and Beiser, D. G., 1995, *Models of Information Processing in the Basal Ganglia*, Cambridge, MA: MIT Press. ♦
- Mink, J. W., 1996, The basal ganglia: Focused selection and inhibition of competing motor programs, *Progr. Neurobiol.*, 50(4):381–425. ♦
- Nicholson, L. F. B., and Faull, R. L. M., 2002, *Basal Ganglia VII*, New York: Plenum Press.
- Redgrave, P., Prescott, T., and Gurney, K., 1999, The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.
- Smith, Y., Bevan, M. D., Shink, E., and Bolam, J. P., 1998, Microcircuitry of the direct and indirect pathways of the basal ganglia, *Neuroscience*, 86:353–387. ♦
- Wickens, J., 1997, Basal ganglia: Structure and computations. *Network-Computation in Neural Systems*, 8(4):R77–R109. ♦
- Wu, Y., Richard, S., and Parent, A., 2000, The organization of the striatal output system: a single-cell juxtacellular labeling study in the rat, *Neurosci. Res.*, 38:49–62.

Bayesian Methods and Neural Networks

David Barber

Introduction

An attractive feature of artificial neural networks is their ability to model highly complex, nonlinear relationships in data. However, choosing an appropriate neural network model for data is compounded by the difficulty of assessing the network's complexity. Since we are rarely certain about either our data measurements or model beliefs, a natural framework is to use probabilities to account for these uncertainties. How can we combine our data observations with these modeling uncertainties in a consistent and meaningful manner? The Bayesian approach provides a consistent framework for formulating a response to these difficulties, and is noteworthy for its conceptual elegance (Box and Tiao, 1973; Berger, 1985; MacKay, 1992). The fundamental probabilistic relationship required for inference is the celebrated Bayes rule, which, for general events A, B, C , is

$$p(A|B, C) = \frac{p(B|A, C)p(A|C)}{p(B|C)} \quad (1)$$

It is convenient to think of different levels of uncertainty in formulating a model. At the lowest level, we may assume that we have the correct model but are uncertain as to the parameter settings θ for this model. This assumption details how observed data are generated, $p(\text{data}|\theta, \text{model})$. The task of inference at this level is to calculate the posterior distribution of the model parameter. Using Bayes's rule, this is

$$p(\theta|\text{data}, \text{model}) = \frac{p(\text{data}|\theta, \text{model}) p(\theta|\text{model})}{p(\text{data}|\text{model})} \quad (2)$$

Thus, if we wish to infer model parameters from data, we need two assumptions: (1) how the observed data are generated under the assumed model, or the *likelihood* $p(\text{data}|\theta, \text{model})$, and (2) beliefs about which parameter values are appropriate before the data

have been observed, or the *prior* $p(\theta|\text{model})$. (The denominator in Equation 2 is the normalizing constant for the posterior and plays a role in uncertainty at the higher model level.) That these two assumptions are required is an inescapable consequence of Bayes's rule, and forces the Bayesian to lay bare all necessary assumptions underlying the model.

Coin Tossing Example

Let θ be the probability that a coin will land heads up. An experiment yields the data, $D = \{h, h, t, h, t, h, \dots\}$, which contains H heads and T tails in $H + T$ flips of the coin. What can we infer about θ from these data? Assuming that each coin is flipped independently, the likelihood of the observed data is

$$p(D|\theta, \text{model}) = \theta^H (1 - \theta)^T \quad (3)$$

A standard approach in the statistical sciences is to estimate θ by maximizing the likelihood, $\theta^{ML} = \arg \max_{\theta} p(D|\theta, \text{model})$. This approach is non-Bayesian, since it does not require the specification of a prior. Consequently, theories that deal with uncertainty in ML estimators are primarily concerned with the data likelihood, and not directly with posterior parameter uncertainty (see LEARNING AND GENERALIZATION: THEORETICAL BOUNDS). In the Bayesian approach, however, we need to be explicit about our prior beliefs $p(\theta|\text{model})$. These are updated by the observed data to yield the posterior distribution

$$p(\theta|D, \text{model}) \propto \theta^H (1 - \theta)^T p(\theta|\text{model}) \quad (4)$$

The Bayesian approach is more flexible than the maximum likelihood approach since it allows (indeed, *instructs*) the user to calculate the effect that the data have in modifying prior assumptions about which parameter values are appropriate. For example, if we believe that the coin is heavily biased, we may express this using

the prior distribution in Figure 1A. The likelihood as a function of θ is plotted in Figure 1B for data containing 13 tails and 12 heads. The resulting posterior (Figure 1C) is bimodal, but less extreme than the prior. It is often convenient to summarize the posterior by either the maximum a posteriori (MAP) value or the mean, $\bar{\theta} = \int \theta p(\theta|D) d\theta$. Such a summary is not strictly required by the Bayesian framework, and the best choice of how to summarize the posterior depends on other loss criteria (Berger, 1985).

Model Comparison and Hierarchical Models

The preceding discussion showed how we can use the Bayesian framework to assess which parameters of a model are a posteriori appropriate, given the data at hand. We can carry out a similar procedure at a higher, model level to assess which models are more appropriate fits to the data. In general, the model posterior is given by

$$p(M|D) = \underbrace{p(D|M)}_{\text{Model likelihood}} \underbrace{p(M)}_{\text{Model prior}} / p(D) \quad (5)$$

If the model is parameterized by some unknown variable θ , we need to integrate this out to calculate the model likelihood

$$p(D|M) = \int p(D|\theta, M) p(\theta|M) d\theta \quad (6)$$

Comparing two competing model hypotheses, M_1 and M_2 , is straightforward:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)} \frac{p(M_1)}{p(M_2)} \quad (7)$$

Bayes factor

In the coin example, we can use this to compare the biased coin hypothesis (model M_1 with prior given in Figure 1A) with a less unbiased hypothesis formed by using a Gaussian prior $p(\theta|M_2)$ with mean 0.5 and variance 0.1² (model M_2). This gives a Bayes factor $p(D|M_1)/p(D|M_2) \approx 0.00018$. If we have no prior preference for either model M_1 or M_2 , the data more strongly favor model M_2 , as

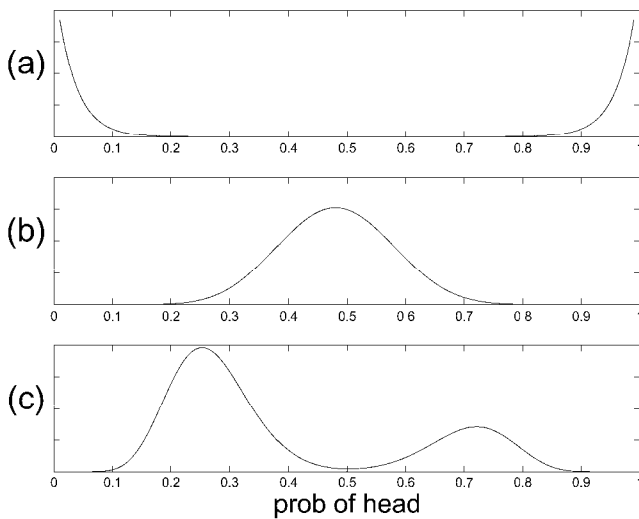


Figure 1. Coin tossing. *A*, The prior: this indicates our belief that the coin is heavily biased. *B*, The likelihood after 13 tails and 12 heads are recorded, $\theta^{ML} = 0.48$. *C*, The posterior: the data have moderated the strong prior beliefs, resulting in a posterior less certain that the coin is biased. $\theta^{MAP} = 0.25$, $\bar{\theta} = 0.39$.

intuition would suggest. If we desired, we could continue in this way, forming a hierarchy of models, each less constrained than the submodels it contains.

Bayesian Regression

Neural networks are often applied to a regression in which we wish to infer an unknown input-output mapping on the basis of observed data $D = \{(\mathbf{x}^\mu, t^\mu), \mu = 1, \dots, P\}$, where (\mathbf{x}^μ, t^μ) represents an input-output pair. For example, fit a function to the data in Figure 2A. Since there is the possibility that each observed output t^μ has been corrupted by noise, we would like to recover the underlying clean input-output function. We assume that each (clean) output is generated from the model $f(\mathbf{x}; \mathbf{w})$, where the parameters \mathbf{w} of the function f are unknown, and that the observed outputs t^μ are generated by the addition of noise η to the clean model output,

$$t = f(\mathbf{x}; \mathbf{w}) + \eta \quad (8)$$

If the noise is Gaussian distributed, $\eta \sim N(0, \sigma^2)$, the model M generates an output t for input \mathbf{x} with probability

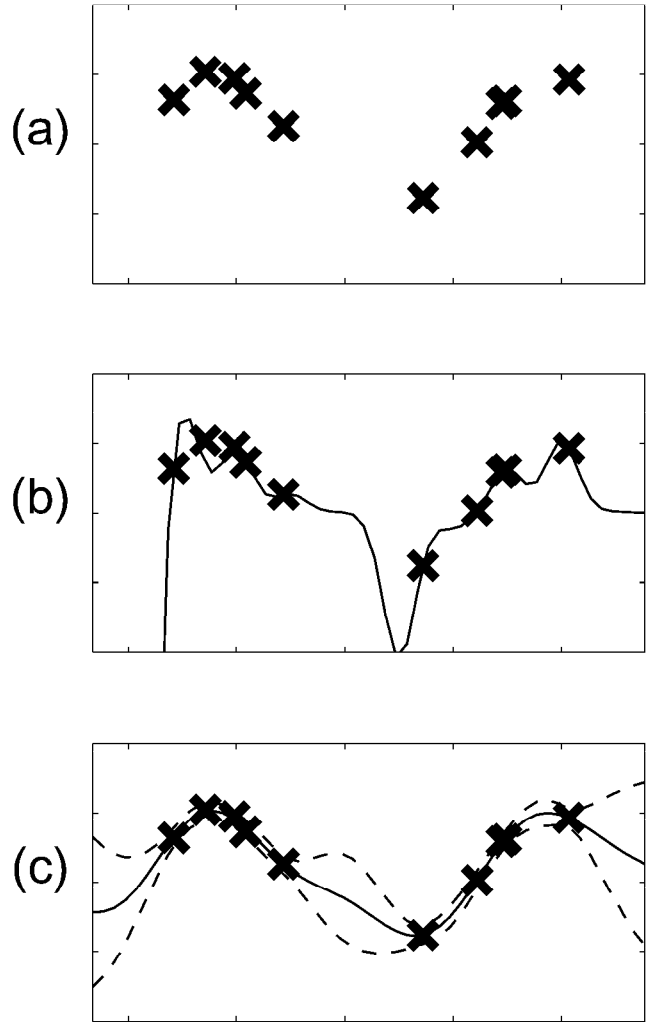


Figure 2. Along the horizontal axis we plot the input x and along the vertical axis the output t . *A*, The raw input-output training data. *B*, Prediction using regularized training and fixed hyperparameters. *C*, Prediction with error bars, using ML-II optimized hyperparameters.

$$p(t|\mathbf{w}, \mathbf{x}, M) = \exp\left[-\frac{1}{2\sigma^2} (t - f(\mathbf{x}; \mathbf{w}))^2\right] / \sqrt{2\pi\sigma^2} \quad (9)$$

If we assume that each data input-output pair is generated identically and independently from the others, the data likelihood is

$$p(D|\mathbf{w}, M) = \prod_{\mu=1}^P p(t^\mu|\mathbf{w}, \mathbf{x}^\mu, M) \quad (10)$$

(Strictly speaking, we should write $p(t^1, \dots, t^P|\mathbf{w}, \mathbf{x}^1, \dots, \mathbf{x}^P, M)$ on the left-hand side of Equation 10. However, since we assume that the training inputs are fixed and non-noisy, it is convenient and conventional to write $p(D|\mathbf{w}, M)$.) The posterior distribution $p(\mathbf{w}|D, M) \propto p(D|\mathbf{w}, M)p(\mathbf{w}|M)$ is

$$\begin{aligned} \log p(\mathbf{w}|D, M) &= -\frac{\beta}{2} \sum_{\mu} (t^\mu - f(\mathbf{x}^\mu; \mathbf{w}))^2 \\ &\quad + \log p(\mathbf{w}|M) + \frac{P}{2} \log \beta + \text{const.} \end{aligned} \quad (11)$$

where $\beta = 1/\sigma^2$. Note the similarity between Equation 11 and the sum square regularized training error used in standard approaches to training neural networks (see GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS and Bishop, 1995). In the Bayesian framework, we can motivate the choice of a sum square error measure as equivalent to the assumption of additive Gaussian noise. Typically, we wish to encourage smoother functions so that the phenomenon of overfitting is avoided. One approach to solving this problem is to use a regularizer penalty term to the training error. In the Bayesian framework, we use a prior to achieve a similar effect. In principle, however, the Bayesian should make use of the full posterior distribution, not just a single weight value. In standard neural network training, it is good practice to use committees of networks, rather than relying on the prediction of a single network (Bishop, 1995). In the Bayesian framework, the posterior automatically specifies a committee (indeed, a distribution) of networks, and the importance attached to each committee member's prediction is simply the posterior probability of that network's weight.

Radial Basis Functions and Generalized Linear Models

Generalized linear models have the form

$$f(\mathbf{x}; \mathbf{w}) = \sum_i w_i \phi_i(\mathbf{x}) \equiv \mathbf{w}^T \Phi(\mathbf{x}) \quad (12)$$

Such models have a linear parameter dependence, but nevertheless represent a nonlinear input-output mapping if the basis functions $\phi_i(\mathbf{x})$, $i = 1, \dots, k$ are nonlinear. Radial basis functions (see RADIAL BASIS FUNCTION NETWORKS) are an example of such a network (Bishop, 1995). A popular choice is to use Gaussian basis functions $\phi_i(\mathbf{x}) = \exp(-(x - \mu^i)^2/(2\lambda^2))$. In this discussion, we will assume that the centers μ^i are fixed, but that the width of the basis functions λ is a hyperparameter that can be adapted. Since the output is linearly dependent on \mathbf{w} , we can discourage extreme output values by penalizing large weight values. A sensible weight prior is thus

$$\log p(\mathbf{w}|\alpha) = -\frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \frac{k}{2} \log \alpha + \text{const.} \quad (13)$$

Under the Gaussian noise assumption, the posterior distribution is

$$\begin{aligned} \log p(\mathbf{w}|\Gamma, D) &= -\frac{\beta}{2} \sum_{\mu=1}^P (t^\mu - \mathbf{w}^T \Phi(\mathbf{x}^\mu))^2 \\ &\quad - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{const.} \end{aligned} \quad (14)$$

where Γ represents the hyperparameter set $\{\alpha, \beta, \lambda\}$. (We drop the fixed model dependency wherever convenient.) The weight posterior is therefore a Gaussian, $p(\mathbf{w}|\Gamma, D) = N(\bar{\mathbf{w}}, \mathbf{S})$, where

$$\begin{aligned} \mathbf{S} &= \left(\alpha \mathbf{I} + \beta \sum_{\mu=1}^P \Phi(\mathbf{x}^\mu) \Phi^T(\mathbf{x}^\mu) \right)^{-1} \\ \bar{\mathbf{w}} &= \beta \mathbf{S} \sum_{\mu=1}^P t^\mu \Phi(\mathbf{x}^\mu) \end{aligned} \quad (15)$$

The mean predictor is straightforward to calculate: $\bar{f}(\mathbf{x}) \equiv \int f(\mathbf{x}; \mathbf{w}) p(\mathbf{w}|D, \Gamma) d\mathbf{w} = \bar{\mathbf{w}}^T \Phi(\mathbf{x})$. Similarly, error bars are straightforward, $\text{var}(f(\mathbf{x})) = \Phi(\mathbf{x})^T \mathbf{S} \Phi(\mathbf{x})$ (predictive standard errors are given by $\sqrt{\text{var}(f) + \sigma^2}$). In Figure 2B, we show the mean prediction on the data in Figure 2A using 15 Gaussian basis functions with width $\lambda = 0.03$ spread out evenly over the input space. We set the other hyperparameters to be $\beta = 100$ and $\alpha = 1$. The prediction severely overfits the data, a result of poor choice of hyperparameters.

Determining Hyperparameters: ML-II

How would the mean predictor be calculated if we were to include the hyperparameters Γ as part of a hierarchical model? Formally, this becomes

$$\begin{aligned} \bar{f}(\mathbf{x}) &= \int f(\mathbf{x}; \mathbf{w}) p(\mathbf{w}, \Gamma|D) d\mathbf{w} d\Gamma \\ &= \int \left\{ \int f(\mathbf{x}; \mathbf{w}) p(\mathbf{w}|\Gamma, D) d\mathbf{w} \right\} p(\Gamma|D) d\Gamma \end{aligned} \quad (16)$$

The term in curly brackets is the mean predictor for fixed hyperparameters. We therefore weigh each mean predictor by the posterior probability of the hyperparameter $p(\Gamma|D)$. Equation 16 shows how to combine different models in an ensemble—each model prediction is weighted by the posterior probability of the model. There are other non-Bayesian approaches to model combinations in which the determination of the combination coefficients is motivated heuristically (see ENSEMBLE LEARNING).

Provided the hyperparameters are well determined by the data, we may instead approximate the above hyperparameter integral by finding the MAP hyperparameters $\Gamma^* = \arg \max_{\Gamma} p(\Gamma|D)$. Since $p(\Gamma|D) = p(D|\Gamma)p(\Gamma)/p(D)$, if the prior belief about the hyperparameters is weak ($p(\Gamma) \approx \text{const.}$), we can estimate the optimal hyperparameters by optimizing the hyperparameter likelihood

$$p(D|\Gamma) = \int p(D|\Gamma, \mathbf{w}) p(\mathbf{w}|\Gamma) d\mathbf{w} \quad (17)$$

This approach to setting hyperparameters is called ML-II (Berger, 1985; Bishop, 1995) and assumes that we can calculate the integral in Equation 17. In the case of GLMs, this involves only Gaussian integration, giving

$$\begin{aligned} 2 \log p(D|\Gamma) &= -\beta \sum_{\mu=1}^P (t^\mu)^2 + \mathbf{d}^T \mathbf{S}^{-1} \mathbf{d} - \log |\mathbf{S}| \\ &\quad + k \log \alpha + P \log \beta + \text{const.} \end{aligned} \quad (18)$$

where $\mathbf{d} = \beta \sum_{\mu} \Phi(\mathbf{x}^\mu) t^\mu$. Using the hyperparameters α, β, λ that optimize the above expression gives the results in Figure 2C, where we plot both the mean predictions and standard predictive error bars. This solution is more acceptable than the previous one in which the hyperparameters were not optimized, and demonstrates that overfitting is avoided automatically. A non-Bayesian approach to model fitting based on minimizing a regularized training error would typically use a procedure such as cross-validation to determine the regularization parameters (hyperparameters). Such approaches require the use of validation data (Bishop, 1995). An advantage of the Bayesian approach is that hyperparameters can be

set without the need for validation data, and thus all the data can be used directly for training.

Relation to Gaussian Processes

The use of GLMs can be difficult in cases where the input dimension is high, since the number of basis functions required to cover the input space fairly well grows exponentially with the input dimension—the so-called *curse of dimensionality* (Bishop, 1995). If we specify n points of interest \mathbf{x}^i , $i \in 1, \dots, n$ in the input space, the GLM specifies an n -dimensional Gaussian distribution on the function values f_1, \dots, f_n with mean $\bar{f}_i = \bar{\mathbf{w}}^T \Phi(\mathbf{x}^i)$ and covariance matrix with elements $c_{ij} = c(\mathbf{x}^i, \mathbf{x}^j) = \Phi(\mathbf{x}^i)^T \Sigma \Phi(\mathbf{x}^j)$ (see GAUSSIAN PROCESSES). The idea behind a GP is that we can free ourselves from the restriction of choosing a covariance function $c(\mathbf{x}^i, \mathbf{x}^j)$ of the form provided by the GLM prior; any valid covariance function can be used instead. Similarly, we are free to choose the mean function $\bar{f}_i = m(\mathbf{x}^i)$. A common choice for the covariance function is $c(\mathbf{x}^i, \mathbf{x}^j) = \exp(-|\mathbf{x}^i - \mathbf{x}^j|^2)$. The motivation is that the function space distribution will have the property that for inputs \mathbf{x}^i and \mathbf{x}^j , which are close together, the outputs $f(\mathbf{x}^i)$ and $f(\mathbf{x}^j)$ will be highly correlated, ensuring smoothness. This is one way of avoiding the curse of dimensionality, since the matrix dimensions depend on the number of training points, and not on the number of basis functions used. However, for problems with a large number of training points, computational difficulties can arise, and approximations again need to be considered.

Multilayer Perceptrons

Consider the case of a single hidden layer neural network

$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^H v_i g(\mathbf{x}^T \mathbf{u}^i + b_i) \quad (19)$$

where $g(x)$ is a nonlinear sigmoidal transfer function, for example $g(x) = 1/(1 + \exp(-x))$. The set of all weights (parameters), including input-hidden weights \mathbf{u}^i , biases b_i , and hidden-output weights v_i , is represented by the vector \mathbf{w} . If the weights are small, the network function f will be smooth, since only the near linear regime of the transfer function g will be accessed. An appropriate prior to control complexity is therefore

$$\log p(\mathbf{w}|\alpha) = -\frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \frac{k}{2} \log \alpha + \text{const.} \quad (20)$$

where $k = \dim(\mathbf{w})$. For the moment, we will assume that we know the value of the parameter α . This gives the weight posterior as

$$\begin{aligned} \log p(\mathbf{w}|\alpha, \beta, D) = & -\frac{\beta}{2} \sum_{\mu=1}^P (t^\mu - f(\mathbf{x}^\mu; \mathbf{w}))^2 \\ & - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{const.} \end{aligned} \quad (21)$$

where $\beta = 1/\sigma^2$. In Figure 3 we show the result of using a six-hidden-unit network to fit the training data in Figure 3. With $\alpha = 0.1$ and $\beta = 1,000$, we drew a number of weight vectors \mathbf{w}^l , $l = 1, \dots, 15$, from the weight posterior $p(\mathbf{w}|D)$, Equation 21 and considered the corresponding functions $f(\mathbf{x}; \mathbf{w}^l)$. The mean and standard error bars calculated from these samples are plotted in Figure 3. How these samples are obtained is discussed later. Note how the error bars automatically increase in regions of low data density.

Monte Carlo Sampling

In general, the posterior distribution $p(\mathbf{w}|\Gamma, D)$ is non-Gaussian, and the integration required over the weight space to find, for example, the mean predictor

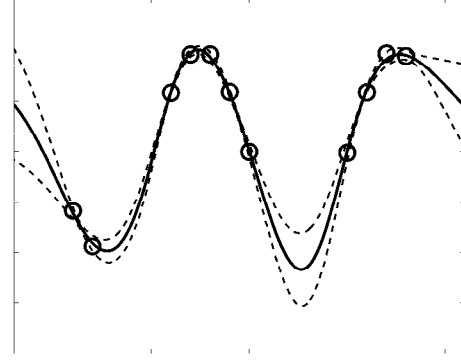


Figure 3. The raw input-output training data, with mean Bayesian MLP predictions (solid curve) and standard error bars (dashed curves). Note how the error bars increase away from the data.

$$\bar{f}(\mathbf{x}) = \int f(\mathbf{x}; \mathbf{w}) p(\mathbf{w}|\Gamma, D) d\mathbf{w} \quad (22)$$

is difficult. An approximate solution is provided by Monte Carlo sampling (Bishop, 1995; Neal, 1996):

$$\int f(\mathbf{x}; \mathbf{w}) p(\mathbf{w}|\Gamma, D) d\mathbf{w} \approx \frac{1}{L} \sum_{i=1}^L f(\mathbf{x}; \mathbf{w}^i) \quad (23)$$

where the sample weights \mathbf{w}^i are drawn from the posterior distribution. In principle, this procedure is exact in the limit $L \rightarrow \infty$. The great difficulty, however, is in constructing a finite, representative set of samples $\{\mathbf{w}_i\}$, and it is easy to remain trapped in unrepresentative parts of the posterior distribution (Neal, 1996).

Consider the problem of drawing samples from a general distribution $p(\mathbf{x}) \propto \psi(\mathbf{x})$ (Figure 4). Let \mathbf{x}^{old} be a sample point from $p(\mathbf{x})$. We propose a new sample point $\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{old}} + \eta$ where each element η_i is sampled from a zero-mean Gaussian distribution with variance τ^2 . We accept \mathbf{x}^{new} if $\psi(\mathbf{x}^{\text{new}}) > \psi(\mathbf{x}^{\text{old}})$, since the new candidate sample point is more likely than the old sample point. However, this does not constitute a valid sampling scheme since we only accept increasingly likely points, targeting therefore only the modes of the distribution. To correct for this, we accept a less likely candidate with probability $\psi(\mathbf{x}^{\text{new}})/\psi(\mathbf{x}^{\text{old}})$. This valid sampling scheme is called the Metropolis method and forms the basis for many generalizations (Neal, 1993, 1996).

In high dimensions, Metropolis sampling can be inefficient, since it is unlikely that testing a new point a long way from the current sample point will result in a more likely point (if you stand on a mountain and jump, it is more likely that you will end up at a point lower than at your current point). Thus, only very small

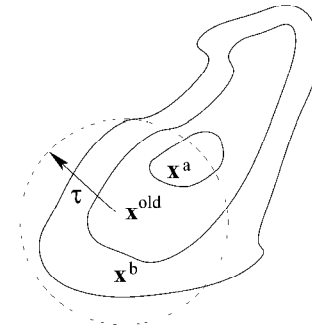


Figure 4. Metropolis Sampling from $p(\mathbf{x}) \propto \psi(\mathbf{x})$. Let \mathbf{x}^{old} be a sample from the distribution $p(\mathbf{x})$. We propose a new candidate \mathbf{x}^{new} by sampling from a Gaussian around \mathbf{x}^{old} with width τ . More likely candidates such as \mathbf{x}^a are accepted. Less likely candidates such as \mathbf{x}^b are accepted with probability $\psi(\mathbf{x}^b)/\psi(\mathbf{x}^{\text{old}})$.

jumps will be accepted in high-dimensional spaces, and many samples are required to form a good representation of the distribution. The hybrid Monte Carlo scheme attempts to improve sampling efficiency and allow larger jumps by exploiting gradient information about the distribution and has been successfully employed in Bayesian neural networks (Neal, 1996).

Laplace's Method

Although sampling techniques can be attractive, convergence to a representative set of samples is difficult to assess and can be very slow. Laplace's method is a perturbation technique motivated by the fact that as the number P of training data points is increased, the posterior distribution typically approaches a Gaussian (Walker, 1969) whose variance goes to zero in the limit $P \rightarrow \infty$ (we leave aside here the issues of inherent network symmetries). In order to calculate this Gaussian approximation, we consider the posterior distribution, Equation 21:

$$p(\mathbf{w}|D, \Gamma) \propto \exp(-\phi(\mathbf{w})) \quad (24)$$

and expand ϕ around a mode of the distribution, $\mathbf{w}_* = \arg \min \phi(\mathbf{w})$,

$$\phi(\mathbf{w}) \approx \phi(\mathbf{w}_*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}_*)^T \mathbf{H} (\mathbf{w} - \mathbf{w}_*) \quad (25)$$

where

$$\mathbf{H} = \nabla \nabla \phi(\mathbf{w})|_{\mathbf{w}_*} \quad (26)$$

is the local Hessian matrix. This local expansion defines a Gaussian approximation

$$p(\mathbf{w}|D, \Gamma) \approx \frac{|\mathbf{H}|^{1/2}}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2} (\mathbf{w} - \mathbf{w}_*)^T \mathbf{H} (\mathbf{w} - \mathbf{w}_*)\right\} \quad (27)$$

The expected value of $f(\mathbf{x}; \mathbf{w})$ as required in Equation 22 can be evaluated by making a further local linearization of the function $f(\cdot, \mathbf{w})$ around the point \mathbf{w}_* . In a practical implementation, a standard nonlinear optimization algorithm such as conjugate gradients is used to find a mode \mathbf{w}_* of the log posterior distribution (Bishop, 1995).

Determining Hyperparameters

So far we have assumed that the hyperparameters of the MLP are fixed. In a fully Bayesian treatment we would define prior distributions of the hyperparameters, and then integrate them out. Since exact integration is analytically intractable, we can use ML-II to estimate specific values for the hyperparameters by maximizing the marginal likelihood $P(D|\Gamma)$ (Equation 17) with respect to Γ . Using MLPs, the integrand in Equation 17 is non-Gaussian and $p(D|\Gamma)$ needs to be approximated. This can be achieved using Laplace's method by locally expanding the integral to second order in the weights. This leads to simple reestimation formulas for the hyperparameters expressed in terms of the eigenvalue/eigenvector decomposition of the Hessian matrix. This treatment of hyperparameters is called the *evidence* framework (MacKay, 1995) and involves alternating the optimization of \mathbf{w} (mode finding) for fixed hyperparameters with reestimation of the hyperparameters by reevaluating the Hessian matrix for the new value of \mathbf{w} . The various approximations involved in this approach improve as the number of data points $P \rightarrow \infty$. However, for a finite data set it can be difficult to assess the accuracy of the method. One obvious limitation is that it only takes account of the behavior of the posterior distribution at the mode.

The KL Variational Approach

The Kullback Leibler divergence is a measure of the difference between two probability distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ (Cover and Thomas, 1991)

$$KL(q, p) = \int \{q(\mathbf{x}) \log q(\mathbf{x}) - q(\mathbf{x}) \log p(\mathbf{x})\} d\mathbf{x} \quad (28)$$

This has the advantageous properties $KL \geq 0$ and $KL = 0$ if and only if $p = q$. Consider the KL divergence

$$KL(q(\mathbf{w}), p(\mathbf{w}|\Gamma, D)) \geq 0 \quad (29)$$

Finding the best distribution $q(\mathbf{w})$ in a restricted set of possible distributions by minimizing $KL(q, p)$ gives the best estimate (in the KL sense) to the posterior distribution. From Equation 29 we immediately have the bound

$$\begin{aligned} \log p(D|\Gamma) &\geq \int -q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w} \\ &\quad + \int q(\mathbf{w}) \log p(D|\Gamma, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \end{aligned} \quad (30)$$

We can make use of this lower bound to carry out an approximate ML-II hyperparameter optimization by the following two-step procedure: First fix the hyperparameters Γ and optimize the bound, Equation 30, with respect to $q(\mathbf{w})$. Then, for fixed $q(\mathbf{w})$, optimize the bound with respect to Γ . This scheme is a generalization of the Expectation-Maximization procedure (see Neal and Hinton in Jordan, 1998) and is also called *ensemble learning* (Barber and Bishop, 1997).

Bayesian Pruning

To this point we discussed the idea of using a prior that encourages smoothness of the input-output mapping. Insofar as neural networks are nonlinear functions of a linear combination of inputs, it is reasonable to use a prior that encourages small weights, $p(\mathbf{w}) \propto \exp(-\mathbf{w}^T \mathbf{A} \mathbf{w}/2)$. Typically, only diagonal matrices \mathbf{A} are considered. We can group weights into clusters containing one or more weights and associate with each cluster c a common hyperparameter α_c . The Bayesian approach results in a posterior distribution over these hyperparameters α_c . Alternatively, we can optimize the hyperparameters using ML-II. If the posterior distribution favors large α_c values, then effectively the weight cluster c is not contributing to the network and may be pruned. A useful choice of clustering is to group all the weights from a single input x_i into the hidden units (note that these weights are different from the weights that fan in to a hidden node). If the hyperparameter α_i (after ML-II optimization) associated with the weights fanning out from input x_i is large, the contribution of input x_i is negligible and can be excluded. This is called *automatic relevance determination* (MacKay, 1995).

The Relevance Vector Machine

In the discussion regarding GLMs, $f(\mathbf{x}) = \sum_i w_i \phi_i(\mathbf{x})$, we fixed the centers of the basis functions ϕ_i . Similarly, in the relevance vector machine we use fixed basis functions (Tipping, 2001). By associating with each weight w_i a regularizing prior $p(w_i) \propto \exp(-\alpha_i w_i^2/2)$, we can perform ML-II to optimize the hyperparameters α_i . After optimization, typically many of the α_i will become very large, effectively removing the basis function ϕ_i from the model. This pruning procedure often results in a much sparser representation of the data in terms of only the "relevant" basis functions; this scheme is therefore particularly useful for compression. This sparseness effect is similar, although not equivalent, to the support vector machine (see SUPPORT VECTOR MACHINES), in which training points

are effectively removed if they do not affect the prediction of the model.

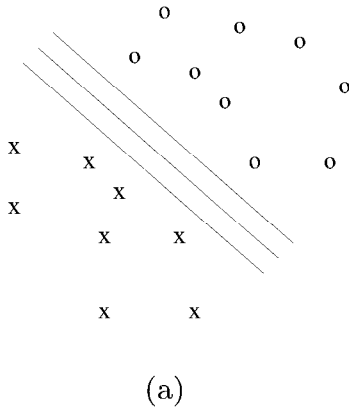
Classification

The previously described methods can be applied to classification, usually with only minor modification. For convenience, we consider here only problems with two classes. The data set is $D = \{(\mathbf{x}^\mu, t^\mu), \mu = 1, \dots, P\}$, where $t^\mu \in \{0, 1\}$. In a probabilistic framework, we use the output of the network $f(\mathbf{x}; \mathbf{w})$ to represent the probability that the input is in class 1. In this case, the likelihood is

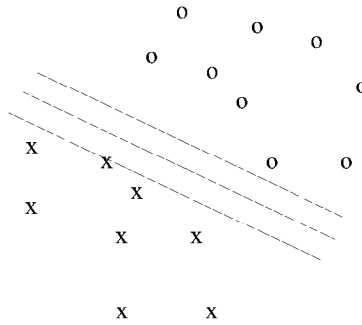
$$p(D|\mathbf{w}) = \prod_{\mu=1}^P f(\mathbf{x}^\mu; \mathbf{w})^{t^\mu} (1 - f(\mathbf{x}^\mu; \mathbf{w}))^{1-t^\mu} \quad (31)$$

For example, we could take $f(\mathbf{x}) = g(\mathbf{w}^T \mathbf{x})$, where $g(x) = 1/(1 + \exp(-x))$ (Bishop, 1995). In the Bayesian approach, we need to specify a prior belief about the weights. As before, a sensible choice is $p(\mathbf{w}) \propto \exp(-\alpha \mathbf{w}^T \mathbf{w}/2)$, since smaller weights will give less certain predictions. This results in a posterior distribution $p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$. For a novel input \mathbf{x} the probability that it belongs to class 1 is

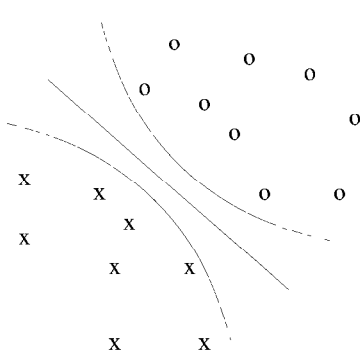
$$\begin{aligned} p(t = 1|\mathbf{x}, D) &= \int p(t = 1|\mathbf{x}, \mathbf{w})p(\mathbf{w}|D)d\mathbf{w} \\ &= \int g(\mathbf{w}^T \mathbf{x})p(\mathbf{w}|D)d\mathbf{w} \end{aligned} \quad (32)$$



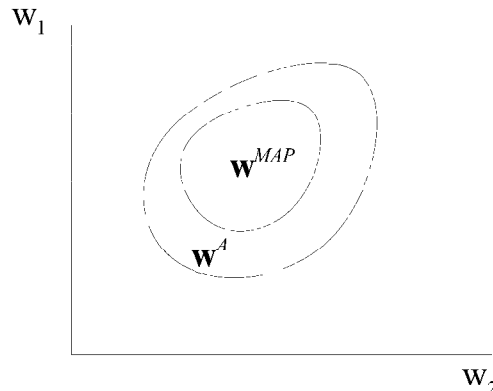
(a)



(b)



(c)



(d)

Consider, for example, fitting the data in Figure 5A. The posterior distribution is given in Figure 5D. The decision boundary ($p(t = 1|\mathbf{x}, \mathbf{w}, D) = 0.5$) for the MAP solution is given in Figure 5B along with the 0.1 and 0.9 decision contours. Another decision boundary associated with the posterior weights \mathbf{w}^A is plotted in Figure 5B. Because the decision boundaries are linear, the predictions of these single networks away from the data remain overly confident. The Bayesian prediction, Equation 32, is plotted in Figure 5C and has decision boundaries that properly account for the uncertainty in the predictions away from the training data.

Since the final integrand in Equation 32 depends only on the weight vector through the “activation” $a = \mathbf{w}^T \mathbf{x}$, we need only know the distribution of this one-dimensional quantity. A reasonable assumption is that the activation will be Gaussian distributed $p(a) = N(\bar{a}, \text{var}(a))$, and the resulting one-dimensional integration $p(t = 1|\mathbf{x}, D) = \int g(a)p(a)da$ can be efficiently performed using quadrature. The statistics of the activation are

$$\bar{a} = \bar{\mathbf{w}}^T \mathbf{x}, \quad \text{var}(a) = \mathbf{x}^T \Sigma \mathbf{x} \quad (33)$$

where $\bar{\mathbf{w}}$ and Σ are the mean and covariance of the weight posterior $p(\mathbf{w}|D)$. It is convenient to approximate these statistics using Laplace’s method.

Discussion

The Bayesian framework deals with uncertainty in a natural, consistent manner by combining prior beliefs about which models are

Figure 5. A, The decision boundary and 0.1, 0.9 decision contours for the most likely predictor \mathbf{w}^{MAP} . B, The predictions for \mathbf{w}^A . C, The posterior averaged predictors. D, The weight posterior distribution.

appropriate with how likely each model would be to have generated the data. This results in an elegant, general framework for fitting models to data, which, however, may be compromised by computational difficulties in carrying out the ideal procedure. There are many approximate Bayesian implementations, using methods such as sampling, perturbation techniques, and variational methods. Often these enable the successful approximate realization of practical Bayesian schemes. An attractive, built-in effect of the Bayesian approach is an automatic procedure for combining predictions from several different models, the combination strength of a model being given by the posterior likelihood of the model. In the case of models linear in their parameters, Bayesian neural networks are closely related to Gaussian processes, and many of the computational difficulties of dealing with more general stochastic nonlinear systems can be avoided.

Bayesian methods are readily extendable to other areas, in particular density estimation, and the benefits of dealing with uncertainty are again to be found (see Bishop in Jordan, 1998). Traditionally, neural networks are graphical representations of functions, in which the computations at each node are deterministic. In the classification discussion, however, the final output represents a stochastic variable. We can consider such stochastic variables elsewhere in the network, and the sigmoid belief network is an early example of a stochastic network (Neal, 1992). There is a major conceptual difference between such models and conventional neural networks. Networks in which nodes represent stochastic variables are called graphical models (see BAYESIAN NETWORKS) and are graphical representations of *distributions* (GRAPHICAL MODELS: PROBABILISTIC INFERENCE). Such models evolve naturally from the desire of incorporating uncertainty and nonlinearity in networked systems.

Bayesian Networks

Judea Pearl and Stuart Russell

Introduction

Probabilistic models based on directed acyclic graphs have a long and rich tradition, beginning with work by the geneticist Sewall Wright in the 1920s. Variants have appeared in many fields. Within statistics, such models are known as *directed graphical models*; within cognitive science and artificial intelligence (AI), they are known as *Bayesian networks*. The name honors the Reverend Thomas Bayes (1702–1761), whose rule for updating probabilities in light of new evidence is the foundation of the approach. The initial development of Bayesian networks in the late 1970s was motivated by the need to model the top-down (semantic) and bottom-up (perceptual) combination of evidence in reading. The capability for bidirectional inferences, combined with a rigorous probabilistic foundation, led to the rapid emergence of Bayesian networks as the method of choice for uncertain reasoning in AI and expert systems, replacing earlier, ad hoc rule-based schemes (Pearl, 1988; Shafer and Pearl, 1990; Jensen, 1996).

The nodes in a Bayesian network represent propositional variables of interest (e.g., the temperature of a device, the sex of a patient, a feature of an object, the occurrence of an event) and the links represent informational or causal dependencies among the variables. The dependencies are quantified by conditional probabilities for each node, given its parents in the network. The network supports the computation of the probabilities of any subset of variables given evidence about any other subset.

Road Map: Learning in Artificial Networks

Related Reading: Bayesian Networks; Gaussian Processes; Graphical Models: Probabilistic Inference; Support Vector Machines

References

- Barber, D., and Bishop, C., 1997, Ensemble learning in Bayesian neural networks, in *Neural Networks and Machine Learning* (C. Bishop, Ed.), NATO ASI Series, New York: Springer-Verlag.
- Berger, J. O., 1985, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., New York: Springer-Verlag. ♦
- Bishop, C. M., 1995, *Neural Networks for Pattern Recognition*, Oxford, Engl.: Oxford University Press. ♦
- Box, G., and Tiao, G., 1973, *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.
- Cover, M., and Thomas, J., 1991, *Elements of Information Theory*, New York: Wiley.
- Jordan, M., Ed., 1998, *Learning in Graphical Models*, Cambridge, MA: MIT Press.
- MacKay, D. J. C., 1992, Bayesian interpolation, *Neural Comput.*, 4(3): 415–447.
- MacKay, D. J. C., 1995, Probable networks and plausible predictions: A review of practical Bayesian methods for supervised neural networks, *Netw. Computat. Neural Syst.*, 6(3). ♦
- Neal, R. M., 1992, Connectionist learning of belief networks, *Artif. Intell.*, 56:71–113.
- Neal, R. M., 1993, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto, Canada.
- Neal, R. M., 1996, *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics 118, New York: Springer-Verlag.
- Tipping, M. E., 2001, Sparse Bayesian learning and the relevance vector machine, *J. Machine Learn. Res.*, no. 1, 211–244. ♦
- Walker, A. M., 1969, On the asymptotic behaviour of posterior distributions, *J. R. Statist. Soc. B*, 31:80–88.

Figure 1 illustrates a simple yet typical Bayesian network. It describes the causal relationships among five variables: the season of the year (X_1), whether it's raining or not (X_2), whether the sprinkler is on or off (X_3), whether the pavement is wet or dry (X_4), and whether the pavement is slippery or not (X_5). Here, the absence of a direct link between X_1 and X_5 , for example, captures our understanding that there is no direct influence of season on slipperiness;

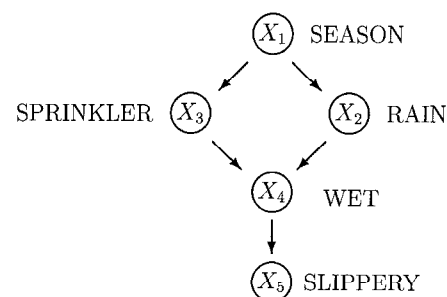


Figure 1. A Bayesian network representing causal influences among five variables. Each arc indicates a causal influence of the “parent” node on the “child” node.

the influence is mediated by the wetness of the pavement. (If freezing is a possibility, then a direct link could be added.)

Perhaps the most important aspect of Bayesian networks is that *they are direct representations of the world, not of reasoning processes*. The arrows in the diagram represent real causal connections and not the flow of information during reasoning (as in rule-based systems and neural networks). Reasoning processes can operate on Bayesian networks by propagating information in any direction. For example, if the sprinkler is on, then the pavement is probably wet (prediction); if someone slips on the pavement, that also provides evidence that it is wet (abduction, or reasoning to a probable cause). On the other hand, if we see that the pavement is wet, that makes it more likely that the sprinkler is on or that it is raining (abduction); but if we then observe that the sprinkler is on, that reduces the likelihood that it is raining (explaining away). It is this last form of reasoning, explaining away, that is especially difficult to model in rule-based systems and neural networks in any natural way, because it seems to require the propagation of information in two directions.

Probabilistic Semantics

Any complete probabilistic model of a domain must, either explicitly or implicitly, represent the *joint distribution*—the probability of every possible event as defined by the values of all the variables. There are exponentially many such events, yet Bayesian networks achieve compactness by factoring the joint distribution into local, conditional distributions for each variable given its parents. If x_i denotes some value of the variable X_i and pa_i denotes some set of values for X_i 's parents, then $P(x_i | pa_i)$ denotes this conditional distribution. For example, $P(x_4 | x_2, x_3)$ is the probability of wetness given the values of sprinkler and rain. The *global semantics* of Bayesian networks specifies that the full joint distribution is given by the product

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i) \quad (1)$$

In our example network, we have

$$\begin{aligned} P(x_1, x_2, x_3, x_4, x_5) \\ = P(x_1)P(x_2 | x_1)P(x_3 | x_1)P(x_4 | x_2, x_3)P(x_5 | x_4) \end{aligned} \quad (2)$$

Provided that the number of parents of each node is bounded, it is easy to see that the number of parameters required grows only linearly with the size of the network, whereas the joint distribution itself grows exponentially. Further savings can be achieved using compact parametric representations, such as noisy-OR models, decision trees, or neural networks, for the conditional distributions. For example, in *sigmoid* networks (see Jordan, 1999), the conditional distribution associated with each variable is represented as a sigmoid function of a linear combination of the parent variables; in this way, the number of parameters required is proportional to, rather than exponential in, the number of parents.

There is also an entirely equivalent *local semantics* that asserts that each variable is independent of its nondescendants in the network given its parents. For example, the parents of X_4 in Figure 1 are X_2 and X_3 , and they render X_4 independent of the remaining nondescendant, X_1 . That is,

$$P(x_4 | x_1, x_2, x_3) = P(x_4 | x_2, x_3)$$

The collection of independence assertions formed in this way suffices to derive the global assertion in Equation 1, and vice versa. The local semantics is most useful in *constructing* Bayesian networks, because selecting as parents *all* the direct causes of a given variable invariably satisfies the local conditional independence conditions (Pearl, 2000, p. 30). The global semantics leads directly to a variety of algorithms for reasoning.

Evidential Reasoning

From the product specification in Equation 1, one can express the probability of any desired proposition in terms of the conditional probabilities specified in the network. For example, the probability that the sprinkler is on, given that the pavement is slippery, is

$$\begin{aligned} P(X_3 = \text{on} | X_5 = \text{true}) &= \frac{P(X_3 = \text{on}, X_5 = \text{true})}{P(X_5 = \text{true})} \\ &= \frac{\sum_{x_1, x_2, x_4} P(x_1, x_2, x_3 = \text{on}, x_4, X_5 = \text{true})}{\sum_{x_1, x_2, x_3, x_4} P(x_1, x_2, x_3, x_4, X_5 = \text{true})} \\ &= \frac{\sum_{x_1, x_2, x_4} P(x_1)P(x_2 | x_1)P(X_3 = \text{on} | x_1)P(x_4 | x_2, X_3 = \text{on})P(X_5 = \text{true} | x_4)}{\sum_{x_1, x_2, x_3, x_4} P(x_1)P(x_2 | x_1)P(x_3 | x_1)P(x_4 | x_2, x_3)P(X_5 = \text{true} | x_4)} \end{aligned}$$

These expressions can often be simplified in ways that reflect the structure of the network itself. The first algorithms proposed for probabilistic calculations in Bayesian networks used a local, distributed message-passing architecture, typical of many cognitive activities (Kim and Pearl, 1983). Initially this approach was limited to tree-structured networks, but it was later extended to general networks in Lauritzen and Spiegelhalter's (1988) method of joint-tree propagation. A number of other exact methods have been developed and can be found in recent textbooks (Jensen, 1996; Jordan, 1999).

It is easy to show that reasoning in Bayesian networks subsumes the satisfiability problem in propositional logic and, hence, is NP-hard. Monte Carlo simulation methods can be used for approximate inference (Pearl, 1988), giving gradually improving estimates as sampling proceeds. (These methods use local message propagation on the original network structure, unlike join-tree methods.) Alternatively, variational methods provide bounds on the true probability (Jordan, 1999).

Uncertainty over Time

Entities that live in a changing environment must keep track of variables whose values change over time. Dynamic Bayesian networks, or DBNs, capture this process by representing multiple copies of the state variables, one for each time step (Dean and Kanazawa, 1989). A set of variables \mathbf{X}_t denotes the world state at time t and a set of sensor variables \mathbf{E}_t denotes the observations available at time t . The *sensor model* $P(\mathbf{E}_t | \mathbf{X}_t)$ is encoded in the conditional probability distributions for the observable variables, given the state variables. The *transition model* $P(\mathbf{X}_{t+1} | \mathbf{X}_t)$ relates the state at time t to the state at time $t + 1$. Keeping track of the world, known as *filtering*, means computing the current probability distribution over world states given all past observations, i.e., $P(\mathbf{X}_t | \mathbf{E}_1, \dots, \mathbf{E}_t)$. Dynamic Bayesian networks include as special cases other temporal probability models, such as hidden Markov models (DBNs with a single discrete state variable) and Kalman filters (DBNs with continuous state and sensor variables and linear Gaussian transition and sensor models). For the general case, exact filtering is intractable, and a variety of approximation algorithms have been developed. The most popular and flexible of these is the family of *particle filtering* algorithms (see Doucet, de Freitas, and Jordan, 2001).

Learning in Bayesian Networks

The conditional probabilities $P(x_i | pa_i)$ can be updated continuously from observational data using gradient-based or Expectation-Maximization (EM) methods that use just local information derived from inference (Binder et al., 1997; Jordan, 1999), in much the same way as weights are adjusted in neural networks. It is also possible to learn the structure of the network, using methods that

trade off network complexity against degree of fit to the data (Friedman, 1998). As a substrate for learning, Bayesian networks have the advantage that it is relatively easy to encode prior knowledge in network form, either by fixing portions of the structure or by using prior distributions over the network parameters. Such prior knowledge can allow a system to learn accurate models from much less data than are required by *tabula rasa* approaches.

Causal Networks

Most probabilistic models, including general Bayesian networks, describe a distribution over possible observed events, as in Equation 1, but say nothing about what will happen if a certain *intervention* occurs. For example, what if I *turn the sprinkler on*? What effect does that have on the season, or on the connection between wetness and slipperiness? A *causal network*, intuitively speaking, is a Bayesian network with the added property that the parents of each node are its direct causes, as in Figure 1. In such a network, the result of an intervention is obvious: the sprinkler node is set to $X_3 = \text{on}$, and the causal link between the season X_1 and the sprinkler X_3 is removed. All other causal links and conditional probabilities remain intact, so the new model is

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 | x_1)P(x_4 | x_2, X_3 = \text{on})P(x_5 | x_4)$$

Notice that this differs from *observing* that $X_3 = \text{on}$, which would result in a new model that included the term $P(X_3 = \text{on} | x_1)$. This mirrors the difference between seeing and doing: after observing that the sprinkler is on, we wish to infer that the season is dry, that it probably did not rain, and so on; an arbitrary decision to turn the sprinkler on should not result in any such beliefs.

Causal networks are more properly defined, then, as Bayesian networks in which the correct probability model after intervening to fix any node's value is given simply by deleting links from the node's parents. For example, *fire* \rightarrow *smoke* is a causal network, whereas *smoke* \rightarrow *fire* is not, even though both networks are equally capable of representing any joint distribution on the two variables. Causal networks model the environment as a collection of stable component mechanisms. These mechanisms may be reconfigured locally by interventions, with correspondingly local changes in the model. This, in turn, allows causal networks to be used very naturally for prediction by an agent that is considering various courses of action (Pearl, 2000).

Functional Bayesian Networks

The networks discussed so far are capable of supporting reasoning about evidence and about actions. Additional refinement is necessary in order to process *counterfactual* information. For example, the probability that "the pavement would not have been slippery had the sprinkler been OFF, given that the sprinkler is in fact ON and that the pavement is in fact slippery" cannot be computed from the information provided in Figure 1 and Equation 1. Such counterfactual probabilities require a specification in the form of functional networks, where each conditional probability $P(x_i | pa_i)$ is replaced by a functional relationship $x_i = f_i(pa_i, \epsilon_i)$, where ϵ_i is a stochastic (unobserved) error term. When the functions f_i and the distributions of ϵ_i are known, all counterfactual statements can be assigned unique probabilities, using evidence propagation in a structure called a "twin network." When only partial knowledge about the functional form of f_i is available, bounds can be computed on the probabilities of counterfactual sentences (Pearl, 2000).

Causal Discovery

One of the most exciting prospects in recent years has been the possibility of using Bayesian networks to discover causal structures

in raw statistical data (Pearl, 2000)—a task previously considered impossible without controlled experiments. Consider, for example, the following *intransitive* pattern of dependencies among three events: A and B are dependent, B and C are dependent, yet A and C are independent. If you ask a person to supply an example of three such events, the example would invariably portray A and C as two independent causes and B as their common effect, namely, $A \rightarrow B \leftarrow C$. (For instance, A and C could be the outcomes of tossing two fair coins, and B could represent a bell that rings whenever either coin comes up heads.) Fitting this dependence pattern with a scenario in which B is the cause and A and C are the effects is mathematically feasible but very unnatural, because it must entail fine tuning of the probabilities involved; the desired dependence pattern will be destroyed as soon as the probabilities undergo a slight change.

Such thought experiments tell us that certain patterns of dependency, which are totally void of temporal information, are conceptually characteristic of certain causal directionalities and not others. When put together systematically, such patterns can be used to infer causal structures from raw data and to guarantee that any alternative structure compatible with the data must be less stable than the one(s) inferred; namely, slight fluctuations in parameters will render that structure incompatible with the data.

Plain Beliefs

In mundane decision making, beliefs are revised not by adjusting numerical probabilities but by tentatively accepting some sentences as "true for all practical purposes." Such sentences, called *plain beliefs*, exhibit both logical and probabilistic character. As in classical logic, they are propositional and deductively closed; as in probability, they are subject to retraction and can be held with varying degrees of strength. Bayesian networks can be adopted to model the dynamics of plain beliefs by replacing ordinary probabilities with nonstandard probabilities, that is, probabilities that are infinitesimally close to either zero or one (Goldszmidt and Pearl, 1996).

Discussion

Bayesian networks may be viewed as normative cognitive models of propositional reasoning under uncertainty. They handle noise and partial information using local, distributed algorithms for inference and learning. Unlike feedforward neural networks, they facilitate local representations in which nodes correspond to propositions of interest. Recent experiments suggest that they accurately capture the causal inferences made by both children and adults (Tenenbaum and Griffiths, 2001). Moreover, they capture patterns of reasoning, such as explaining away, that are not easily handled by any competing computational model. They appear to have many of the advantages of both the "symbolic" and the "subsymbolic" approaches to cognitive modeling, and are now an essential part of the foundations of computational neuroscience (Jordan and Sejnowski, 2001).

Two major questions arise when we postulate Bayesian networks as potential models of actual human cognition. First, does an architecture resembling that of Bayesian networks exist anywhere in the human brain? At the time of writing, no specific work has been done to design neurally plausible models that implement the required functionality, although no obvious obstacles exist. Second, how could Bayesian networks, which are purely propositional in their expressive power, handle the kinds of reasoning about individuals, relations, properties, and universals that pervade human thought? One plausible answer is that Bayesian networks containing propositions relevant to the current context are constantly being assembled, as needed, from a more permanent store of knowledge. For example, the network in Figure 1 may be assembled to help

explain why this particular pavement is slippery right now, and to decide whether this can be prevented. The background store of knowledge includes general models of pavements, sprinklers, slipping, rain, and so on; these must be accessed and supplied with instance data to construct the specific Bayesian network structure. The store of background knowledge must utilize some representation that combines the expressive power of first-order logical languages (such as semantic networks) with the ability to handle uncertain information. Substantial progress has been made on constructing systems of this kind (Koller and Pfeffer, 1998), but as yet no overall cognitive architecture has been proposed.

Road Maps: Artificial Intelligence; Learning in Artificial Networks

Related Reading: Bayesian Methods and Neural Networks; Decision Support Systems and Expert Systems; Graphical Models: Probabilistic Inference

References

- Binder, J., Koller, D., Russell, S., and Kanazawa, K., 1997, Adaptive probabilistic networks with hidden variables, *Machine Learn.*, 29:213–244.
- Dean, T., and Kanazawa, K., 1989, A model for reasoning about persistence and causation, *Computat. Intell.*, 5:142–150.
- Doucet, A., de Freitas, J., and Gordon, N., 2001, *Sequential Monte Carlo Methods in Practice*, Berlin: Springer-Verlag.
- Friedman, N., 1998, The Bayesian structural EM algorithm, in *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference* (G. F. Cooper and S. Moral, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 129–138.
- Goldschmidt, M., and Pearl, J., 1996, Qualitative probabilities for default reasoning, belief revision, and causal modeling, *Artif. Intell.*, 84:57–112.
- Jensen, F. V., 1996, *An Introduction to Bayesian Networks*, New York: Springer-Verlag. ♦
- Jordan, M. I., Ed., 1999, *Learning in Graphical Models*, Cambridge, MA: MIT Press. ♦
- Jordan, M. I., and Sejnowski, T. J., Eds., 2001, *Graphical Models: Foundations of Neural Computation*, Cambridge, MA: MIT Press.
- Kim, J. H., and Pearl, J., 1983, A computational model for combined causal and diagnostic reasoning in inference systems, in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, San Mateo, CA: Morgan Kaufmann, pp. 190–193.
- Koller, D., and Pfeffer, A., 1998, Probabilistic frame-based systems, in *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Menlo Park, CA: AAAI Press, pp. 580–587.
- Lauritzen, S. L., and Spiegelhalter, D. J., 1988, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. R. Statist. Soc.*, series B, 50:157–224.
- Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann. ♦
- Pearl, J., 2000, *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. ♦
- Shafer, G., and Pearl, J., Eds., 1990, *Readings in Uncertain Reasoning*, San Mateo, CA: Morgan Kaufmann.
- Tenenbaum, J. B., and Griffiths, T. L., 2001, Structure learning in human causal induction, in *Advances in Neural Information Processing Systems 13*, Cambridge, MA: MIT Press.

Biologically Inspired Robotics

Noel E. Sharkey

Introduction

At the beginning of the twenty-first century, living organisms have still not been successfully replicated by machines. Computers are much faster at number crunching than humans and can even beat the greatest at chess, and other machines can perform routine physical tasks faster than us and with a precision that we cannot approach. However, animals exhibit such remarkable capacities for flexible adaptation to novel circumstances that roboticists can only gaze in wonder. It is thus an important goal of modern robotics to learn from the way organisms are constructed biologically, and how this creates adaptive behaviors.

Biologically inspired robotics, also known as biomimetic robotics or biorobotics, refers to robotics research in which the life sciences, including biology, psychology, ethology, neuroscience, and evolutionary theory, play a key role in motivating the research. It is necessarily broad because the field is just beginning to emerge as a unified discipline, and so it still has fuzzy boundaries. Biorobotics research ranges from modeling animal sensors in hardware for guiding robots in target environments to investigating the interaction between neural learning and evolution in a variety of robot tasks. There are, however, common themes that will be explored here.

In the following sections, some of the major issues in biorobotics research and the aims of this approach are examined. First we briefly consider the historical roots of the core ideas. The seminal work of Grey Walter (1953) sets the scene and introduces some of the key elements of biologically inspired robotics. The re-introduction and development of the ideas in the 1980s occurred with Braitenberg's *synthetic psychology* and Brook's *behavior-based* robotics. In summarizing the breadth of the current work, we attempt a threefold classification of biologically inspired robotics.

The Roots of Biologically Inspired Robotics

The roots of biologically inspired robotics date back to the early twentieth century, when Hammond constructed a heliotrope based on the biologist Loeb's tropism theory of animal behavior. Loeb proposed that animals are attracted and repelled by stimuli in the environment in a way similar to the phototropic responses of plants. Although Hammond's heliotrope did not model an animal, its mechanized movement toward light was sufficient to satisfy Loeb that his theory has physical plausibility (cf. Sharkey and Ziemke in Ziemke and Sharkey, 1998, pp. 361–392, for an account).

There were a number of robot learning studies during the first half of the twentieth century, before the birth of artificial intelligence (AI). However, the prototypical biorobotics work was conducted by Grey Walter (1953). He went far beyond Hammond in testing the mechanistic plausibility of animal tropism. His aim was to create a self-sustaining artificial life form that could adapt. This required the development of a robot that could seek out a source to recharge its batteries on demand.

Grey Walter used electromechanical robots equipped with two input "receptors": a photo-electric cell for sensitivity to light, and an electrical contact as a touch receptor. The controller, between sensors and motors, was a small artificial nervous system built from miniature valves, relays, condensers, batteries, and small electric motors—no computer. There was a hutch where a robot could drive in to have the battery automatically recharged.

Behavior resulted from the interaction of the internal states of the robot (battery level) and the intensity of light sources, as well as other environmental factors such as obstacles. When the battery levels were high, the robot was repelled by the bright light of the hutch and attracted by the moderate light in the room, where it

"explored." With low battery levels, the robot was attracted to the bright light of the hutch for an automatic recharge. In this way Grey Walter demonstrated that mechanical tropism could work as a means of exploration and maintaining energy.

Grey Walter (1953) also investigated adaptation and showed how a simple learning mechanism could extend the behavior of a robot using the conditioned reflex analog (CORA) with a microphone for auditory input.

Biorobotics more or less died when Grey Walter moved on to other research in the 1950s. With the rise of AI and computing, the focus was on providing robots with human-inspired perception and cognition. The new robots had a series of modules, such as visual processing, planning, and reasoning, through which sensory information passed serially. Typically, a decision-making module controlled the output to the actuators. This was in contrast to the more direct control approach of Grey Walter, in which the only mediation between sensing and moving was provided by an artificial neural net consisting of two hardware neurons. Another difference was that whereas AI robotics focused on human cognition, Grey Walter focused on the question of how seemingly complex animal-like behavior could arise from simple mechanisms such as tropisms and reflexes.

Today the term *taxis* is used instead of tropism to refer to the movement of an animal directed by a stimulus, either negatively or positively. Examples of such stimulus-directed activity include chemotaxis (chemical taxis), geotaxis (gravity), phototaxis (light), and phonotaxis (auditory). Although Grey Walter worked only on individual taxes, biologists at the time (e.g., Fraenkel and Gunn; cf. Sharkey and Ziemke in Ziemke and Sharkey, 1998, pp. 361–392) proposed that the behavior of many organisms could be explained by a combination of taxes working together and in opposition. They cited Fraenkel's study of the coastal slug, *Littorina neritoides*. *Littorina* combines positive and negative phototaxis with negative geotaxis to feed and survive. Subsequently, combinations of taxes have been used as powerful explanations of many animal behaviors, from bacteria feeding to insect pheromone trailing to fish breeding and feeding.

These ideas began to emerge in a new wave of robotics during the 1980s as a result of two major influences. First, the neuroanatomist Valentino Braitenberg showed how a number of complex behaviors could emerge from a combination of very simple neural networks encoding different taxes (Braitenberg, 1984). Second, Rodney Brooks's development of subsumption architecture allowed autonomous control by a combination of taxes, and drove home the effectiveness of behavior-based robotics. His major papers from this period are reprinted in Brooks (1999). In this style of robotics, each behavior-producing module, such as *avoid obstacles* or *move toward light*, is encoded as a separate program module such that each is directly under the control of environmental circumstances rather than a central controller. For example, when there is light on the sensors, the *move toward light* module will be active until the light is occluded by an obstacle, at which point the *avoid obstacles* module takes over.

Current Directions in Biorobotics

A large emerging body of research in robotics is making the connection between sensing and moving simple, and the relationship between robot and world tightly coupled. It was the dramatic increase in robotics research, riding on the back of the new behavior-based approach, that enabled biologically inspired robotics to flourish. Since the behavior-based approach grew directly from ideas in the life sciences, it was only natural that once the tools and techniques of the approach had been developed, they would be turned back to work on the source of inspiration.

In this article, biorobotics is divided into three main classes. Although these classes are mutually supportive and their paths often cross, the distinctions between them are nonetheless useful.

- The *generalized* approach follows from the lineage of ideas that inspired Grey Walter to use robots to investigate and extend general mechanistic theories of animal behavior and adaptation. This includes research using neural network adaptation through learning and/or evolutionary methods (see REACTIVE ROBOTIC SYSTEMS).
- The *specific* approach uses methods from the generalized approach to investigate specific species or organisms. The research can range from studies of the physical plausibility of a simple neural explanation for some target behavior pattern to the physical modeling of a particular animal or some of its senses. Models can be evaluated by observing the target behavior of the robot interacting with the environment through sensing and moving. One of the main goals of specific biorobotics is to develop new methods for scientific modeling.
- The *theoretical* division is a mixed bag that provides an examination of the implications of the research for a number of disciplines. The issues range widely, from discussions of robot embodiment to the nature of life. Although all biorobotics has a theoretical component, the theoretical approach is distinct in not requiring empirical work.

The idea was to include only work that at least touched base with the life sciences with respect to the type of controllers and the method of adaptation used. Each of the classes is dealt with in more depth in the following three subsections.

Generalized Biorobotics

The research impetus is to use broad notions derived from the life sciences for robot control. Many of these notions are in the form of implicit assumptions, such as deriving complex behavior from the simplest possible mechanisms or using the ideas of taxis or tropism for automated control. In this sense, Grey Walter's research was prototypical generalized biorobotics. His work on classical conditioning with the CORA architecture also foresaw the modern focus on adaptive techniques in robotics. The biological currency in the generalized biorobotics community mostly consists of abstract models of neural network learning, animal learning, or evolutionary processes, or a mixture. In the next two subsections, the main trends will be discussed.

Evolution. Evolutionary methods have been used for many applications since the 1950s when the first Genetic Algorithm (GA; see EVOLUTION OF ARTIFICIAL NEURAL NETWORKS) was developed by Friedman for his master's thesis on evolving control circuits for autonomous robots. These methods are particularly useful for constraining search in very large search spaces. However, from the perspective of biorobotics, the most important reason for employing evolutionary methods is that they are abstractly related to the Darwinian principle of natural selection and may be seen as analogous to real evolutionary theory; i.e., there is a fitness function to decide how fit a particular program is in the context of the problem it is to solve, and there are mutation and crossover to operate on the computer equivalent of gene strings. Given the intended relationship between the behavior of biorobots and natural biological behavior, the development of an *evolutionary robotics* is a very important step.

A fairly typical example of evolutionary methods for single robots is Nolfi's garbage collector (in Sharkey, 1997, pp. 187–198). The connection weights were evolved to control a miniature robot equipped with distance sensors and a gripper. The task was to

“clean” an arena by picking up objects and dropping them off outside. To do this, the robot had to move around the arena, avoid obstacles, locate an object, pick it up, move toward the walls, and release the object outside the arena. After 1,000 generations, robot controllers were evolved that performed the cleaning task to a high degree of accuracy.

Most evolutionary robotics research relies on using a fixed neural network architecture on which the weights are evolved. Another interesting approach is to let the evolutionary method decide on the type of connectivity between the units in the net; i.e., the pattern of connectivity is “genetically” represented (see Husbands et al. in Ziemke and Sharkey, 1998, pp. 185–210).

An important research area in biorobotics is concerned with how the environment and other species co-evolve with a given organism, resulting in an *evolutionary arms race*. This issue has been taken up in the simple form of evolving two competing robot controllers at the same time. For example, Floreano and Nolfi (1997) co-evolved the controllers for predator and prey behavior for two different “species” as part of each other’s environment. One of the main problems was that in one generation the predators would win but in the next generation the prey would win because a counterstrategy was evolved. This instability has been overcome by introducing neural network learning during the lifetime of the individuals. In this way the predators were able to adapt to the new evolved strategies of the prey.

The approach of combining the two adaptive techniques of evolutionary methods and neural network learning is proving to be a very effective adaptation technique that has a naturalistic flavor. Much of the research on combining has focused on how learning can help guide evolution—the *Baldwin effect*. The idea is that if the genotype of an individual is close to an optimal combination of genes, learning can allow that individual to increase its suitability for its environment, thereby increasing its probability of survival and reproduction. This could lead to a larger “basin” of fitness around optimal genotypes, channeling evolution toward optimal solutions (see Nolfi and Floreano, 1999, for a review).

Learning. One of the most widely used learning techniques in biorobotics is reinforcement or reward learning (RL) (see, e.g., Krose, 1995). RL has been studied psychologically since the beginning of the twentieth century. An advantage of RL techniques in robotics is that the learner needs only occasional reinforcement. RL is therefore unlike supervised learning, which requires a trainer to provide the learner with an exact target action in every time step, suited for use in unknown environments or tasks (but see Sharkey, 1998, on the use of innate controllers for training supervised learning).

More recently there has been a move toward using the *operant conditioning* techniques developed in the 1940s for studies of animal learning. Operant conditioning involves the shaping of pre-given behaviors. In particular, animals can be trained to produce an experimenter-required behavior when they are rewarded for successive approximations to that behavior. For example, to begin training a rat to press a bar for food, rewards are given for any reaching movement. Then successive approximations to the goal are rewarded until the target behavior is observed. In robotics, this has also been called behavior editing by Dorigo and Colombetti (1998), who have conducted most of the experimental work on this technique. An extension of this work to include incremental shaping is discussed by Urzelai et al. (in Ziemke and Sharkey, 1998, pp. 341–360).

In a realistic approach, Saksida et al. (in Sharkey, 1997, pp. 231–249) successfully used operant conditioning to modify the interaction between behaviors that had been preprogrammed into a robot. This departure from using reinforcement learning as a trial-and-error approach to modify existing behaviors is a step toward

real animal training. Furthermore, unlike most RL work, the training was conducted by a human trainer rather than a programmed reinforcer. Initially, the robot has three categories of objects: a bright orange jacket, green and pink plastic dog toys, and blue plastic recycling bins. One of its innate behaviors was to approach the plastic dog toys and pick them up. Successful (fast) shaping was shown for a number of new behaviors, including *Follow the Trainer*, *Recycling*, and *Playing Fetch*.

Specific Biorobotics

One of the attractions of robotics is that there is strong potential for testing the relationship between a model and some hypothesized behavioral consequences in the physical world. When Hammond built his heliotrope in the early twentieth century to test Loebian theory, it was essentially the physical plausibility of the theory that was under scrutiny. This was generalized biorobotics in that the hypotheses were about all animals. One of the goals of specific biorobotics is to extend such physical testing to test specific hypotheses about specific species. The motivation is that mathematical specification and computer simulation provide only a weak test of a model in that the inputs are typically chosen by the researcher and the outputs are designed to be interpretable as data points or graphs. The central idea of specific biorobotics is to test the model by situating the robot in a physical environment that provides the main features of the world of the target species.

There are a number of dangers with this approach, and a number of wrinkles will have to be ironed out before such modeling reaches maturity as a test methodology in biology and psychology. For example, with complex neural networks such as brains, it is not always possible to isolate a mechanism and test its behavioral consequences. Although robotics can offer a window on the possible behaviors resulting from particular models, a model cannot generally be used directly as a robot controller; a number of “gaps” between the sensors, the model, and motor output have to be filled in. This can be advantageous in forcing the theorist to extend the theoretical mechanisms, but care must be taken to ensure that mechanisms outside the theoretical framework do not play a causal role in the robot behavior.

Robotic modeling of living systems has taken a number of different forms, from behavioral modeling (see, e.g., Webb in Gausnier, 1996, pp. 117–134, on cricket phonotaxis; Grasso et al. in Chang and Gadiano, 2000, pp. 115–131 on lobster chemotaxis) to neuroscientific modeling (e.g., Burgess et al. in Ziemke and Sharkey, 1998, pp. 291–300, and Recce et al. in Sharkey, 1997, pp. 393–406, on the rat hippocampus; van der Smagt in Ziemke and Sharkey, 1998, pp. 301–320, on the human cerebellum for arm control) to modeling animal sensing (e.g., Lambrinos et al. in Chang and Gadiano, 2000, pp. 39–64, on ant solar compass sensing; Blanchard et al. in Chang and Gadiano, 2000, pp. 17–38, on locust sensing of approach; Rucci in Chang and Gadiano, 2000, pp. 181–193, on localization of auditory and visual structures in the barn owl) to biomechanics (e.g., Delcomyn and Nelson in Chang and Gadiano, 2000, pp. 5–15, and Quinn and Ritzman in Ziemke and Sharkey, 1998, pp. 239–254, on hexapod walking in the cockroach).

One of the most successful attempts at behavioral modeling has been the work of Webb and her associates (e.g., Webb in Gausnier, 1996, on mate selection in the female cricket). A wheeled robot was used to physically model a female cricket locating a conspecific male by following its calls. The robot was equipped with an auditory system capable of selectively localizing the sound of a male cricket stridulating (rubbing its wings together rapidly to produce a sound that attracts potential mates).

A similar approach has been taken by Lambrinos et al. (in Chang and Gadiano, 2000, pp. 39–64) for modeling the sensors of the

desert ant *Cataglyphus*, which maintains its heading across a largely featureless desert using polarized light sensing. Lambrinos et al. built special-purpose polarized light sensors based on what is known about the neural mechanisms of polarization that the honey bee *Apis mellifera*, the field cricket *Gryllus campestris*, and the desert ant *Cataglyphus bicolor* use to determine the position of the sun. The sensors were mounted on a wheeled robot and used to test different models of how *Cataglyphus* maintains its heading with polarized light. The research has been successfully conducted on a mobile robot in the ant's natural habitat with a homing performance similar to that of the ant.

As in Webb's work, the "ant robot" was used to model only a small part of the whole process of finding the direction to the nest. It did not, for example, accommodate the movement of the sun across the sky during the day (although this information was used to make corrections to the data). The sun moves relative to Earth at an average of 15° per hour (this figure varies greatly according to the time of day). In the early part of the twentieth century, this fact was used to show that ants both memorized the position of the sun and compensated for its movement. When the ants are imprisoned in a dark box for 2½ hours and released, they deviate from their original bearing by approximately the same number of degrees as the sun moved during their imprisonment. These findings reveal that *Cataglyphus* keeps track of the azimuth during the day and uses this information in maintaining a course.

Another important aspect of robotics used for modeling concerns legged locomotion. This leads to a two-way interaction between model testing and engineering. A number of researchers have turned to insect locomotion as a way to find a type of gait for a legged robot. Quinn and Ritzmann (in Ziemke and Sharkey, 1998, pp. 239–254) have designed and built a hexapod robot based on detailed neurobiological and kinematic observations of the locomotion of the death's head cockroach, *Blaberus discoidalis*. As a result, the robot's kinematics are remarkably similar to those of the real cockroach, and issues addressed in controlling the artificial cockroach have actually lead to new understanding of its natural counterpart.

Moving onto the mammalian nervous system, Burgess, Donnett, and O'Keefe (in Ziemke and Sharkey, 1998, pp. 291–300) used a miniature mobile robot equipped with a camera to test a neuronal model of how internal and external sensory information contribute to the firing of place cells in the rat hippocampus, and how these cells contribute to rat navigation behavior. They tested hypotheses based on their earlier neurophysiological work on the rat hippocampus using single-cell recording techniques. The robot experiments showed that the information provided by the robot's on-board video, odometry, and proximity sensors was sufficient to allow reasonably accurate return to an unmarked goal location. Similar robot modeling work has also been carried out by Recce et al. (in Sharkey, 1997, pp. 393–406) using the hippocampus as a method of absolute localization.

Research in specific biorobotics is gathering momentum as robot and sensing technology continues to improve. There are still many modeling issues to be worked out in conjunction with biology. The next step would be to get the morphology of robots to more accurately model the bodies and movement of the target species and to work continuously toward the goal of modeling whole animals, rather than installing patches to cover the missing bits. Like computational modeling, great care must be taken to ensure that the patches do not have a causal role in the target behavior.

Theoretical Biorobotics

Theoretical biorobotics is the most abstract level of biologically inspired robotics. Essentially, theoretical biorobotics is an all-

encompassing category for work that does not involve implementation on a robot but rather addresses metaquestions about robotics. Although wide-ranging, the main theoretical focus of biologically inspired robotics concerns biological and psychological issues. Many of these issues draw on detailed philosophical reasoning. Here we will confine ourselves to setting out some of the main points and referencing more detailed works in the literature.

A strong impetus for the new wave in biologically inspired robotics was the way it differed from traditional AI. Rodney Brooks, one of the prime movers in the mid-1980s, was concerned with the inadequacy of the prevailing methods used in AI for robotics (see, e.g., Brooks, 1999). Based mainly on the cognitivist conception of human intelligence, the sensory input to robots went through a number of strategic stages such as perception, planning, and reasoning before each move. All of the information was presented to a central controller, which decided how to act. This slowed performance to a single small move about every 15 minutes.

Rejecting cognitivism, theoretical biorobotics views intelligence as *embodied* in the machine and in its interactions with the world in which it is situated. Extreme cognitivists hold that mind is essentially a computer program, a language of thought, that could be run on any machine capable of running it. Mind is simply linked to the machine running it and the external world through transducers. Extreme biorobotists might claim that mind is inseparable from the individual machine and the more encompassing environmental machine of which the individual machine is a part. That is, the robot is *situated* in the world and is an *embodied* or *physically grounded* intelligence.

Varela, Thompson, and Rosch (1991) provide an insightful discussion of the details of embodiment in robots and its relation to life and mind. These authors are primarily interested in how living systems are embodied and how they are situated in their interactions with the world. Their purpose is to urge cognitive science to reject the vacuity of ungrounded thought. However, Sharkey and Ziemke (in Ziemke and Sharkey, 1998), while going along with some of the account by Varela et al. of living systems, argue for a weak embodiment in robotics, i.e., that robots can be used to *model* embodiment without themselves being embodied (see PHILOSOPHICAL ISSUES IN BRAIN THEORY AND CONNECTIONISM).

Another idea that has received considerable attention is that of *emergence* or *emergent behavior*. This is the notion that we can get something for nothing (or very little). One analogy is that from a collection of many molecules of water a cloud emerges that is greater than the sum of the parts. Perhaps a better example is the emergence of collective behavior in insects when each insect carries out very simple behaviors. For example, it is argued that the extraordinary structures that termites build in the desert emerge from very simple behaviors. Clark (1997) provides an in-depth discussion of emergent behavior and describes the two rules required by the termites: "If not carrying anything and you bump into a wood chip, pick it up"; and "If carrying a wood chip and you bump into another one, put it down." The resultant piling behavior emerges from the interplay between simple rules and the constraints of the environment.

The idea, then, is that coherent behavior emerges from a collection of simple taxes working together at the same time. This was an outright rejection of the notion of a central controller for action that was prevalent in AI. The idea in AI was to provide the robot with a model of the world, whereas one of the favorite slogans of the new roboticists is "the world is its own model." Nonetheless, Sharkey and Ziemke (2001) caution that even the taxes are emergent in the sense that they are in the eye of the beholder; i.e., they are distal descriptions of behavior.

One of the healthiest signs in the field is that some mainstream biologists and psychologists have begun to write about the relationship between specific biological findings and robotics. Navi-

gation, for example, is an important topic in both biology and robotics, and biologists (e.g., Collett in Ziemke and Sharkey, 1998, pp. 255–270; Etienne in Ziemke and Sharkey, 1998, pp. 271–290; Franz and Mallot in Chang and Gadiano, 2000, pp. 133–153) have discussed the relation between different aspects of navigation from an insect and mammalian perspective. Moreover, psychologists are beginning to take robot studies using animal learning techniques seriously enough to write detailed discussions of the relationship between the natural and the metallic (e.g., Savage in Ziemke and Sharkey, 1998, pp. 321–340).

Conclusions

The field of biologically inspired robotics has been classified into the three separate subfields of generalized, specific, and theoretical. General and theoretical biorobotics has a long but patchy history that is now a considerable and growing field. Specific biorobotics has gradually emerged from the other two and is fast making headway toward the goal of accurately modeling specific animal species. With the ever-increasing improvements in materials, sensors, and computing equipment, we can look forward to many exciting new developments over the coming decade and the transfer of the findings into engineering.

Road Map: Robotics and Control Theory

Related Reading: Arm and Hand Movement Control; Neuroethology, Computational; Potential Fields and Neural Networks; Reactive Robotic Systems

References

- Braitenberg, V., 1984, *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA: MIT Press. ♦
- Brooks, R., 1999, *Cambrian Intelligence: The Early History of the New AI*, Cambridge, MA: MIT Press.
- Chang, C., and Gadiano, P., Eds., 2000, *Biomimetic Robotics, Robot. Auton. Syst.*, 31(1–2):1–218 (special issue).
- Clark, A., 1997, *Being There: Putting Brain, Body and World Together Again*, Cambridge, MA: MIT Press.
- Dorigo, M., and Colombetti, M., 1998, *Robot Shaping: An Experiment in Behavior Engineering*, Cambridge, MA: MIT Press.
- Floreano, D., and Nolfi, S., 1997, Adaptive behaviour in competing co-evolving species, in *Proceedings of the Fourth European Conference on Artificial Life* (P. Husbands and I. Harvey, Eds.), Cambridge, MA: MIT Press.
- Gaussier, P., Ed., 1996, *Moving the Frontiers Between Robotics and Biology, Robot. Auton. Syst.*, 16:107–362 (special issue).
- Grey Walter, W., 1953, *The Living Brain*, New York: Norton. ♦
- Krose, B., Ed., 1995, Special issue on reinforcement learning and robotics. *Robot. Auton. Syst.*, 15:233–340.
- Nolfi, S., and Floreano, D., 1999, Learning and evolution, *Auton. Robots*, 7:89–113.
- Sharkey, N., Ed., 1997, *Robot Learning: The New Wave, Robot. Auton. Syst.*, 22(3–4):135–274 (special issue). ♦
- Sharkey, N., 1998, Learning from innate behaviors: A quantitative evaluation of neural network controllers, *Auton. Robots*, 5:317–334.
- Sharkey, N., and Ziemke, T., 2001, Mechanistic vs. phenomenal embodiment: Can robot embodiment lead to strong AI? *Cognit. Syst. Res.*, 2:251–262.
- Varela, F., Thompson, E., and Rosch, E., 1991, *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MA: MIT Press.
- Ziemke, T., and Sharkey, N., Eds., 1998, *Biorobotics, Connect. Sci.*, 10(3–4):161–360 (special issue).

Biophysical Mechanisms in Neuronal Modeling

Lyle J. Graham

Introduction

Models of single neurons span a wide range, with more or less fidelity to biological facts (see PERSPECTIVE ON NEURON MODEL COMPLEXITY; SINGLE-CELL MODELS; MECHANISMS IN NEURONAL MODELING). So-called biophysically detailed compartmental models of single neurons typically aim to quantitatively reproduce membrane voltages and currents in response to some sort of “synaptic” input. We may think of them as Hodgkin-Huxley-Rall models, based on the hypothesis of the neuron as a dynamical system of nonlinear membrane channels (e.g., conductances described by Hodgkin-Huxley kinetics; see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION; AXONAL MODELING) distributed over an electrotonic cable skeleton (e.g., as described by Rall dendritic cable theory; see DENDRITIC PROCESSING).

Such models can incorporate as much biophysical detail as desired (or practical), but, in general, all include some explicit assortment of voltage-dependent and transmitter-gated (synaptic) membrane channels. Many Hodgkin-Huxley-Rall models also include some system for describing intracellular Ca^{2+} dynamics, for example, to account for the gating of Ca^{2+} -dependent K^+ channels. Modeling these dynamics involves not only Ca^{2+} channels but often associated buffer systems and membrane pumps as well.

This article summarizes the application of the more common mathematical models of these basic biophysical mechanisms (Borg-Graham, 1999; Koch, 1999). The models for each of these mechanisms are at an intermediate level of biophysical detail, appropriate for describing macroscopic variables (e.g., membrane

currents, ionic concentrations) on the scale of the entire cell or anatomical compartments thereof.

First, we will discuss general issues regarding model formulations, and data interpretation for constructing models of biophysical mechanisms. We will then describe models for nonlinear channel properties, including Hodgkin-Huxley and Markov kinetic descriptions of voltage and second-messenger-dependent ion channels. Similar models aimed particularly for synaptic mechanisms are covered in SYNAPTIC INTERACTIONS. We will then discuss concentration systems, including models of membrane pumps and concentration buffers. Finally, examples of model definitions are illustrated using the Surf-Hippo Neuron Simulation System (Graham, 2002), pointing out an essential and minimal syntax that facilitates model documentation and analysis.

General Issues for Constructing Biophysical Models

Phenomenological and Mechanistic Models

A first consideration in choosing a mathematical model for a given cellular mechanism is whether the model is intended only to capture an empirical relationship between an independent variable (the input or signal) and a dependent variable (the output or response), or whether the model represents an explicit mechanistic hypothesis. Phenomenological models may be instantiated by a function with few (e.g., a low-dimensional polynomial fit) or many (e.g., look-up table) degrees of freedom, depending on the nature of the problem.

Of course, a mechanistic model can also have few or many parameters, but explanatory power tends to diminish with the number of parameters. The mechanistic and phenomenological model alternatives are not mutually exclusive, since the former may incorporate the latter, and in some ways the distinction between them is rather ad hoc.

Static (Instantaneous) and Dynamic (Kinetic) Models

Another basic consideration is whether the relation between signal and response is instantaneous on the time scale relevant to the entire system at hand. In some cases an instantaneous mechanism may permit analysis by exploiting separation of variables, for example assuming instantaneous activation for Na^+ currents relative to K^+ currents during spiking (see OSCILLATORY AND BURSTING PROPERTIES OF NEURONS). For cellular models that are solved by explicit integration over time, however, instantaneous relationships between state variables can introduce troublesome numerical instabilities unless there are intervening kinetics with slow time constants (relative to the time scale of the integration) that serve to “decouple” element dynamics at the faster time scale.

Deterministic and Stochastic Models

A stochastic component, or “noise,” in experimental measures is ubiquitous, for example, in the trial-to-trial variability of spike responses to deterministic stimuli, or in membrane voltage or membrane current fluctuations (especially in vivo). There is accumulating experimental and theoretical evidence that noise places an important constraint on information processing under some conditions while conversely serving a useful computational role in others.

Some system noise can be traced to the inherent stochasticity of molecular kinetics at the cellular level, and there is increasing interest in analyzing single-neuron models that explicitly consider this contribution. For example, simulations with stochastic Hodgkin-Huxley-type channel models can show functional dynamics that would be completely missed by deterministic models. A deterministic approximation should be valid when the number of channels is very large, the usual assumption, but the actual number in a local region of the neuron membrane may be rather low, considering both realistic estimates of channel densities and, especially, the small number of open channels near spike threshold (Schneidman, Freedman, and Segev, 1998).

The Experimenter’s Model Versus the Theorist’s Model

Every model is based on some empirical data set, but an often overlooked point is how the theorist’s model relates to, or rather is constrained by, that of the experimentalist. It may be a bit surprising to discover there is such a thing as an experimentalist’s model (which is not the same thing as an *experimental* model). However, in reality, experimental data are never arbitrary but reflect the experimenter’s explicit or implicit notion of either the necessary and sufficient parameters of a phenomenon, the functionally relevant mappings between signal and response, what is experimentally tractable (no one is able to do his or her “dream” experiment!), or some combination of all three. The first issue, in particular, is essentially equivalent to assuming some hypothetical model, but importantly, the associated experiments are not normally designed for *testing* that hypothesis (since it is taken as *a priori*). Examples include electrophysiological reports on whole-cell current kinetics, which usually focus on voltage-dependent activation and inactivation characteristics according to the classical model of Hodgkin and Huxley (described below). However, this paradigm, while practical, may miss crucial functional characteristics, basically by

not sufficiently characterizing certain important dynamical trajectories. We shall return to this point later in discussing Markov channel models.

Channel Models

Membrane channels underlie both intrinsic neuronal excitability and the direct postsynaptic action of synaptic transmission. The channel current I , assuming some permeant ion X , may be expressed as the product of a conduction term $f(V, \Delta[X])$ and a gating term $h(V, t, \dots)$:

$$I = f(V, \Delta[X])h(V, t, \dots)$$

where V is the membrane voltage, t is time, and $\Delta[X]$ represents the concentration gradient of X across the cell membrane. The ellipsis in the argument of $h()$ stands for the various ligand-dependent processes, for example, Ca^{2+} dependence or the action of synaptic neurotransmitters.

Ohmic and Permeation Conduction Models

The two common models of the conduction term $f()$ are the ohmic model (thermodynamic equilibrium conduction) and the constant-field permeation model (nonequilibrium conduction). In the ohmic model, current is proportional to the difference of the membrane voltage and the reversal potential for I :

$$f(V, \Delta[X]) = \bar{g}_X(V - E_X)$$

where \bar{g}_X is the maximum conductance. The reversal potential E_X for the ion X is given by the Nernst equation:

$$E_X = \frac{-RT}{zF} \log \frac{[X]_{out}}{[X]_{in}}$$

where R is the gas constant, F is Faraday’s constant, and T is temperature in degrees Kelvin. $[X]_{in}$ and $[X]_{out}$ are the intracellular and extracellular concentrations, and z is the valence of the permeant ion X . For more than one permeant ion (all with the same valence) and under some assumptions, the similar Goldman-Hodgkin-Katz voltage equation (e.g., Hille, 2002) may be used. Note that if the effect of channel current on $[X]$ is considered (see section on concentration integration), then the ohmic $f()$ is in fact implicitly nonlinear.

As the conducting ion moves farther from equilibrium (specifically the case for Ca^{2+}), the ohmic model becomes less accurate. A widely used nonequilibrium model is the constant field model, described by the Goldman-Hodgkin-Katz current equation. In this equation (Jack, Noble, and Tsien, 1983; Hille, 2002), the nonlinearity of permeation is explicit:

$$f(V, \Delta[X]) = \bar{p}_X \frac{Vz^2F^2}{RT} \frac{[X]_{in} - [X]_{out} \exp(-zFV/RT)}{1 - \exp(-zFV/RT)}$$

where \bar{p}_X is the permeability (*not* the conductance) of the channel (typically in cm^3/s). Note that at membrane potentials far from the reversal point (e.g., < -20 mV for Ca^{2+} channels) the Goldman-Hodgkin-Katz current equation becomes linear, and thus the ohmic model may suffice if the model voltages are appropriately bounded.

Channel Gating à la Hodgkin and Huxley: Independent Voltage-Dependent Gating Particles

Hodgkin and Huxley (1952) described channel gating as an interaction between independent two-state (open and closed) elements or “particles,” all of which must be in the open state for channel conduction (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION; AXONAL MODELING). The state dynamics of each particle are described with first-order kinetics:



where x_C and x_O represent the closed and open states of gating particle x , respectively. $\alpha(V)$ and $\beta(V)$ are the forward and backward rate constants of the particle as a function of voltage, respectively.

An Extended Hodgkin and Huxley Model

While Hodgkin and Huxley hypothesized that the steady-state behavior of each particle fit a Boltzmann distribution, their underlying rate equations were essentially ad hoc fits to the experimental data. Although taken as a canonical form by countless cell models, one consequence is that there is not an obvious relationship between the equations' parameters and the more "observable" steady-state, $x_\infty(V)$, and time-constant, $\tau_x(V)$, functions associated with Equation 1.

The Hodgkin-Huxley model can be recast in more explicit form by considering parameters of a single-barrier kinetic model for each particle (Jack et al., 1983; Borg-Graham, 1991, 1999). In its basic form this formulation has five parameters for each particle, compared to six parameters in the Hodgkin-Huxley model. Nevertheless, this formulation may be readily fitted to the original Hodgkin-Huxley equations of squid axon I_{Na} and I_K ; the error is comparable to the error between the original equations and the data to which they were fit (cf. Figures 4, 7, and 9 in Hodgkin and Huxley, 1952).

We first derive the expressions the forward, $\alpha'_x(V)$, and backward, $\beta'_x(V)$, rate constants of the single-barrier transition. The parameter z (dimensionless) is the *effective* valence of the gating particle: when positive (negative), the particle opens (closes) with depolarization; thus it is an "activation" ("inactivation") particle. The effective valence is the product of the actual valence of the particle and the proportion of the membrane thickness that the particle moves through during state transitions. γ (dimensionless, between 0 and 1) is the asymmetry of the gating particle voltage sensor within the membrane (symmetric when $\gamma = 0.5$). K is the leading rate coefficient of both $\alpha'_x(V)$ and $\beta'_x(V)$. This term can be described in terms of Eyring rate theory, but here we just take K as a constant. $V_{1/2}$ is the voltage for which $\alpha'_x(V)$ and $\beta'_x(V)$ are equal. The final equations for $\alpha'_x(V)$ and $\beta'_x(V)$ are then:

$$\alpha'_x(V) = K \exp \left(\frac{z\gamma(V - V_{1/2})F}{RT} \right)$$

$$\beta'_x(V) = K \exp \left(\frac{-z(1 - \gamma)(V - V_{1/2})F}{RT} \right)$$

An additional parameter, τ_0 (*not* the passive membrane time constant), is crucial for fitting the expressions to the original Hodgkin-Huxley equations. τ_0 represents a rate-limiting step in the state transition, for example, "drag" on the particle conformation change (similar considerations have been explored for other, more general, kinetic schemes; e.g., Patlak, 1991), and may be incorporated directly into the expression for the time constant $\tau_x(V)$. $x_\infty(V)$, however, is not affected by τ_0 :

$$\tau_x(V) = \frac{1}{\alpha'_x(V) + \beta'_x(V)} + \tau_0$$

$$x_\infty(V) = \frac{\alpha'_x(V)}{\alpha'_x(V) + \beta'_x(V)}$$

Two additional parameters, α_0 and β_0 , may be considered in some cases, although they are not necessary in reproducing the original Hodgkin-Huxley equations. These parameters are voltage-independent forward and backward rate constants, respectively, of parallel state transitions. If considered, these transitions will change the final forms of $\tau_x(V)$ and $x_\infty(V)$.

The parameters of this form have clear relationships to the corresponding $x_\infty(V)$ and $\tau_x(V)$ functions. Thus, the $V_{1/2}$ parameter gives the midpoint and z sets the steepness of the $x_\infty(V)$ sigmoid. The symmetry parameter γ determines the skew of $\tau_x(V)$: $\gamma = 0.5$ gives a symmetric bell-shaped curve for $\tau_x(V)$, which otherwise bends to one side or the other as γ approaches 0 or 1. z sets the width of $\tau_x(V)$, unless γ is equal to either 0 or 1, in which case $\tau_x(V)$ becomes sigmoidal and thus z sets the steepness as for $x_\infty(V)$.

With this scheme a particle with a voltage-independent rate constant can be represented by setting $1/K \ll \tau_0$ (and $\alpha_0 = \beta_0 = 0$), thus making τ_0 the effective time constant. Likewise, both the time constant and steady state become voltage independent by setting $K = 0$ and choosing the appropriate α_0 and β_0 .

Determining the Number of Particles in Hodgkin-Huxley Models

The Hodgkin-Huxley paradigm includes the possibility of multiple gating particles of a given type associated with a given channel. Some experimental papers report fitting integer powers of hypothetical gating particles to the observed kinetics, but more typically steady-state activation or inactivation data are simply the observed macroscopic behavior (that is, reflecting the steady state of the ensemble of particles). Thus, gating particle powers for channel models can often be considered as a free parameter.

Gating particle powers greater than 1 have several kinetic consequences, including a sigmoidal "delayed" time course of activation (Hodgkin and Huxley, 1952), a more rapid approach to 0 in the steady-state characteristic as a function of voltage, and a shift in either the peak (when $0 < \gamma < 1$) or inflection point (when $\gamma = 0$ or 1) of $\tau_x(V)$ in the direction of voltage for which $x_\infty(V)$ tends to 0.

Channel Gating as Dynamical Systems à la Markov Models

The independence and simplicity of the Hodgkin-Huxley gating particle models have at least two advantages: model kinetics can be predicted in an intuitive way, and their numerical evaluation is efficient (Hines, 1984). In addition, as mentioned, electrophysiological measures of whole-cell currents are often guided by this model. The two-state gating model can also be readily adapted to include factors such as intracellular $[Ca^{2+}]$, by using the appropriate functions for α and β (see below).

On the other hand, the independence of the two-state Hodgkin-Huxley particles constrains the equivalent state space description (e.g., allowed state transitions) given by the more general Markovian model (see SYNAPTIC INTERACTIONS). General Markov kinetic models are standard for detailed biophysical analysis of single-channel kinetics, but there have been relatively few applications in the neural modeling literature. One practical limitation is that Markov models are often much more computationally expensive than the Hodgkin-Huxley model. Nevertheless, the richer dynamics of Markov models may prove necessary for capturing functional properties of some channel types, including subthreshold steady-state Na^+ channel rectification (Figure 4), delay of activation for Na^+ and K^+ currents, and the coupling between opening of K^+ channels by Ca^{2+} entering during the action potential and subsequent inactivation (Borg-Graham, 1999).

Although the Markovian framework puts no restrictions on the functions that define state transitions (other than the no-memory condition), the form presented above of the $\alpha(V)$ and $\beta(V)$ functions for the extended Hodgkin-Huxley model is convenient and very general. Another form is the following squeezed exponential formula for the transition rate $\alpha_{ij}(V)$ from state i to state j :

$$\alpha_{ij}(V) = \left(\tau_{min} + \left((\tau_{max} - \tau_{min})^{-1} + \exp\left(\frac{(V - V_{1/2})}{k}\right) \right)^{-1} \right)^{-1} \quad (2)$$

where the inverse of τ_{min} (analogous to τ_0 in the extended Hodgkin-Huxley model) and τ_{max} put upper and lower bounds, respectively, on the rate constant $\alpha_{ij}(V)$. Note that there is an implicit coefficient of the exponential term of 1/ms (same units as either $1/\tau_{min}$ or $1/\tau_{max}$) in this equation.

Ca^{2+} -Dependent Gating

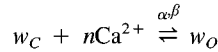
Neural models have used a variety of explicit relationships between the concentration of some second messenger and the activation state of the target mechanism. Here we consider a range of examples that have been used to describe Ca^{2+} -dependent K^+ channels.

A simple instantaneous model for Ca^{2+} -dependent gating is given by a static rectified power function of concentration with some threshold θ_{Ca} , reminiscent of firing rate models:

$$w = K \times \sigma([Ca^{2+}]^n - \theta_{Ca})$$

where $\sigma(x) = 0$ for $x < 0$, and $\sigma(x) = x$ otherwise, for the gating variable w .

A simple kinetic model for Ca^{2+} -dependent gating can be described by the following reaction. Assume that w_C and w_O represents the closed and open probabilities, respectively, of a Ca^{2+} -dependent gating particle w with forward and backward rate constants α and β , respectively:



Note that the open state w_O is bound to n Ca^{2+} ions. One can also imagine a similar but reverse reaction for the description of Ca^{2+} -dependent inactivation, as has been reported for some Ca^{2+} channels. In the more general Markovian framework, w_C and w_O refer to two adjacent states out of the entire state space. This scheme assumes that binding with Ca^{2+} ions is cooperative: either all binding sites are occupied or none are. We may also consider a τ_0 parameter as in the extended Hodgkin-Huxley model. If we assume that the binding of Ca^{2+} in this reaction does not appreciably change $[Ca^{2+}]$, then the steady-state value for w , w_∞ , and the time constant for the kinetics, τ_w , are given by:

$$w_\infty = \frac{\alpha}{\alpha + \beta[Ca^{2+}]_in^{-n}}$$

$$\tau_w = \frac{1}{\alpha[Ca^{2+}]_in^n + \beta} + \tau_0$$

An important distinction is whether or not a Ca^{2+} -dependent channel is also dependent on voltage. For example, in recordings of the large-conductance Ca^{2+} -dependent K^+ (BK) channel, Barrett, Magleby, and Pallotta (1982) found an approximate third-power relationship between channel open times and $[Ca^{2+}]$ that was strongly facilitated by depolarization. At a membrane voltage of 10 mV, channels became open with a $[Ca^{2+}]$ threshold of about 1 μ M.

If the dependences are separable, it may be convenient to consider a product of voltage-only and Ca^{2+} -only gating terms, for example according to the formulations presented earlier. Otherwise, a single gating “particle” must take into account both voltage and Ca^{2+} . A direct voltage dependence of the simple kinetic scheme above can be added in a number of ways, for example by adding a voltage-dependent term to the forward rate constant, now defined as $\alpha(V, [Ca^{2+}]_in)$:

$$\alpha(V, [Ca^{2+}]_in) = \alpha_V(V) \times \alpha[Ca^{2+}]_in^n$$

where, e.g., $\alpha_V(V)$ is the squeezed exponential function of voltage in Equation 2, such that the forward reaction speeds up with depolarization.

Moczydlowski and Latorre (1983) proposed a detailed Markovian kinetic scheme for the Ca^{2+} - and voltage-dependent gating of the BK channel, which has been interpreted in several neuron models. The essential dynamics are captured by a two-state scheme as in Equation 1, with rate constants dependent on both voltage and Ca^{2+} , thus:

$$\beta(V, [Ca^{2+}]_in) = \beta_0 \left(1 + \frac{k_1(V)}{[Ca^{2+}]_in} \right)^{-1}$$

$$\alpha(V, [Ca^{2+}]_in) = \alpha_0 \left(1 + \frac{[Ca^{2+}]_in}{k_4(V)} \right)^{-1}$$

where

$$k_i(V) = k_i(0) \times \exp\left(\frac{-V\delta_i FZ}{RT}\right)$$

Ionic Concentration Dynamics

An inevitable consequence of channel currents is that the concentrations on either side of the membrane will change as a function of electrical activity. In addition to the negative feedback on channel currents already mentioned (as a result of a reduction in driving force), such changes can have a variety of other functional consequences. These include the activation of intracellular or extracellular receptors, the most important being those that underlie the myriad Ca^{2+} -dependent pathways (including the Ca^{2+} -dependent channel gating just described). We may also consider the role of the membrane pumps, which tend to maintain ionic gradients (see MECHANISMS IN NEURONAL MODELING), and of the intracellular buffer systems. In the following discussion we emphasize Ca^{2+} dynamics, but similar considerations are relevant for other ions.

Concentration Integrators

Most neuron models that consider concentration changes rely on some partition of the extracellular and intracellular space into a set of well-mixed compartments (e.g., “shells”), with or without an “infinite” compartment with a fixed concentration. Simple diffusion is normally assumed between compartments, according to the geometry of the partitioning and assumptions about the diffusion coefficient for the free ion, D . Compartments adjacent to the cell membrane also take into account ion flow across the membrane, e.g., due to channels and pumps. The physical partitioning into compartments depends on the question being addressed, with the simplest system being a single intracellular compartment (extracellular concentration being assumed constant).

Any model of $[Ca^{2+}]$ must take into account not only the influx of Ca^{2+} but also some mechanism for the removal of Ca^{2+} . The simplest method is to include a steady-state term in the differential equation(s) describing $[Ca^{2+}]$. In the general case this value is associated with a second parameter corresponding to the time constant for concentration decay.

Membrane Pump Models

More explicit models of ion removal includes mechanisms, such as membrane-bound pumps, that transport Ca^{2+} , K^+ , Na^+ , and other ions against their respective concentration gradients. A general pump model may be described with a Michaelis-Menton mechanism, assuming no appreciable change in the extracellular $[Ca^{2+}]$:

$$J_{Ca^{2+}} = V_{max} \frac{[Ca^{2+}]_{in}}{K_d + [Ca^{2+}]_{in}} - J_{leak}$$

where $J_{Ca^{2+}}$ is the removal rate of Ca^{2+} per unit area, V_{max} is the maximum flux rate per unit area, and K_d is the half-maximal $[Ca^{2+}]_{in}$. J_{leak} compensates for the resting pump rate and is typically adjusted so that there is no net pump current at rest, given some resting activation of Ca^{2+} channels.

For example, the spine model by Zador, Koch, and Brown (1990) included two Ca^{2+} pumps with Michaelis-Menton kinetics: one high-affinity, low-capacity, corresponding to a Ca ATPase-driven mechanism, and the other low-affinity, high-capacity, corresponding to a Ca^{2+}/Na^{+} exchange mechanism (see also Koch, 1999). These pumps were treated as separate currents in the $[Ca^{2+}]$ differential equation. Other models have incorporated a pump that binds to intra- and extracellular Ca^{2+} with various rate constants. These reactions are then solved simultaneously with another binding reaction between $[Ca^{2+}]_{in}$ and a buffer.

Buffer Models

Endogenous intracellular Ca^{2+} buffers have a strong effect on free intracellular $[Ca^{2+}]_{in}$. Cell models that consider Ca^{2+} dynamics have incorporated buffer mechanisms of varying complexities, including solving the dynamical equations for the buffer- Ca^{2+} reaction during the course of the simulation. Note that in some models an explicit (instantaneous) buffer mechanism is replaced by adjusting Ca^{2+} sensitivities of Ca^{2+} -dependent mechanisms, such as Ca^{2+} -dependent K^{+} channels.

A simple way to treat intracellular buffering of Ca^{2+} is to assume a nonsaturated buffer (i.e. $[Bu] \gg [Ca^{2+}]_{in}$, where $[Bu]$ is the concentration of buffer binding sites) with instantaneous kinetics. The key parameter, β_{Bu} , in this mechanism equals the ratio of the concentration of bound Ca^{2+} and free Ca^{2+} :

$$\beta_{Bu} = \frac{[Ca^{2+}]_{in}^{bound}}{[Ca^{2+}]_{in}^{free}}$$

and thus is a function of $[Bu]$. This mechanism implies that the measured $[Ca^{2+}]_{in}$ is equal to the total $[Ca^{2+}]_{in}$ divided by $(\beta_{Bu} + 1)$. For nondiffusional models of $[Ca^{2+}]_{in}$ (e.g., where there is one Ca^{2+} compartment per electrical compartment), this is the only role of the instantaneous buffer. For multiple-compartment systems, the effective diffusion constant D' applied to the difference in $[Ca^{2+}]$ between compartments must also be adjusted to take into account the instantaneous buffer, by setting D' equal to $D/(\beta_{Bu} + 1)$. A variation on this scheme would be to assume that β_{Bu} is a function of each compartment. In this case the diffusion equation between compartments would reference the original D , with the concentration difference between any two compartments determined by the difference of the total concentrations, weighted by the appropriate β_{Bu} s.

Practical Aspects of Coding Biophysical Models

The translation of experimental data on some biophysical mechanism into simulator code, within the framework of a given mathematical model, and the reverse process (which is a necessary step in formulating experimental predictions from a model) have received little attention. However, these steps have many practical aspects, not the least of which is that as models become more and more complex, the opportunity for errors becomes more and more serious. For this reason it is useful to consider model *syntax* (see NEUROSIMULATION: TOOLS AND RESOURCES; GENESIS SIMULATION SYSTEM; NEURON SIMULATION ENVIRONMENT; NSL NEURAL SIMULATION LANGUAGE).

In most situations, it is desirable that a simulator program act essentially as a “black box,” so that model analysis concerns only the input (some collection of model definition files) and the output (numerical data, usually time sequences). Thus, when composing the model definition, one should ideally be able to focus on the model algorithms and their parameters, rather than on their implementation. Model syntax should therefore allow the expression of mathematical (and symbolic, if appropriate) relationships in as close to a “natural” syntax as possible. In other words, one should be able to simply “write down the equations” defining the model. Certainly a practical consequence of such a syntax is that the learning curve for the simulator is reduced, but more important over the long term is simply that if model definitions are easier to read, they are also easier to verify, document, and change.

To illustrate these ideas, here we present examples of biophysical model definitions taken from the Surf-Hippo Neuron Simulation System (Graham, 2002). This system is written in Lisp, an important point, since the system exploits many advantages of this truly high-level language that are well known to the AI community (at the same time having a numerical performance on par with languages such as C). In particular, Lisp supports an emphasis on more

```
(CHANNEL-TYPE-DEF
  (NA-HH
    (GBAR-DENSITY . 1200) ; pS/um2
    (E-REV . 50) ; mV
    (V-PARTICLES . ((M-HH 3) (H-HH 1))))))

(PARTICLE-TYPE-DEF
  (M-HH
    (CLASS . :HH)
    (ALPHA . (LAMBDA (VOLTAGE)
      (/ (* -0.1 (- VOLTAGE -40))
        (1- (EXP (/ (- VOLTAGE -40) -10)))))))
    (BETA . (LAMBDA (VOLTAGE)
      (* 4 (EXP (/ (- VOLTAGE -65) -18)))))))
```

Figure 1. The Surf-Hippo definitions of the classical Hodgkin-Huxley model of the squid axon Na^{+} channel and the associated M-activation gating particle. The plethora of parentheses may seem daunting; however, all formatting (including indentation) is done automatically by Lisp-savvy editors (such as Emacs). The last line in the CHANNEL-TYPE-DEF form specifies three M-HH and one H-HH gating particles. Surf-Hippo model definitions allow concise inclusion of arbitrary functions, in this case for the ALPHA and BETA rate constants (refer to equations of the Hodgkin-Huxley Na^{+} channel in AXONAL MODELING, q.v.) in the PARTICLE-TYPE-DEF form. The LAMBDA symbol denotes the beginning of a function definition. In the context of gating particle definitions, Surf-Hippo assumes only that the rate functions take a single VOLTAGE argument (in millivolts), and return a rate value (in ms^{-1}). Comments are indicated with a semicolon. Expression precedence is unambiguous with the prefix notation of Lisp. The first element of each (parenthesized) list defines the operation applied to the rest of the list, and in nested lists everything is evaluated from the inside out: e.g., $(+ A (* B C) D)$ is $A + BC + D$.

```
(PARTICLE-TYPE-DEF
  (M-HH-FIT
    (CLASS . :HH-EXT)
    (VALENCE . 2.7)
    (GAMMA . 0.4)
    (BASE-RATE . 1.2) ; 1/ms
    (V-HALF . -40) ; mV
    (TAU-0 . 0.07))) ; ms
```

Figure 2. The Surf-Hippo definition for the extended Hodgkin-Huxley model for the M-activation particle type of the squid axon Na^{+} channel.

```

(PARTICLE-TYPE-DEF
  \ (NA-X-HPC
    (CLASS . :MARKOV)
    (STATES . (0 I C1 C2))
    (OPEN-STATES . (0))
    (STATE-TRANSITIONS .
      ((O I 3)
        (O C1 (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -51 :K -2 :TAU-MIN 1/3))
        (C1 O (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -42 :K 1 :TAU-MIN 1/3))
        (O C2 (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -57 :K -2 :TAU-MIN 1/3))
        (C2 O (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -51 :K 1 :TAU-MIN 1/3))
        (I C1 (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -53 :K -1 :TAU-MAX 100 :TAU-MIN 1))
        (C1 C2 (SQUEEZED-EXPONENTIAL VOLTAGE :V-HALF -60 :K -1 :TAU-MAX 100 :TAU-MIN 1))))))

```

Figure 3. The Surf-Hippo definition of a Markovian gating particle for a hippocampal pyramidal cell Na^+ channel model (Borg-Graham, 1999). Transition rates between states, in ms^{-1} , are defined with either constants (for example, 3 ms^{-1} for the transition from state O to state I) or functions

of voltage (as indicated by the “dummy” variable VOLTAGE). The SQUEEZED-EXPONENTIAL function (Equation 2) is built into Surf-Hippo (when :TAU-MAX is not specified, then the minimum rate is 0).

declarative descriptions (emphasizing *what* kind of model is desired), rather than imperative ones (emphasizing *how* to construct a model). Thus, model syntax in Surf-Hippo is designed to minimize the actual code for mechanism specification: correspondingly, these examples illustrate the necessary and sufficient parameters for each mechanism, avoiding “overhead” code that would be simulator specific. Surf-Hippo also includes automatic generation of mechanism definition code, for example, allowing capture of the “state” of a given mechanism model that has been modified on-line during automatic or manual parameter exploration. This capability, which is facilitated by both the minimal requirements for model specification and the relative ease with which Lisp programs may be able to write Lisp code, is important for avoiding errors when documenting model results.

Figure 1 illustrates the definitions of the classical Hodgkin-Huxley model of the squid axon Na^+ channel and the associated M-activation gating particle, showing in particular how arbitrary functions are represented. Once the basic syntax of Lisp is grasped, the human readability of this format is enhanced because it includes only the essential kinetic parameters. As a comparison, the equivalent source code in similar simulation systems such as GENESIS and NEURON is about two to three times larger.

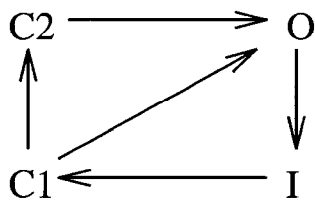


Figure 4. State diagram of a hypothetical Markov gating model used for I_{Na} in a model of hippocampal pyramidal cells (Borg-Graham, 1999). From the single inactivated state I, the two closed states C_i are reached with increasing hyperpolarization. The $C_i \rightarrow O$ transitions implement in effect distinct thresholds, occurring at progressively lower potentials with increasing i . Likewise, the $I \rightarrow C_1$ and $C_1 \rightarrow C_2$ transitions occur at voltages hyperpolarized to the associated $C_1 \rightarrow O$ and $C_2 \rightarrow O$ transitions, respectively, somewhat like a ratchet mechanism. The $O \rightarrow I$ transition is voltage independent. The arrows denote the dominant transitions during spike depolarization/repolarization. One important aspect of this model is that the inactivation state is reached only after channel opening, as reported from studies of single Na^+ channels (Patlak, 1991; Hille, 1992). Such coupling contradicts the central assumption of independent activation and inactivation kinetics in the Hodgkin-Huxley model.

Several parameterized models of biophysical mechanisms are included in Surf-Hippo, including those discussed in this article. For example, Figure 2 shows the definition for the extended Hodgkin-Huxley model of the Na^+ channel M-gating particle. Markov models for particle gating are also readily represented in this system. As an example, Figure 3 illustrates the definition of a Markovian gating particle for a hippocampal pyramidal cell Na^+ channel model (Borg-Graham, 1999); the state diagram is shown in Figure 4.

Discussion

Biophysical details are likely to be crucial for understanding neural computation (see MECHANISMS IN NEURONAL MODELING). This process entails an informed trade-off between incorporating every known experimental nuance of a given cellular mechanism and the practical application of abstractions and simplifications that capture essential dynamic relationships between biological molecules and various neuronal signals. In this article we have presented some of the more commonly used mathematical descriptions for these relationships. We note in closing that an increasingly important (and unavoidable) problem with complicated neural models that rely on these sorts of mechanisms is the lack of formal or analytic verification. This situation calls for alternative methods, in particular the cross-validation of numerical results using several tools of similar capability (e.g., NEURON, GENESIS, Surf-Hippo). Practical aspects of coding biophysical models, such as the minimal model syntax discussed here, should facilitate such efforts.

Road Map: Biological Neurons and Synapses

Related Reading: Activity-Dependent Regulation of Neuronal Conductances; Axonal Modeling; GENESIS Simulation System; Ion Channels: Keys to Neuronal Specialization; NEURON Simulation Environment; Neurosimulation: Tools and Resources; NSL Neural Simulation Language; Oscillatory and Bursting Properties of Neurons; Perspective on Neuron Model Complexity; Single-Cell Models; Synaptic Interactions

References

- Barrett, J. N., Magleby, K. L., and Pallotta, B. S., 1982, Properties of single calcium-activated potassium channels in cultured rat muscle, *J. Physiol.*, 331:211–230.
- Borg-Graham, L., 1991, Modelling the non-linear conductances of excitable membranes, in *Cellular Neurobiology: A Practical Approach* (J. Chad and H. Wheal, Eds.), New York: IRL/Oxford University Press, chap. 13.

- Borg-Graham, L., 1999, Interpretations of data and mechanisms for hippocampal pyramidal cell models, *Cereb. Cortex*, 13:19–138. ♦
- Graham, L., 2002, The Surf-Hippo Neuron Simulation System, available: <http://www.cnrs-gif.fr/iaf/iaf9/surf-hippo.html>, v3.0.
- Hille, B., 2002, *Ionic Channels of Excitable Membranes*, 3rd ed., Sunderland, MA: Sinauer. ♦
- Hines, M., 1984, Efficient computation of branched nerve equations, *Int. J. Biomed. Comput.*, 15:69–76.
- Hodgkin, A. L., and Huxley, A. F., 1952, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.*, 117:500–544.
- Jack, J. J. B., Noble, D., and Tsien, R. W., 1983, *Electric Current Flow in Excitable Cells*, Oxford, Engl.: Clarendon Press. ♦
- Koch, C., 1999, *The Biophysics of Computation: Information Processing in Single Neurons*, Oxford, Engl.: Oxford University Press. ♦
- Moczydlowski, E., and Latorre, R., 1983, Gating kinetics of Ca^{2+} -activated K^+ channels from rat muscle incorporated into planar lipid bilayers, *J. Gen. Physiol.*, 82:511–542.
- Patlak, J., 1991, Molecular kinetics of voltage-dependent Na^+ channels, *Physiol. Rev.*, 71:1047–1080.
- Schneidman, E., Freedman, B., and Segev, I., 1998, Ion-channel stochasticity may be critical in determining the reliability and precision of spike timing, *Neural Computat.*, 10:1679–1703.
- Zador, A., Koch, C., and Brown, T. H., 1990, Biophysical model of a Hebbian synapse, *Proc. Natl. Acad. Sci. USA*, 87:6718–6722.

Biophysical Mosaic of the Neuron

Lyle J. Graham and Raymond T. Kado

Introduction

In this article we broadly review the biophysical mechanisms of neurons that are likely to be relevant to computational function (Table 1). These mechanisms operate within the complex three-dimensional anatomy of the single neuron and are manifested by electrical and chemical interactions between ions on either side of the cell membrane and the diverse proteins and other molecules

Table 1. Neuronal Biophysical Mechanisms Relevant to Computational Function

Ion channels
Control by intrinsic signals (membrane voltage, intracellular molecules)
Control by extrinsic signals (extracellular molecules, e.g., released from presynaptic terminals)
Receptors
Ionotropic: Direct control of ion channels
Metabotropic: Indirect control of ion channels and other internal systems
External binding sites (synaptic and pancreatic)
Internal binding sites associated with neuronal and organelle membranes
Control of internal biochemical systems
Enzymes (kinases and phosphatases, which determine the state of most proteins; others)
Gap junctions
Pumps, transporters, exchangers (electrogenic and nonelectrogenic)
Organelles
Ca^{2+} sequestering and release (endoplasmic reticulum, mitochondria)
Protein synthesis and metabolism
Maintenance and modulation of three-dimensional structure (spines, dendritic morphology)
Transmitter sequestering and release (synaptic vesicles)
Cytoplasmic biochemical systems
Transmitter synthesis and degradation
G proteins: Initiate second messenger release after receptor activation
Second messengers: Diffusible molecules linking various stages of internal biochemical systems
Effectors: Targets of second messengers, including channels and enzymes (kinases, phosphatases)
Three-dimensional structure
Macroscopic anatomy (soma, dendritic and axonal trees)
Microscopic anatomy (spines, synaptic junctions, variations in dendritic or axonal dimensions)
“Electrotonic” anatomy (modulated by state of local membrane)
Geometrical synaptic and channel distribution
Functional synaptic localization (e.g., retinotopic, tonotopic)
Computational synaptic localization (e.g., on-the-path interactions, coincidence detection)

embedded in the membrane and within the cytoplasm (Figure 1). The signals mediating these interactions may be defined by the voltage across the cell membrane or by concentrations of specific molecules in specific conformational or metabolic states. The first case relies on the voltage sensitivity of various membrane proteins; the second relies on a vast multitude of receptor proteins that link the functional state of neuronal proteins with the external or internal concentrations of ions and molecules.

It may be noted that none of these cellular mechanisms is unique to neurons. For example, essentially all the mechanisms discussed in this article may be relevant when considering a possible computational role of the neuroglia network (Laming et al., 2000). By the same token, neurons (and glial cells) include the essential mosaic of biochemical systems found in all cells required for metabolism, reproduction, growth, and repair. The complexity is daunting. Here we focus on the better-known elements most clearly linked to the reception, processing, and transmission of neuronally represented information. It may seem that there are so many such elements, and an even larger number of unknown relationships, that it would not be possible for a theory to take all of the actual dynamic behaviors into consideration. Nevertheless, it also seems likely that an oversimplification of these interactions—for example, in the extreme case by describing single neuron function as an abstracted trigger device—may put fundamental limits to the explanatory and predictive power of any neural model. The challenge remains, then, to develop a description of single neuron function that can serve as the foundation for a practical yet sufficient neural theory.

We start with a metaphor, the mosaic neuron. A *mosaic* is a collection of discrete parts, each with unique properties, fitted together in such a way that an image emerges from the whole in a nonobvious way. Similarly, the neuronal membrane is packed with a diversity of receptors and ion channels and other proteins with a recognizable distribution. In addition, the cytoplasm is not just water with ions, but a mosaic of interacting molecular systems that can directly affect the functional properties of membrane proteins. Whether for the developing or for the mature neuron, this mosaic is not stationary. To begin with, neuronal proteins are constantly recycled, as is the case for all cells. Furthermore, on both long and short time scales, most mechanistic theories for learning and memory implicate physical changes in various cellular constituents. On time scales of seconds or less, different signaling systems impinging on the neuron from the network or present in the cytoplasm can modify the properties of the mosaic elements, and in some cases their distribution within the cell (see ACTIVITY-DEPENDENT REGULATION OF NEURONAL CONDUCTANCES). Thus, just as a mo-

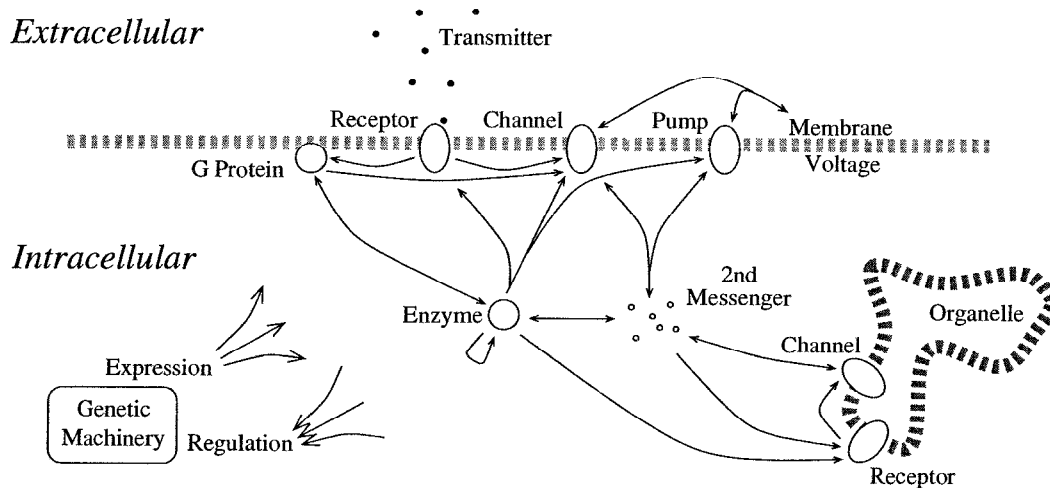


Figure 1. Sketch of the molecular circuit underlying the neuron's biophysical mosaic. This caricature outlines the many interrelated control paths among the molecular elements that determine the functional properties of the neuron. For example, activation of an ionotropic synapse starts with the binding of extracellular transmitter to the membrane receptor, which then directly turns on an associated ion channel. Alternatively, activation of a metabotropic synapse is initiated by transmitter binding to a receptor, which

then activates a G protein, which turns on an enzyme, which raises the concentration of a second messenger, which, finally, activates a target ion channel. Many control pathways are immediately bidirectional: current through an ion channel changes the membrane voltage, which in turn can control the gating of that same channel. Some elements can even control themselves, for example autophosphorylating enzymes.

saic painting provokes perception of a complete image out of a maze of individually diversified tiles, current thinking holds that a given neuron performs a well-defined computational role that depends not only on the network of cells in which it is embedded but also to a large extent on the dynamic distribution of macromolecules throughout the cell.

The Minimal Essential Model and the Biophysical Mosaic

It remains an open question as to what constitutes a minimal neuron model for reproducing functional neuronal computation (Meunier and Segev, 2000; see also CANONICAL NEURAL MODELS). This is in part because to date, only a handful of neuron circuit models come close to predicting known experimental data in any nontrivial way. The question of finding a minimal model is hardly an academic one, as can be appreciated by reviewing the dimensionality of the mosaic neuron (Table 2).

Whatever the minimal essential model turns out to be, a detailed knowledge of neuronal biophysics is most likely necessary for understanding the system behavior (even if this understanding is not sufficient). The clearest evidence for this point of view comes from psychopharmacology: although we lack a clear understanding of the mechanisms, we know that adding certain chemicals to the brain parenchyma can qualitatively alter cognitive behavior. We know that the direct action of psychotropic drugs is probably to change one or more biophysical properties at the microscopic cellular level, such as blocking an ion channel, altering the binding kinetics of a receptor, modulating a biochemical pathway, and so on, rather than acting at a more macroscopic systems level, such as cleanly disconnecting a circumscribed subcircuit from the entire network. We know that physical access to the brain is necessary for this action (preventing a drug from crossing the blood-brain barrier eliminates its effect), and we also know, in many cases, that some neurons have highly specific membrane receptors for a given psychotropic molecule that are often localized in very restricted areas at the level of brain substructures and even at the single-cell level. Often there is direct evidence of a drug's effect in electro-

physiological measurements of single cells, when a change in intrinsic response properties or synaptic dynamics is seen after a given chemical is added to the fluid bathing the nervous tissue.

The Mosaic's Tiles

We will now review the major proteins that compose the neuron mosaic and discuss some basic implications of their diversity and complexity. These macromolecules include ion channels, receptors (described along with the molecules that activate them), enzymes, gap junctions, pumps, exchangers, and transporters. Note that these classifications can sometimes overlap. For example, an ionotropic receptor is a protein multimer that includes both a receptor part and a channel part.

Several texts may be consulted for more detail on these mechanisms (Johnston and Wu, 1995; Weiss, 1996; Koch and Segev, 1998; Fain, 1999; Hille, 2002). In particular, the textbook by Koch (1999) provides an explicit foundation for the computation/algorithm/implementation trinity that is fundamental for understanding brain function.

Ion Channels

Ion channels are membrane-spanning proteins that, owing to their conformational states which allow the passage of ionic current, are the primary basis for the dynamical electrical behavior of neurons (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION). The permeability of the conducting states and the kinetics governing state transitions (generally referred to as channel gating) can be affected by a variety of factors, principally the membrane voltage and the intra- and extracellular concentrations of the permeable ions and other specific molecules. Sensitivity to extracellular molecules is generally mediated by either direct action on the channel or various receptor proteins (e.g., in response to neurotransmitters), as discussed below. Molecules that affect channel gating from the inside include second messengers. The kinetic relationship between membrane voltage, the concentration of neurotransmitters, second messengers, and a channel's conductance state can be quite complex, a point we return to later.

Table 2. Quantitative Summary of the Neuron's Biophysical Mosaic Relevant to Computational Function

Spatial scales (voltage and concentration transients): <1 to thousands of microns
Temporal scales
Kinetics of gating, binding, and diffusion: <1 ms to seconds
Gene expression: days to years?
Anatomy
Tens to hundreds of dendritic and axonal branches
Models of electrotonic structure can require thousands of compartments
Synapses and channels
Thousands of pre- and postsynaptic sites
Five major categories of charge carriers: Na^+ , K^+ , Ca^{2+} , Cl^- , other (e.g., "cationic," proton, etc.)
Tens of types for each category, several of which may be expressed in a single neuron
Receptors and associated agonists (neurotransmitters, second messengers, etc.)
Approximately 30 major types, several of which may be expressed in a single neuron
Possibly tens of identified subtypes for some major receptor types

A brute force map of a single neuron would be very large indeed. For example, compartmental models of the dendritic tree can require hundreds or thousands of coordinates, corresponding to spatial scales relevant for representing gradients of membrane voltage or of the concentration of intracellular molecules. A given neuron may have thousands of synaptic inputs, each associated with one of several types of receptors, and the cell membrane can include tens of types of ion channels. Furthermore, the different synaptic receptor and channel types are typically scattered inhomogeneously over the neuron. It is also important to consider the stationarity of the map. For computations over hundreds of milliseconds or less, the map may properly be thought to be static, but at longer time scales it may be necessary to consider a dynamic layout of the mosaic. For another *carte du monde* of computational cellular mechanisms and their spatial and temporal scales, see Figure 21.2 in Koch, 1999.

Since channels are the most direct mechanism determining the basic firing properties of the cell (e.g., regular adapting, bursting, fast spiking), and since channels are subject to functional modulation on a variety of time scales, it is not surprising that a given neuron can exhibit more than one "stereotypical" firing behavior, depending on the conditions (see NEOCORTEX: BASIC NEURON TYPES).

Receptors and Their Agonists and Antagonists: Neurotransmitters, Neuromodulators, Neurohormones, and Second Messengers

Receptors are membrane proteins whose functional action is triggered by the reversible binding of specific molecules called ligands (Cooper, Bloom, and Roth, 1996; see NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK MODELS). A given molecule may be a ligand for more than one kind of receptor, with very different or even opposite functional effects; likewise, a given receptor may be able to be activated by more than one endogenous (or artificial, that is, experimental or pharmaceutical) ligand. A ligand that tends to upregulate the functional activity of a receptor protein is called an agonist for that receptor. Conversely, antagonists are molecules that inhibit the activity of a receptor.

There are two basic types of receptors, ionotropic and metabotropic. Ionotropic receptors are directly associated with an ion channel whose gating is controlled by the presence of the receptor agonist. The action of metabotropic receptors is more complex: upon binding to an agonist, these receptors activate a G protein (so named because their action involves the conversion between guan-

osine diphosphate and guanosine triphosphate), which may directly control channel gating or may initiate a biochemical cascade mediated by second messengers. The end point of this "chain reaction" can be, for example, the opening of a channel, or the phosphorylation of a receptor by the activation of a kinase.

Agonists are properly called neurotransmitters when released by the presynaptic terminal of an axon (or possibly a dendrite) arising from another neuron (see NEOCORTEX: CHEMICAL AND ELECTRICAL SYNAPSES). Extracellular agonists also include neuromodulators and neurohormones, with the latter distributed through the vasculature as well as the perineuronal space (see NEUROMODULATION IN MAMMALIAN NERVOUS SYSTEMS AND NEUROMODULATION IN INVERTEBRATE NERVOUS SYSTEMS). From a functional viewpoint, the main difference between these agonists and neurotransmitters is that neurotransmitters generally mediate synaptic communication between two specific pre- and postsynaptic cells, whereas the release of a neuromodulator or neurohormone into the extracellular space mediates *pancrinic* transmission, affecting a local region of tissue rather than a single postsynaptic site. Another, somewhat arbitrary, difference is that neuromodulators and neurohormones tend not to overtly excite or inhibit their targets, but rather shape the response of a neuron to classical synaptic transmitters in various and subtle ways (Kaczmarek and Levitan, 1987). Note that a given molecule can be assigned more than one of these roles (e.g., neurotransmitter versus neuromodulator), depending on the cell type or region in the nervous system.

Intracellular second messengers are called such because their concentration is often subsequent to the message delivered by neurotransmitters (e.g., after activation of a metabotropic receptor). Second messengers may have direct actions or, as mentioned, may participate in more complicated reaction schemes. Depending on the complexity of the reaction, the functional action of second messengers can be quite delayed and last for minutes if not longer. In addition, the more complicated the biochemical cascade, the more opportunities there are for interactions with modulatory pathways.

The most well-known second messenger is the Ca^{2+} ion, which modulates various membrane channels and biochemical cascades, including many neurotransmitter release systems, and whose intracellular concentration is mediated by a variety of Ca^{2+} -permeable channels, pumps, buffers, and intracellular stores (involving as well the extensive endoplasmic reticulum network, which may support regenerative intracellular Ca^{2+} waves [Berridge, 1998]).

There is a vast array of receptor types, some of which are associated with classical point-to-point synaptic transmission, others that mediate *pancrinic* transmission, and still others that function as links along intracellular pathways. Presynaptic membrane may also express extracellular receptors whose agonist is either the transmitter released by the same terminal (and thus implementing an immediate feedback loop) or another substance, which then may modulate the presynaptic terminal properties. A given neuron may express many different types of receptors in response to the signaling molecules released from other cells, normally in a nonuniform distribution over its surface. In contrast, the number of neuroactive compounds that a single neuron releases itself is usually one, probably (according to current knowledge) at most two or three.

Enzymes

Among the wide variety of enzymes distributed in the neuron's cytoplasm, the most important types for signal processing include kinases and phosphatases, as well as those involved in the metabolism of signaling molecules (e.g., synthases and lipases). The kinases and phosphatases respectively phosphorylate (add a phosphate group) and dephosphorylate specific target proteins, as a result modifying the functional properties of the target. This is the

most common mechanism of regulating the activity of neuronal proteins, for example, by altering the responsiveness of a receptor to an agonist, or the voltage dependency or conductance of an ion channel.

Gap Junctions

Gap junctions are membrane proteins that form a direct electrical path between two neurons, essentially as a nonlinear, nonselective ion channel (see NEOCORTEX: CHEMICAL AND ELECTRICAL SYNAPSES). Thus, on the one hand, these connections are like conventional synapses in that they mediate information flow from cell to cell, but on the other hand, they are quite unlike conventional synapses in that this flow is (more or less) reciprocal and instantaneous. As with essentially all the other neuronal elements, gap junctions can be functionally modulated, typically by Ca^{2+} or other second messengers.

Pumps, Exchangers, and Transporters

Pumps, exchangers, and transporters are membrane proteins responsible for the active maintenance of concentration gradients of different ions and molecules crucial for neural signal processing, and thus are able to modify the membrane potential, either directly or indirectly.

For example, the enzyme Na/K ATPase maintains the characteristic Na^+ and K^+ gradients across all cell membranes; related proteins include the calcium and proton pumps. The action of these pumps depends on the hydrolysis of adenosine triphosphate (ATP) to adenosine diphosphate (ADP), and thus they are tightly coupled to the metabolic machinery of the neuron. Since these cations directly or indirectly contribute to the membrane potential, and since the kinetics of the pump can be modulated, a pump can set the neuron's long-term electrical behavior.

In addition to driving channel currents, the Na^+ and K^+ gradients across the cell membrane also provide the energy for exchangers and transporters. Exchanger proteins move ions such as Ca^{2+} and protons out of the neuron, against their gradients, in exchange for Na^+ moving down its gradient. The exchangers react faster than pumps and thus provide early protection against excessive accumulation of various ions. Transporter proteins move molecules such as glutamate and GABA (respectively the principal excitatory and inhibitory neurotransmitters in the central nervous system) back into the neuron (and into surrounding glia as well) after being released into the extracellular space during synaptic transmission.

The activity of some of these proteins is electrogenic. For example, the Na/K pump cycles two K^+ ions in for three Na^+ ions out, and therefore directly generates a net outward current that can cause a hyperpolarization of many millivolts, depending on conditions. Although not always inherently electrogenic, there is an indirect link between the activity of exchangers and transporters and the membrane potential. Since they are driven by the inward movement of Na^+ , an increase in exchanger or transporter activity leads to an increase in the cytoplasmic concentration of Na^+ , which will then be countered by increased Na/K ATPase activity and its attendant electrogenic effect.

Implications of Neuronal Macromolecule Diversity and Complexity

Channels, receptors, pumps, enzymes, and so on are comprised of one or several individual proteins, called subunits, each of which is coded by a specific gene. For any given type of channel (etc.) there may be many variations of the complete ensemble, or *multimer*, as one subunit substitutes for another, which often imparts different peculiarities to the functional properties of the multimeric

protein (binding sites, effect on kinetics, etc.—in fact, the same sort of properties that may be affected by protein phosphorylation). Thus, a particular Ca^{2+} channel type, for example, may have ten or so identified variants or subtypes (with the strong likelihood that more remain to be discovered). There are as yet but few demonstrations, either by explicit functional studies or by model prediction, that these differences between subtypes are relevant for neural computation. Nevertheless, correlations are increasingly being found between particular disease states and subtle functional alterations of cellular elements, or, in the opposite sense, functional (e.g., behavioral) expressions of genetic manipulation (e.g., knock-out) protocols. Thus, the reality of subtype diversity suggests an important limitation for models that employ a single stereotypical kinetic model of a given type of neural protein.

Subunit substitution in a receptor, channel, or other neural protein can, among other things, determine different endogenous modulatory agonists or antagonists. Since there are many candidates for pancretic pathways at most neurons, this mechanism is important for understanding circuitry dynamics in the intact brain. This functional diversity also has extremely important implications for clinical pharmacology: different subunits can also impart sensitivities to different exogenous compounds, allowing the eventual possibility of targeting very specific synapses or other cellular elements with the appropriately chosen (or designed) drug.

Individual proteins are comprised of contorted chains of thousands of amino acids. This fundamental complexity allows for, in principle, several mechanisms by which a protein may be influenced by its local environment. Thus, there may be an important location dependence of the functional properties of a particular kind of protein, reflecting subtle variations in the protein's microenvironment. For the same reason, it is not surprising that the behavior of a channel, for example, may be modified by the membrane voltage or by binding with a signaling molecule. In this context, we may note that quantitative experimental measurements of a given channel or receptor type in different cell types are inevitably different, beyond what would be expected from experimental variability. Sometimes such differences are seen even between different locations of a single cell type (in particular somatic versus dendritic). Thus, there are at least two possible explanations for such differences: they may be intrinsic to the neural protein under investigation (i.e., a difference in subunit composition), or they may reflect how different local environments, specific to different cell types or location within a single cell, can influence the protein's behavior.

Neuron Models and the Biophysical Mosaic

We now return to the question of neuron models and how they might relate to cellular details. In the most general sense, a single neuron provides a dynamic mapping from a spatiotemporal pattern of pulsed inputs impinging on its dendrites and soma, into a single sequence of output spikes at the axon hillock, which may then be further altered by distinct mechanisms in the axonal tree and presynaptic boutons. Overall, the neuron models employed by theorists describe the time-varying three-dimensional biophysical mosaic underlying this complex signal processing to varying degrees (see PERSPECTIVE ON NEURON MODEL COMPLEXITY and SINGLE-CELL MODELS).

At the simplest level, an extreme abstract model might be a point integrator whose output is passed through a static sigmoid transfer function, where the scalar output is analogous to the firing rate of a spiking neuron. Here the biophysical basis is essentially limited to the resistive nature of the neuron membrane and the spike threshold. As a next step, the basic temporal characteristics of neuronal function may be represented by a leaky integrate-and-fire model that captures the resistive-capacitive nature of the neuron mem-

brane and the action potential-based point process communication between neurons (see INTEGRATE-AND-FIRE NEURONS AND NETWORKS). Among other things, this scheme allows for encoding by both firing rate and higher-order statistics of spike trains, as well as a more tractable analysis of generalized stochastic mechanisms (see ADAPTIVE SPIKE CODING, RATE CODING AND SIGNAL PROCESSING, and SENSORY CODING AND INFORMATION TRANSMISSION).

A more explicit description of biophysical mechanisms might start with the characteristics of membrane channels and dendritic cables (see DENDRITIC PROCESSING). For example, a single neuron model may include transmitter-gated synaptic conductance inputs distributed on a linear (or “passive”) cable tree topology, with conductance-based (i.e., voltage-dependent channels) spike generation at a central somatic node. An anatomically based dendritic cable structure provides an explicit basis for synaptic weights via different coupling impedances to the soma, as well as cable-dependent (e.g., “on-the-path”) nonlinear synaptic interactions. Simple channel models can capture basic spike firing properties such as absolute and relative refractory period, adaptation, or non-zero minimum firing rates.

A model with increased biophysical realism could include voltage-dependent membrane properties distributed throughout the cell (Stuart, Spruston, and Häusser, 1999; see DENDRITIC PROCESSING). Intrinsic and synaptic mechanisms can be modeled with less or more sophisticated kinetic descriptions, either deterministic or stochastic (see BIOPHYSICAL MECHANISMS IN NEURONAL MODELING; TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES, SYNAPTIC INTERACTIONS; and SYNAPTIC NOISE AND CHAOS IN VERTEBRATE NEURONS). Further details of functional properties may require descriptions of the microphysiology of extra- and intracellular systems, and thus explicit modeling of biochemical dynamics, including Ca^{2+} diffusion, buffering, sequestration, and release; protein conformations; and enzyme activation/inactivation. Finally, the most faithful cellular model would require a four-dimensional construct whose biophysical properties vary with both space and time, in particular depending on past activity, or “experience” (see HEBBIAN SYNAPTIC PLASTICITY).

State Variables and Functional Compartments of the Mosaic Neuron

The many cellular elements we have described suggest a similar number of variables that characterize the functional state of a neuron as a signal processing device, each of which may be thought of as representing information. The most classical variable, of course, is the membrane voltage, which defines the immediate integration of synaptic input onto the dendritic tree and soma and, eventually, the action potential output of the cell. However, it may be argued that for predicting spike output, the first derivative of the membrane voltage may be nearly as important as the actual value of the voltage, a behavior that is easily predicted by Hodgkin-Huxley-type models (see AXONAL MODELING; BIOPHYSICAL MECHANISMS IN NEURONAL MODELING; and ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION). Other variables that may be important include the concentration of ions and various neuroactive molecules (e.g., transmitters and second messengers) both inside and outside the cell, and the metabolic or conformational state of various membrane and intracellular proteins. Finally, it may be useful to consider structural or anatomical parameters of the single neuron as functional state variables, such as number and distribution of spines or postsynaptic sites.

All of these state variables are determined by complex relationships between the cellular constituents. For example, the membrane voltage at any given point in the neuron is determined by the spatial distribution of electrically conducting membrane channels and their

reversal potentials, the membrane capacitance, and the electrical coupling to the rest of the cell as determined by the three-dimensional branching cable structure and cytoplasmic resistivity. In turn, the ion concentration gradients that underlie channel reversal potentials are determined by an interplay between the currents through the appropriate channels (which tend to reduce the gradients) and membrane pumps, exchangers, transporters, and intracellular buffer and sequestering systems (which in general tend to maintain the gradients). Finally, in some cases ions passing through a given membrane channel can subsequently bind with and then modulate the conduction state of either the same or possibly other types of channels. Clearly, feedback pathways are the rule rather than the exception in the mosaic’s interactions (see Figure 1).

The state variables of a neuron can be associated with a variety of functional compartments, with spatial scales that range from less than a micron to the entire cell. These compartments may correspond to the spatial gradients of voltage (e.g., as determined by dendritic cable properties) or second messenger concentration (e.g., determined by diffusion constant and geometry of the intracellular space), or to the localization of a given biochemical pathway, or to an explicit subcellular structure (e.g., a dendritic spine, an organelle, the cell nucleus). In summary, for most state variables the neuron is far from a classical “well-mixed” system. Rather, an extreme internal heterogeneity is usually the case: a single cell thus becomes a cell of cells.

Emergent State Properties of Intracellular Chemical Systems

Recent work has demonstrated the possibility of various stable arrangements between the concentrations of certain intracellular molecules or the metabolic states of certain proteins, all of which in turn can participate in various biochemical pathways, including those regulating membrane properties and plasticity (Weng, Bhalla, and Iyengar, 1999). From an information processing viewpoint, these combinations can be thought of as essentially distinct states that partially define the functional input-output properties of the neuron. It has also been proposed that changes in cellular properties on even longer time scales may be due to self-stabilizing conformational states of proteins such as CAM kinase II or others. Of course, any mechanism underlying a long-lasting modification of the neuronal transfer function is a candidate for the molecular basis of memory (see INVERTEBRATE MODELS OF LEARNING: *APLYSIA* AND *HERMISSENDA*).

Discussion: How Much Biophysics Needs to Be Known for a Compelling Brain Theory?

Where does the complexity of the mosaic neuron leave us in terms of formulating a theory for the brain? In particular, how detailed does a model of the neuron have to be, and how does the power of current methods compare with the computational complexity inherent at various levels of biophysical description? These open questions have a very practical importance, since there are few opportunities for formal analyses of these nonlinear dynamical systems. Furthermore, the increasing experimental knowledge of neuronal cellular mechanisms is daunting. The details seem to have an almost fractal quality; no matter what level is being examined, underneath any given mechanism there is another Pandora’s box of parameters waiting to be described. Today, so many biophysical properties are known to be present in the neuron membrane that the biggest risk is to choose to include only those that will give the model the properties desired. This leads to a model that is unlikely to fail, and therefore unlikely to teach us anything that we did not know before.

The brute force strategy is to construct a bottom-up, biophysically detailed cell model in order to cover as much as possible the

high-order interactions between various mechanisms. Once the map has been laid out, sufficiently rich deterministic or stochastic kinetic equations for the membrane elements and intracellular dynamics may then be assigned. Such maps can be evaluated, at least in principle, since all known biophysical mechanisms can be described by nonlinear partial differential equations, and thus are amenable to standard numerical integration techniques.

However, it may appear that at some point there will be too many equations with too many free parameters for practical evaluation or, more important, for true *understanding*. (On the other hand, the rapid evolution of computing power suggests that the threshold for “impractical” is hard to define.) An optimistic view is that once a sufficiently elaborate biophysical cellular model is constructed, its behavior may be well enough understood so that a more abstract model that captures the functional essentials with considerably less computational expense may be derived. But until there are some formal criteria for establishing what exactly “essential” means, this process will inevitably be an iterative one, moving back and forth between an analysis of detailed cell models in isolation and in nontrivial networks (limited most probably by the available tools) and an analysis of more abstract neural networks. The fundamental challenge to theorists, then, is to go beyond this sort of “reverse engineering” approach and develop a method that is at once commensurate with the complexity of the brain, yet can produce a *bona fide* theory whose detail avoids that of the actual biological reality.

Road Map: Biological Neurons and Synapses

Related Reading: Activity-Dependent Regulation of Neuronal Conductances; Adaptive Spike Coding; Axonal Modeling; Dendritic Processing; Hebbian Synaptic Plasticity; Ion Channels: Keys to Neuronal Specialization; Neocortex: Chemical and Electrical Synapses; NMDA Receptors: Synaptic, Cellular, and Network Models; Rate Coding and Signal

Processing; Synaptic Interactions; Synaptic Noise and Chaos in Vertebrate Neurons; Temporal Dynamics of Biological Synapses

References

- Berridge, M. J., 1998, Neuronal calcium signaling, *Neuron*, 21:13–26.
- Borg-Graham, L., 1999, Interpretations of data and mechanisms for hippocampal pyramidal cell models, *Cereb. Cortex*, 13:19–138. ♦
- Cooper, J. R., Bloom, F. E., and Roth, R. H., 1996, *The Biochemical Basis of Neuropharmacology*, 7th ed., Oxford, Engl.: Oxford University Press. ♦
- Hain, G. L., 1999, *Molecular and Cellular Physiology of Neurons*, Cambridge, MA: Harvard University Press. ♦
- Hille, B., 2002, *Ionic Channels of Excitable Membranes*, 3rd ed., Sunderland, MA: Sinauer. ♦
- Johnston, D., and Wu, S. M.-S., 1995, *Foundations of Cellular Neurophysiology*, Cambridge, MA: MIT Press. ♦
- Kaczmarek, L. K., and Levitan, I. B., Eds., 1987, *Neuromodulation: The Biochemical Control of Neuronal Excitability*, Oxford, Engl.: Oxford University Press.
- Koch, C., 1999, *The Biophysics of Computation: Information Processing in Single Neurons*, Oxford, Engl.: Oxford University Press. ♦
- Koch, C., and Segev, I., Ed., 1998, *Methods in Neuronal Modeling*, 2nd ed., Cambridge, MA: MIT Press. ♦
- Laming, P. R., Kimelberg, H., Robinson, S., Salm, A., Hawrylak, N., Müller, C., Roots, B., and Ng, K., 2000, Neuronal-glial interactions and behaviour, *Neurosci. Biobehav. Rev.*, 24:295–340.
- Meunier, C., and Segev, I., 2000, Neurons as physical objects: Structure, dynamics and function, in *Handbook of Biological Physics* (F. Moss and Gielen S., Eds.), New York: Elsevier.
- Stuart, G., Spruston, N., and Häusser, M., Eds., 1999, *Dendrites*, Oxford, Engl.: Oxford University Press. ♦
- Weiss, T. F., 1996, *Cellular Biophysics* (2 vol.), Cambridge, MA: MIT Press. ♦
- Weng, G., Bhalla, U. S., and Iyengar, R., 1999, Complexity in biological signaling systems, *Science*, 284:92–96.

Brain Signal Analysis

Jeng-Ren Duann, Tzyy-Ping Jung, and Scott Makeig

Introduction

Artificial neural networks (ANNs) have now been applied to a wide variety of real-world problems in many fields of application. The attractive and flexible characteristics of ANNs, such as their parallel operation, learning by example, associative memory, multifactorial optimization, and extensibility, make them well suited to the analysis of biological and medical signals. In this study, we review applications of ANNs to brain signal analysis, for instance, for analysis of the electroencephalogram (EEG) and magnetoencephalogram (MEG) or electromyogram (EMG), and as applied to computed tomographic (CT) images and magnetic resonance (MR) brain images, and to series of functional MR brain images (i.e., fMRI).

Most ANNs are implemented as sets of nonlinear summing elements interconnected by weighted links, forming a highly simplified model of brain connectivity. The basic operation of such artificial neurons is to pass a weighted sum of their inputs through a nonlinear hard-limiting or soft “squashing” function. To form an ANN, these basic calculating elements (artificial neurons) are most often arranged in interconnected layers. Some neurons, usually those in the layer farthest from the input, are designated as output neurons. The initial weight values of the interconnections are usually assigned randomly.

The operation of most ANNs proceeds in two stages. Rules used in the first stage, training (or learning), can be categorized as supervised, unsupervised, or reinforced. During training, the weight values for each interconnection in the network are adjusted either to minimize the error between desired and computed outputs (supervised learning) or to maximize differences (or minimize similarities) between the output categories (unsupervised or competitive learning). In reinforced learning, an input-output mapping is learned during continued interaction with the environment so as to maximize a scalar index of performance (Haykin, 1999). The second stage is recall, in which the ANN generates output for the problem the ANN is designed to solve, based on new input data without (or sometimes with) further training signals.

Because of their multifactorial character, ANNs have proved suitable for practical use in many medical applications. Because most medical signals of interest usually are not produced by variations in a single variable or factor, many medical problems, particularly those involving decision making, must involve a multifactorial decision process. In these cases, changing one variable at a time to find the best solution may never reach the desired objective (Dayhoff and DeLeo, 2001), whereas multifactorial ANN approaches may be more successful. In this article, we review recent applications of ANNs to brain signal processing, organized ac-

cording to the nature of brain signals to be analyzed and the role that ANNs play in the applications.

Roles of ANNs in Brain Signal Processing

To date, ANNs have been applied to brain data for the following purposes:

- *Feature extraction, classification, and pattern recognition.* ANNs in this application serve mainly as nonlinear classifiers. The inputs are preprocessed so as to form a feature space. ANNs are used to categorize the collected data into distinct classes. In other cases, inputs are not subjected to preprocessing but are given directly to an ANN to extract features of interest from the data.
- *Adaptive filtering and control.* In this application, ANNs operate within closed-loop systems to process changing inputs, adapting their weights “on the fly” to filter out unwanted parts of the input (adaptive filtering), or mapping their outputs to parameters used in on-line control (adaptive control).
- *Linear or nonlinear mapping.* In this application, ANNs are used to transform inputs to outputs of a desired form. For example, an ANN might remap its rectangular input data coordinates to circular or more general coordinate systems.
- *Modeling.* ANNs can be thought of as function generators that generate an output data series based on a learned function or data model. ANNs with two layers of trainable weights have proved capable of approximating any nonlinear function.
- *Signal separation and deconvolution.* In this application, ANNs separate their input signals into the weighted sum or convolution of a number of underlying sources using assumptions about the nature of the sources or of their interrelationships (e.g., their independence).
- *Texture analysis and image segmentation.* Image texture analysis is becoming increasingly important in image segmentation, recognition, and understanding. ANNs are being used to learn spatial or spatial-frequency texture features and, accordingly, to categorize images or to separate an image into subimages (image segmentation).
- *Edge detection.* In an image, an edge or boundary between two objects can be mapped to a dark band between two lighter areas (objects). By using the properties of intensity discontinuity, ANNs can be trained to recognize these dark bands as edges, or can learn to draw such edges based on contrast and other information.

Application Areas

In this section, we describe some applications of ANNs to brain signals by means of examples involving neurobiological time series and brain images. Neurobiological signals of clinical interest recorded noninvasively from humans include electroencephalographic (EEG), magnetoencephalographic (MEG), and electromyographic (EMG) data. Research in brain imaging includes the analysis of structural brain images, mainly focused on the extraction of three-dimensional (3D) structural information, from various kinds of brain images (e.g., magnetic resonance images, or MRIs), as well as the analysis of functional brain imaging series that mainly reveal changes in the brain state during cognitive tasks using medical imaging techniques (e.g., functional MRI, or fMRI, and positron emission tomography, or PET). These examples by no means cover all the applications in the field.

Neurobiological Signals

EEG and MEG. EEG provides a noninvasive measure of brain electrical activity recorded as changes in the potential difference

between two points on the scalp. MEG is the magnetic counterpart of EEG. In accordance with the assumption that the ongoing EEG can be altered by stimulus or event to form respectively the event-related potential (ERP) or the evoked potential (EP), these changes, though tiny, can be recorded through the scalp. It is possible for researchers to apply pattern recognition algorithms to search for the differences in brain status while the brain is performing different tasks. Thus, Peters, Pfurtscheller, and Flyvbjerg (2001) applied an autoregressive model to four-channel EEG potentials to obtain features that were used to train an ANN using a backpropagation algorithm to differentiate the subject's intention to move the left or right index finger or right foot. They suggested that the framework might be useful for designing a direct brain-computer interface. In the study of Zhang et al. (2001), ANNs were trained to determine the stage of anesthesia based on features extracted from the middle-latency auditory-evoked potential (MLAEP) plus other physiological parameters. By combining power spectral estimation, principal component analysis (PCA), and ANNs, Jung et al. (1997) demonstrated that continuous, accurate, noninvasive, and near real-time estimation of an operator's global level of alertness is feasible using EEG measures recorded from as few as two scalp sites. The results of their ANN-based estimation compared favorably with results obtained using a linear regression model applied to the same PCA-reduced EEG power spectral data.

Sun and Scialabassi (2000) employed an ANN as a linear mapping device to transform the EEG topography obtained from a forward solution in a simple spherical model to a more realistic spheroidal model whose forward solution was difficult to compute directly. In that study, a backpropagation learning algorithm was used to train an ANN to convert spatial locations between spherical and spheroidal models. Instead of computing the infinite sums of the Legendre functions required in the asymmetric spheroidal model, the calculations were carried out in the spherical model and then converted by the ANN to the more realistic model for display and evaluation.

Recently, ANNs have made a substantial contribution to the analysis of EEG/MEG by separating the problem of EEG/MEG source identification from that of source localization, a mathematically underdetermined problem: any scalp potential distribution can be produced by a limitless number of potential distributions within the head. Because of volume conduction through cerebrospinal fluid, skull, and scalp, EEG and MEG data collected from any point on the scalp may include activity arising in multiple locally synchronous but relatively independent neural processes within a large brain volume. This has made it difficult to relate EEG measurements to underlying brain processes and to localize the sources of EEG and MEG signals. Progress has been made by several groups in separating and identifying the distinct brain sources from their mixtures in scalp EEG or MEG recordings, assuming only their temporal independence and spatial stationarity (Makeig et al., 1997; Jung et al., 2001), using a class of independent component analysis (ICA) or blind source separation algorithms.

Muscle and movement signals. From recordings of muscle stretching (mainly the EMG), it is possible to predict the intent of subjects to perform actions such as hand or finger movements, or to judge the disability of a specific bundle of muscle cells. For example, Khalil and Duchene (2000) used wavelet coefficients obtained from uterine EMG to train ANNs to separate the inputs into four labeled categories: uterine contractions, fetal movements, Alvarez waves, and long-duration low-frequency band waves. They reported that the system was useful for maintaining preterm births. On the other hand, Stites and Abbas (2000) used an ANN as a pattern shaper to refine the output patterns of a functional neuromuscular stimulation system that served as a pattern generator of control signals for cyclic movements to help the paraplegic patient stand using functional neuromuscular stimulation.

Brain Images

Structural images. In structural brain image analysis, ANNs may play roles in image segmentation, image labeling, or edge detection. Image segmentation is the first and probably the most important step in digital image processing. Segmentation may be a labeling problem in which the goal is to assign, to each voxel in a gray-level image, a unique label that represents its belonging to an anatomical structure. The results of image segmentation can be used for image understanding and recognition, 3D reconstruction, visualization, and measurements, including brain volume changes in developmental brain diseases such as Alzheimer's disease and autism. The rapid pace of development of medical imaging devices such as MRI and computed tomography permits a better understanding of anatomical brain structure without, prior to, or even during neurosurgery. However, the results are highly dependent on the quality of the image segmentation processes.

Here we give some examples using ANNs in image segmentation: Dawant et al. (1991) presented a backpropagation neural network approach to the automatic characterization of brain tissues from multimodal MR images. The ability of a three-layer backpropagation neural network to perform segmentation based on a set of MR images (T1-weighted, T2-weighted, and proton density-weighted) acquired from a patient was studied. The results were compared with those obtained using a maximum likelihood classifier. Dawant et al. found no significant difference in the results obtained by the two methods, although the backpropagation neural network gave cleaner segmentation images. Using the same analysis strategy, Reddick et al. (1997) first trained a self-organizing map (SOM) on multimodal MR brain images to efficiently extract and convert the 3D inputs (from T1-, T2- and proton density-weighted images) into a feature space and utilized a backpropagation neural network to separate them into classes of white matter, gray matter, and cerebrospinal fluid. Their work demonstrated high intraclass correlation between the automated segmentation and classification of tissues and standard radiologist identification, as well as high intrasubject reproducibility.

Functional images. Today, not only are medical imaging devices able to provide impressive spatial resolution and details of the fine structure of the human brain, they can also reveal changes in brain status in awake subjects who are performing a task or even daydreaming, by measuring ongoing metabolic changes, including cerebral blood flow, cerebral blood volume (by PET), and blood oxygenation level-dependent signal levels (by fMRI). We will give some examples, mainly from fMRI analysis.

Functional brain imaging emerged in the early 1990s, based on the observation that increases in local neuronal activity are followed by local changes in oxygen concentration. Changing the amount of oxygen carried by hemoglobin changes the degree to which hemoglobin disturbs a magnetic field, as a result of which in vivo changes in blood oxygenation could be detected by MRI (Ogawa et al., 1992). The subsequent changes in the MRI signal became known as the blood oxygenation level-dependent or BOLD signal. This technique was soon applied to normal humans during functional brain activation (the subjects performed cognitive tasks), giving birth to the rapid growing field of fMRI.

Theoretically, the fMRI BOLD signal from a given brain voxel can be interpreted as a linear combination of different sources with distinguishable time courses and spatial distributions, including use-dependent hemodynamic changes, blood, or CSF flows, plus subject movement and machine artifacts. Recently, ANNs (especially those using ICA), applied to fMRI data, have proved to be a powerful method for detecting and separating task-related activations with either known or unanticipated time courses (McKeown et al., 1998) that could not be detected using standard

hypothesis-driven analyses. Duann et al. (2002) have given further details of applying ICA to the fMRI BOLD signal. They showed that the hemodynamic response to even widely spaced stimulus presentations may be dependent on the trial, site, stimulus, and subject. Thus, the standard regression-based method of applying a fixed hemodynamic response model to find stimulus- or task-related BOLD activations needs to be reconsidered.

Discussion

The use of ANNs as classifiers currently dominates their applications in the field of brain signal analysis. This includes classification of brain or related signals as exhibiting normal or abnormal features or processes. Not surprisingly, published studies report promising results.

If the measurements can be modeled as an additive mixture of different sources, including task-related signals and artifacts, applying blind source separation prior to the further processing, visualization, or interpretation may better reveal the underlying physical phenomena (such as different brain processes), which in the raw data could be contaminated or overwhelmed by other processes of no interest.

A survey of relevant papers shows that the most popular architecture for ANNs is the multilayer perceptron (MLP). The MLP architecture is both simple and straightforward to implement and use. In MLPs, information flows in one direction except during training, when error terms are backpropagated. Backpropagation updates network weights in a supervised manner. Although it cannot guarantee a globally minimal solution, backpropagation at least arrives at a local minimum through gradient descent. Various techniques have been derived in an attempt to avoid overfitting to a local minimum. Once the network weights have been learned and fixed, feedforward networks can be implemented in hardware and made to run in real time. All of these characteristics make the backpropagation algorithm most popular in biomedical applications.

In some applications, target outputs may not be available or may be too expensive to acquire. In these cases, unsupervised learning algorithms may be used. Among unsupervised learning algorithms, self-organizing maps (SOMs) are the most popular for biomedical applications. During training, SOMs attempt to assign their input patterns to different output regions. Often SOMs may converge after only few learning cycles.

Application Issues

Although most published papers have concluded that ANNs are appropriate for their domain of interest, many issues still have to be resolved before ANNs can be claimed to be the general method of choice. Unfortunately, most published studies have not gone beyond demonstrating application to a very limited amount of data. As with any type of method, ANNs have their limitations that should be carefully considered:

- Every study should provide a rationale for the data chosen as input. For example, ANN-based computer-aided diagnostic systems may give misleading results if the ANNs are not given adequately representative features and sufficient naturally occurring data variations in their training data. With ANNs, any input may yield some sort of output, correct and useful or not ("garbage in, garbage out"). Therefore, the keys to success of ANN applications are not only to pick an appropriate architecture or learning algorithm, but also to choose the right data and data features to train the network.
- Although methods of applying ANNs to biomedical signals have already shown great promise, great care must be taken to examine

the results obtained. The issue of trust in the outputs of ANNs always deserves informed as well as statistical consideration. Since medical diagnosis is nearly always a multifactorial and multidisciplinary problem, medical experts should always evaluate network outputs in light of other direct or indirect convergent evidence before making final decisions affecting the health of patients.

- Before practical implementation is planned, ANN methods should be compared to more direct ways of obtaining the same answers, as these might sometimes prove more accurate or cost effective.

Model Mining

Since the first wave of popularization of backpropagation networks nearly two decades ago, an ever greater number and variety of ANN models have been devised to tackle an ever greater variety of problems. The overall insight that ANNs both embody and exemplify is perhaps that our human intelligence is multifactorial and highly adaptable to using whatever forms of information are available to us. In this spirit, we suggest that researchers always attempt to interpret the physiological meaning both of the features of their input data and of the data models that their trained ANNs represent. Too often ANNs have been treated like black boxes. We believe it is time to open the black boxes and interpret what is happening inside them. Such interpretations might even yield new insights into the nature of the biomedical signals, or suggest new or more efficient ways to look at the input data. It is also possible that the ANN models and methods might suggest more efficient methods to collect input data. Such “model mining” might even prove to be the most rewarding result of applying ANNs. Researchers who simply recount classification accuracy may ignore nuggets of novel information about brain processes hidden in the ANN models that they and the data have jointly constructed.

Road Map: Implementation and Analysis

Related Reading: EEG and MEG Analysis; Muscle Models; Neuroinformatics; Statistical Parametric Mapping of Cortical Activity Patterns

References

- Dawant, B. M., Ozkan, M., Zijdenbos, A., and Margolin, R., 1991, A computer environment for 2D and 3D quantitation of MR images using neural networks, *Magn. Reson. Imaging*, 20:64–65.
- Dayhoff, J. E., and DeLeo, J. M., 2001, Artificial neural networks: Opening the black box, *Cancer*, 91:1615–1635. ♦
- Duann, J. R., Jung, T. P., Kuo, W. J., Yeh, T. C., Makeig, S., Hsieh, J. C., and Sejnowski, T. J., 2002, Single-trial variability in event-related BOLD signals, *NeuroImage*, 15:823–835. ♦
- Haykin, S., 1999, *Neural Network: A Comprehensive Foundation*, Englewood Cliffs, NJ: Prentice Hall. ♦
- Jung, T.-P., Makeig, S., Stensmo, M., and Sejnowski, T. J., 1997, Estimating alertness from the EEG power spectrum, *IEEE Trans. Biomed. Eng.*, 44:60–69.
- Jung, T.-P., Makeig, S., McKeown, M. J., Bell, A. J., Lee, T.-W. and Sejnowski, T. J., 2001, Imaging brain dynamics using independent component analysis, *Proc. IEEE*, 89:1107–1122. ♦
- Khalil, M., and Duchene, J., 2000, Uterine EMG analysis, a dynamic approach for change detection and classification, *IEEE Trans. Biomed. Eng.*, 47:748–756.
- Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D., and Sejnowski, T. J., 1997, Blind separation of auditory event-related brain responses into independent components, *Proc. Natl. Acad. Sci. USA*, 94:10979–10984. ♦
- McKeown, M. J., Jung, T.-P., Makeig, S., Brown, G. G., Kindermann, S. S., Lee, T.-W., and Sejnowski, T. J., 1998, Spatially independent activity patterns in functional MRI data during the Stroop color-naming task, *Proc. Natl. Acad. Sci. USA*, 95:803–810. ♦
- Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., and Ugurbil, K., 1992, Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging, *Proc. Natl. Acad. Sci. USA*, 89:5951–5959. ♦
- Peters, B. O., Pfurtscheller, G., and Flyvbjerg, H., 2001, Automatic differentiation of multichannel EEG signals, *IEEE Trans. Biomed. Eng.*, 48:111–116.
- Reddick, W. E., Glass, J. O., Cook, E. N., Elkin, T. D., and Deaton R. J., 1997, Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks, *IEEE Trans. Med. Imaging*, 16:911–918.
- Sites, E. C., and Abbas, J. J., 2000, Sensitivity and versatility of an adaptive system for controlling cyclic movements using functional neuromuscular stimulation, *IEEE Trans. Biomed. Eng.*, 47:1287–1292. ♦
- Sun, M., and Scialabassi, R. J., 2000, The forward EEG solution can be computed using artificial neural networks, *IEEE Trans. Biomed. Eng.*, 47:1044–1050. ♦
- Zhang, X.-S., Roy, R. J., Schwender, D., and Dauberer, M., 2001, Discrimination of anesthetic states using mid-latency auditory evoked potential and artificial neural networks, *Ann. Biomed. Eng.*, 29:446–453.

Brain-Computer Interfaces

José del R. Millán

Introduction

There is a growing interest in the use of physiological signals for communication and operation of devices for the severely motor disabled as well as for able-bodied people. Over the last decade evidence has accumulated to show the potential of analyzing brainwaves on-line to derive information about the subject's mental state that is then mapped into some external action such as selecting a letter from a virtual keyboard or moving a robotics device. A *brain-computer interface (BCI)* is an alternative communication and control channel that does not depend on the brain's normal output pathway of peripheral nerves and muscles (Wolpaw et al., 2000).

Most BCI systems use electroencephalogram signals (EEG and MEG ANALYSIS) measured from scalp electrodes that do not require invasive procedures. Although scalp EEG is a simple way to record brainwaves, it does not provide detailed information on the

activity of single neurons (or small clusters of neurons) that could be recorded by implanted electrodes in the cortex (PROSTHETICS, NEURAL). Such a direct measurement of brain activity may, in principle, enable faster recognition of mental states and even more complex interactions.

A BCI may monitor a variety of brainwave phenomena. Some groups exploit evoked potentials generated in response to external stimuli (see Wolpaw et al., 2000, for a review). Evoked potentials are, in principle, easy to pick up but are constrained by the fact that the subject must be synchronized to the external machinery. A more natural and practical alternative is to rely on components associated with spontaneous mental activity. Such spontaneous components range from slow cortical potentials of the EEG (e.g., Birbaumer et al., 1999), to variations of EEG rhythms (e.g., Wolpaw and McFarland, 1994; Kalcher et al., 1996; Anderson, 1997;

Roberts and Penny, 2000; Millán et al., 2002b), to the direct activity of neurons in the cortex (e.g., Kennedy et al., 2000; Wessberg et al., 2000).

Direct Brain-Computer Interfaces

Direct BCIs involve invasive procedures to implant electrodes in the brain (PROSTHETICS, NEURAL). Apart from ethical concerns, a major difficulty is to obtain reliable long-term recordings of neural activity. Recent advances have made it possible to develop direct BCIs with animals and even human beings.

Kennedy and colleagues (2000) have implanted a special electrode into the motor cortex of several paralyzed patients. These electrodes contain a neurotrophic factor that induces growth of neural tissue within the hollow electrode tip. With training, patients learn to control the firing rates of the multiple recorded neurons to some extent. One of them is able to drive a cursor and write messages.

Wessberg et al. (2000) have recorded the activity of ensembles of neurons with microwire arrays implanted in multiple cortical regions involved in motor control, as monkeys performed arm movements. From these signals they have obtained accurate real-time predictions of arm trajectories and have been able to reproduce the trajectories with a robot arm. Although these experiments do not describe an actual BCI, they support the feasibility of controlling complex prosthetic limbs directly by brain activity. In addition, earlier work by Nicolelis and colleagues showed that neural predictors can be derived for rats implanted with the same kind of microelectrodes (see Nicolelis, 2001, for details and reference). The rats were trained to press a bar to move a simple device delivering water, and later learned to operate this device through neural activity only.

For a more detailed analysis and prospects of this area, see Nicolelis (2001).

Noninvasive Brain-Computer Interfaces

Noninvasive BCIs are based on the analysis of EEG phenomena associated with various aspects of brain function. Thus, Birbaumer et al. (1999) measure slow cortical potentials (SCP) over the vertex (top of the scalp). SCP are shifts in the depolarization level of the upper cortical dendrites and indicate the overall preparatory excitation level of a cortical network. Other groups look at local variations of EEG rhythms. The most commonly used rhythms are related to the imagination of movements and are recorded from the central region of the scalp overlying the sensorimotor cortex. In this respect, there exist two main paradigms. Pfurtscheller's team works with event-related desynchronization (ERD, EEG and MEG ANALYSIS) computed at fixed time intervals after the subject is commanded to imagine specific movements of the limbs (Kalcher et al., 1996; Obermaier, Müller, and Pfurtscheller, 2001). Alternatively, Wolpaw and co-workers analyze continuous changes in the amplitudes of the μ (8–12 Hz) or β (13–28 Hz) rhythms (Wolpaw and McFarland, 1994). Finally, in addition to motor-related rhythms, Anderson (1997) and Millán et al. (2002b) also analyze continuous variations of EEG rhythms, but not only over the sensorimotor cortex and in specific frequency bands. The reason is that a number of neurocognitive studies have found that different mental tasks—such as imagination of movements, arithmetic operations, or language—activate local cortical areas to different extents. The insights gathered from these studies guide the placement of electrodes to obtain more relevant signals for the different tasks to be recognized. In this latter case, rather than looking for predefined EEG phenomena as in the previous paradigms, the approach aims at discovering EEG patterns embedded in the continuous EEG signal associated with different mental states.

Most of the existing BCIs are based on synchronous experimental protocols in which the subject must follow a fixed repetitive scheme to switch from one mental task to the next (Wolpaw and McFarland, 1994; Kalcher et al., 1996; Wolpaw and McFarland, 1994; Birbaumer et al., 1999; Obermaier et al., 2001). A trial consists of two parts. A first cue warns the subject to get ready and, after a fixed period of several seconds, a second cue tells the subject to undertake the desired mental task for a predefined time. The EEG phenomena to be recognized are time locked to the last cue and the BCI responds with the average decision over the second period of time. In these synchronous BCI systems, the shortest trial lengths that have been reported are 4 s (Birbaumer et al., 1999) and 5 s (Obermaier et al., 2001). This relatively long time is necessary because the EEG phenomena of interest, either SCP or ERD, need some seconds to recover. On the contrary, other BCIs rely on more flexible asynchronous protocols where the subject makes self-paced decisions on when to stop doing a mental task and start immediately the next one (Roberts and Penny, 2000; Millán et al., 2002b). In this second case, the time of response of the BCI goes from 0.5 s (Millán et al., 2002b) to several seconds (Roberts and Penny, 2000).

EEG signals are characterized by a poor signal-to-noise ratio and spatial resolution. Their quality is greatly improved by means of a Surface Laplacian (SL) derivation, which requires a large number of electrodes (normally 64–128). The SL estimate yields new potentials that represent better the cortical activity originated in radial sources immediately below the electrodes (for details see McFarland et al., 1997; Babiloni et al., 2001; and references therein). The superiority of SL-transformed signals over raw potentials for the operation of a BCI has been demonstrated in different studies (e.g., McFarland et al., 1997). Although significant progress has been obtained (and will still continue) with studies using a high number of EEG electrodes (from 26 to 128), today's practical BCI systems should have a few electrodes (no more than 10) to allow their operation by laypersons, as the procedure of electrode positioning is time consuming and critical. Most groups have developed BCI prototypes with a limited number of electrodes that, however, do not benefit from SL transformations. On the contrary, Babiloni et al. (2001) and Millán et al. (2002b) compute SL derivations from a few electrodes, using global and local methods, respectively.

Wolpaw and McFarland (1994) as well as Birbaumer et al. (1999) have demonstrated that some subjects can learn to control their brain activity through appropriate, but lengthy, training in order to generate fixed EEG patterns that the BCI transforms into external actions. In both cases the subject is trained over several months to modify the amplitude of either the SCP or μ rhythm, respectively. A few other groups follow machine learning approaches to train the classifier embedded in the BCI. These techniques range from linear classifiers (Babiloni et al., 2001; Obermaier et al., 2001), to compact multi-layer perceptrons and Bayesian neural networks (Anderson, 1997; Roberts and Penny, 2000), to variations of LVQ (Kalcher et al., 1996), to local neural classifiers (Millán, 2002; Millán et al., 2002b). Most of these works deal with the recognition of just two mental tasks (Roberts and Penny, 2000; Babiloni et al., 2001; Obermaier et al., 2001), or report classification errors bigger than 15% for three or more tasks (Kalcher et al., 1996; Anderson, 1997). An exception is Millán's approach that achieves error rates below 5% for three mental tasks, but correct recognition is 70% (Millán, 2002; Millán et al., 2002b). Obermaier et al. (2001) reports on a single disabled person who, after several months of training, has reached a performance level close to 100%. It is also worth noting that some of the subjects who follow Wolpaw's approach are able to control their μ rhythm amplitude at four different levels. These classification rates, together with the number of recognizable tasks and duration of the trials, yield bit rates from approximately 0.15 to 2.0.

Some approaches are based on a mutual learning process in which the user and the brain interface are coupled together and adapt to each other (Roberts and Penny, 2000; Obermaier et al., 2001; Millán, 2002; Millán et al., 2002b). This should accelerate the training time. Thus, Millán's approach has allowed subjects to achieve good performances in just a few hours of training (Millán, 2002; Millán et al., 2002b). Analysis of learned EEG patterns confirms that for a subject to operate a personal BCI satisfactorily, the BCI must fit the individual features of the user (Millán et al., 2002a).

Another important concern in BCI is the incorporation of rejection criteria to avoid making risky decisions for uncertain samples. This is extremely important from a practical point of view. Roberts and Penny (2000) apply Bayesian techniques for this purpose, while Millán et al. (2002b) use a confidence probability threshold. In this latter case, more than ten subjects have experimented with their BCI (Millán, 2002; Millán et al., 2002b). Most of them were trained for a few consecutive days (from three to five). Training time was moderate, around half an hour daily. Experimental results show that, at the end of training, the correct recognition rates were 70% (or higher) for three mental tasks. This figure is more than twice random classification. This modest rate is largely compensated by two properties: wrong responses were below 5% (in many cases even below 2%) and decisions were made every half-second. Some other subjects have undertaken consecutive training sessions (from four to seven) in a single day. None of these subjects had previous experience with BCIs and, in less than two hours, all of them reached the same excellent performance as noted previously. It is worth noting that one of the subjects was a physically impaired person suffering from spinal muscular atrophy.

Brain-Actuated Applications

These different BCI systems are being used to operate a number of brain-actuated applications that augment people's communication capabilities, provide new forms of education and entertainment, and also enable the operation of physical devices. There exist virtual keyboards for selecting letters from a computer screen and writing a message (Birbaumer et al., 1999; Obermaier et al., 2001; Millán, 2002). Using these three different approaches, subjects can write a letter every 2 minutes, every 1 minute, and every 22 seconds, respectively. Wolpaw's group has also its own virtual keyboard (Wolpaw, personal communication). A patient who has been implanted with Kennedy and colleagues' special electrode has achieved a spelling rate of about three letters per minute using a combination of neural and EMG signals (Kennedy et al., 2000).

On the other hand, it is also possible to make a brain-controlled hand orthosis open and close (see references in Wolpaw et al., 2000; Obermaier et al., 2001) and even guide in a continuous manner a motorized wheelchair with on-board sensory capabilities (Millán, 2002). In this latter case, the key idea is that users' mental states are associated with high-level commands that the wheelchair executes autonomously (ROBOT NAVIGATION). Another critical aspect for the control of the wheelchair is that subjects can issue high-level commands at any moment, as the operation of the BCI is self-paced and does not require waiting for specific events.

Finally, Millán (2002) illustrates the operation of a simple computer game, but other educational software could have been selected instead.

Discussion

Despite recent advancements, BCI is a field still in its infancy and several issues must be addressed to improve the speed and performance of BCI. One of them is the exploration of local components of brain activity with fast dynamics that subjects can consciously

control. For this we will need increased knowledge of the brain (where and how cognitive and motor decisions are made) as well as the application of more powerful digital signal processing (DSP) methods than those commonly used currently. In addition, extraction of more relevant features, by means of these DSP methods, together with the use of more appropriate classifiers, will improve BCI performance in terms of classification rates and the number of recognizable mental tasks. It may be possible to apply recurrent neural networks to exploit temporal dynamics of brain activity. However, a main limitation in scaling up the number of recognizable mental tasks is the quality—signal-to-noise ratio (SNR)—and resolution of the measured brain signals. This is especially true in the case of EEG-based BCIs, in which the SNR is very poor and we cannot get detailed information on the activity of small cortical areas unless we use a large number of electrodes (64, 128, or more). It is thus crucial to develop better electrodes that are also easy to position, thereby enabling the use of a large number of electrodes even by laypersons. Finally, another key concern is to keep the BCI constantly tuned to its owner. This requirement arises because, as subjects gain experience, they develop new capabilities and change their EEG patterns. In addition, brain activity changes from one session (with which data the classifier is trained) to the next (where the classifier is applied). The challenge here is to adapt the classifier on-line while the subject operates a brain-actuated application, even if the subject's intention is not known until later. In this respect, local neural networks are better suited for on-line learning (STATISTICAL MECHANICS OF ON-LINE LEARNING AND GENERALIZATION) than other methods, due to their robustness against catastrophic interference. This list of topics is not exhaustive, but space limits prevent further discussion (see Wolpaw et al., 2000, for additional details on these and other issues).

Although the immediate application of BCI is to help physically impaired people, its potentials are extensive. Ultimately, they may lead to the development of truly adaptive interactive systems that, on the one hand, augment human capabilities by giving the brain the possibility to develop new skills and, on the other hand, make computer systems fit the pace and individual features of their owners. Most probably, people will use BCI in combination with other sensory interaction modalities (e.g., speech, gestures) and physiological signals (e.g., electromyogram, skin conductivity). Such a multimodal interface will yield a higher bit rate of communication with better reliability than would occur if only brainwaves were utilized. On the other hand, the incorporation of other interaction modalities highlights a critical issue in BCI, namely the importance of filtering out from the recorded brain signals non-CNS artifacts originated by movements of different parts of the body. INDEPENDENT COMPONENT ANALYSIS (q.v.) is a method of detecting and removing such artifacts.

Road Map: Applications

Related Reading: Event-Related Potentials; Kalman Filtering; Neural Implications; Prosthetics, Motor Control; Prosthetics, Sensory Systems

References

- Anderson, C. W., 1997, Effects of variations in neural network topology and output averaging on the discrimination of mental tasks from spontaneous EEG, *J. Intell. Syst.*, 7:165–190.
- Babiloni, F., Cincotti, F., Bianchi, L., Pirri, G., Millán, J. del R., Mourino, J., Sallinari, S., and Marciani, M. B., 2001, Recognition of imagined hand movements with low resolution surface Laplacian and linear classifiers, *Med. Eng. & Physics*, 23:323–328.
- Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H., 1999, A spelling device for the paralysed, *Nature*, 398:297–298.
- Kalcher, J., Flotzinger, D., Neuper, C., Göll, S., and Pfurtscheller, G., 1996, Graz brain-computer interface II, *Med. Biol. Eng. Comput.*, 34:382–388.

- Kennedy, P. R., Bakay, R. A. E., Moore, M. M., Adams, K., and Goldwithe, J., 2000, Direct control of a computer from the human central nervous system, *IEEE Trans. Rehab. Eng.*, 8:198–202.
- McFarland, D. J., McCane, L. M., David, S. V., and Wolpaw, J. R., 1997, Spatial filter selection for EEG-based communication, *Electroenceph. Clin. Neurophysiol.*, 103:386–394.
- Millán, J. del R., 2002, Adaptive brain interfaces, *Comm. of the ACM*, to appear. ♦
- Millán, J. del R., Franzé, M., Mouríño, J., Cincotti, F., and Babiloni, F., 2002a, Relevant EEG features for the classification of spontaneous motor-related tasks, *Biol. Cybern.*, 86:89–95.
- Millán, J. del R., Mouríño, J., Franzé, M., Cincotti, F., Varsta, M., Heikonen, J., and Babiloni, F., 2002b, A local neural classifier for the recognition of EEG patterns associated to mental tasks, *IEEE Trans. on Neural Networks*, 11:678–686.
- Nicolelis, M. A. L., 2001, Actions from thoughts, *Nature*, 409:403–407. ♦
- Obermaier, B., Müller, G., and Pfurtscheller, G., 2001, “Virtual Keyboard” controlled by spontaneous EEG activity, in *Proceedings of the International Conference on Artificial Neural Networks*, Heidelberg: Springer-Verlag.
- Roberts, S. J., and Penny, W. D., 2000, Real-time brain-computer interfacing: A preliminary study using Bayesian learning, *Med. Biol. Eng. Computing*, 38:56–61.
- Wessberg, J., Stambaugh, C. R., Kralik, J. D., Beck, P. D., Laubach, M., Chapin, J. K., Kim, J., Biggs, S. J., Srinivassan, M. A., and Nicolelis, M. A. L., 2000, Real-time prediction of hand trajectory by ensembles of cortical neurons in primates, *Nature*, 408:361–365.
- Wolpaw, J. R., and McFarland, D. J., 1994, Multichannel EEG-based brain-computer communication, *Electroenceph. Clin. Neurophysiol.*, 90:444–449.
- Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M., 2000, Brain-computer interface technology: A review of the first international meeting, *IEEE Trans. on Rehab. Eng.*, 8:164–173. Special Section on Brain-Computer Interfaces. ♦

Canonical Neural Models

Frank Hoppensteadt and Eugene M. Izhikevich

Introduction

Mathematical modeling is a powerful tool for studying the fundamental principles of information processing in the brain. Unfortunately, mathematical analysis of a certain neural model could be of limited value, because the results might depend on the particulars of that model: various models of the same neural structure could produce different results. For example, if an investigator obtains results with a Hodgkin-Huxley-type model (see AXONAL MODELING) and then augments the model by adding more parameters and variables to take into account more neurophysiological data, would similar results emerge? A reasonable way to circumvent this problem is to derive results that are largely independent of the model and that can be observed in a class or a family of models.

Having understood the importance of considering families of neural models instead of a single model, we carry out this task by reducing an entire family of Hodgkin-Huxley-type models to a canonical model (for precise definitions, see Section 4.1 in Hoppensteadt and Izhikevich, 1997). Briefly, a model is *canonical* for a family if there is a continuous change of variables that transforms any other model from the family into this one, as we illustrate in Figure 1. For example, the entire family of weakly coupled oscillators of the form in Equation 1 can be converted into the canonical phase model described by Equation 6, where H_{ij} depend on the particulars of the functions f_i and g_{ij} . The change of variables does not have to be invertible, so the canonical model is usually lower dimensional, simple, and tractable. Yet it retains many important features of the family. For example, if the canonical model has multiple attractors, then each member of the family has multiple attractors.

The major advantage to considering canonical models is that one can study universal neurocomputational properties that are shared by all members of the family, since all such members can be put into the canonical form by a continuous change of variables. Moreover, an investigator need not actually present such a change of variables explicitly, so that derivation of canonical models is possible even when the family is so broad that most of its members are given implicitly, e.g., in the abstract form of Equation 1. For example, the canonical phase model in Equation 6 reveals universal computational abilities (e.g., oscillatory associative memory) that are shared by all oscillatory systems, regardless of the nature of

each oscillator or the particulars of the equations that describe it. Thus, the canonical model approach provides a rigorous way to obtain results when only partial information about neuron dynamics is known. Many examples are given subsequently in this article.

The process of deriving canonical neural models is more an art than a science, because a general algorithm for doing so is not known. However, much success has been achieved when we consider weakly connected networks of neurons whose activity is near a bifurcation, a situation that often occurs when the membrane potential is near the threshold value (see DYNAMICS AND BIFURCATION IN NEURAL NETS and PHASE-PLANE ANALYSIS OF NEURAL NETS). We review such bifurcations and corresponding canonical

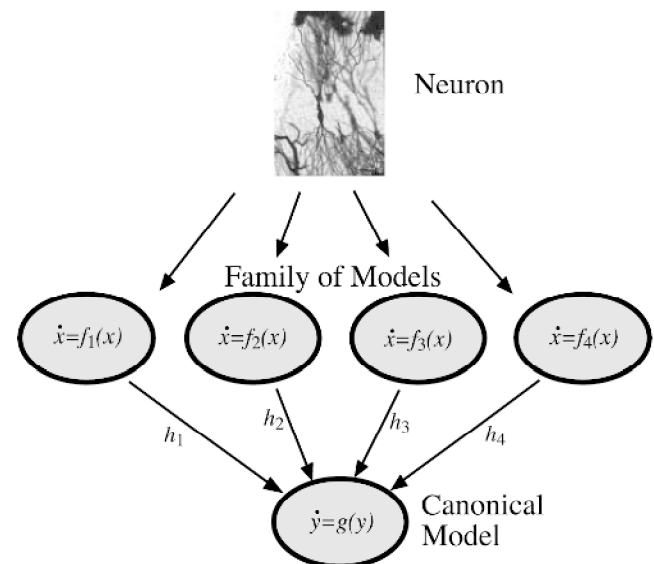


Figure 1. Dynamical system $\dot{y} = g(y)$ is a canonical model for the family $\{f_1, f_2, f_3, f_4\}$ of neural models $\dot{x} = f(x)$ because each such model can be transformed into the form $\dot{y} = g(y)$ by the continuous change of variables h_i .

models. Their rigorous derivation and detailed analysis can be found in Hoppensteadt and Izhikevich (1997).

Weakly Connected Neural Networks

The assumption of weak neuronal connections is based on the observation that the typical size of a postsynaptic potential is less than 1 mV, which is small in comparison with the mean size necessary to discharge a cell (around 20 mV) or the average size of the action potential (around 100 mV) (for a detailed review of relevant electrophysiological data, see Hoppensteadt and Izhikevich, 1997, chap. 1). From the mathematical point of view, this results in neural models of “weakly connected” form

$$\dot{x}_i = f(x_i, \lambda_i) + \varepsilon \sum_{j=1}^n g_{ij}(x_i, x_j, \varepsilon) \quad (1)$$

where each vector $x_i \in \mathbb{R}^m$ describes membrane potential, gating variables, and other electrophysiological variables of the i th neuron (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION). Each vector $\lambda_i \in \mathbb{R}^l$ denotes various biophysical parameters of the neuron. The function f describes the neuron’s dynamics, and the functions g_{ij} describe connections between the neurons. The dimensionless parameter $\varepsilon \ll 1$ is small, reflecting the strength of connections between neurons.

Bistability and Hysteresis

Bistable and hysteretic dynamics are ubiquitous in neural models, and they may play important roles in biological neurons. The cusp bifurcation depicted in Figure 2 is one of the simplest bifurcations leading to such dynamics. For example, the sigmoidal neuron

$$\dot{x} = -x + aS(x), \quad S(x) = 1/(1 + e^{-x})$$

is at a cusp bifurcation point $x = 0.5$ when $a = 4$. It is bistable when $a > 4$. If each neuron in the weakly connected network described by Equation 1 is near a supercritical cusp bifurcation, then the entire network can be transformed into the canonical form (Hoppensteadt and Izhikevich, 1997)

$$y_i' = r_i - y_i^3 + \sum_{j=1}^n s_{ij} y_j \quad (2)$$

where each scalar $y_i \in \mathbb{R}$ describes rescaled dynamics of the i th neuron. Particulars of the functions f and g_{ij} and the value of the

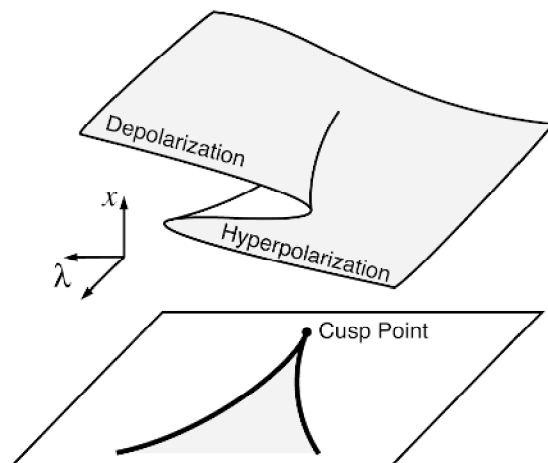


Figure 2. Cusp surface.

parameters λ_i do not affect the form of the canonical model. They only affect the parameters r_i and s_{ij} . Thus, by studying the canonical model in Equation 2 one can gain some insight into the neurocomputational behavior of any neural model near a cusp bifurcation, whether it is a simple sigmoidal neuron or a biophysically detailed conductance-based (Hodgkin-Huxley-type) neuron.

The canonical model in Equation 2 is quite simple: each equation has only one nonlinear term, namely, y_i^3 , and two internal parameters, r_i and s_{ij} . Still, the Cohen-Grossberg-Hopfield convergence theorem applies, which means that the canonical model has the same neurocomputational properties as the standard Hopfield network (see COMPUTING WITH ATTRACTORS).

Theorem 1 (Cohen-Grossberg-Hopfield Convergence Theorem)

If the connection matrix $S = (s_{ij})$ is symmetric, then the canonical neural network of Equation 2 is a gradient system.

One can easily check that

$$E(y) = -\sum_{i=1}^n (r_i y_i - \frac{1}{4} y_i^4) - \frac{1}{2} \sum_{i,j=1}^n s_{ij} y_i y_j$$

is a potential function for Equation 2 in the sense that $y_i' = -\partial E / \partial y_i$ (see also ENERGY FUNCTIONALS FOR NEURAL NETWORKS).

Small-Amplitude Oscillations

Many biophysically detailed neural models can exhibit small-amplitude (damped) oscillations of the membrane potential, especially when the system is near transition from rest state to periodic activity. In the simplest case this corresponds to the supercritical

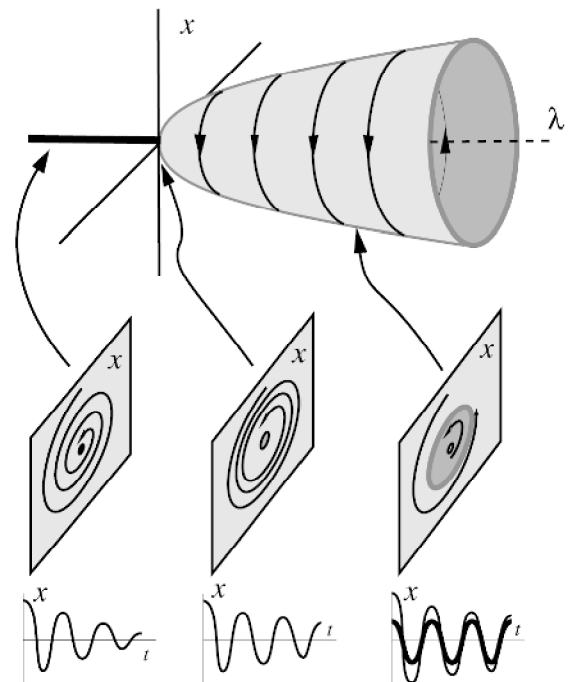


Figure 3. Supercritical Andronov-Hopf bifurcation in $\dot{x} = f(x, \lambda)$. On the left, the rest state is stable. In the middle, the rest state is losing stability, giving birth to a stable limit cycle corresponding to periodic activity. On the right, the system exhibits periodic activity.

Andronov-Hopf bifurcation in Figure 3. Many weakly connected networks (in the form of Equation 1) of such neurons can be transformed into the canonical model

$$z'_i = (r_i + i\omega_i)z_i - z_i|z_i|^2 + \sum_{j=1}^n c_{ij}z_j \quad (3)$$

by a continuous change of variables (Aronson, Ermentrout, and Kopell, 1990). Here $i = \sqrt{-1}$, and each complex variable $z_i \in \mathbb{C}$ describes oscillatory activity of the i th neuron. Again, particulars of the form of the functions f and g_{ij} in Equation 1 affect only the values of the parameters r_i and ω_i and the complex-valued synaptic coefficients $c_{ij} \in \mathbb{C}$.

Even though the canonical model in Equation 3 exhibits oscillatory dynamics, one can still prove the following analogue of the Cohen-Grossberg convergence theorem, which implies that the canonical model in Equation 3 has oscillatory associative memory; that is, it can memorize and retrieve complex oscillatory patterns (Hoppensteadt and Izhikevich, 1996) (Figure 4).

Theorem 2 (Synchronization Theorem for Oscillatory Neural Networks)

If in the canonical neural network in Equation 3 all neurons have equal frequencies $\omega_1 = \dots = \omega_n$ and the connection matrix $C = (c_{ij})$ is self-adjoint, i.e.,

$$c_{ij} = \bar{c}_{ji} \quad \text{for all } i \text{ and } j \quad (4)$$

then the network always converges to an oscillatory phase-locked pattern; that is, the neurons oscillate with equal frequencies and constant, but not necessarily identical, phases. There could be many such phase-locked patterns corresponding to many memorized images.

The proof follows from the existence of an orbital energy function

$$E(z) = -\sum_{i=1}^n (r_i|z_i|^2 - \frac{1}{2}|z_i|^4) - \sum_{i,j=1}^n c_{ij}\bar{z}_i z_j$$

for Equation 3 (see ENERGY FUNCTIONALS FOR NEURAL NETWORKS).

The self-adjoint synaptic matrix arises naturally when one considers complex Hebbian learning rules (Hoppensteadt and Izhikevich, 1996):

$$c_{ij} = \frac{1}{n} \sum_{s=1}^k \xi_i^s \bar{\xi}_j^s \quad (5)$$

where each vector $\xi^s = (\xi_1^s, \dots, \xi_n^s) \in \mathbb{C}^n$ denotes a pattern of phase relations between neurons to be memorized (see also HEBBIAN SYNAPTIC PLASTICITY). Notice that the problem of negative (mirror) images does not arise in oscillatory neural networks, since both ξ^k and $-\xi^k$ result in the same phase relations.

The key difference between the Hopfield-Grossberg network and the oscillatory network (Equation 3) is that memorized images correspond to equilibrium (point) attractors in the former and to limit cycle attractors in the latter. Pattern recognition by an oscillatory neural network involves convergence to the corresponding limit cycle attractor, which results in synchronization of the network activity with an appropriate phase relation between neurons, as in Figure 4 (see also COMPUTING WITH ATTRACTORS).

Large Amplitude Oscillations

Suppose that neurons in the weakly connected network described by Equation 1 exhibit periodic spiking (Figure 5; see also CHAINS OF OSCILLATORS IN MOTOR AND SENSORY SYSTEMS, COLLECTIVE BEHAVIOR OF COUPLED OSCILLATORS, and PHASE-PLANE ANALYSIS OF NEURAL NETS). If they have nearly equal frequencies, then the network can be transformed into the phase canonical model

$$\varphi'_i = \omega_i + \sum_{j=1}^n H_{ij}(\varphi_j - \varphi_i) \quad (6)$$

where each $\varphi_i \in \mathbb{S}^1$ is a one-dimensional (angle) variable that describes the phase of the i th oscillator along the limit cycle attractor corresponding to its periodic spiking (see Figure 5), and each H_{ij} is a function that depends on f and g_{ij} that can be explicitly computed using Malkin's theorem (Theorem 9.2 in Hoppensteadt and Izhikevich, 1997).

The phase canonical model (Equation 6) describes frequency locking, phase locking, and synchronization properties of the original system in Equation 1. Therefore, to understand these and other nonlinear phenomena that might take place in oscillating neural

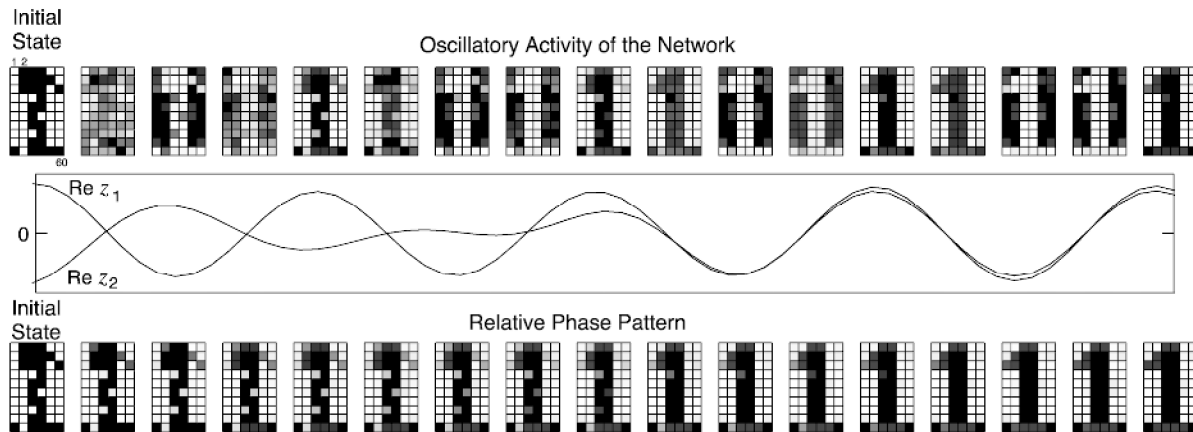


Figure 4. Pattern recognition via phase locking by the oscillatory canonical model (Equation 3). Complex Hebbian learning rule (5) was used to memorize patterns “1,” “2,” and “3.” When the distorted pattern “1” is presented

as an initial state, the neurons synchronize with the phase relations corresponding to the memorized pattern “1.”

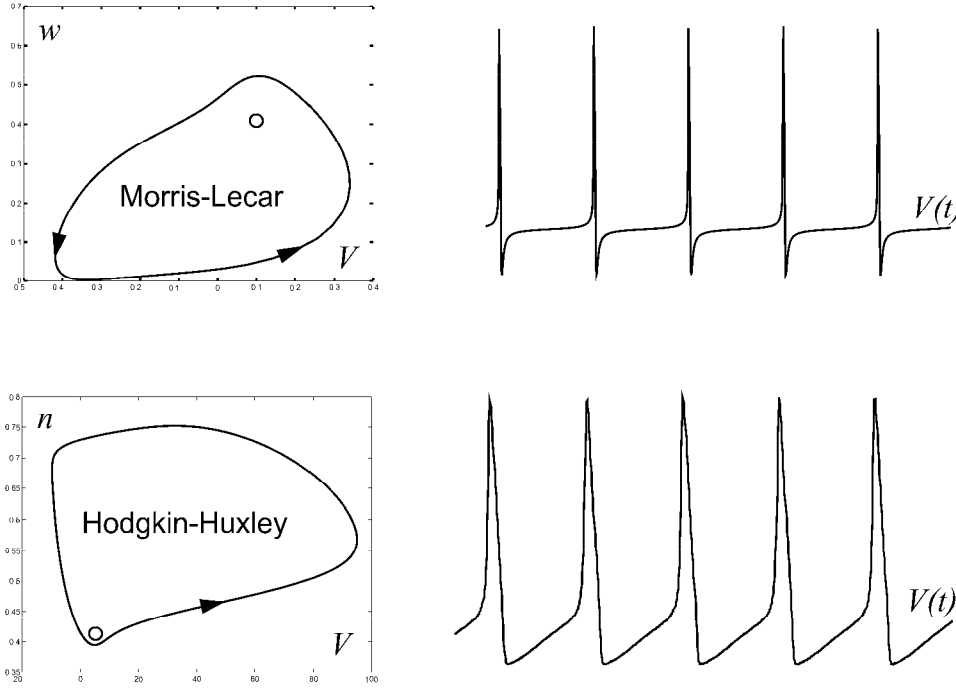


Figure 5. Examples of large-amplitude limit cycle attractors corresponding to periodic spiking in two biophysically detailed neural models, Morris and Lecar (1981) and Hodgkin and Huxley (1952).

networks, it usually suffices to consider the phase model. In particular, one can glimpse the universal computational abilities that are shared by all oscillatory systems, regardless of the nature of each oscillator or the particulars of the equations that describe it. Indeed, one can prove the following analogue of Theorem 2.

Theorem 3 (Synchronization Theorem for Oscillatory Neural Networks)

If all oscillators in Equation 6 have equal frequencies, i.e., $\omega_1 = \dots = \omega_n$, and the connection functions H_{ij} have pairwise odd form, i.e.,

$$H_{ij}(-\psi) = -H_{ij}(\psi) \quad (7)$$

for all i and j , then the canonical phase model in Equation 6 converges to a phase-locked pattern $\varphi_i(t) \rightarrow \omega_1 t + \phi_i$ for all i , so the neurons oscillate with equal frequencies (ω_1) and constant phase relations ($\varphi_i(t) - \varphi_j(t) = \phi_i - \phi_j$). In this sense the network dynamic is synchronized. There could be many stable synchronized patterns corresponding to many memorized images.

The proof is based on the observation that the phase canonical model in Equation 6 has the energy function

$$E(\varphi) = \frac{1}{2} \sum_{i,j=1}^n R_{ij}(\varphi_j - \varphi_i)$$

where R_{ij} is the antiderivative of H_{ij} ; that is, $R'_{ij} = H_{ij}$ (see Theorem 9.15 in Hoppensteadt and Izhikevich, 1997, and ENERGY FUNCTIONALS FOR NEURAL NETWORKS).

For example, Kuramoto's (1984) model

$$\varphi'_i = \omega_i + \sum_{j=1}^n s_{ij} \sin(\varphi_j + \psi_{ij} - \varphi_i) \quad (8)$$

has such an oscillatory associative memory when $\omega_1 = \dots = \omega_n$:

$$s_{ij} = s_{ji} \quad \text{and} \quad \psi_{ij} = -\psi_{ji}$$

for all i and j . If we denote $c_{ij} = s_{ij} e^{i\psi_{ij}}$, then these conditions have the form shown by Equation 4. The energy function for Kuramoto's model is

$$E(\varphi) = -\frac{1}{2} \sum_{i,j=1}^n s_{ij} \cos(\varphi_j + \psi_{ij} - \varphi_i)$$

There are various estimates of the storage capacity of the network, as discussed by Vicente, Arenas, and Bonilla (1996). In particular, those authors found a time scale during which oscillatory networks can have better performance than Cohen-Grossberg-Hopfield-type attractor neural networks.

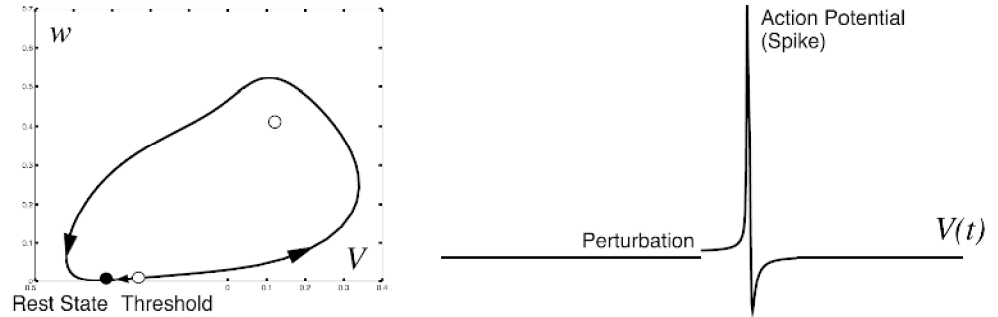
Since neither the form of the functions f and g_{ij} nor the dimension of each oscillator in Equation 1 were specified, one could take the above result to the extreme and claim that *anything that can oscillate can also be used for computing*, as for associative pattern recognition, etc. The only problem is how to couple the oscillators so that Equation 7 is satisfied.

Neural Excitability

An interesting intermediate case between rest and periodic spiking behavior is when a neural system is *excitable*; that is, it is at rest, but can generate a large-amplitude spike in response to a small perturbation (Figure 6) (see PHASE-PLANE ANALYSIS OF NEURAL NETS and OSCILLATORY AND BURSTING PROPERTIES OF NEURONS). A simple but useful criterion for classifying neural excitability was suggested by Hodgkin (1948), who stimulated cells by applying currents of various strengths. When the current is weak the cell is quiet. When the current is strong enough, the cell starts to fire repeatedly. He suggested the following classification according to the emerging frequency of firing (Figure 7):

- **Class 1 neural excitability:** Action potentials can be generated with arbitrarily low frequency. The frequency increases as the applied current is increased.
- **Class 2 neural excitability:** Action potentials are generated in a certain frequency band that is relatively insensitive to changes in the strength of the applied current.

Figure 6. Neural excitability in Morris and Lecar (1981) neuron having fast Ca^{2+} and slow K^{+} voltage-gated currents. The rest state (solid circle) is stable, but small perturbations can push the voltage beyond the threshold (open circle), thereby causing a large-amplitude excursion, or action potential. The voltage variable changes slowly near the rest states but fast during the generation of action potentials.



Their class of excitability influences neurocomputational properties of cells (see review in Izhikevich, 2000). For example, class 1 neural systems have a well-defined threshold manifold for their state variables, beyond which they generate a large-amplitude spike. They generate an all-or-none response, and they act as *integrators*, meaning that the higher the frequency of the incoming pulses, the sooner they fire. In contrast, class 2 neural systems may not have a threshold manifold. They could generate spikes of arbitrary intermediate amplitude, and they could act as *resonators*. That is, they respond to certain resonant frequencies of the incoming pulses. Increasing the incoming frequency may delay or even terminate their response.

A canonical model for class 1 excitable systems is described in the next section. The canonical model for class 2 systems has yet to be found.

Class 1 Excitable Systems

Class 1 excitable systems are understood relatively well (Rinzel and Ermentrout, 1989; Ermentrout, 1996; Hoppensteadt and Izhikevich, 1997; Izhikevich, 2000).

The transition from rest to periodic spiking in such systems occurs via a saddle node on invariant circle bifurcation, as shown in Figure 8 (see also DYNAMICS AND BIFURCATION IN NEURAL NETS AND OSCILLATORY AND BURSTING PROPERTIES OF NEURONS). A weakly connected network of such neurons can be transformed into a canonical model, which can be approximated by

$$\vartheta'_i = 1 - \cos \vartheta_i + (1 + \cos \vartheta_i) \left(r_i + \sum_{j=1}^n s_{ij} \delta(\vartheta_j - \pi) \right) \quad (9)$$

where $\vartheta_i \in \mathbb{S}^1$ is the phase of the i th neuron along the limit cycle corresponding to the spiking solution. Again, particulars of the functions f and g_{ij} in Equation 1 do not affect the form of the canonical model in Equation 9, but only affect the values of the parameters r_i and s_{ij} , which can be computed explicitly (Hoppensteadt and Izhikevich, 1997, chap. 8). Notice that the canonical model in Equation 9 is pulse coupled, whereas the original weakly coupled network in Equation 1 is not. The qualitative reason for pulse coupling is that the voltage changes are extremely slow most

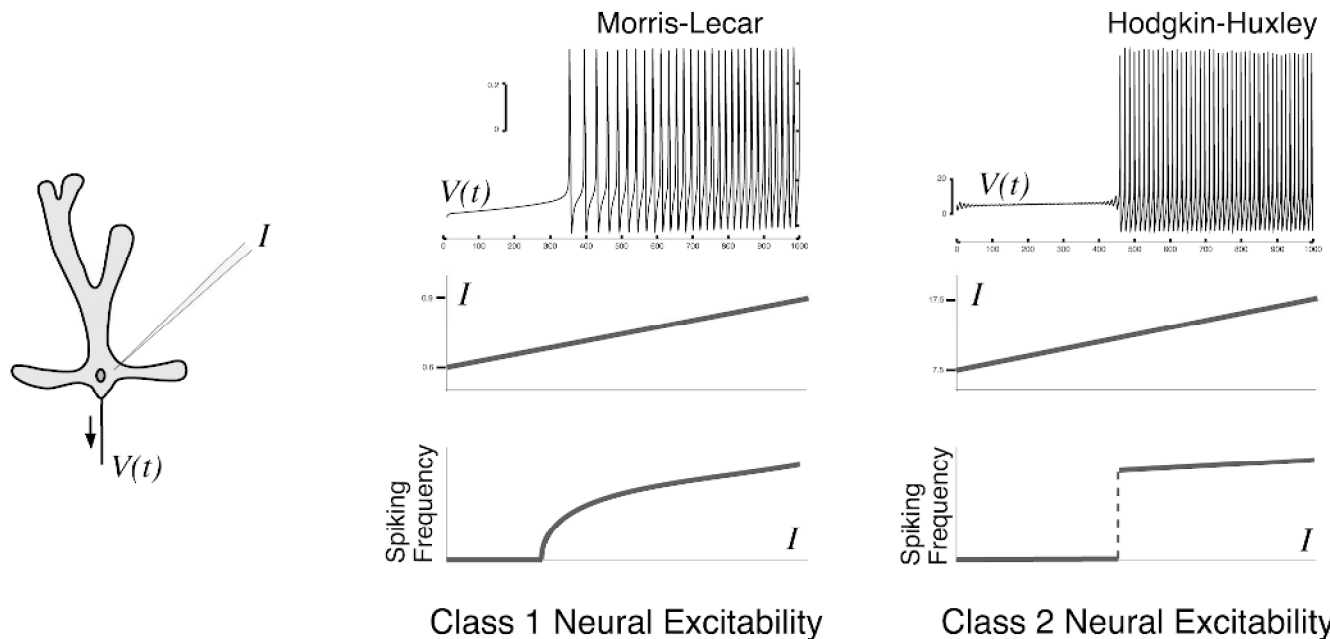


Figure 7. Transition from rest to repetitive spiking in two biophysical models when the strength of applied current, I , increases. The neural excitability is classified according to the frequency of emerging spiking.

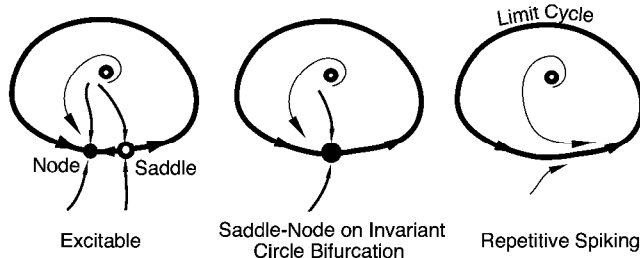


Figure 8. Class 1 neural excitability via saddle node on invariant circle bifurcation: The threshold state (saddle) approaches the rest state (node), they coalesce and annihilate each other leaving only limit cycle attractor. The oscillation on the attractor has two time scales: slow transition through the “ghost” of the saddle-node bifurcation and fast rotation along the rest of the limit cycle.

of the time because of the proximity to the rest state, but they are relatively instantaneous during the generation of an action potential. Hence the duration of coupling looks infinitesimal on the slow time scale.

The neuron is quiescent when $r_i < 0$ (Figure 8, left) and fires periodically when $r_i > 0$ (Figure 8, right). It fires a spike exactly when ϑ_i crosses the value π , which results in a step-like increase in the phases of other neurons. Hence, the canonical model in Equation 9 is a *pulse coupled neural network* (Izhikevich, 1999). It has many important physiological features, including absolute and relative refractory periods (Figure 9). Indeed, the effect of every incoming pulse depends on the internal state of the neuron, since it is multiplied by the function $(1 + \cos \vartheta_i)$. The effect is maximal when the neuron is near rest, since $(1 + \cos \vartheta_i) \approx 2$ when $\vartheta_i \approx 0$. It is minimal when the neuron is generating a spike, since $(1 + \cos \vartheta_i) \approx 0$ when $\vartheta_i \approx \pi$.

A canonical model for *slowly* connected class 1 excitable neurons with spike frequency adaptation has the form (Izhikevich, 2000):

$$\begin{aligned}\vartheta'_i &= 1 - \cos \vartheta_i + (1 + \cos \vartheta_i) \left(r_i + \sum_{j=1}^n s_{ij} w_j \right) \\ w'_i &= \delta(\vartheta_i - \pi) - \eta w_i\end{aligned}$$

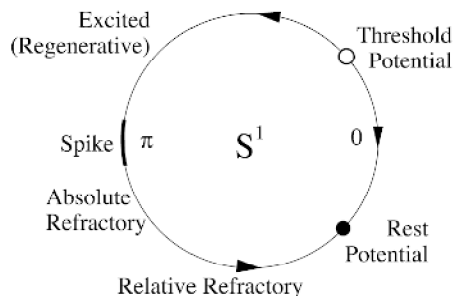


Figure 9. Diagram of the canonical model in Equation 9 for class 1 neural excitability. (From Hoppensteadt and Izhikevich, 1997.)

where w_i describes slow synaptic processes. The term $s_{ij}w_j$ denotes not a self-synapse but a slow spike frequency adaptation ($s_{ii} < 0$) or facilitation ($s_{ii} > 0$) process.

Discussion

The canonical model approach to computational neuroscience provides a rigorous way to derive simple yet accurate models that describe single-cell or network dynamics (see SINGLE-CELL MODELS). Such a derivation is possible even when no assumptions are made regarding the detailed form of equations describing neural activity. Indeed, we specify neither f nor g_{ij} in Equation 1. The only assumptions we make are those concerning the dynamics of each neuron—whether it is quiescent, excitable, periodic spiking, and so on. Nevertheless, any such neural system can be transformed into a canonical model by a piecewise continuous change of variables.

The derivation of canonical models can be a daunting mathematical task. However, once found, the canonical models provide invaluable information about universal neurocomputational properties shared by a large family of neural systems. For example, studying the canonical model in Equation 9 sheds light on the behavior of *all* class 1 excitable systems and their networks, regardless of the details of equations describing their dynamics.

Road Map: Dynamic Systems

Background: Dynamics and Bifurcation in Neural Nets; Phase-Plane Analysis of Neural Nets

Related Reading: Axonal Modeling; Chains of Oscillators in Motor and Sensory Systems; Collective Behavior of Coupled Oscillators; Computing with Attractors; Cooperative Phenomena; Energy Functionals for Neural Networks; Pattern Formation, Neural

References

- Aronson, D. G., Ermentrout, G. B., and Kopell, N., 1990, Amplitude response of coupled oscillators, *Physica D*, 41:403–449.
- Ermentrout, G. B., 1996, Type I membranes, phase resetting curves, and synchrony, *Neural Computat.*, 8:979–1001. ♦
- Hodgkin, A. L., 1948, The local electric changes associated with repetitive action in a non-medulated axon, *J. Physiol.*, 107:165–181.
- Hodgkin, A. L., and Huxley, A. F., 1952, A quantitative description of membrane current and application to conduction and excitation in nerve, *J. Physiol.*, 117:500–544.
- Hoppensteadt, F. C., and Izhikevich, E. M., 1996, Synaptic organizations and dynamical properties of weakly connected neural oscillators: II. Learning of phase information, *Biol. Cybern.*, 75:129–135. ♦
- Hoppensteadt, F. C., and Izhikevich, E. M., 1997, *Weakly Connected Neural Networks*, New York: Springer-Verlag.
- Izhikevich, E. M., 1999, Class 1 neural excitability, conventional synapses, weakly connected networks, and mathematical foundations of pulse-coupled models, *IEEE Trans. Neural Netw.*, 10:499–507.
- Izhikevich, E. M., 2000, Neural excitability, spiking, and bursting, *Int. J. Bifurcat. Chaos*, 10:1171–1266. ♦
- Kuramoto, Y., 1984, *Chemical Oscillations, Waves, and Turbulence*, New York: Springer-Verlag.
- Morris, C., and Lecar, H., 1981, Voltage oscillations in the barnacle giant muscle fiber, *Biophys. J.*, 35:193–213.
- Rinzel, J., and Ermentrout, G. B., 1989, Analysis of neural excitability and oscillations, in *Methods in Neuronal Modeling* (C. Koch and I. Segev, Eds.), Cambridge, MA: MIT Press. ♦
- Vicente, C. J., Arenas, A., and Bonilla, L. L., 1996, On the short-term dynamics of networks of Hebbian coupled oscillators, *J. Phys. A*, L9–L16.

Cerebellum and Conditioning

Jeffrey S. Grethe and Richard F. Thompson

Introduction

For many years, psychologists and neurobiologists have been searching for the substrates underlying learning and memory. Aristotle hypothesized that learning involved the association of ideas with one another. Pavlov (1927, cited in Gormezano, Kehoe, and Marshall, 1983) combined the concepts of learning and association with the production of reflexes (see **CONDITIONING**). Classical conditioning (CC) has been extremely useful in examining the substrates underlying learning and memory. The standard CC paradigm (Figure 1) consists of pairing a neutral stimulus, the conditioned stimulus (CS), with an aversive stimulus, the unconditioned stimulus (US). At the beginning of training, the only response to the stimuli is the unconditioned response (UR) to the US. Over repeated pairings of the stimuli, an association is formed between the CS and US, resulting in the performance of a conditioned response (a response that resembles the UR).

By varying stimulus parameters, a large number of behavioral-conditioning phenomena can be observed (Gormezano et al., 1983). Eye-blink conditioning can occur with a CS-US interstimulus interval (ISI) ranging from 100 ms to well over 1 s. The rate of learning and asymptotic response level are optimal at an ISI of about 250 ms. After 100–200 training trials at the optimal ISI, rabbits give CRs on more than 90% of trials. The CR onset initially develops near the US onset, gradually begins earlier in the trial over the course of training, and generally peaks near the US onset (Figure 1). More complex conditioning phenomena can also be observed (see **CONDITIONING**).

Cerebellar Substrates of Classical Eye-Blink Conditioning

Current evidence from extensive anatomical, lesion, and physiological studies argues very strongly that the essential memory trace for the classically conditioned nictitating membrane response is formed and stored in the cerebellum (Thompson et al., 1997). This research has identified much of the network subserving classical conditioning (Figure 2). Information regarding the US is transmit-

ted by the dorsal accessory olive to the cerebellum through the climbing fibers. Stimulation of the dorsal accessory olive, in place of a corneal airpuff, has been shown to be an effective US. If the dorsal accessory olive is lesioned before paired presentations of the CS and US, learning of the conditioned response is prevented. In addition, the interpositus provides inhibitory feedback to the inferior olive, and over the course of learning, the inhibitory feedback increases, thereby decreasing the output of the inferior olive. This evidence points to the olivary climbing fiber system as being the essential reinforcing pathway that transmits an error signal to the cerebellum. Information regarding the CS is transmitted to the cerebellum from the pontine nuclei by the mossy fibers. Stimulation of the pontine nuclei or mossy fibers has been shown to be an effective CS. The CR pathway consists of the anterior interpositus nucleus, magnocellular red nucleus, and finally the accessory abducens nucleus and the facial (seventh) nucleus. This circuit points to the cerebellum as the site of convergence where associative learning may take place. Experimental evidence supports this view, since lesions of the cerebellum, including the critical regions of the interpositus nucleus, completely abolish the conditioned response without affecting the UR.

One of the first theories to detail how associative memories could be formed in the cerebellum was proposed by Albus (1971). The most striking aspect of this model was that the parallel fiber synapses on Purkinje cells were modifiable. This theory of parallel fiber–Purkinje cell plasticity now has considerable experimental support. In addition, Albus predicted the occurrence of long-term depression at this synapse (see **CEREBELLUM: NEURAL PLASTICITY**). Furthermore, Albus predicted that climbing fiber spikes are the US and that mossy fiber activity patterns are the CS, predictions that are now supported by the conditioning literature. However, this theory does not account for the temporal dynamics of the CR. One of the more interesting features of classical conditioning is that of the well-timed response. Over the course of training, the timing of the response to the CS becomes more precise (i.e., the CR predicts the onset of the US). Most models of the cerebellum and its involvement in the classically conditioned eye-blink response focus on the production of this well-timed CR.

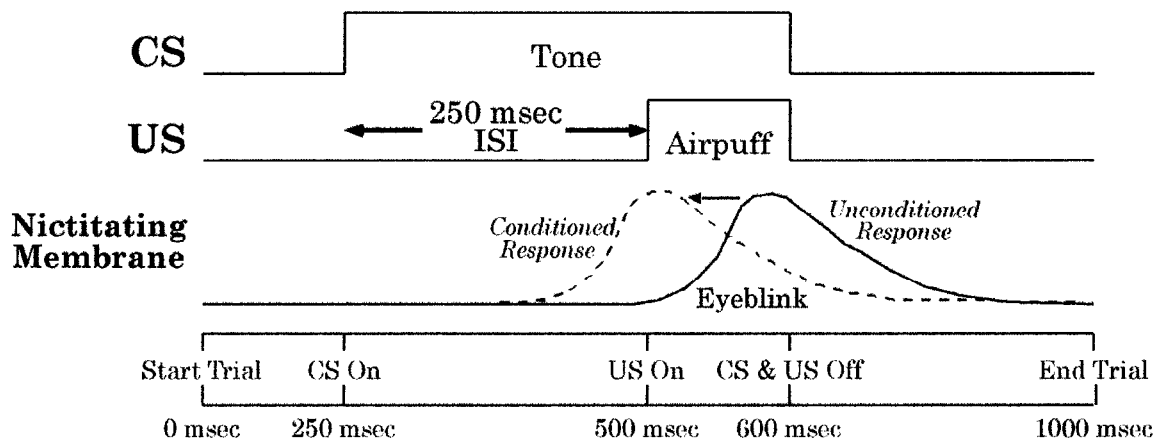


Figure 1. Standard delay conditioning paradigm. The CS is a 350-ms tone and the US is a 100-ms airpuff directed at the cornea. Early in training, the animal responds to the airpuff (UR). Over time, the animal learns to as-

sociate the tone with the airpuff, and produces a defensive blink that coincides with the airpuff (CR).

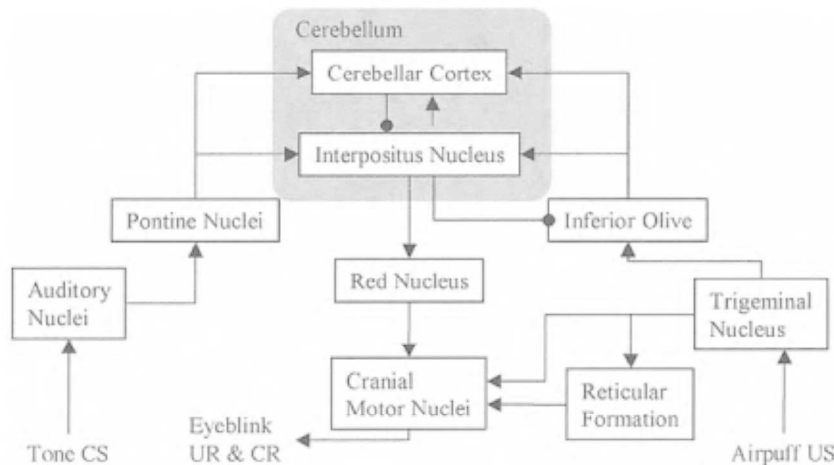


Figure 2. Essential circuitry for the classically conditioned nictitating membrane response.

Modeling the Role of the Cerebellum in Classical Conditioning

One of the first theories as to how the cerebellum could control movement timing was proposed by Braitenberg (1961). He suggested that parallel fibers could act as tapped delay lines, with the conduction time of the parallel fibers yielding movement timing. However, this theory cannot account for movements with time scales on the order of hundreds of milliseconds; it can only account for delays on the order of a few milliseconds. Many current models, however, still rely on this notion.

Moore, Desmond, and Berthier (1989) constructed a model of the cerebellum based on an earlier adaptive network model. In order to temporally associate the CS and US, the CS inputs are not discrete events but rather stimulus traces that persist for some time after the CS is gone (see *CONDITIONING*). Learning is then allowed to occur when the CS trace and the US coincide. In the cerebellar implementation the adaptive unit is the Purkinje cell, whereas the Golgi cell learns to gate an image of the CR from the brainstem to the Purkinje cell. Purkinje and Golgi cell plasticity, combined with the model's tapped delay lines, yields anticipatory CRs for delay and trace conditioning at all ISIs. The learning rules follow the form of Rescorla and Wagner's (1972) model and thus account for the same stimulus context effect. One problem with this model is that the tapped delay lines must be on the order of hundreds of milliseconds, and no physiologically plausible mechanism has been suggested for this. Another problem is the existence of the CR image: the model does not include learning of this hypothesized image. Moore et al. have hypothesized that this image is located around the trigeminal nucleus. Learning-induced models of the CR are present in the region bordering the trigeminal; however, evidence has shown that this model is relayed to the trigeminal from the interpositus via the red nucleus. It is interesting to note that a cerebellar implementation of the Sutton and Barto tapped delay model (Moore and Choi, 1997) also requires feedback from extracerebellar structures. In order to properly determine the difference between the predicted reinforcement and actual reinforcement, the model uses an efference copy of the conditioned response from the red nucleus and spinal trigeminal nucleus.

Jaffe (1990) proposed a model that moved the tapped delay lines from the network and placed them at the level of the neuron. The model proposed that single interpositus neurons can generate the full range of delays for CRs by adjusting their input weights and exploiting the phenomenon of delayed inhibitory rebound. The most significant problem with this model is that the interpositus

cell must be quiescent at CS onset and cannot fire through the delay period. However, most neurons in the interpositus are spontaneously active, and it seems unlikely that a few neurons silent during the delay could account for the production of the CR.

Fiala, Grossberg, and Bullock (1996) also developed a neuron-centered tapped delay line model of the metabotropic glutamate receptor (mGluR) second-messenger system (see *CEREBELLUM: NEURAL PLASTICITY*), which is responsible for the well-timed CR. Temporal correlation between the CS (parallel fiber-induced mGluR response) and US causes a phosphorylation of AMPA receptors and calcium-dependent potassium channels. Phosphorylation of the calcium-dependent potassium channels results in a reduction in Purkinje cell firing during the CS-US interval. This model is very interesting in that it explores possible biochemical mechanisms for production of the well-timed response. However, for this model to produce responses across the full range of timing, the variety in the density of mGluR receptors on dendritic spines must be highly variable, which may not be physiologically plausible.

In addition to tapped delay line models, researchers began investigating how dynamical network processes within the cerebellum could yield precise timing. Buonomano and Mauk (1994) developed a semirealistic population model of the cerebellar cortex. The CS activates a subset of mossy fibers, which in turn excites a population of granule cells. Both the mossy fibers and the granule cells excite a population of Golgi cells. The resulting Golgi inhibition of the granule cells produces a varying pattern of granule cell activity over time in which different subsets of granule cells are active at different times. Learning of the CR would occur through weakening of the parallel fiber synapses that were active around the occurrence of the US. An extension of this model (Mauk and Donegan, 1997) includes two sites of plasticity within the cerebellum, the cerebellar cortex and interpositus nucleus. More important, the model showed that long-term depression of the parallel fiber–Purkinje cell synapses, coupled with recurrent projections between the interpositus and inferior olive, produces a stable learning system. The most pressing problem with these models is extreme sensitivity to noise. If there is a substantial amount of noise in the network or if the input pattern varies during the CS, the Purkinje cell timing would be disrupted due to the changes in the granule cell activity pattern.

Bartha (1992) developed a network simulation of the cerebellum and associated circuitry that stressed the input and output representation of the cerebellum. The input representation was constrained by known properties of the mossy fibers and the output

representation was modeled through detailed information on the CR pathway and oculomotor plant. The cerebellar model consists of two populations of granule cells, one responsive to the tone CS and one unresponsive. The Golgi cell's inhibitory influence on the tone-unresponsive granule cells is to produce a variety of firing patterns so that the granule cells display differing periods of depression. The Purkinje cell then selects the granule cells (through long-term depression) that display the proper time interval of depression to produce a properly timed eye-blink. With realistic single-neuron parameters, the model is able to reproduce many aspects of CR timing and form. One concern with the model lies with the Golgi cells. For the Golgi cell to be able to produce the proper spectrum of delays for short ISIs, the time constant of the Golgi cells' influence on the granule cells must fall within the 100 ms to 250 ms range, which can be considered physiological. However, to produce properly timed blinks of longer latencies, the time constant of this inhibitory effect must be considerably longer, which does not seem physiologically plausible.

Grethe (2000) developed a model that focuses on the cerebellar microcomplex as a Hebbian cell assembly (a highly interconnected group of neurons that forms a reverberatory circuit that can sustain activity). Each microcomplex contains a set of Purkinje cells and their related nuclear neurons. The cerebellar cortex receives mossy fiber input from both pontine nuclei and recurrent projections from the interpositus nucleus. This basic architecture allows the recurrent excitation in the excitatory interpositus cells to be modulated by the Purkinje cells in the cerebellar cortex. The architecture of the model tries to preserve the topographic projections between the cerebellar cortex and the deep nuclei as well as the beam-like organization of Purkinje cells receiving input from the parallel fibers. Local reverberations in the cerebellum, between the cerebellar cortex and deep nuclei, are responsible for the response topography and timing. Long-term depression at the parallel fiber–Purkinje cell synapse is responsible for the precise timing of the response. The reverberations are controlled by the anatomical connectivity, the bistability of Purkinje neurons, and the process of synaptic fatigue. One interesting aspect of the model is the effect of stimulation intensity on the response timing of the assembly. Since the timing of the model is inherently dependent on the recurrent projections, this timing process can be sped up by increasing the stimulation intensity of the CS, which has been found experimentally.

Gluck and associates (2001) developed a connectionist-level model of classical eye-blink conditioning incorporating basic features of the essential cerebellar circuitry, based on the Rescorla-Wagner model, a most successful trial-level behavioral formulation of classical conditioning. The Rescorla-Wagner model assumes that the change in association between a neutral CS and a response-evoking US is a function of the difference between the US and an animal's *expectation* of the US, given all CSs present in the trial. Because the discrepancy, or "error," between the animal's expectations and what actually occurs drives learning in this theory, the theory is referred to as an "error-correcting" learning procedure.

Thirty years after its publication, the Rescorla-Wagner model still stands as the most influential and powerful model in psychology for describing and predicting animal learning behavior in conditioning studies. Moreover, its influence has extended far beyond animal conditioning. The model's basic error correction principle has been rediscovered within cognitive psychology and cognitive neuroscience in the form of connectionist network models, many of which rely on the same principle. In addition, the most commonly used connectionist learning procedure, back propagation (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION), along with its simpler predecessor, the delta rule (see CONDITIONING), both are generalizations of the Rescorla-Wagner model.

Although the Gluck et al. model does not attempt to model neuronal processes, it includes two key features of cerebellar circuitry:

positive feedback via the pontine nuclei and negative feedback via inhibition of the inferior olive. This rather simple connectionist model successfully predicts a wide range of behavioral phenomena of classical conditioning, including, in particular, adaptive timing of the CR and blocking. In the model, adaptive timing depends on positive feedback; if this feedback is blocked, adaptive timing no longer occurs. The model circuitry predicts that inhibition of the inferior olive-climbing fiber input system is necessary for the behavioral phenomenon of blocking. This prediction was empirically verified by Kim, Krupa, and Thompson (1998), who showed that blocking interpositus-evoked GABA inhibition of the inferior olive during compound-stimulus training completely prevented behavioral blocking.

Discussion

Currently, no model accounts for all the data on cerebellar neurobiology and conditioning. However, models of conditioning that take into account the neurobiology of the cerebellum can pose questions that researchers may be able to examine. For example, in order to experimentally test some of the network models presented, single-unit recordings from the granule would be necessary. These recordings are difficult because granule cells are small and closely spaced. Moreover, it would be difficult to obtain these recordings while the task is being performed. The predictions from several of the models regarding granule cells are all different:

- Buonomano and Mauk predict that a subset of granule cells will exhibit nonperiodic activity.
- Bartha predicts that the subset of granule cells will primarily show depressions in activity at all conditionable intervals after the presentation of a CS.
- Grethe predicts that activation of the granule cells in the cell assembly should occur in a wave-like fashion as the cell assembly's reverberating activity increases.

The future will bring new experiments that test these and future models, which in turn will lead to further experiments and models.

Road Map: Mammalian Brain Regions

Related Reading: Cerebellum and Motor Control; Cerebellum: Neural Plasticity; Conditioning

References

- Albus, J. S., 1971, A theory of cerebellar function, *Math. Biosci.*, 10:25–61.
- Bartha, G. T., 1992, A computer model of oculomotor and neural contributions to conditioned blink timing, Ph.D. diss., University of Southern California.
- Braitenberg, V., 1961, Functional interpretation of cerebellar histology, *Nature*, 190:539–540.
- Buonomano, D. V., and Mauk, M. D., 1994, Neural network model of the cerebellum: Temporal discrimination and the timing of motor responses, *Neural Computat.*, 6:38–55.
- Fiala, J. C., Grossberg, S., and Bullock, D., 1996, Metabotropic glutamate receptor activation in cerebellar Purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response, *J. Neurosci.*, 16:3760–3774.
- Gluck, M. A., Allen, M. T., Myers, C. E., and Thompson, R. F., 2001, Cerebellar substrates for error-correction in motor conditioning, *Neurobiol. Learn. Mem.*, 76:314–341.
- Gormezano, I., Kehoe, E. J., and Marshall, B. S., 1983, Twenty years of classical conditioning research with the rabbit, *Prog. Psychobiol. Physiol. Psychol.*, 10:197–275. ♦
- Grethe, J. S., 2000, Neuroinformatics and the cerebellum: Towards an understanding of the cerebellar microzone and its contribution to the well-timed classically conditioned eyeblink response, Ph.D. diss., University of Southern California.

- Jaffe, S., 1990, A neuronal model for variable latency response, in *Analysis and Modeling of Neural Systems* (F. H. Eeckman, Ed.), Boston: Kluwer, pp. 405–410.
- Kim, J., Krupa, D., and Thompson, R. F., 1998, Inhibitory cerebello-olivary projections and blocking effect in classical conditioning, *Science*, 27:570–573.
- Mauk, M. D., and Donegan, N. H., 1997, A model of pavlovian eyelid conditioning based on the synaptic organization of the cerebellum, *Learn. Mem.*, 4:130–158.
- Moore, J. W., and Choi, J.-S., 1997, The TD model of classical conditioning: Response topography and brain implementation, in *Neural-Network Models of Cognition* (J. Donahoe and V. Dorsel, Eds.), North-Holland: Elsevier, pp. 387–405.
- Moore, J. W., Desmond, J. E., and Berthier, N. E., 1989, Adaptively timed conditioned responses and the cerebellum: A neural network approach, *Biol. Cybern.*, 62:17–28.
- Rescorla, R. A., and Wagner, A. R. A., 1972, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement, in *Classical Conditioning: II. Current Research and Theory* (A. H. Black and W. F. Prokasy, Eds.), New York: Appleton-Century-Crofts.
- Thompson, R. F., Bao, S., Berg, M. S., Chen, L., Cipriano, B. D., Grethe, J. S., Kim, J. J., Thompson, J. K., Tracy, J., and Krupa, D. J., 1997, Associative learning, in *The Cerebellum and Cognition* (J. Schmammann, Ed.), *Int. Rev. Neurobiol.*, 41:151–189 (special issue). ♦

Cerebellum and Motor Control

Mitsuo Kawato

Introduction

Fast, smooth, and coordinated movements cannot be achieved by basic feedback control alone because delays associated with feedback loops are long (about 200 ms for visual feedback and 100 ms for somatosensory feedback) and feedback gains are low. Additionally, feedback controllers such as the commonly used PID (proportional, integral, and derivative) controllers do not incorporate predictive dynamic or kinematic knowledge of controlled objects or environments. Two major feedforward control schemes have been proposed: the equilibrium point control hypothesis and the inverse dynamics model hypothesis (see EQUILIBRIUM POINT HYPOTHESIS and MOTOR CONTROL, BIOLOGICAL AND THEORETICAL). Some versions of the former scheme advocate that the central nervous system (CNS) can avoid inverse dynamics computation by relying on the spring-like properties of muscles and reflex loops. For this mechanism to work efficiently, the mechanical and neural feedback gains, which can be measured as mechanical stiffness in perturbation experiments, must be quite high. The low stiffness of the arm, which was measured during visually guided point-to-point multijoint movements (Gomi and Kawato, 1996), suggests the necessity of inverse dynamics models in these well-practiced and relaxed movements. The internal models in the brain must be acquired through motor learning in order to accommodate the changes that occur with the growth of controlled objects such as hands, legs, and torso, as well as the unpredictable variability of the external world.

Where in the brain are internal models of the motor apparatus likely to be stored? First, the locus should exhibit a remarkable adaptive capability, which is essential for acquisition and continuous update of internal models of the motor apparatus. A number of physiological studies have suggested important functional roles of the cerebellum in motor learning and remarkable synaptic plasticity in the cerebellar cortex (Ito, 1984, 2001). Second, the biological objects of motor control by the brain, such as the arms, speech articulators, and the torso, possess many degrees of freedom and complicated nonlinear dynamics. Correspondingly, neural internal models should receive a broad range of sensory inputs and possess a capacity high enough to approximate complex dynamics. Extensive sensory signals carried by mossy fiber inputs and an enormous number of granule cells in the cerebellar cortex seem to fulfill these prerequisites for internal models. (See Figure 1 for cerebellar circuitry and its connection to the cerebellar nucleus in the case of the lateral cerebellum.) Finally, the cerebellar symptoms usually classified as the “triad” of hypotonia, hypermetria, and intention tremor could be understood as degraded performance when

control is forced to rely solely on primitive feedback control after internal models are destroyed or cannot be updated. This is because precise, fast, coordinated movements can be executed if accurate internal models of the motor apparatus can be utilized during trajectory planning, coordinate transformation, and motor control, while primitive feedback controllers with long feedback delays and small gains can attain only poor performance in these computations and usually lead to oscillatory instability for forced fast movements.

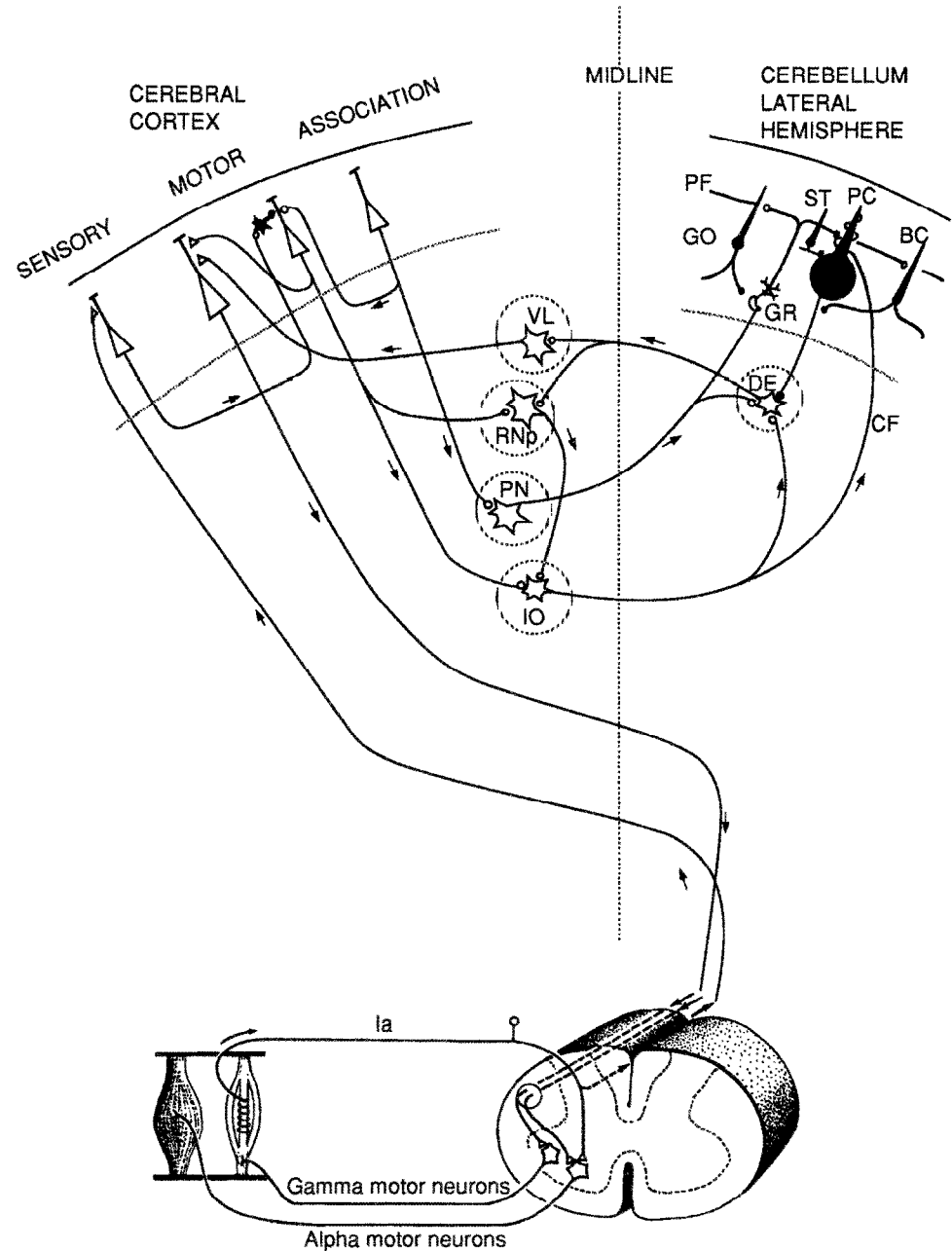
Miall et al. (1993) grouped into classes many theories regarding the role of the cerebellum (such as coordination, studied by Flour-ens; comparators, studied by Holmes; gain controllers; associative learning). The most complete class of theories, these authors suggested, comprises theories in which the cerebellum forms an internal model of the motor system; such theories can encompass all the alternative theories, while fitting many of the known facts of cerebellar organization. These theories require that the cerebellum be an adaptive system capable of learning and of updating a model as the behavior of the motor system changes. They also require that the cerebellum store relevant parameters of the motor system, as these parameters form part of the description of the motor system behavior. Another requirement is timing capabilities: the motor system is dynamic, so a useful model will also need dynamic (i.e., time-dependent) behavior. How might internal models be acquired in the cerebellum through motor learning?

Marr-Albus Model and Synaptic Plasticity

Purkinje cells are the only output neurons from the cerebellar cortex. They receive two major synaptic inputs, from parallel fibers and from climbing fibers (Figure 1). Waveforms of neuronal spikes generated by the two kinds of synaptic inputs are different and can be discriminated even on extracellular recordings: simple spikes are triggered by parallel fibers and complex spikes are triggered by climbing fibers. Modularity is the basic design principle in the cerebellum. In the spinocerebellum, somatotopic fractured maps have been identified for parallel fiber inputs to the cerebellar cortex. Furthermore, microzones as long as several centimeters along the longitudinal axis of the cerebellar folia and as wide as 0.2 mm have been identified for climbing fiber inputs (Ito, 1984).

Marr (1969) and Albus (1971) proposed a detailed model of the cerebellum, according to which the cerebellum can form associative memories between particular patterns on parallel fiber inputs and Purkinje cell outputs. The basic idea is that the parallel fiber–Purkinje cell synapses can be modified by input from the climbing

Figure 1. Schematic diagram of a neural circuit for voluntary movement learning control by a cerebrocerebellar communication loop (see Ito, 1984, for details). Although the lateral part of the cerebellum hemisphere is shown and the input and output of other cerebellar regions are vastly different, the neural circuit of the cerebellar cortex is rather the same and uniform. It must be emphasized that only one climbing fiber makes contact with a single Purkinje cell, whereas parallel fibers make 200,000 synapses on a single Purkinje cell. The number of granule cells, the origin of parallel fibers, is about 10^{11} . CF, climbing fiber; BC, basket cell; GO, Golgi cell; GR, granule cell; MF, Mossy fiber; PC, Purkinje cell; PF, parallel fiber; ST, stellate cell; DE, dentate nucleus; IO, inferior olivary nucleus; PN, pontine nuclei; RNp, parvocellular red nucleus; VL, ventrolateral nucleus of the thalamus.



fibers. In the perceptron models, the efficacy of a parallel fiber–Purkinje cell synapse is assumed to change when there exists a parallel fiber and climbing fiber input conjunction. The presence of the putative heterosynaptic plasticity of Purkinje cells was demonstrated as a long-term depression (LTD) (Ito, 2001).

The original Marr-Albus model did not take into account the dynamic and temporal characteristics of sensorimotor integration and was inappropriate in proposing simple associative memories for motor control problems. More satisfactory dynamic modeling in cerebellar learning was started by Fujita (1982). An associative LTD found in Purkinje cells can be modeled as the following heterosynaptic plasticity rule (Fujita, 1982): the rate of change of the synaptic efficacy of a single parallel fiber synapse is proportional to the negative product of the firing rate of that synapse's input and

the increment of the climbing fiber firing rate from its spontaneous level:

$$\tau dw_i/dt = -x_i(F - F_{\text{spont}}) \quad (1)$$

where τ is the time constant, w_i is the synaptic weight of the i th parallel fiber–Purkinje cell synapse, x_i is the firing frequency of the i th parallel fiber–Purkinje cell synapse, F is the firing frequency of the climbing fiber input, and F_{spont} is its spontaneous level. This single rule reproduces both the LTD and the long-term potentiation (LTP) found in Purkinje cells. When the climbing fiber and the parallel fiber are stimulated simultaneously, the parallel fiber synaptic efficacy decreases. In contrast, the parallel fiber synaptic efficacy increases when only the parallel fiber is stimulated (that is, when the climbing fiber firing frequency is lower than its sponta-

neous level). From a computational viewpoint, if the Purkinje cell output is the linear weighted summation of the parallel fiber inputs, Equation 1 can be regarded as the steepest descent of the error function defined as the square of the second factor. That is, this equation could provide a supervised learning rule if and only if the climbing fiber firing rate encodes an error signal.

Although early-day cerebellar models (Marr, 1969; Albus, 1971) were epoch-making in proposing Purkinje cell plasticity as a basis of cerebellar learning at the hardware level, they were not satisfactory at the representational and computational theory levels. What is actually learned and stored in the cerebellum? What neural representations are used in inputs and outputs of the Purkinje cells? Recent models and experimental efforts point to answers entirely different from those suggested by the early-day models.

Models of Limb Motor Control in the Cerebellum

Boylls (1975) proposed that the spatiotemporal neural firing patterns formed by the excitatory loop due to the cerebellar reverberating circuit and the inhibitory loop via the Purkinje cells are computationally beneficial for the generation of rhythmic interlimb coordination patterns in locomotion. In Boylls's theory, the purpose of cerebellar computation is to create synergically meaningful excitation profiles on a cerebellar nucleus, whose profiles are subsequently transmitted via an "output nucleus" to spinal levels. Boylls's model was later extended by Houk and Barto (1992) to accommodate motor learning in the cerebellum as an adjustable pattern generator (APG) model of the cerebellum. Temporal patterns of movement are acquired through motor learning, based on the LTD of Purkinje cells in combination with the reverberating circuit. Artificial neural network models with recurrent connections that can learn and generate arm trajectories were the computational bases of their model. The learning scheme proposed is mathematically based on associative reward-penalty learning (see REINFORCEMENT LEARNING IN MOTOR CONTROL). One of its attractive features is that a temporal movement pattern is selected and generated, which is impossible with a simple internal forward or inverse model. Correspondingly, however, learning is more difficult.

Cerebellar Feedback-Error-Learning Model

Internal models can be largely classified into forward models and inverse models (see SENSORIMOTOR LEARNING). Forward models predict the sensory consequences of movements from an efference copy of issued motor commands. Inverse models compute necessary feedforward motor commands from desired movement information. Both kinds of internal models are assumed to be located in the cerebellum (Kawato, 1999). However, the evidence for the forward models is much more circumstantial than is the evidence for the inverse models (see MOTOR CONTROL, BIOLOGICAL AND THEORETICAL). Learning forward models are generally much easier than inverse models because actual sensory feedback can be utilized as a teaching signal in the supervised learning equation (Equation 1), except for the difficulty associated with the delay in sensory feedback. Miall et al. (1993) proposed that the cerebellum forms two types of forward internal models. One model is a forward model of the motor apparatus. The second is a forward model of the transport time delays in the control loop (due to receptor and effector delays, axonal conductance, and cognitive processing delays). The second model delays the copy of prediction made by the first model so that it can be compared in temporal registration with actual sensory feedback from the movement. The second model resolves the difficulty of a temporal mismatch between the sensory signal delayed by the feedback loop and the output calculated by forward internal models.

Acquiring an inverse dynamics model through motor learning is computationally difficult because the necessary teaching signal for the desired motor command, which is the output of the inverse dynamics model, is not available. Several computational learning schemes to resolve this difficulty have been proposed (see SENSORIMOTOR LEARNING). Kawato and colleagues (Kawato, 1999) proposed a cerebellar feedback-error-learning model (CBFELM; Figure 2), which turned out to be the most biologically plausible of the various proposals as a model of the cerebellum. In this model, simple spikes (SS) represent feedforward motor commands and the parallel fiber inputs represent the desired trajectory as well as the sensory feedback of the current status of the controlled object. A microzone of the cerebellar cortex constitutes (part of) an inverse model of a specific controlled object such as an eye or an arm. Most important climbing fiber inputs are assumed to carry a copy of the feedback motor commands generated by a crude feedback control circuit. Thus, the complex spikes (CS) of Purkinje cells activated by climbing fiber inputs are predicted to be sensory error signals already expressed in motor command coordinates. The supervised learning equation (Equation 1) allows an interpretation that a crude feedback controller could generate approximation to the necessary error signal in motor commands. Stability and convergence of the CBFELM have been proved mathematically in the recent control theory literature.

Experimental Supports for the Cerebellar Feedback-Error-Learning Model

The CBFELM model was directly supported by neurophysiological studies in the ventral paraflocculus (VPFL) of monkey cerebellum

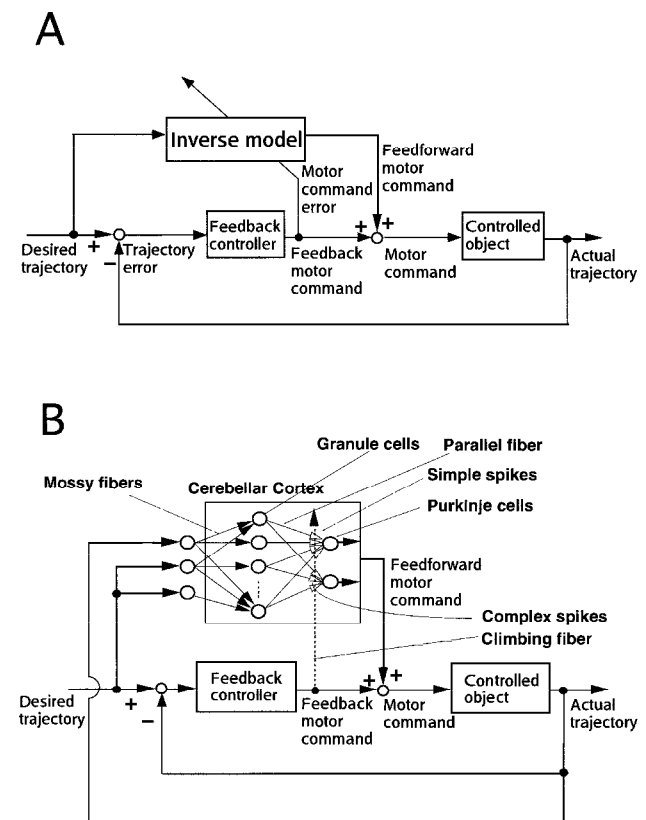


Figure 2. A, The general feedback-error-learning model. B, The cerebellar feedback-error-learning model (CBFELM) (Kawato, 1999). The "controlled object" is a physical entity that needs to be controlled by the CNS, such as the eyes, hands, legs, or torso.

during ocular following responses (OFRs) (Shidara et al., 1993; Kawato, 1999; Yamamoto et al., 2002). OFRs are tracking movements of the eyes evoked by movements in a visual scene and are thought to be important for the visual stabilization of gaze. The phylogenetically older, crude feedback circuit of the CBFELM is comprised of the retina, the accessory optic system (AOS), and the brainstem. The phylogenetically newer, more sophisticated feed-forward pathway and the inverse dynamics model of the CBFELM correspond to the cerebral and cerebellar cortical pathway and the cerebellum, respectively.

During OFRs, the temporal waveforms of SS firing frequency of VPFL Purkinje cells show complicated patterns. However, they (Figure 3, thin curve) were quite accurately reconstructed by using an inverse dynamics representation of the eye movement (Figure 3, thick curve; Shidara et al., 1993). The model fit was good for the majority of the neurons studied under a wide range of visual stimulus conditions. The same inverse dynamics analysis of firing frequency was applied to neurons in the area MST and dorsolateral pontine nucleus (DLPN), which provide visual mossy fiber inputs to the VPFL. In this area neural firing patterns were not well reconstructed. Taken together, these data suggest that the VPFL is the major site of the inverse dynamics model of the eye for OFRs.

The CBFELM model assumes that motor commands, which are conveyed by SS, are directly modified and acquired through synaptic plasticity by motor command errors, which are conveyed by climbing fiber inputs. For this to work, the motor commands and climbing fiber inputs must have comparable temporal and spatial characteristics, but the ultra-low discharge rates of the latter (1–2 spikes/s) would appear to rule this out. This apparently discrete and intermittent nature of climbing fiber inputs once suggested a reinforcement learning type of theory of cerebellar learning. However, if thousands of trials were averaged to compute firing frequencies of the climbing fiber inputs, the firing rates actually conveyed very accurate and reliable temporal waveforms of motor command error (Figure 3Cb). Because the LTD is a rather slow process of several tens of minutes of time constants, the averaging over many trials can actually be conducted by the LTD dynamics itself. Consequently, the firing probability of climbing fiber inputs aligned with the stimulus motion onset had high-frequency temporal dynamics matching those of the dynamic command signals. Thus, the most critical assumption of the CBFELM model was satisfied.

The preferred directions of MST and DLPN neurons were evenly distributed over 360 degrees. Thus, the visual coordinates for OFRs are uniformly distributed over all possible directions. On the other hand, the spatial coordinates of the extraocular muscles lie in either the horizontal or vertical directions, and are entirely different from the visual coordinates. The preferred directions of Purkinje cell SS were either downward or ipsilateral, and at the site of each recording, electrical stimulation of a Purkinje cell elicited eye movement toward the preferred direction of the SS of that Purkinje cell. This observation indicates that the SS coordinate framework is already that of the motor commands. Thus, at the parallel fiber–Purkinje cell synapse, a drastic visuomotor coordinate transformation occurs. Hence, the neural representation dramatically changes from population coding in MST and DLPN to firing rate coding of Purkinje cells at the parallel fiber–Purkinje cell synapse. What is the origin of this drastic transformation? According to the CBFELM model, the CS and eventually the AOS are the source of this motor command spatial framework. The preferred directions of pretectum neurons are upward, and those of nucleus of optic tract neurons are contralateral, and they are propagated to the inferior olive neurons and the CS of Purkinje cells. Yamamoto et al. (2002) reproduced all these experimental findings of Purkinje cell firing

characteristics during OFRs based on CBFELM, thus providing quite strong evidence for the theory.

Although direct and rigorous support of the CBFELM model was limited to a small portion of the cerebellum and to only several types of eye movements (OFR, VOR, OKR, smooth pursuit), because the neural circuit of different parts of the cerebellum is uniform and LTD is ubiquitous, we believe that the computational principle and the neural architecture demonstrated are common to all parts of the cerebellum. Recent physiological and brain imaging experiments provided further support of the CBFELM model in visually guided arm reaching movements (Kitazawa, Kimura, and Yin, 1998) and in the learning of a new tool (Imamizu et al., 2000).

Discussion

Humans can manipulate a vast number of tools, and exhibit an almost infinite number of behaviors in different environments. Given this multitude of contexts for sensorimotor control, there are two qualitatively distinct strategies to motor control and learning. The first is to use a single controller that uses all the contextual information in an attempt to produce an appropriate control signal. However, such a controller would have to be enormously complex to allow for all possible scenarios. If this controller were unable to encapsulate all the contexts, it would need to adapt every time the context of the movement changed before it could produce appropriate motor commands. This would produce transient but possibly large performance errors. Alternatively, a modular approach could be used in which multiple controllers coexist, with each controller suitable for one or a small set of contexts. Depending on the current context, only those appropriate controllers should be active to generate the motor command.

The modular approach has several computational advantages over the nonmodular approach. First, by using multiple inverse models, each of which might capture the motor commands necessary when interacting with a particular object or within a particular environment, we could achieve an efficient coding of the world. In other words, the large set of environmental conditions in which we are required to generate movement requires multiple behaviors or sets of motor commands, each embodied within a module. Second, the use of a modular system allows individual modules to adapt through motor learning, without affecting the motor behaviors already learned by other modules. Third, many situations that we encounter are derived from combinations of previously experienced contexts, such as novel conjoints of manipulated objects and environments. By modulating the contribution to the final motor command of the outputs of the inverse modules, an enormous repertoire of behaviors can be generated. With as few as 32 inverse models, in which the output of each model either contributes or does not contribute to the final motor command, we have 2^{32} behaviors—sufficient for a new behavior for every second of one's life. Therefore, multiple internal models can be regarded conceptually as *motor primitives*, which are the building blocks used to construct intricate motor behaviors with enormous vocabulary (see MOTOR PRIMITIVES).

Based on the benefits of a modular approach and the experimental evidence for modularity in observed behaviors, Wolpert and Kawato (1988) have proposed that the problem of motor learning and control is best solved using multiple controllers—that is, inverse models. At any given time, one or a subset of these inverse models will contribute to the final motor command (Figure 4 gives the details of the model). However, if there are multiple controllers, then there must also be some scheme to select the appropriate controller or controllers at each moment in time. The basic idea is that multiple inverse models exist to

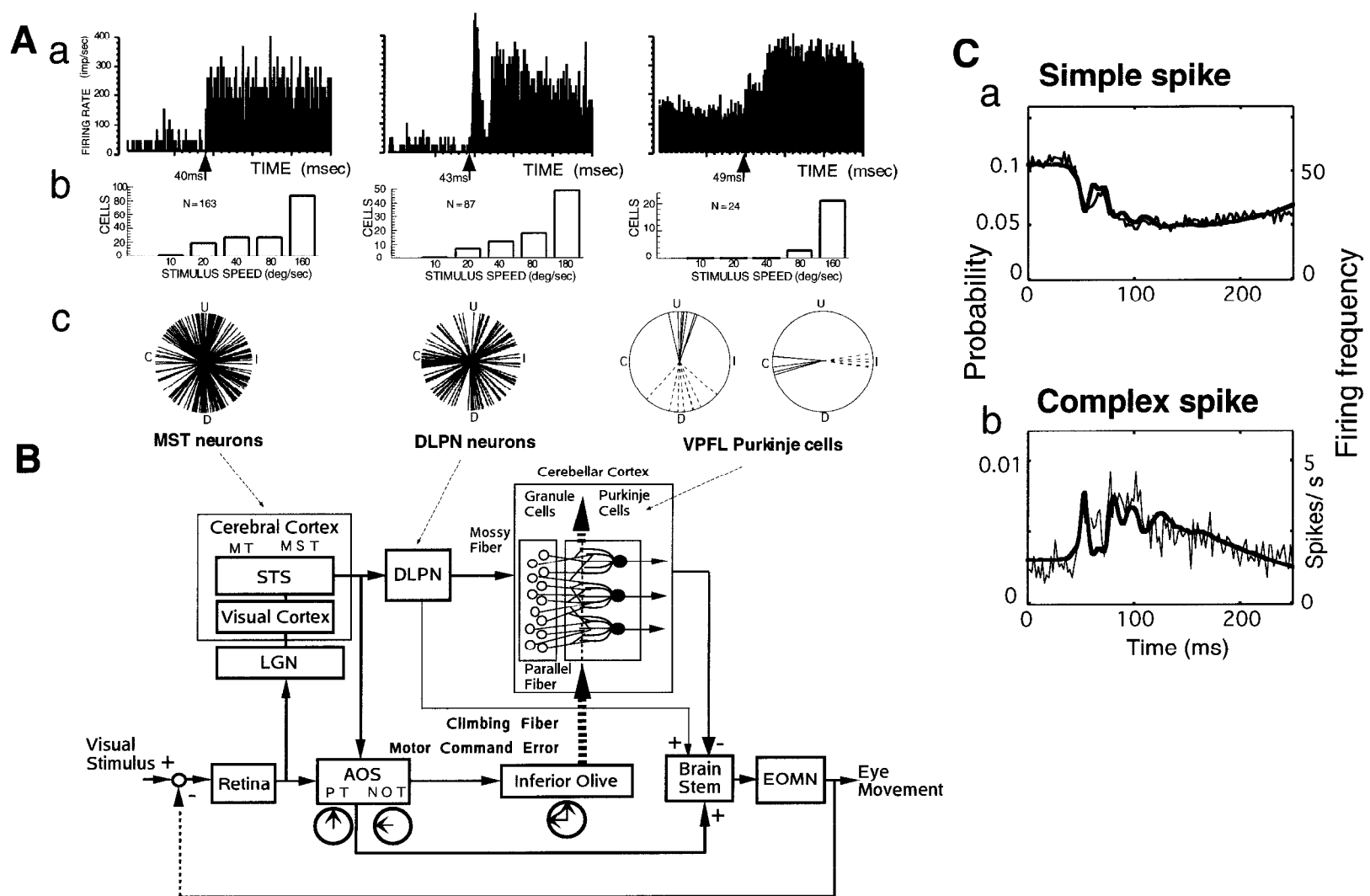


Figure 3. Change of neural codes and learning of inverse dynamics model in the cerebellum for ocular following responses (OFRs) (see Yamamoto et al., 2002, for a model reproduction of this experiment). **A**, The firing characteristics of MST, DLPN, and VPFL neurons. **B**, A schematic neural circuit for OFR. **C**, The temporal firing patterns of VPFL Purkinje cells in upward eye movements in response to upward visual motion. In **A**, the left, middle, and right columns are for MST, DLPN, and VPFL neurons. In **a**, post-stimulus-time histograms of the firing rates of a typical neuron in each of the three columns are shown. The origin of time is the onset of visual stimulus motion. In **b**, histograms of a number of cells within a given range of the optimum stimulus speeds are shown. In **c**, polar plots of optimum stimulus directions are shown. U, C, D, and I indicate upward, contralateral, downward, and ipsilateral, respectively. VPFL Purkinje cells were classified into two groups, vertical cells and horizontal cells, based on simple spike (dotted line) and complex spike (solid line) optimum directions. **B**, The circuit can be divided into two main pathways. The upper part shows the corticocerebellar pathway (from the cerebral cortex to the cerebellar cortex), which corresponds to the feedforward arc of the feedback-error-learning model. The lower part shows the phylogenetically older feedback pathway (from the accessory optic system, which corresponds to a crude feedback controller in the feedback-error-learning scheme). **C**, Temporal firing patterns of nine Purkinje cells accumulated (thin curves) and their reconstruction based on an inverse-dynamics model (bold curves). The model predicts that the temporal firing patterns of simple spikes (**a**) and complex spikes (**b**) should be mirror images of each other, and this was confirmed experimentally. MST, medial superior temporal area; DLPN, dorsolateral pontine nucleus; VPFL, ventral paraflocculus; AOS, accessory optic system; PT, pretectal nucleus of optic tract; MT, middle temporal area; STS, superior temporal sulcus; LGN, lateral geniculate nucleus; EOMN, extraocular motor neurons.

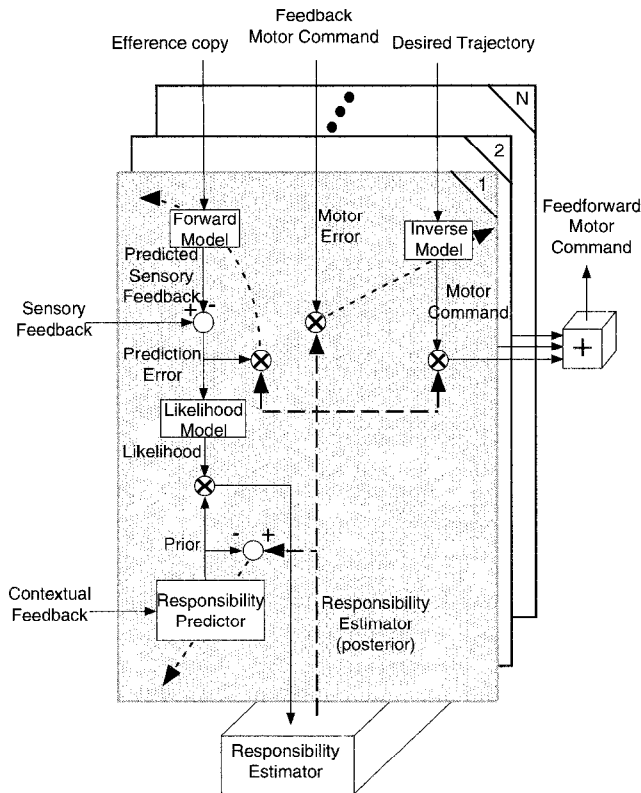


Figure 4. A schematic of the MOSAIC model (Wolpert and Kawato, 1998). N -paired modules are shown as stacked sheets (the dotted lines represent training signals and three-signal multiplication). The details of the first module are shown. Interactions between modules take place through the responsibility estimator. Each module consists of three interacting parts. The first two, the forward model and the responsibility predictor, are used to determine the responsibility of the module. This responsibility signal reflects the degree to which the module captures the current context and should, therefore, participate in control.

control the system, and each is augmented with a forward model that determines the responsibility each controller should assume during movement. This responsibility signal reflects, at any given time, the degree to which each pair of forward and inverse models should be responsible for controlling the current behavior. Within each module, the inverse and forward internal models are tightly coupled during their acquisition, through motor learning. This ensures that the forward models learn to divide up experience so that at least one forward model can predict the consequence of actions performed in any given context. By coupling the learning of the forward and inverse models, the inverse models learn to provide appropriate control commands in contexts in which their paired forward model produces accurate predictions. The model was once called the multiple paired forward and inverse models, but it was later renamed MOSAIC (MODular Selection And Identification Control). MOSAIC is a version of the mixture-of-experts architecture (see MODULAR AND HIERARCHICAL LEARNING SYSTEMS).

Recent human brain imaging studies have started to accumulate evidence supporting multiple internal models of tools in the cerebellum (Imamizu et al., 2000, and successive studies). Other imaging studies suggest forward models in the cerebellum as well as inverse models. Each modular internal model in MOSAIC could have good anatomical correspondence with micro-zones. MOSAIC is capable of learning to produce appropriate motor commands in a variety of contexts and can switch rapidly between controllers as the context changes. These features are important for a full model of motor control and motor learning, as it is clear that the human motor system is capable of very flexible, modular adaptation. Furthermore, MOSAIC has the potential to explain many human cognitive capabilities such as thinking, communication, and language. This point is intriguing, since the cerebellum was shown to be involved in these uniquely human cognitive activities.

Road Maps: Mammalian Brain Regions; Mammalian Motor Control; Neural Plasticity

Background: Motor Control, Biological and Theoretical

Related Reading: Cerebellum and Conditioning; Imaging the Grammatical Brain; Sensorimotor Learning

References

- Albus, J. S., 1971, A theory of cerebellar functions, *Math. Biosci.*, 10:25–61.
- Boylls, C. C., 1975, *A Theory of Cerebellar Function with Applications to Locomotion: I. The Physiological Role of Climbing Fiber Inputs in Anterior Lobe Operation*, COINS Technical Report, Amherst: University of Massachusetts, Computer and Information Science.
- Fujita, M., 1982, Adaptive filter model of the cerebellum, *Biol. Cybern.*, 45:195–206.
- Gomi, H., and Kawato, M., 1996, Equilibrium-point control hypothesis examined by measured arm stiffness during multi-joint movement, *Science*, 272:117–120.
- Houk, J. C., and Barto, A. G., 1992, Distributed sensorimotor learning, in *Tutorial in Motor Behavior II* (G. E. Stelmach and J. Requin, Eds.), Amsterdam: Elsevier, pp. 71–100.
- Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Puetz, B., Yoshioka, T., and Kawato, M., 2000, Human cerebellar activity reflecting an acquired internal model of a new tool, *Nature*, 403:192–195. ◆
- Ito, M., 1984, *The Cerebellum and Neural Control*, New York: Raven Press.
- Ito, M., 2001, Long-term depression: Characterization, signal transduction, and functional roles, *Physiol. Rev.*, 81:1143–1195. ◆
- Kawato, M., 1999, Internal models for motor control and trajectory planning, *Curr. Opin. Neurobiol.*, 9:718–727. ◆
- Kitazawa, S., Kimura, T., and Yin, P., 1998, Cerebellar complex spikes encode both destinations and errors in arm movements, *Nature*, 392:494–497.
- Marr, D., 1969, A theory of cerebellar cortex, *J. Physiol.*, 202:437–470.
- Miall, R. C., Weir, D. J., Wolpert, D. M., and Stein, J. F., 1993, Is the cerebellum a Smith predictor? *J. Motor Behav.*, 25:203–216.
- Shidara, M., Kawano, K., Gomi, H., and Kawato, M., 1993, Inverse dynamics model eye movement control by Purkinje cells in the cerebellum, *Nature*, 365:50–52.
- Wolpert, D., and Kawato, M., 1998, Multiple paired forward and inverse models for motor control, *Neural Netw.*, 11:1317–1329. ◆
- Yamamoto, K., Kobayashi, Y., Takemura, A., Kawano, K., and Kawato, M., 2002, Computational studies on acquisition and adaptation of ocular following responses based on cerebellar synaptic plasticity, *J. Neurophysiol.*, 87:1554–1571. ◆

Cerebellum: Neural Plasticity

Hervé Daniel and Francis Crepel

Introduction

The participation of the cerebellum in motor learning was postulated by Brindley as early as 1964 and then formalized by Marr and Albus (see *CEREBELLUM AND MOTOR CONTROL* for details) around 1970. Purkinje cells (PCs) are the only output neurons of the cerebellar cortex. Each PC receives two excitatory synaptic inputs displaying distinct characteristics. The first and most powerful one-to-one synaptic input corresponds to climbing fibers (CFs) that originate from neurons in the contralateral inferior olive and make multiple synapses on primary and secondary dendrites of PCs. The second, weaker excitatory input corresponds to parallel fibers (PFs) that originate from cerebellar granule cells; 80,000 PFs converge onto tertiary dendrites of each PC, where each PF makes few synapses. PFs and CFs are likely to use the excitatory amino acid glutamate (Glu) as a neurotransmitter.

Given the dual arrangement of these excitatory synaptic inputs and taking into account current models of memory, it has been proposed that during motor learning, the gain of synaptic transmission at PF-PC synapses changes if and only if these synapses are repetitively activated at low rates in conjunction with CF impinging on the same PC. In this scheme, the CF acts as an external teacher to instruct PF-PC synapses to change their gain to adapt the cerebellar cortex output to the motor command (see *CEREBELLUM AND MOTOR CONTROL*). Thus, Marr's theory of motor learning predicts that repeated coincident activation of CFs and PFs leads to a long-term potentiation (LTP) of the PF synaptic inputs. However, Albus proposed instead that a long-term depression (LTD) occurs rather than LTP at PF-PC synapses during motor learning, to avoid saturation of neuronal networks.

The first experimental support in favor of this theory was provided by Ito and co-workers in a series of *in vivo* experiments on rabbits (Ito, Sakurai, and Tongroach, 1982). They demonstrated that LTD is associative, since only concomitant stimulation of PFs at low frequency (1–4 Hz) and of the CF impinging on the same PC leads to LTD of synaptic transmission at PF-PC synapses. In contrast, activation either of the CF or of PFs alone has no long-term effect. In addition, this long-term change in synaptic strength is restricted to PF-PC synapses activated in conjunction with CFs and is thus considered to be input specific. This is a crucial point for deciphering the mechanisms involved in motor learning (Ito et al., 1982). With the experimental advantages of *in vitro* brain slices and culture preparations, more recent studies have elucidated, at least in part, the complex processes of glutamate receptor activation and subsequent second-messenger cascades controlling the induction and the expression of this form of synaptic plasticity. The present article presents a detailed review of the findings of many experimentalists. Since space does not allow a comprehensive bibliography, the reader is referred to Daniel, Levenes, and Crepel (1998) and Daniel et al. (1999) for a full bibliography and citation of these many researchers.

Glutamatergic Receptors Involved in LTD Induction

We first examine the components in Figure 1. In marked contrast to most other neurons in the brain, PCs in the mature brain do not bear functional *N*-methyl-D-aspartate (NMDA) ionotropic receptors (see *NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK MODELS*). As such, fast excitatory synaptic transmission at PF-PC synapses is entirely mediated by non-NMDA ionotropic receptors, mostly of the AMPA type. These synapses also possess

type-1 α metabotropic glutamatergic receptors (mGluR1) coupled to phospholipase C, the activation of which leads to the production of inositol 1,4,5-triphosphate (IP₃) and of diacylglycerol (DAG, a protein kinase C (PKC) activator). There is now wide agreement that LTD induction of PF-mediated synaptic responses in PCs in acute slices and of Glu-induced currents in cultured PCs requires the activation of these two receptor groups. In cultured PCs, evidence for AMPA receptor participation in LTD has been provided by Linden and co-workers (1991). They have shown that iontophoretic application of quisqualate (an agonist of both mGluRs and AMPA receptors) induces LTD of quisqualate-mediated excitatory currents when combined with PC depolarization sufficient to produce Ca²⁺ entry through voltage-gated Ca²⁺ channels (VGCCs). This LTD is blocked by bath application of CNQX (a selective antagonist of AMPA receptors) during the pairing protocol (Linden et al., 1991). Likewise, using acute slices, other workers demonstrated that application of CNQX during a classical pairing protocol (PC depolarization/PF stimulation) prevents induction of LTD of PF-EPSCs but not its expression.

Concerning mGluRs, it was shown that application of mGluR1-inactivating antibodies or of the mGluR antagonist MCPG completely blocks LTD induction, while LTD induced by a classical pairing protocol is significantly impaired in knockout (KO) mice lacking functional mGluR1s. Moreover, LTD in mGluR1-null mutant mice can be rescued either by bypassing the disrupted mGluR1s with direct pharmacological activation of downstream intracellular cascades (activation of IP₃ signal transduction pathway by photolytic release of IP₃) (Daniel et al., 1999) or by transfecting

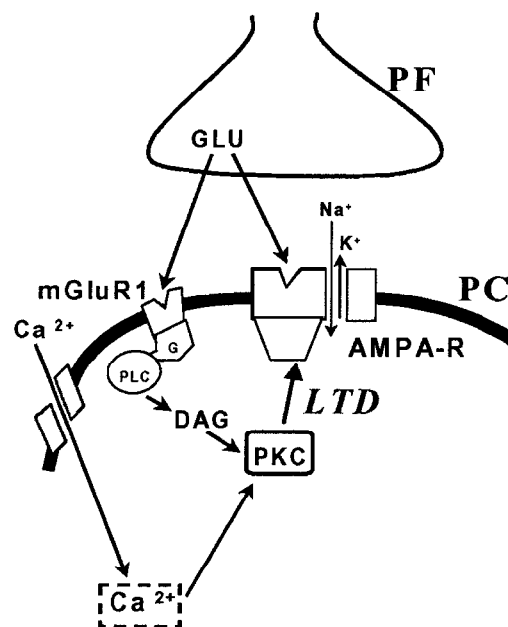


Figure 1. Schematic diagram of signal transduction processes involving the glutamatergic receptors proposed to participate in LTD induction: AMPA receptor (AMPA-R), and G (G)-protein-coupled-type 1-metabotropic glutamate receptor (mGluR1), which is positively coupled to phospholipase C (PLC) and can therefore lead to the production of 1,2-diacylglycerol (DAG), a protein kinase C (PKC) activator. Other abbreviations: GLU: Glutamate, PC: Purkinje cell, PF: parallel fiber.

functional mGluR1s in these KO mice under the control of the PC-specific L7 promoter (Ichise et al., 2000). Thus, these studies demonstrate the key role of mGluR1 in PCs for LTD induction, as they rule out the possibility that impairment of LTD in mGluR1 KO mice is due to some indirect developmental defects.

Involvement of Increases in Intracellular Ca^{2+} in LTD Induction

The crucial role of increases in Ca^{2+} resulting from VGCC activation for LTD induction was initially demonstrated in acute slices maintained in vitro with induction of LTD of PF-mediated synaptic responses by concomitant stimulation of PFs and CF impinging on the same PC prevented when PC intracellular free Ca^{2+} is buffered by 1,2-bis(2-aminophenoxy)ethane- N,N,N',N' -tetra-acetic acid (EGTA). The involvement of Ca^{2+} increases following VGCC activation in LTD induction was also evidenced in this same preparation, where LTD of PF-mediated responses was consistently induced by pairing these synaptic responses with direct depolarization of PCs sufficient to produce Ca^{2+} entry through VGCCs, thus mimicking activation of CFs. Likewise, as shown by Crepel and Krupa (1988) in acute slices and by Linden et al. (1991) in cultured PCs, LTD of responses elicited in PCs by iontophoretic application of Glu in their dendritic fields is also induced when these responses are combined with membrane depolarization, giving rise to Ca^{2+} spike firing. Finally, in patch-clamped PCs in acute slices, combining recordings of synaptic currents with fluorometric measurements of intracellular Ca^{2+} concentration directly demonstrates that PF stimulation paired with depolarization-induced Ca^{2+} transients is sufficient to induce LTD (see Figure 2).

In addition to Ca^{2+} entry through VGCCs, it was hypothesized that the cascade of events leading to LTD in PCs may involve Ca^{2+} release from IP3- or ryanodine-sensitive internal stores (Daniel et

al., 1998). However, in acute slices, the extent to which this process is involved seems to depend crucially on experimental conditions. Two groups have shown that both inhibition of Ca^{2+} release from IP3-sensitive stores with heparin (an antagonist of IP3 receptors) or with a specific antibody against IP3-receptor type 1 block LTD induced by a classical pairing protocol. Likewise, slices prepared from mice with a disrupted IP3-receptor type 1 gene, which is predominantly expressed in PCs, lack LTD induced by the same pairing protocol. Finally, and as mentioned before, the impairment of LTD induced by a classical pairing protocol in mGluR1-deficient PCs is pharmacologically rescued by photolysis of a caged IP3-compound. Along the same line, LTD can be rescued in transgenic mice with defective IP3-mediated Ca^{2+} signaling in spines, by local photolysis of a caged Ca^{2+} compound. Taken together, these data seem to establish that Ca^{2+} release from internal stores are critical for LTD induction in slices, at least in certain experimental conditions, but they do not reveal the precise mechanisms of their involvement in more physiological conditions. Indeed, synaptically driven Ca^{2+} mobilization from internal stores by PF stimulation has been now detected in PC spines and dendritic microdomains with confocal microscopy as well as with the use of two-photon laser scanning microscopy.

The problem of the participation of Ca^{2+} release from internal stores in LTD induction has been recently elucidated in an elegant series of experiments using two-photon laser scanning microscopy. In particular, it has been shown that with conjunctive activation of CF and of small number of PFs, both Ca^{2+} signals and LTD are confined to the activated synapses and require Ca^{2+} release from IP3-sensitive stores. In contrast, with co-activation of CF and of a large number of PFs, Ca^{2+} signals spread to dendritic shafts, and LTD induced in these conditions is now entirely mediated by VGCC-dependent Ca^{2+} entry. These results may reconcile previous conflicting findings regarding the involvement of Ca^{2+} release from internal stores in LTD induction in acute slices. Indeed, depletion of internal Ca^{2+} stores with thapsigargin blocks induction of LTD of PF-EPSPs induced by PC depolarization/bath application of mGluR agonist (1S-3R ACPD) without PF stimulation but does not prevent the induction of LTD by a classical pairing protocol (PC depolarization/PF stimulation).

In cultured PCs, the results are also puzzling, since certain studies have reported that inhibition of Ca^{2+} release from ryanodine or IP3-sensitive internal Ca^{2+} stores blocks LTD induced by Glu/depolarization conjunctive stimulation. In contrast, others have shown that LTD induction appears to be independent from Ca^{2+} release from IP3-sensitive stores. Thus, the potential involvement of Ca^{2+} release from internal stores in LTD induction in these reduced preparations is unclear, especially in the light of a recent study demonstrating that Ca^{2+} release from internal stores in PCs maintained in culture is impaired (Womack, Walker, and Khodakhah, 2000).

Second Messengers and LTD Induction: Cascades Involving PKC Activation and Nitric Oxide (NO)-cGMP-Dependent Protein Kinase (PKG) Activation

PKC Pathway

We now examine other aspects of the pathway shown in Figure 2. It is now well accepted that the second-messenger cascades following rises in internal Ca^{2+} involve PKC activation. Since PKC γ is abundantly expressed in PCs (Daniel et al., 1998) and mGluR1 activation also results in DAG production (see above), it was tempting to postulate that the cascade of events leading to LTD involves the activation of PKC by Ca^{2+} entry through VGCCs and DAG produced by mGluR activation following Glu release from PFs (Crepel and Krupa, 1988). Indeed, studies by Crepel and Krupa

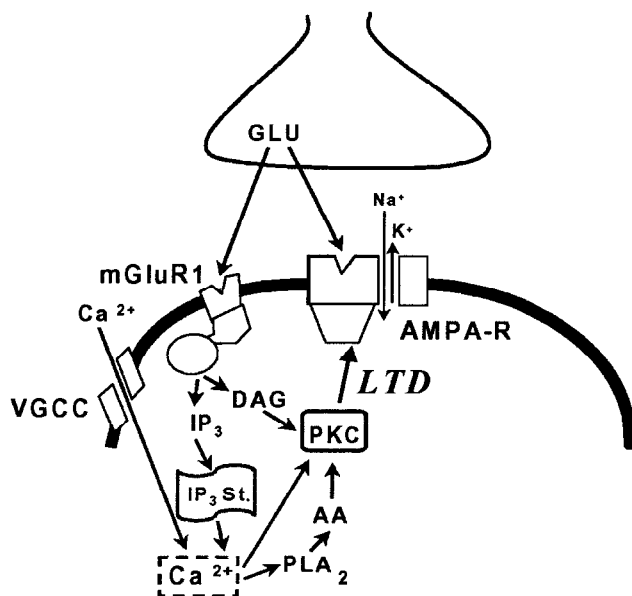


Figure 2. Schematic diagram of signal transduction processes proposed to participate in LTD induction involving increase in intracellular calcium, resulting from voltage-gated calcium channel (VGCC) activation and/or from calcium release from Inositol-1,4,5-trisphosphate (IP3)-internal stores (IP3 St.), and involving the full PKC activation which requires in addition to the intracellular cascade leading to DAG production, the activation of the calcium-dependent enzyme phospholipase A₂ (PLA₂) leading to arachidonic acid (AA) production. Other abbreviations as in Figure 1.

(1988) in acute slices and later studies in cultures showed that direct activation of PKC by phorbol esters induces LTD of the responsiveness of PCs to exogenously applied Glu or quisqualate, whereas inactive analogs are without effect. Moreover, studies in acute cerebellar slices have shown that the potent protein kinase inhibitor polymixin B or the selective PKC inhibitory peptide PKC[19-36] abolishes LTD induction by conventional pairing protocols (PC depolarization/PF stimulation). Likewise, PKC[19-36] also blocks LTD induction by Glu/depolarization conjunction in cultured PCs. Moreover, in this reduced preparation, full activation of PKC is required to induce LTD and involves activation of the calcium-dependent enzyme phospholipase A_2 (PLA $_2$) leading to arachidonic acid (AA) production, in addition to the intracellular cascade leading to DAG production. Such an involvement of PLA $_2$ in LTD induction was also evidenced more recently in acute slices by Reynolds and Hartell (2001). Finally, the crucial role of PKC activation in the cellular mechanisms leading to LTD induction was recently supported by the introduction of an additional player in the game, that is, the corticotropin-releasing factor that is contained in CFs and is critical for LTD induction in cerebellar slices, probably acting by enhancing activation of PKC.

However, and surprisingly enough, mice with null mutation in protein kinase $C\gamma$ exhibit apparently normal cerebellar LTD and this LTD is still abolished by specific PKC inhibitors. These puzzling results suggest that compensatory processes involving other subtypes of PKC are activated in these mutants to sustain LTD or, as was recently proposed, PKC α and/or PKC β_1 but not PKC γ are involved in LTD induction (Hirono et al., 2001). More straightforward results were obtained by using transgenic mice selectively expressing the pseudosubstrate PKC inhibitor, PKC[19-31] in PC, since no LTD could be produced in these cells either in cultures or (Goossens et al., 2001) in acute slices. Thus, altogether, there is now a wide agreement that induction of cerebellar LTD requires PKC activation in *in vitro* brain slices and in culture preparations.

NO-cGMP-PKG Pathway

Further mechanisms are summarized in Figure 3. Inhibition of post-synaptic protein phosphatase activity through a cascade involving nitric oxide (NO), cGMP production, and cGMP-dependent protein kinase (PKG) activation is also required for LTD induction. In neurons, increases in intracellular Ca^{2+} can induce the formation of NO from arginine, by activating calmodulin-dependent NO-synthase (NO-S). As early as 1990, it was shown, in acute slices, that LTD of PF-mediated EPSPs induced by a conventional pairing protocol is prevented by bath application of NO-S inhibitors and furthermore that this blocking effect is reversed by addition of an excess of arginine in the bath. In the same year, it was shown that, again in acute slices, bath application of NO-S inhibitor prevents enduring desensitization of AMPA receptors of PCs resulting from successive exposures of slices to quisqualate. Moreover, it was shown later on that an LTD-like phenomenon can be induced by bath application of NO donors or by dialyzing such NO donors in the recorded cell through the recording patch pipette. However, neuronal nitric oxide synthase (nNO-S) has not been identified in PCs, even following RT-PCR analysis of mRNAs directly harvested from these neurons during patch-clamp experiments. In contrast, nNO-S is highly expressed in neighboring elements such as PFs and basket cells. Further studies, with NO-sensitive electrodes inserted into the molecular layer of cerebellar slices, have shown that protocols that are known to induce LTD lead to NO release, most probably originating from PFs. Finally, it has been shown that photolytic release of NO and Ca^{2+} inside PCs could replace the conjunctive PF stimulation and depolarization, respectively, required for LTD induction, suggesting that NO release is sufficient to replace PF stimulation. If this finding underlines the potential

role of NO in LTD induction, it is somewhat puzzling, as it contradicts previous experiments demonstrating that activation of AMPA receptors and of mGluRs are necessary for LTD induction (see the previous section). This apparent discrepancy could be due to the use of the photolytic release of NO, which is a powerful pharmacological tool but may exaggerate the participation of this NO pathway in LTD induction, which otherwise might be much less prominent in physiological conditions. Nevertheless, taken together, data obtained in acute slices are entirely consistent with a role for NO in LTD induction in PCs.

The target for NO is likely to be the soluble enzyme guanylate cyclase (sGC), which is abundantly expressed in PCs. Indeed, it is now indirectly established (by measurements of cGMP-dependent activation of phosphodiesterases) that NO, either pharmacologically applied or endogenously released following electrical stimulation of the molecular layer, triggers cGMP production in Purkinje cells (Hartell et al., 2001). Numerous studies have examined the potential role of cGMP on LTD induction. A long-lasting depression of PF-mediated responses is induced by direct dialysis of cGMP into the recorded PCs through the recording patch-pipette or by bath application of 8-bromo-cGMP (a membrane-permeative cGMP analog). In addition, this cGMP-induced LTD-like phenomenon partially occludes subsequent induction of LTD by classical pairing protocols. Moreover, LTD induced by a classical pairing protocol is totally inhibited by intracellular injection of the selective and potent inhibitor of sGC, 1H-(1,2,4)oxadiazolo(4,3-a)quinoxalin-1-one (ODQ), confirming that the sGC of PCs is the target of NO.

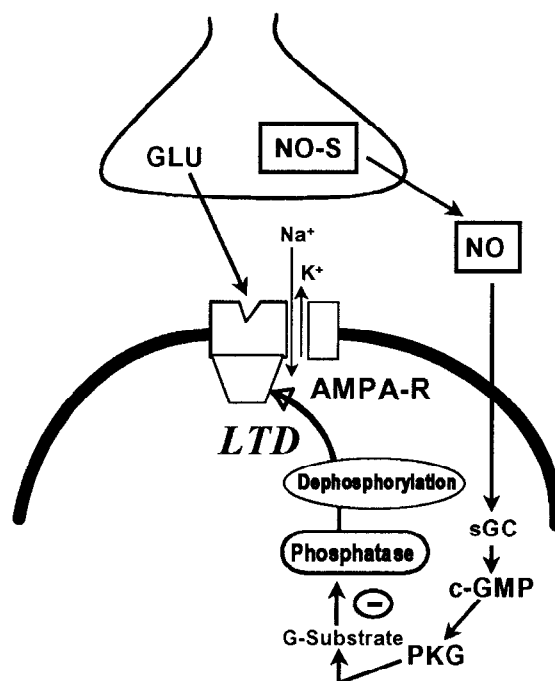


Figure 3. Schematic diagram of signal transduction processes involving the NO-cGMP-PKG pathway proposed to participate in LTD induction: Nitric oxide (NO), produced by NO-synthase (NO-S), activates synthesis of cyclic guanosine monophosphate (cGMP) by the catalytic action of the soluble guanylate cyclase (sGC); cGMP production in turn activates Protein kinase G (PKG), a cGMP-dependent protein kinase, thereby allowing phosphorylation of its specific endogenous "G-substrate," which is, in the phosphorylated state, a powerful inhibitor of protein phosphatase. Other abbreviations as in Figure 1.

In keeping with these findings, Ito and Karachot suggested in 1992 that cGMP production would in turn activate a cGMP-dependent protein kinase (PKG), thereby allowing phosphorylation of its specific endogenous G-substrate, which is, in the phosphorylated state, a powerful inhibitor of protein phosphatases. In keeping with such a hypothesis, it has been shown that selective inhibition of PKG blocks LTD induced by co-activation of PFs and CFs or LTD induced by PF stimulation paired with PC depolarization (Reynolds and Hartell, 2001). In addition, it has been reported that bath application of the protein phosphatase inhibitor calyculin A induces LTD in acute slices.

All these experiments support the view that the NO-cGMP-PKG pathway participates in cellular events leading to LTD. Because of its ability to diffuse over large distances, the role of NO in LTD cannot be to bring synapse specificity for this change in synaptic strength. Rather, this compound might allow recruitment of additional synapses into the pool of those exhibiting LTD following co-activation of PFs and CFs, thus increasing the signal-to-noise ratio of this process.

While the involvement of NO-cGMP signaling in LTD induction in cerebellar slices is now firmly established, the induction of LTD of glutamate-evoked currents in cultured PCs is unaffected by treatments that stimulate or inhibit this signaling pathway. This discrepancy might be due to the fact that in reduced preparations, putative NO donors such as PFs and basket cells are more scarcely represented. Alternatively, it is also conceivable that mechanisms of LTD induction are different in neurons grown in dissociated cultures versus those in acute brain slices from young adult animals.

Changes in the Functional Characteristics of Postsynaptic AMPA Receptors During LTD Expression

In early *in vivo* experiments, Ito and co-workers showed that co-activation of PCs by CF stimulation and iontophoretic Glu application into their dendritic fields induced a long-lasting decrease in their responsiveness to this agonist (Ito et al., 1982). This result led Ito to propose that induction of LTD might ultimately lead to a long-term desensitization of PC postsynaptic ionotropic Glu receptors. Consistent with this view, experimental evidence based on the use of the coefficient of variation (CV) applied to PF-EPSCs suggests that LTD of PF-EPSCs, induced either by a classical pairing protocol or by bath application of NO donors in acute slices, is entirely expressed at a postsynaptic level. Moreover, a true modification of Glu receptor properties during LTD expression has been also suggested in acute slices. Indeed, it has been shown that aniracetam, a compound that is known to markedly reduce desensitization of AMPA receptors, has a larger potentiating effect on PF-EPSCs after induction of LTD than before, which suggests that LTD involves a true desensitization of postsynaptic AMPA receptors at PF-PC synapse. These results support the view that this change in long-term synaptic efficacy involves a genuine change in the functional characteristics of postsynaptic AMPA receptors.

Redistribution of AMPA Receptors and LTD Expression

In addition to functional modifications of AMPA receptors, recent studies have shown that trafficking of these receptors could play a crucial role in the expression of LTD (see Xia et al., 2000). Indeed, LTD expression in cultured PCs is associated with a rapid endocytotic-dependent decrease in the number of GluR2-containing synaptic AMPA receptors. This internalization requires clathrin-complex-mediated endocytosis, since peptides that inter-

fere with this complex block LTD. Additionally, this process depends on the intracellular carboxy-terminal domain of AMPA receptor subunit GluR2(3), which has been identified as a likely candidate region to interact with proteins of the postsynaptic density, allowing their intracellular anchoring. These latest studies have made substantial advance toward a molecular understanding of LTD expression by establishing a potential link between the internalization of AMPA receptors and PKC activation (see above). Indeed, phosphorylation of Ser 880 in the C-terminal domain of GluR2(3) by pharmacological PKC activation is accompanied by a reduction in the binding affinity of this subunit to an anchoring protein of the postsynaptic density termed "glutamate receptor interacting protein" (GRIP) (Figure 4), thereby leading to significant disruption of postsynaptic GluR2 clusters (Matsuda et al., 2000). In addition, treatments that disrupt the interaction between the carboxy-terminal of GluR2(3) and PICK1 (another protein component of the postsynaptic density interacting with protein kinase C) (Figure 4) strongly attenuated both LTD induced by glutamate/depolarization pairing and a LTD-like phenomenon produced by exogenous application of PKC-activating phorbol ester (Xia et al., 2000). Taken together, these findings demonstrate that expression

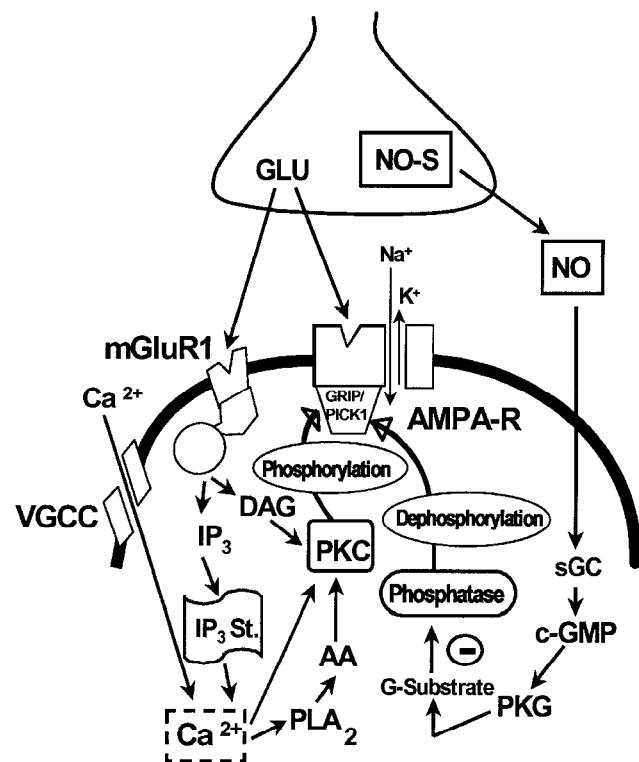


Figure 4. General schematic diagram of signal transduction processes leading to the phosphorylation and internalization of AMPA receptors, proposed to participate in LTD induction (glutamate receptor interacting protein (GRIP) and protein interacting with protein kinase C (PICK1) are proteins of the postsynaptic density interacting with AMPA receptors). Other abbreviations as in previous figures (Abbreviations: AA: arachidonic acid, AMPA-R: AMPA receptor, cGMP: cyclic guanosine monophosphate, DAG: 1,2-diacylglycerol, G: G protein, GLU: glutamate, GRIP: glutamate receptor interacting protein, IP₃: inositol-1,4,5-trisphosphate, IP₃ St.: inositol-1,4,5-trisphosphate internal Ca²⁺ stores, mGluR1: type-1 metabotropic glutamate receptor, NO: nitric oxide, NOS: nitric oxide synthase, PC: Purkinje cell, PF: parallel fiber, PICK1: protein interacting with C-kinase 1, PKC: protein kinase C, PKG: protein kinase G, PLA₂: phospholipase A₂, PLC: phospholipase C, sGC: soluble guanylate cyclase.)

of cerebellar LTD requires PKC-regulated interaction between the carboxy terminal of GluR2(3) and specific proteins of the postsynaptic density. Ultimately, internalization and/or changes in the properties of AMPA receptors could then create a disturbance in the postsynaptic density, which might in turn initiate a cascade of events that leads to a reorganization of the dendritic cellular cytoskeleton, thereby creating a permanent imprint of LTD.

Transcription Factors, Protein Synthesis, and LTD Expression

Recent findings have demonstrated that the mechanisms involved in the maintenance phase of LTD (which may underlie long-term memory) include the expression of transcription factors and require protein synthesis in addition to modifications of AMPA receptor properties and/or changes in their distribution. The experimental protocols that are known to induce LTD trigger expression of the immediate-early genes *c-Fos* and *Jun-B*, suggesting that the expression of these genes may help to establish cerebellar long-term depression. Moreover, translational inhibitors that were perfused immediately after the induction protocol abolished LTD of glutamate responsiveness in cultured PCs only after a delay of 45 minutes, suggesting that postsynaptic protein synthesis is specifically involved in the establishment of the late phase of LTD. However, Karachot et al. (2001) have reported more recently that translational inhibitors entirely abolish the LTD induced by conjunctive stimulation of PFs and CF in acute slices, including its early phase, even if this early phase is less sensitive to these inhibitors.

Discussion

After two decades of research, the understanding of the cellular and molecular mechanisms pathways underlying LTD has made rapid progress. Although part of the signal transduction pathways remains obscure, the stable phosphorylation of AMPA receptors appears to play a key step for expressing LTD (Figure 4). Thus, taking into account the experimental data available so far, Kuroda and co-workers have recently established a very interesting computational model that links kinetics of phosphorylation of AMPA receptors with different phases of LTD expression (Kuroda, Schweighofer, and Kawato, 2001).

Despite remaining uncertainties in the current understanding of the cellular and molecular mechanisms underlying the induction and expression of LTD, this associative and input specific form of synaptic plasticity has been proposed as the cellular basis for error-driven learning and memory in the motor system (see MOTOR CONTROL, BIOLOGICAL AND THEORETICAL). In particular, recent in vivo experiments point toward a role of LTD in the adaptation of the vestibulo-ocular reflex (VOR). This conclusion has been strengthened by the observations that selective expression of PKC inhibitors in PCs not only blocks cerebellar LTD without affecting PC excitability in alert transgenic mice, but also is accompanied by a lack of VOR adaptation in these animals (Goossens et al., 2001). Such a correlation between impairment of LTD and deficiency of associative learning of eyeblink conditioning has also been consistently found in other types of transgenic mice, such as mice that are deficient in mGluR1 (Aiba et al., 1994, in Daniel et al., 1998) or GFAP (Shibuki et al., 1996, in Daniel et al., 1998). Thus, there is now a growing consensus to support the notion that the associ-

ative nature of LTD makes it an attractive candidate for the cellular substrate underlying motor learning in the cerebellum. Moreover, the timing conditions for LTD induction may account for the temporal specificity of cerebellar motor learning. In this respect, an important future development in the field will be to study developmental aspects of LTD in relation to acquisition of motor skills.

Road Map: Neural Plasticity

Related Reading: Cerebellum and Conditioning; Cerebellum and Motor Control; NMDA Receptors: Synaptic, Cellular, and Network Models

References

- Crepel, F., and Krupa, M., 1988, Activation of protein kinase C induces a long-term depression of glutamate sensitivity of cerebellar Purkinje cells: An in vitro study, *Brain Res.*, 458:397–401. ♦
- Daniel, H., Levenes, C., and Crepel, F., 1998, Cellular mechanisms of cerebellar LTD, *Trends Neurosci.*, 21:401–407.
- Daniel, H., Levenes, C., Fagni, L., Conquet, F., Bockaert, J., and Crepel, F., 1999, Inositol-1,4,5-trisphosphate-mediated rescue of cerebellar long-term depression in subtype 1 metabotropic glutamate receptor mutant mouse, *Neuroscience*, 92:1–6. ♦
- Goossens, J., Daniel, H., Rancillac, A., van der Steen, J., Oberdick, J., Crepel, F., De Zeeuw, C. I., and Frens, M. A., 2001, Expression of protein kinase C inhibitor blocks cerebellar long-term depression without affecting Purkinje cell excitability in alert mice, *J. Neurosci.*, 21:5813–5823.
- Hartell, N. A., Furuya, S., Jacoby, S., and Okada, D., 2001, Intercellular action of nitric oxide increases cGMP in cerebellar Purkinje cells, *NeuroReport*, 12:25–28.
- Hirono, M., Sugiyama, T., Kishimoto, Y., Sakai, I., Miyazawa, T., Kishio, M., Inoue, H., Nakao, K., Ikeda, M., Kawahara, S., Kirino, Y., Katsuki, M., Horie, H., Ishikawa, Y., and Yoshioka, T., 2001, Phospholipase C β 4 and protein kinase C α and/or protein kinase C β 1 are involved in the induction of long term depression in cerebellar Purkinje cells, *J. Biol. Chem.*, 276:45236–45242.
- Ichise, T., Kano, M., Hashimoto, K., Yanagihara, D., Nakao, K., Shigemoto, R., Katsuki, M., and Aiba, A., 2000, mGluR1 in cerebellar Purkinje cells essential for long-term depression, synapse elimination, and motor coordination, *Science*, 288:1832–1835.
- Ito, M., Sakurai, M., and Tongroach, P., 1982, Climbing fiber induced depression of both mossy fiber responsiveness and glutamate sensitivity of cerebellar Purkinje cells, *J. Physiol.*, 324:113–134. ♦
- Karachot, L., Shirai, Y., Vigot, R., Yamamori, T., and Ito, M., 2001, Induction of long-term depression in cerebellar Purkinje cells requires a rapidly turned over protein, *J. Neurophysiol.*, 86:280–289.
- Kuroda, S., Schweighofer, N., and Kawato, M., 2001, Exploration of signal transduction pathways in cerebellar long-term depression by kinetic simulation, *J. Neurosci.*, 21:5693–5702.
- Linden, D. J., Dickinson, M. H., Smeyne, M., and Connor, J. A., 1991, A long-term depression of AMPA currents in cultured cerebellar Purkinje neurons, *Neuron*, 7:81–89. ♦
- Matsuda, S., Launey, T., Mikawa, S., and Hirai, H., 2000, Disruption of AMPA receptor GluR2 clusters following long-term depression induction in cerebellar Purkinje neurons, *EMBO J.*, 19:2765–2774.
- Reynolds, T., and Hartell, N. A., 2001, Roles for nitric oxide and arachidonic acid in the induction of heterosynaptic cerebellar LTD, *NeuroReport*, 12:133–136.
- Womack, M. D., Walker, J. W., and Khodakhah, K., 2000, Impaired calcium release in cerebellar Purkinje neurons maintained in culture, *J. Gen. Physiol.*, 115:339–346.
- Xia, J., Chung, H. J., Wihler, C., Haganir, R. L., and Linden, D. J., 2000, Cerebellar long-term depression requires PKC-regulated interactions between GluR2/3 and PDZ domain-containing proteins, *Neuron*, 28:499–510.

Chains of Oscillators in Motor and Sensory Systems

Nancy Kopell and G. Bard Ermentrout

Introduction

Collections of oscillators that send signals to one another can phase-lock with many patterns of phase differences. This article discusses a set of examples that illustrate how those phases emerge from the oscillator interactions. Much of the work was motivated by spatiotemporal patterns in networks of neurons that govern undulatory locomotion. The original experimental preparation to which this work was applied is the lamprey central pattern generator (CPG) for locomotion (see SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS). However, the mathematics is considerably more general, and can be used to gain insight into other systems. In addition to the lamprey CPG, we will briefly discuss related pattern generators in the crayfish swimmeret system (Skinner, Kopell, and Mulloney, 1997) and the leech network of swimming (Friesen and Pearce, 1993), as well as waves of activity in other oscillatory neural tissue (Kleinfeld et al., 1994). For relationships to other patterns in the nervous system, see SYNCHRONIZATION, BINDING AND EXPECTANCY; OLFACTORY BULB; and OLFACTORY CORTIX.

Though much has been written about chains of oscillators in the context of physical problems, very little of that literature is relevant to chains of biological oscillators. Much of the physics literature depends on the existence of certain conserved quantities that are irrelevant in a biological setting. Biological oscillators have other structure that is important to the behavior of collections of such oscillators. For example, unlike models of mechanical oscillators, models of biological oscillators have stable limit cycles (see PHASE-PLANE ANALYSIS OF NEURAL NETS). In the context of locomotion, each "oscillator" is likely to be a local subnetwork of neurons that produces rhythmic patterns of membrane potentials. Since the details of the oscillators often are not known and difficult to obtain, the object of the mathematics is to find the consequences of what is known, and to generate sharper questions to motivate further experimentation.

Physical Models and Phase Models

Each of the examples we describe below can be considered, roughly, as an array of oscillators. In the absence of knowledge of details, it is desirable to keep the oscillator description as general as possible. Thus, we assume only that the local network can be described by some (possibly high-dimensional) system of ordinary differential equations having a stable limit cycle, with no concern for the mechanistic origin of that oscillation.

In general, the behavior of interacting oscillators can be arbitrarily complex (Guckenheimer and Holmes, 1983, chap. 6). However, in some circumstances, the network behaves like a much more simply described collection of units. For example, suppose that the coupling is (sufficiently) weak. Then the collection of oscillators acts as if there were a well-defined "phase" to each oscillator, and the signals between the oscillators become dependent only on the differences of the phases (Kopell, 1987). The form of the equations is then

$$\dot{\theta}_j = \omega_j + \sum H_{jk}(\theta_k - \theta_j) \quad (1)$$

Here θ_j is the phase of the j th oscillator and ω_j is its frequency when uncoupled. The interaction functions $H_{jk}(\phi)$ measure how much the j th oscillator is sped up or slowed down by the interaction with the k th oscillator; they can be computed from the original, more complicated equations by averaging the effects of the cou-

pling terms over a cycle, and depend on the properties of the oscillators as well as the coupling (Guckenheimer and Holmes, 1983, chap. 4). There are also other computational ways of producing (at least the relevant parts of) the coupling functions, including a method that can in principle be done in the absence of explicit knowledge of the equations (see SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS).

The existence of this reduction procedure allows us to come back to questions about how details of the oscillators or the coupling can affect the network behavior via their effects on the functions H_{jk} . The simple description in Equation 1 can also be valid in some circumstances relevant to CPGs in which the coupling is strong. For example, if each oscillator is itself a composite of cells and emits coupling signals several times per cycle, the system can behave like the averaged one (see Cohen et al., 1992, for reference).

Mechanisms for the Production of Traveling Waves

We now consider arrays of oscillators, of which nearest-neighbor coupled chains are the simplest example. When the above reduction is valid, the equations have the form

$$\dot{\theta}_j = \omega_j + H_A(\theta_{j+1} - \theta_j) + H_D(\theta_{j-1} - \theta_j) \quad (2)$$

Here H_A and H_D are the functions that represent the coupling in the ascending and descending directions of the chain. There are at least two different mechanisms that can produce waves in Equation 2. One of these relies on differences in natural frequency along the chain and the other on properties of the coupling.

The first mechanism is easy to illustrate, using the simple choice of coupling functions $H_A(\phi) = H_D(\phi) = \sin(\phi)$, where ϕ denotes the relevant phase difference. Such functions arise from the reduction procedure when the differential equations producing the oscillations have, for example, odd symmetry and the interaction is via the standard mathematical description of diffusion between compartments (Ermentrout and Kopell, 1984). In this case, if the natural frequencies are all equal, the synchronous solution ($\theta_j = \theta_k$ for all j, k) is the stable output of the chain. If the frequencies are not all the same, however, other behavior is produced. For example, a gradient in frequencies produces a traveling wave of activity from the oscillator with the highest frequency to the one with the lowest (Cohen, Holmes, and Rand, 1982), but not one with constant phase lags (i.e., independent of position along the chain).

Waves induced by frequency differences are important in the electrical activity produced by smooth muscle of the intestines, where there is a linear gradient in the uncoupled frequencies (Ermentrout and Kopell, 1984). They are also known to exist in the leech CPG for swimming (Friesen and Pearce, 1993) and appear to be the mechanism underlying waves in *Limax* (Ermentrout, Flores, and Gelperin, 1998). There is no evidence for such a gradient in the lamprey or crustacean swimmeret system.

A variation on the frequency gradient idea has an oscillator at one or both ends of the chain with a different frequency than the ones in the middle. Even with symmetric coupling such as $H_A(\phi) = H_D(\phi) = \sin(\phi)$, there can be waves of activity. In this case, the waves can have constant phase lags.

Another mechanism that produces traveling waves in chains of locally coupled oscillators relies on the properties of the coupling, allowing the frequencies to be identical. The essential property for H_A or H_D is that, if there is one-way coupling using either of these functions, the oscillators would lock with a non-zero phase difference. A simple example of such a coupling function is $H_A(\phi) =$

$\alpha \sin \phi + \beta \cos \phi$, with $\beta \neq 0$, while $H_B(\phi) = 0$. We note that for almost all kinds of coupling between two oscillators (in particular for models of chemical synapses), $H(0) \neq 0$. An important special exception is the mathematical description of simple diffusion across a membrane, which is a standard model of electrical synapses.

To understand how this mechanism works, consider a chain of oscillators coupled only in one direction, e.g., the ascending direction. Then the equations take the form

$$\begin{aligned}\theta_j' &= \omega + H_A(\phi_j); j \neq N \\ \theta_N' &= \omega\end{aligned}\quad (3)$$

where $\phi_j = \theta_{j+1} - \theta_j$. The hypothesis about H_A implies that $H_A(0) \neq 0$, i.e., that the zero ϕ_0 of H_A satisfies $\phi_0 \neq 0$. Since all equations with $j \neq N$ are the same, we get equal phase lags $\phi_j = \phi_0$. This corresponds to a wave traveling at constant speed determined by ϕ_0 . The direction of the wave depends on the sign of ϕ_0 ; if $\phi_0 > 0$, the wave travels in the ascending direction; if $\phi_0 < 0$, the wave travels in the opposite direction. Thus, the direction of the coupling does not determine the direction of the wave.

If the coupling is in both directions, and/or if the frequencies are not uniform, the outcome can still be calculated (Kopell and Ermentrout, 1986). We will say more about this later in the context of particular applications. One of the applications (to *Limax*) uses a generalization of the ideas of coupling-induced waves. In that example, there is a gradient of coupling strength, so the equations have the form

$$\theta_j' = \omega_j + s_j H_A(\theta_{j+1} - \theta_j) + s_j H_D(\theta_{j-1} - \theta_j) \quad (4)$$

Here s_j denotes the strength of the connections to the j th cell; a gradient in coupling strength corresponds to the s_j changing monotonically with j .

The Lamprey Central Pattern Generator for Swimming

This CPG is described in more detail in SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS (q.v.). The most striking aspect of the lamprey CPG for undulatory locomotion is the linear geometry of the network. That is, the network can be portrayed crudely as a chain of oscillators (Cohen et al., 1982). The linear geometry provides important constraints on the behavior of the network, making it possible to use the theory described above. It is also important that the number of oscillators associated with the circuit be fairly large: there are approximately 100 segments, and small numbers (at least two) of isolated segments are known to be able to oscillate. The general theory in (Kopell and Ermentrout, 1986) is addressed to understanding the behavior of such long chains of oscillators; as we will see in the section on crayfish, CPGs that correspond to small numbers of oscillators must have a different construction in order to have a similar behavior. Although long fibers are known to exist in the lamprey cord, we consider first only local coupling; this was remarkably successful in explaining the behavior of the spinal cord. For references to the long fibers, see Cohen et al. (1982).

The first issue to be addressed about this preparation is the mechanism for producing the traveling wave that is observed in the spinal cord (see SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS). This wave has a constant speed (i.e., the phase lag between adjacent oscillators is independent of position along the chain). For the large class of equations for which the representation in Equation 2 is valid, the general theory of Kopell and Ermentrout (1986) tells us what properties of the oscillators and coupling are necessary to produce this constant speed wave.

In general, if there are frequency differences among the oscillators, the wave speed is not constant along the chain. (This can be

mitigated by creating coupling that connects each oscillator to multiple neighbors; see Cohen et al., 1992.) Thus, frequency gradients do not provide the basis for waves in the lamprey. Chains with differences in frequency for the first and last oscillators do produce constant speed waves. However, small sections from any portion of the isolated lamprey cord produce local phase differences identical to the differences produced in the whole cord, invalidating the hypothesis that the ends of the cord are intrinsically different from the rest.

The edges of a piece of cord *are* different, not because of their intrinsic properties but because of their placement, and this can produce waves driven by edge effects. We saw in the previous section that one-way coupling provides constant speed waves if the uncoupled frequencies of the oscillators are equal; in that mechanism, the relevant edge is the oscillator that gets no input.

The theory of Kopell and Ermentrout (1986) proves the surprising fact that, even with two-way coupling, a long but finite chain in general behaves like one with only one-way coupling. (For this, the coupling must be asymmetric.) That is, for the purpose of determining the phase lags between any two points on the chain, one of the two directions of coupling is dominant over the other; the output is almost as if the other coupling were silent. The phase lags among the oscillators are affected by the nondominant coupling only near one end of the chain, where there is a "boundary layer" of phase lags different from those in the rest of the chain. (This occurs only for long chains; when the chain is short, as in the swimmeret system discussed later, the boundary layer can be most of the chain.) This dominance can be understood from a linearized version of the equations (see SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS); however, the linearized equations do not reproduce other effects elicited by experimental perturbations. An important result of the theory is that the lags depend only on the coupling functions; changes in the frequency uniformly along the chain do not change the expression of phase lags.

Three sets of experiments have been done with the isolated lamprey spinal cord to test the ideas described above. Theory based on Kopell and Ermentrout (1986) and other mathematical work was used in each case to predict an effect that should be seen only if one direction of the coupling was dominant, and to determine the dominant direction. In one of these experiments, a small motor was attached to one end of an otherwise pinned-down cord and used to wag the cord periodically at measured frequencies; the predicted effect concerned the range of frequencies over which the cord could be entrained by the forcing at either end of the cord. In another set of experiments, the local frequency of the intrinsic circuits was manipulated to be different in different parts of the cord; the mathematics makes predictions about changes in phase lags due to the frequency perturbations. The third set of experiments concerned the existence of "boundary layers" in phase lags at one end of the chain. In each case, the predicted effect was found, and the data revealed that the dominant direction of coupling is caudal to rostral (ascending), though the wave itself is descending.

For more information on the the experiments and interpretation, see SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS and Cohen et al. (1992) and the references therein; the Cohen article has the references to the mathematical work. We note that if the edge effects needed to create the waves are produced by changing the frequencies of the first and last oscillators in a chain that would otherwise display synchrony, the mathematics says that the consequences of the experimental perturbation would not be as observed.

The central question that motivated the lamprey work concerned the constancy of the phase lags as swimming speed is increased. In vivo, the swimming speed of an animal is proportional to the frequency of the local oscillators of the spinal cord. In the isolated spinal cord, the frequency can be manipulated directly; changing the oscillatory frequency uniformly does not change the phase lags

along the cord. The theory described above provides a first step toward explaining this constancy; it says that changing the frequencies uniformly along the cord should not change the phase lags, provided that the zero of the dominant coupling does not change. In the lamprey spinal cord, changes of frequency may be achieved in ways that also change the coupling. Thus, the theory raises the sharper question of how the local network is constructed so that the excitation that increases the frequency leaves unchanged the relevant zero crossing.

Swimmeret Systems of Crayfish

The swimmeret system of crayfish is another example of modular control of motor behavior via oscillatory components. The swimmerets are four paired, jointed limbs in the abdomen of the animal that move rhythmically to propel the animal through the water. Each swimmeret is controlled by a local oscillatory neural circuit in a ganglion containing motor neurons that innervate the swimmeret and interneurons that produce the rhythm (Skinner and Mullen, 1998).

During a bout of swimming, the circuits all oscillate with the same frequency and maintain a constant phase relationship. The left and right sides are synchronized and there is approximately a quarter-cycle phase difference between neighboring circuits, with a posterior circuit leading its anterior neighbor. Thus, the chain of circuits produces a traveling wave whose direction is opposite to that of the lamprey. As in the lamprey, the wave produced has a wavelength approximately that of the body length, and the phase lag between adjacent segments is the same between any pair of segments. Other parallels with the lamprey swimming CPG are (1) there is no evidence for frequency differences between uncoupled local circuits, and (2) pharmacologically induced uniform changes in frequency of the circuits do not affect the phase lags.

In the swimmeret system, intersegmental coordination between the local circuits is accomplished by interneurons projecting to neighboring circuits. The details of connectivity and how the rhythms are produced are not known. Thus, as in the lamprey investigation, it is useful to consider the network behavior from a more abstract point of view, to understand what constraints on the bidirectional connectivity and local oscillators are necessary to reproduce key experimental observations. Based on recordings from isolated ganglia bathed in nicotinic agonists to produce swim-like patterns, some observations are as follows: (1) when anterior or posterior ganglia are selectively excited, a significant change in phase lag occurs at the excitation boundary, and (2) the frequency of the swimmeret pattern increases with an increase in the number of ganglia excited.

A model having the same form as Equation 2 can reproduce the basic findings about normal phases and perturbations in phase and frequency due to experimental perturbations (see Skinner and Mullen, 1998, for references). However, in order for this to be so, there are critical differences from the lamprey work in assumptions about the coupling. For the lamprey, the length of the chain of oscillators allows one direction of coupling to dominate; the other direction makes its presence known (at least with respect to the phase lags) only in a small boundary layer. For a short chain (four oscillators), both directions of coupling contribute significantly, and must be coordinated in order for the phase lags to be appropriate. The mathematical theory suggests that the two coupling directions must be constructed so that each, by itself, would produce the appropriate lag of 90 degrees. Since the lags are measured in a fixed direction (e.g., $(j + 1)$ -st to j th ganglion), this requires a zero crossing for the ascending coupling to be 90 degrees, and that for the descending coupling to be -90 degrees. That is, the coupling in the two directions must be asymmetric in a very particular way. If they have the same strength and that asymmetry, all the experimental observations follow Skinner et al. (1997).

The Leech Central Pattern Generator for Swimming

The 21 segmental ganglia of the leech are responsible for producing the traveling wave that enables the animal to swim. A small number of neurons in each ganglion have been identified as responsible for producing the rhythms. These can be crudely divided into three groups according to the phase in the cycle when they fire: (1) -10° to 70° , (2) 130° to 180° , and (3) 220° to 260° . Most of the interactions are inhibitory, and the three groups essentially divide the cycle into 120° subgroups. The fact that the coupling signals occur at different phases in the cycle motivated the idea of temporal averaging that was used to create the phase models described in the preceding sections for the lamprey circuits.

In the leech, the coupling appears to be important in the creation of the oscillations: an isolated ganglion is able to produce a crude oscillatory rhythm, but the rhythm is much more regular in the intact chain. The coupling within the leech CPG has many long-range connections. While the lamprey also has such connections, the success of the simple nearest-neighbor models in explaining many of the details of the lamprey swim pattern suggests that the long-range connections are not as important as they are in the leech. The length of the coupling is important in the formation of the phase lags for the leech CPG; unlike the lamprey CPG, the intersegmental phase lags are smaller for the intact cord than for pieces of the cord. This can be understood from mathematical work showing that coupling to multiple neighbors in general decreases the phase lag (see Cohen et al., 1992, for a reference).

In this CPG, there is a systematic gradient in the intrinsic frequency of each of the ganglia; the gradient is opposite the direction of the wave, which is rostral-caudal. The general mathematical theory in Kopell and Ermentrout (1986) shows that there can be waves in locally coupled chains with both frequency gradients and asymmetric coupling. However, the theory does not address the effects of the significant conduction delays in the leech cord, or the effects of coupling that might extend for a large fraction of the cord. Thus, the creation of the phase lags is less understood in the leech than in the other two examples presented above.

The Procerebral Lobe of *Limax*

Waves in the nervous system occur not only in locomotor activities but in sensory processing as well. A well-studied example occurs in the procerebral (PC) lobe of the *Limax*, or common garden slug, that is responsible for the olfactory processing of the slug. The PC lobe can be isolated from the animal along with the sensory apparatus. Under resting conditions, optical imaging of the lobe reveals a regular slow oscillatory wave that travels from the distal region of the lobe to the basal region (Kleinfeld et al., 1994). The frequency is about a half a cycle per second and the velocity is about $250 \mu\text{m}$ per second. The lobe itself is about $500 \mu\text{m}$, so that, as in the lamprey, at any given point in time, there is only one wave on the lobe. When the animal is exposed to a novel odor, the oscillations quickly synchronize for several cycles and then return to the traveling waves. The reasons for the waves and the synchronization are not known, although it is believed that they may be crucial in odor learning and recognition. Blocking the oscillations impairs the animal's ability to be conditioned to odor stimuli.

In a recent paper, Ermentrout, Flores, and Gelperin (1998) developed a minimal model for the oscillations and waves in the lobe under a variety of experimental conditions. As in the lamprey work, the intent was to suggest a general model that was not dependent on biophysical details that are not known for the PC lobe. Since the lobe consists of many coupled intrinsically oscillating cells, modeling it as a network of coupled oscillators is natural.

The lobe consists of two layers, the cell layer and the neuropil. The source of the waves appears to be intrinsically oscillatory bursting cells in the cell layer. There are many interneurons that

send processes into the neuropil and are likely important in the maintenance of the wave and the switch to synchrony. The double layer of cells is an important part of the anatomical structure to be retained in the model. One layer represents the locally connected bursting neurons and the other represents the diffusely coupled interneurons. Interneuron connections, believed to be an important part of the odor associative memory, are globally coupled, while the bursting neurons are coupled locally. Each layer consists of an $N \times m$ array of oscillators, with N representing the apical-basal axis (the direction of wave propagation) and m representing the axis parallel to the wave fronts. Each bursting cell influences only the interneuron in its immediate neighborhood, while each interneuron influences every bursting cell. The interneuron coupling is very weak during the resting state; the odor stimulus induces greater activity in the interneurons, which is modeled by an increase in their coupling strength.

In the PC lobe, the mechanism for wave production was hypothesized by Ermentrout et al. (1998) to be a gradient in the coupling strength, with the neurons in the basal end having stronger coupling than the neurons in the apical end. The reason for this choice is that the basal end is thicker and broader than the apical end, so that there could be more neurons involved in the rhythm at the thicker end. The interaction function for each layer was

$$H(\phi) = C + K \sin(\phi - \zeta) \quad (5)$$

with slightly different values of C and K for the two layers. The layer with the locally connected bursting neurons is responsible for producing the waves, while the other layer has connections that collapse the wave to synchrony when there is an odor stimulant. Considering a slice in the direction of the waves, the relevant form of the full set of equations for the layer of oscillators is given by Equation 4.

As in the simpler Equation 2, without a gradient in coupling strengths, there is a dominant direction of coupling that determines the phase lags. The major role of the gradient is to determine which coupling direction is dominant. (It is the one with coupling from the larger strengths to the weaker ones if the ascending and descending coupling is the same.) In particular, the coupling strengths do not change the phase lags. This can be deduced from the general equations in Kopell and Ermentrout (1990), and can be seen easily from equations for one-way coupling, as in Equation 3, with coupling strengths included.

An interesting feature of a model in which the dominance is determined by coupling gradients is that it mimics one feature of a frequency gradient while still producing constant speed waves: If a chain of oscillators has a frequency gradient, splitting the chain at some point produces two smaller chains having different locked frequencies; this is also true if the chain has a gradient in coupling strengths. This feature is seen in the *Limax* preparation.

Long-Range Coupling

The theory discussed in the previous sections deals only with local coupling (nearest neighbor and multiple neighbor, but still short compared to the length of the chain). There have been few studies so far on the effects of long-range coupling (see Cohen et al., 1992, for some references). One of these (Ermentrout and Kopell, 1994a) was motivated by trying to understand why the wavelength of the traveling wave in lamprey (and other anguilliform species that swim by undulation) is approximately one body length. Although the theory discussed above provides a framework for understanding how the wavelength can stay constant when the swimming speed changes, it does not provide any clues to why that wavelength should be related to the body length.

Ermentrout and Kopell showed that long coupling between the ends and an interior region, adjusted so that the coupling produces antiphase between the oscillators directly coupled, can create trav-

eling waves that have the correct wavelength. With the same architecture, but changes in the relative strengths of the ends-to-middle coupling versus middle-to-ends coupling, a new, stable pattern can emerge: the chain displays a standing wave, characteristic of the spinal cord of amphibia during a trotting gait. Thus, this architecture is a substrate for both swimming movements (traveling wave) and trotting movements in an amphibian (Ermentrout and Kopell, 1994a). The long fibers cannot be essential to the production of the phase lags in adult lamprey, since the waves form in small portions of the animal. It is possible that a mechanism involving long fibers is responsible for creating the appropriate phase lags in young animals, and that learning mechanisms then enable this information to be encoded in local synapses, which can then work without global connectivity (Ermentrout and Kopell, 1994b).

Phase Oscillators and Relaxation Oscillators

Equation 1 is derived from a physical description of oscillators and their interactions by a reduction procedure described in a previous section. When the reduction procedure is not valid, the emergent behavior of the collection of oscillators can be very different from that of the chains described in the preceding sections.

One such situation in which the reduction procedure is not generally valid involves relaxation oscillators coupled using models of excitatory chemical synapses (see PHASE-PLANE ANALYSIS OF NEURAL NETS). Such oscillators have critical differences in the mechanisms by which the phase lags are determined. For interactions mimicking fast chemical synapses, Somers and Kopell (1995) described a coupling scheme they called fast threshold modulation, or FTM. In that mathematical description, at sufficiently high or low voltages of the presynaptic cell, the synaptic conductance saturates, so that the postsynaptic cell receives a current that is (on the high or low branch) independent of the trajectory of the presynaptic cell. Thus, the postsynaptic cell is essentially uncoupled from the presynaptic one, but has its voltage equation (and so its effective threshold for firing) changed during the input.

The different mechanisms lead to some important contrasts in properties. One is that variations in frequency or in inputs among relaxation oscillators coupled by FTM need not lead to phase differences among the oscillators, as occurs for Equation 2 (Somers and Kopell, 1995). This provides a potential mechanism to be used in constructing a network whose phase lags are well regulated even if the frequencies are not. A second difference concerns speed of locking: long chains of oscillators of the form of Equation 2 can take many, many cycles to approach the locked solution, with the time increasing with the length of the chain. By contrast, a long chain of relaxation oscillators can lock within a couple of cycles (Somers and Kopell, 1995).

It is difficult to obtain traveling waves in a chain of such oscillators; the usual outcome is approximate synchrony unless the coupling is very weak (in which case the averaging procedure is relevant). Indeed, changes in gating of synapses or voltage range of the presynaptic cell can change the mechanism of the interaction from that of phase oscillator-like behavior to that of FTM, changing waves into almost synchronous cell assemblies. Thus, a generalized signal in the network that changes, e.g., the threshold or sharpness of the synapses can provide a stimulus-induced change in mechanism, and hence of emergent behavior (Somers and Kopell, 1995). We saw in the *Limax* example how a stimulus produced change from waves to synchrony; this provides another kind of mechanism for such a change.

In a recent paper, Izhikevich (2000) has shown that it is possible to derive coupled-phase models derived from relaxation oscillators. The corresponding interaction functions, $H(\phi)$, have a discontinuity at the origin. This implies that a reciprocally coupled pair will synchronize even if there is a difference in the intrinsic frequencies. Thus, in theory one can connect fast-threshold modulation with

phase models and obtain the properties of the former by using the easily analyzed formalism of the latter.

Chains of relaxation oscillators have also featured in models of a single dendrite of the basal ganglia dopaminergic neuron. Based on the experiments and models of Wilson and Callaway, Medvedev and Kopell modeled the compartments of the dendrite as a chain of coupled relaxation oscillators. (For a reference to the work of Wilson and Callaway, see Medvedev and Kopell, 2001.) The aim of the work was to understand the origin of long transients in frequency observed in slice preparations and biophysical models. The model was able to explain how the electrical coupling between the compartments can produce a spiking rate decreasing over many cycles; this decrease is analogous to spike frequency adaptation, but does not use any unusual currents normally associated with adaptation.

More complicated local circuits, with several different elements within a local circuit, may have some features of each type. The properties of such composite oscillators were studied in the context of the swimmeret CPG system (Skinner and Mulloney, 1998). In that system, the theory using averaged coupling is successful in understanding phase lags in normal and perturbed settings. A model based on the biophysical underpinnings of the component cells allows one to address finer questions relating to control of burst durations and frequency.

Road Maps: Dynamic Systems; Motor Pattern Generators

Related Reading: Collective Behavior of Coupled Oscillators; Half-Center Oscillators Underlying Rhythmic Movements; Spinal Cord of Lamprey: Generation of Locomotor Patterns

References

- Cohen, A., Ermentrout, G. B., Kiemel, T., Kopell, N., Sigvardt, K., and Williams, T., 1992, Modelling of intersegmental coordination in the lamprey central pattern generator for locomotion, *Trends Neurosci.*, 15:434–438. ♦
- Cohen, A. H., Holmes, P. J., and Rand, R. H., 1982, The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model, *J. Math. Biol.*, 13:345–369.
- Ermentrout, G. B., Flores, J., and Gelperin, A., 1998, Minimal model of oscillations and waves in the *Limax* olfactory lobe with tests of the model's predictive power, *J. Neurophysiol.*, 79:2677–2689.
- Ermentrout, G. B., and Kopell, N., 1984, Frequency plateaus in a chain of weakly coupled oscillators: I, *SIAM J. Math. Anal.*, 15:215–237.
- Ermentrout, G. B., and Kopell, N., 1994a, Inhibition-produced patterning in chains of coupled nonlinear oscillators, *SIAM J. Appl. Math.*, 54:478–507.
- Ermentrout, G. B., and Kopell, N., 1994b, Learning of phase-lags in coupled neural oscillators, *Neural Computat.*, 6:225–241.
- Friesen, W. O., and Pearce, R. A., 1993, Mechanisms of intersegmental coordination in leech locomotion, *Semin. Neurosci.*, 5:41–47. ♦
- Guckenheimer, J., and Holmes, P., 1983, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, New York: Springer-Verlag. ♦
- Izhikevich, E., 2000, Phase equations for relaxation oscillators, *SIAM J. Appl. Math.*, 60:1789–1805.
- Kleinfeld, D., Delaney, K. R., Fee, M. S., Flores, J. A., Tank, D. W., and Gelperin, A., 1994, Dynamics of propagating waves in the olfactory network of a terrestrial mollusk: An electrical and optical study, *J. Neurophysiol.*, 72:1402–1419.
- Kopell, N., 1987, Toward a theory of modeling central pattern generators, in *Neural Control of Rhythmic Movements in Vertebrates* (A. H. Cohen, S. Grillner, and S. Rossignol, Eds.), New York: Wiley, pp. 369–413. ♦
- Kopell, N., and Ermentrout, G. B., 1986, Symmetry and phaselocking in chains of weakly coupled oscillators, *Commun. Pure Appl. Math.*, 39:623–660.
- Kopell, N., and Ermentrout, G. B., 1990, Phase transitions and other phenomena in chains of coupled oscillators, *SIAM J. Appl. Math.*, 50:1014–1052.
- Medvedev, G., and Kopell, N., 2001, Synchronization and transient dynamics in chains of FitzHugh-Nagumo oscillators with strong electrical coupling, *SIAM J. Appl. Math.*, 61:1762–1801.
- Skinner, F., and Mulloney, B., 1998, Intersegmental coordination of limb movements during locomotion: Mathematical models predict circuits that drive swimmeret beating, *J. Neurosci.*, 18:3831–3842.
- Somers, D., and Kopell, N., 1995, Waves and synchrony in networks of oscillators of relaxation and non-relaxation type, *Physica D*, 88:1–14.
- Skinner, F., Kopell, N., and Mulloney, B., 1997, How does the crayfish swimmeret system work: Insights from nearest neighbor-coupled models, *J. Comp. Neurosci.*, 4:151–160.

Chaos in Biological Systems

Leon Glass

Introduction

Chaotic dynamics, i.e., *aperiodic dynamics in deterministic systems displaying sensitivity to initial conditions*, has emerged from a relatively obscure topic of mathematics to an area that is of great current interest among scientists as well as the general public. The important characteristics of chaos are the apparent irregularity of time traces and the divergence of the trajectories over time (starting from two nearby initial conditions) in a system that is deterministic. Although the rhythm is irregular, the underlying deterministic equations can lead to structure in the dynamics (Figure 1).

Deterministic chaotic dynamics is different in principle from what is commonly known as random dynamics. In random dynamics, prediction is intrinsically impossible, except in a statistical sense. A natural system believed to be random is the radioactive decay of isotopes, leading to random time intervals between decay events.

Chaotic dynamics is now well-documented in a variety of different mathematical models, physical systems, and biological systems (Cvitanovic, 1989). Introductions to chaotic dynamics suit-

able for biology include Glass and Mackey (1988), Kaplan and Glass (1995), Strogatz (1994), and Liebovitch (1998). Research in biology has extended from studies of dynamics in well-controlled model systems to analysis of spontaneous dynamics in organisms. Although all agree that spontaneous activity is very complicated, reflecting the interactions of the intrinsic physiological control mechanisms with the fluctuating environment, the extent to which concepts developed in simpler chaotic systems can be applied is controversial. The word *chaos* has been used with so many different nuances of meaning and operational definitions that the original technical meaning (i.e., the highlighted phrase in the opening sentence of this article) can no longer be assumed. However, in this article, I will stick to this original technical definition of chaos.

Identifying Chaos

Chaos in Deterministic Systems

In a deterministic system there is a definite equation governing the dynamics. Biological systems have been modeled by difference

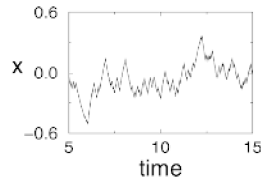


Figure 1. Dynamics from a deterministic high-dimensional ordinary differential equation representing a highly interconnected gene or neural network. The extremely irregular dynamics reflect deterministic chaos rather than random noise, but it would be difficult to develop criteria to determine this without huge amounts of data. The network is of the sort discussed in Mestl et al. (1997) with 32 model “neurons,” each of which has 15 randomly chosen inputs. The graph shows the values of one of those variables as a function of time.

equations, differential equations, delay-differential equations, and partial differential equations that display chaotic dynamics. As long as one is dealing with deterministic equations, the signature of deterministic chaos is clear. It is easy to numerically integrate the equations starting from two nearby initial conditions and to watch the evolution. Chaotic systems display aperiodic dynamics with the same statistical properties (such as density histograms) for both trajectories. In addition, the trajectories should diverge but remain in a bounded region. Many equations also show characteristic bifurcations (changes in qualitative dynamics) as parameters change that are representative of well-studied routes to chaos. Perhaps the best known of these, the “period doubling route to chaos,” has been found in a variety of simple equations and experimental systems. As a parameter of an equation changes, the period of an oscillation undergoes successive doublings. For example, if the period of an oscillation was initially 1 s, as a parameter changed the period would have doubled to 2 s, then as the parameter changed a bit more the period would have doubled to 4 s, and so forth. However, the parameter range over which each successively doubled period persists decreases geometrically, so that in practice it is very difficult to observe experimentally more than two or three of the successive doublings. The period doublings lead eventually to chaotic dynamics, which arise after a parameter crosses a threshold (Cvitanovic, 1989).

When dealing with dynamics in experimental systems or in natural systems, it is much more problematic to figure out the dynamic origins of complicated rhythms that are not periodic. Such rhythms will usually arise from the interaction between a deterministic system and random noise, and it is difficult to dissect out the relative effects of the determinism and the noise (Ruelle, 1994). Since there is no clear operational definition for chaos in experimental data, analyses generally focus on other properties of time series that can be defined operationally. However, if chaos has property C , and a time series has a property C , it does not follow that the time series is chaotic. Several books and web sites deal with time-series analysis methods and their pitfalls (Ott, Sauer, and Yorke, 1994; Abarbanel, 1996; Kantz and Schreiber, 1997; Goldberger et al., 2000).

I now summarize many features that are assessed in analyses of time series.

Power spectrum. The power spectrum of a time series can be easily determined using fast Fourier transform algorithms. However, since one can always construct random models that have the same power spectrum as a deterministic model, the power spectrum is not a good measure for defining chaos.

Dimension. Grassberger and Procaccia (1983) defined the correlation dimension of a set of points embedded in d dimensions. The set of points could be generated in a variety of ways, including

discrete time processes (e.g., time intervals between neural spikes or outputs of a difference equation) or continuous processes (e.g., recording from an electroencephalogram or a continuous differential equation). Call r the distance away from a given point. The point correlation function, $C_i(r)$ is proportional to the number of points lying within a radius r of a point i . The spatial correlation function, $C(r)$, is an average of the $C_i(r)$ over all points. Then the correlation dimension is ν , if $C(r) = kr^\nu$ for small values of r . For random points distributed in d dimensions, $C(r) = kr^d$, so that the dimension is an integer equal to the dimension of the space in which points are distributed. Thus a line has dimension 1, a plane has dimension 2, and so forth. However, some sets of points, called *fractals*, have a fractional dimension. Some workers take a fractional correlation dimension as a definition for deterministic chaos, since many chaotic dynamical systems have attractors with a fractional dimension. However, there are many pitfalls associated with determining the dimension, and not all chaotic systems have a fractional dimension. Consequently, a fractional correlation dimension cannot be accepted as a definition for chaos.

Lyapunov exponent. The Lyapunov exponent measures the rate at which trajectories diverge. A negative Lyapunov exponent indicates convergence of trajectories and a positive Lyapunov exponent indicates divergence of trajectories. Therefore, for a time series a positive Lyapunov exponent is a necessary condition for chaotic dynamics. However, the numerical algorithms for determination of the Lyapunov exponent may yield a positive exponent for periodic oscillations with large first derivatives, or in noisy systems. Therefore, a positive Lyapunov exponent found using numerical methods from naturally occurring data cannot be taken as a definition for chaos.

Poincaré map. Certain difference equations display period doubling bifurcations and chaos. An example is the quadratic map in which the value of a variable at one time is a quadratic function of the value at the previous time. The period doubling route to chaos is observed as the height of the quadratic function is increased. Some experimental data can be plotted so that the value of a variable at one time is plotted as a function of its value at the preceding time. For example, given a sequence of interspike intervals, it is possible to plot one interspike interval as a function of the preceding interval. If data plotted in this fashion clearly fall on a one-dimensional curve that is known to give chaotic dynamics, then this is good evidence for chaotic dynamics in the data. However, most biological examples in which data are plotted in this fashion do not give such clear-cut results, but rather give clouds of points.

Determinism and prediction. Since chaotic dynamics are generated by deterministic equations, in chaotic systems it should be possible either to demonstrate determinism or to predict dynamics that will occur in the proximate future. Several methods have been developed to try to make predictions based on past dynamics without necessarily knowing the equations of motion. A confounding factor of prediction tests is that random inputs to linear and nonlinear systems lead to correlations for short times that enable prediction even though the system is not deterministic and hence not chaotic. Random signals in which there are brief randomly inserted segments, all of which have the same waveform, would also be expected to show short-term predictability, but the underlying signal would not be chaotic.

Surrogate data. None of the above measures can be simply applied to determine if a given data set displays deterministic chaos. Consequently, several people have advocated generating “surrogate” data sets that have similar statistical properties to a given data set but are generated from a random process (Kantz and Schreiber, 1997). For example, it is a simple matter to generate a random time

series that has an identical power spectrum to a given time series. Statistical analyses are then carried out using the time series under investigation and surrogates. The results of the analysis of the surrogate data are compared with the analysis of the original data. The origins of significant differences are then analyzed. They may be due to deterministic chaos, although other hypotheses, such as failure to capture some important aspect in the surrogate data set, need to be entertained. Application of this technique has led to the recognition that many earlier claims for deterministic chaos were in error.

Chaos at Subcellular and Cellular Levels

Carefully controlled experiments, in which it is possible to generate large amounts of high-quality data while varying critical control parameters, have demonstrated chaotic dynamics in chemical, electronic, and hydrodynamic systems (Cvitanovic, 1989). Observations of chaotic dynamics in biological systems are more problematic. I will briefly review some of the areas in which there have been claims for chaotic dynamics.

Subcellular Dynamics: Ion Channels

The basis for neural activity lies in the changes of conductance of specialized protein molecules called ion channels that allow ions to pass between the intracellular and the extracellular medium. For example, action potentials of nerve cells are usually associated with the opening of sodium ion channels. Classical intracellular electrophysiological techniques record the average activity over the thousands of ion channels present on a single nerve cell. Macroscopic ionic models describe this average activity and its dependence on the membrane voltage. An important technological advance, called the patch clamp, enabled researchers to measure directly the conductance of a single ion channel during its open state. Experimental recording of a single ion channel typically shows a complex switching behavior in which the channel makes transitions between open and closed states in a seemingly random manner. Most researchers believe that this switching is random and have developed appropriate theoretical descriptions. However, deterministic chaotic models may show the same statistical features as the random models (Liebovitch, 1998). At the current time, new information about the structural properties of ion channels is accumulating, and this should yield new insights into the molecular basis of the conformational changes that underlie the channel activity. Independent of the outcome of that endeavor, the observation that subcellular ion channel events are following a random (or perhaps chaotic) dynamics raises a conceptual problem: Why is it possible to study cellular or supercellular processes using deterministic ionic models? Presumably, the thousands of channels in a single cell, and the millions and billions in collections of cells, enable an averaging, so that macroscopic equations are valid, with fluctuations being small.

Cellular Activity in Single Cells

The electrical activity of single cells can be easily measured using intracellular or extracellular microelectrodes, either in vitro or in vivo. The activity can be measured during spontaneous activity or in response to manipulation of environmental parameters. In my opinion, the most convincing demonstrations of chaotic dynamics in biological systems have been carried out in this scale of preparation. Periodic stimulation of oscillatory or excitable biological systems, such as nerve or cardiac tissue, has identified chaotic dynamics over restricted ranges of stimulation parameters (Glass and Mackey, 1988). Over other stimulation ranges there are regular rhythms in which there are repeating sequences of stimuli and action potentials. This work is placed on a firm footing by extensive theoretical analyses and numerical simulation mathematical mod-

els. Depending on the stimulation parameters, the theoretical models display periodic behavior or deterministic chaos in agreement with experimental observations. Moreover, in the chaotic region, plotting the stimulus phase as a function of the preceding stimulus phase to generate the Poincaré map can give a plot that is similar to the quadratic map (Glass and Mackey, 1988), further supporting the identification of chaos. However, there is no clear functional role for chaos at a cellular level.

Complex Dynamics in Biological Networks

Chaos in Complex Networks

One of the outstanding characteristics of biological systems is that they are characterized by complex networks of interacting elements. Thus, schematic diagrams of biological systems controlling metabolism, hormone secretion, gene activation, hormone secretion, motor activity, and heart rate are notable for being incredibly complex, with multiple feedbacks. Mathematical models developed for complex biological networks often demonstrate chaotic dynamics (Arbib, Érdi, and Szentágothai, 1997; Goldbeter, 1997; Mestl, Bagley, and Glass, 1997; Glass and Mackey, 1988; Sreenivasan, Pradhan, and Rapp, 1999). However, because there are no well-established operational tests for chaos, analyses that claim chaos in experimental or clinical data are often subject to alternative interpretations. Thus, the extent to which data such as heart rate variability or electroencephalographic data reflect deterministic chaos has been hotly debated. In recent years, there have been suggestions that many biological time series that had previously been called chaotic are not chaotic but may be better described using other adjectives ($1/f$ long-range scaling, fractal, multifractal). However, these terms do not have clear implications with regard to the underlying mechanisms. Independent of the detailed mechanisms of fluctuations, a variety of studies have suggested implications for the study of health and disease.

Chaos and Health

Although early studies of biological control networks often emphasized that they served to maintain homeostasis—a relatively constant internal environment—there is now increasing recognition that normal, healthy individuals can be expected to show complex fluctuations in important physiological functions (Goldberger, 1996). Many diseases can be characterized by qualitative changes in dynamics in biological control systems. Since these changes may be associated with bifurcations in appropriate mathematical models of the physiological systems, Mackey and I proposed the term “dynamical disease” to capture the dynamic aspects of disease (Glass and Mackey, 1988).

There are many examples of altered physiological dynamics associated with disease. Diminished heart rate variability is associated with a higher risk of sudden cardiac death. However, since patients do not die from reduced heart rate variability, it is important to clarify the underlying mechanisms. One hypothesis is that impaired cardiac function leads to increased sympathetic nervous activity associated with elevated levels of circulating chemicals (catecholamines). This in turn will increase the heart rate and lead to less variability. Further, since circulating catecholamines are pro-arrhythmic, there can be an increased risk of arrhythmia. In addition, patients with impaired heart function are also more likely to be taking drugs that blunt fluctuations in heart rate, or may even have implanted artificial pacemakers that might lead to a more regular rhythm. Clinical studies that associate higher risk for sudden cardiac death with decreased heart rate variability do not always address these potentially confounding factors. Independent of the mechanism, since low heart rate variability is associated with a higher risk for sudden cardiac death, the analysis of heart rate variability in the clinic has practical utility.

In a similar vein, in neurology, there have been claims that time-series analysis techniques can be useful in predicting the occurrence of an epileptic episode (Schiff, 1998). These types of claims are often hard to validate, since the data sets and the analysis algorithms often are not readily available. However, the recently established Research Resource for Complex Physiologic Signals (Goldberger et al., 2000) may provide an important avenue for publishing both data sets and data analysis packages.

Another practical direction is to apply to biological systems (Moss and Gielen, 2000) methods that can successfully control chaotic dynamics in mathematical models and in physical systems (Ott et al., 1994). Control methods have been applied to control comparatively simple rhythms associated with deterministic dynamics in model experimental systems. However, in cases in which the rhythms are more complex and not definitely associated with deterministic dynamics, further research is needed to validate the utility of chaos control methods.

Discussion

Theoretical studies have demonstrated chaotic dynamics in mathematical models of biological systems ranging from the subcellular level to the organismal level. The most convincing experimental observations of chaotic dynamics in biological systems are associated with the response of biological systems in vitro to periodic stimulation of comparatively well-defined cells or assemblies of cells. The high dimensions of biological systems, the fluctuating environment (leading to nonstationarity), and the subtleties of the numerical methods are factors that make convincing demonstration of chaos in vivo a difficult matter. Therefore, although it might not be surprising that clear demonstrations of deterministic chaos in biology are rare, it is surprising that exaggerated claims founded on weak experimental evidence have been widely accepted without sufficient critical analysis.

Insofar as real biological systems (as opposed to computer models of these systems) are exposed to random thermal and other fluctuations and are composed of large numbers of interacting elements (i.e., they are high-dimensional), it is not easy to decide the appropriate class of theoretical model for a given system. Another difficult issue is to understand the biological significance, if any, of the observed fluctuations. Consider complex fluctuations in the timing between spikes of a neuron: does the neuron function well despite these irregularities, or are the irregularities essential to its task? And if the irregularities are essential for the task, is there any reason to expect that deterministically chaotic irregularities would be better than random ones? Although there have been several interesting speculations about an association of chaos with health and a possible role for chaos in information processing (for a summary, see Arbib et al., 1997), convincing demonstrations of the function of chaos are not yet available. Probably what drives many in brain theory is to understand the neural correlates of the human mind: higher cognitive function, originality, and free will. Humans be-

have neither in a random manner nor in a totally predictable manner. Although it seems inevitable that the mathematical concept of chaos will help us interpret the human brain, definite results pointing in that direction are still meager.

Road Map: Dynamic Systems

Background: I.3. Dynamics and Adaptation in Neural Networks

Related Reading: Chaos in Neural Systems; Stochastic Resonance; Synaptic Noise and Chaos in Vertebrate Neurons

References

- Arbib, M. A., Érdi, P., and Szentágothai, J., 1997, *Neural Organization: Structure, Function, and Dynamics*, Cambridge, MA: MIT Press, Sect. 4.3.
- Abarbanel, H. D. I., 1996, *Analysis of Observed Chaotic Data*, Berlin: Springer-Verlag. ♦
- Cvitanovic, P., Ed., 1989, *Universality in Chaos*, 2nd ed., Bristol: Adam Hilger.
- Glass, L., and Mackey, M. C., 1988, *From Clocks to Chaos: The Rhythms of Life*, Princeton, NJ: Princeton University Press. ♦
- Goldberger, A. L., 1996, Non-linear dynamics for clinicians: Chaos theory, fractals, and complexity at the bedside, *The Lancet*, 347:1312–1314.
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E., 2000, PhysioBank, PhysioTools, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, *Circulation*, 101:e215–e220. [Circulation electronic pages; <http://circ.ahajournals.org/cgi/content/full/101/23/e215>]. See also <http://www.physionet.org> for online data collections.
- Goldbeter, A., 1997, *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*, Cambridge, Engl.: Cambridge University Press.
- Grassberger, I., and Procaccia, I., 1983, Measuring the strangeness of strange attractors, *Physica D*, 9:189–208.
- Kantz, H., and Schreiber, T., 1997, *Nonlinear Time Series Analysis*, Cambridge, Engl.: Cambridge University Press. Software available at: <http://www.mpi-pks-dresden.mpg.de/tisean>. ♦
- Kaplan, D., and Glass, L., 1995, *Understanding Nonlinear Dynamics*, New York: Springer-Verlag. ♦
- Liebovitch, L. S., 1998, *Fractals and Chaos Simplified for the Life Sciences*, New York: Oxford University Press. ♦
- Ott, E., Sauer, T., and Yorke, J. A., Eds., 1994, *Coping with Chaos*, New York: Wiley.
- Mestl, T., Bagley, R. J., and Glass, L., 1997, Common chaos in arbitrarily complex feedback networks, *Phys. Rev. Lett.*, 79:653–656.
- Moss, F., and Gielen, S., Eds., 2000, *Handbook of Biological Physics*, vol. 4, *Neuro-informatics, Neural Modelling*, Amsterdam: Elsevier, chaps. 6 and 7.
- Ruelle, D., 1994, *Physics Today*, 47:24–30.
- Schiff, S. J., 1998, Forecasting brain storms, *Nature Med.*, 4:1117–1118.
- Sreenivasan, R., Pradhan, N., and Rapp, P. E., Eds., 1999, *Nonlinear Dynamics and Brain Functioning*, Nova Science Publishers.
- Strogatz, S. H., 1994, *Nonlinear Dynamics and Chaos: With Applications in Physics, Biology, Chemistry, and Engineering (Studies in Nonlinearity)*, Cambridge, MA: Perseus. ♦

Chaos in Neural Systems

Kazuyuki Aihara

Introduction

From the viewpoint of dynamical systems theory, biological neurons and neural networks can be understood as nonlinear dynamical systems. Therefore, research on neural systems is a combination of brain science and nonlinear science.

The progress of nonlinear science in the twentieth century deepened our understanding of interesting dynamical phenomena called

deterministic chaos, or simply chaos, although the research history of chaos can be traced back at least to Poincaré's work on the three-body problem in celestial mechanics 100 years ago. Chaotic phenomena, in which a deterministic law generates complicated, non-periodic, and unpredictable behavior, exist in many real-world systems and mathematical models. Chaos has many intriguing characteristics, such as a sensitive dependence on initial conditions,

or the so-called butterfly effect (i.e., the influence of a small perturbation in the initial condition diverges with time), and fractal geometrical structure (i.e., chaotic behavior is represented by a strange attractor with a noninteger dimension in the state space) (Ott, Sauer, and Yorke, 1994; see also CHAOS IN BIOLOGICAL SYSTEMS; DYNAMICS AND BIFURCATION IN NEURAL NETS).

In this article, the nonlinear behavior of neural systems is considered with respect to deterministic chaos. Although many of the neural systems that have been studied so far for neural computation have simple dynamics that converge into a steady state corresponding to an equilibrium point or a limit cycle, neural systems with chaotic dynamics produce much more dynamical behavior that may be functional from the viewpoint of information processing (see also COMPUTING WITH ATTRACTORS).

Neurodynamics and Chaos

The nonlinear dynamics of a single neuron can naturally generate deterministic chaos under some conditions. For example, nerve membranes of squid giant axons respond to periodic forcing, such as sinusoidal and periodic pulse stimulation, not only periodically but also chaotically, depending on parameter values of the frequency and the strength of the force and states of nerve membranes (Aihara, 2002). Figure 1 demonstrates a chaotic response observed experimentally in squid giant axon. Chaotic and various other neuronal responses, which provide possible mechanisms of synaptic coding, are also observed in biological neurons stimulated through synapses (Segundo et al., 1998; see also SYNAPTIC NOISE AND CHAOS IN VERTEBRATE NEURONS). Thus, the behavior of a single neuron is essentially very dynamical, although some computational aspects of excitable neurodynamics can be reduced to static and logical calculations typically modeled by the McCulloch-Pitts neuron. Furthermore, chaotic dynamics in neural systems may also be discussed both in the more microscopic level of ion channels and in the mesoscopic and more macroscopic levels of neural networks (Elbert et al., 1994; Freeman, 2000; see also CHAOS IN BIOLOGICAL SYSTEMS).

Although chaotic behavior appears extremely irregular and complicated, it differs entirely from stochastic randomness because chaos is generated according to a definite deterministic rule. This

deterministic dynamics enables control, short-term prediction, and change through bifurcation of chaotic behavior. On the other hand, because real neural systems operate in noisy environments, it is an important and delicate problem to clarify the effects of noise on chaotic neurodynamics (Freeman, 2000; Kaneko and Tsuda, 2000; Tsuda, 2001).

Models of Chaotic Neural Networks

Can deterministic chaos in neural systems play a useful role? Biologically based answers to this question are few (Glass, 2001; see also CHAOS IN BIOLOGICAL SYSTEMS), but possible functions of chaos have been explored using theoretical models of chaotic neural networks.

Various models of neurons and neural networks with chaotic dynamics, in both discrete time and continuous time, have been proposed (Aihara, Takabe, and Toyoda, 1990; Lewis and Glass, 1992; Nara et al., 1995; van Vreeswijk and Sompolinsky, 1996; Rabinovich and Abarbanel, 1998; Freeman, 2000; Kaneko and Tsuda, 2000; Tsuda, 2001). For example, chaos in nerve membranes can be well described by continuous-time nerve equations like the Hodgkin-Huxley and FitzHugh-Nagumo equations (Aihara, 2002). It is an important research subject to extend such biologically realistic neuronal models by considering difference of excitability types (see CANONICAL NEURAL MODELS) and input effects through both chemical and electrical synapses (see NEOCORTEX: CHEMICAL AND ELECTRICAL SYNAPSES; TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES).

On the other hand, simpler neuron models are desirable for elucidating possible principles of information processing with chaotic dynamics in large-scale neural networks both mathematically and numerically. For this purpose, a simple discrete-time model of chaotic neural networks has been proposed in which the neuronal elements qualitatively reproduce chaotic responses experimentally observed in squid giant axons (Aihara et al., 1990; Aihara, 2002). This model, based on the earlier Caianiello's neuronic equations and the Nagumo-Sato model with refractoriness, considers spatiotemporal external inputs from outside the network, spatiotemporal feedback inputs through mutual interactions among constituent neurons in the network, and refractory effects as a general discrete-

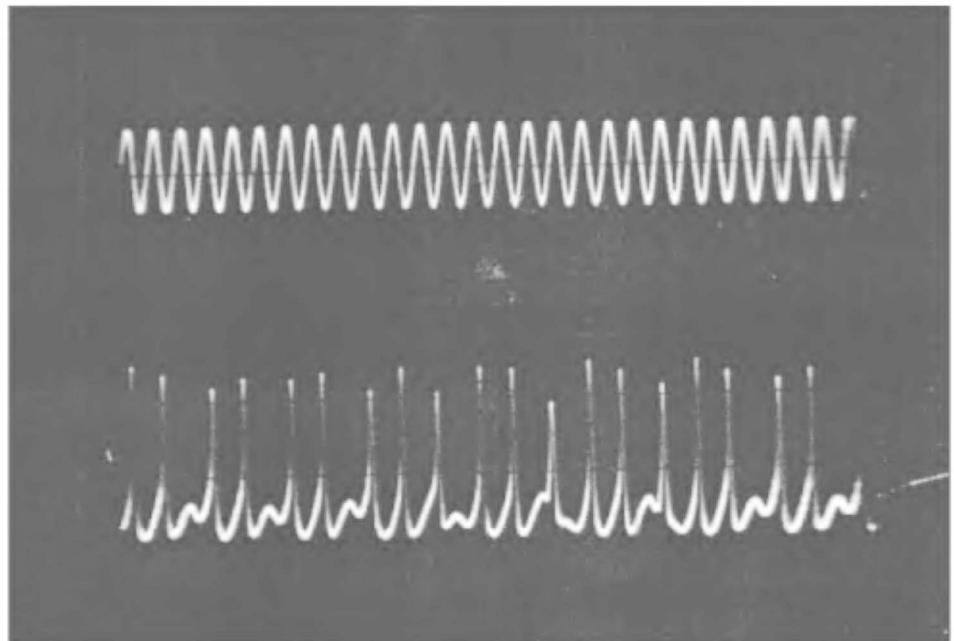


Figure 1. An example of chaotic response (below) of squid giant axon in a state of self-sustained oscillation to stimulation of a sinusoidal current (above).

time model of neural networks composed of neurons with their own chaotic dynamics (Aihara et al., 1990; Adachi and Aihara, 1997; Aihara, 2002). In fact, by adjusting the values of parameters, the model of chaotic neural networks can include many conventional discrete-time neuron models, such as the McCulloch-Pitts neurons with the Heaviside output function, analog neurons with the logistic output function used in backpropagation neural networks, and the Nagumo-Sato neurons with response characteristics of complete devil's staircases. Thus, with the model of chaotic neural networks, chaotic dynamics can be introduced into conventional discrete-time neural networks to explore possible functions and computational roles of spatiotemporal chaos in neural information processing by comparing the performance of the chaotic neural networks with that of the conventional neural networks in the common framework.

There is another approach to modeling chaos in neural systems. Rather than connecting chaotic neuronal elements, network dynamics can produce chaos by devising interactions through feedback connections among simpler neuronal elements, as follows: piecewise linear models (Lewis and Glass, 1992), balanced excitation and inhibition with sparsely connected strong synapses (van Vreeswijk and Sompolinsky, 1996), reduction of asymmetric connectivity storing cyclic memories by a pseudo-inverse method (Nara et al., 1995), coupled mesoscopic oscillators with different delayed feedback loops (Freeman, 2000), and two recurrent neural networks with excitatory and inhibitory feedback connections and stochastic renewal of dynamics representing synaptic noise (Kaneko and Tsuda, 2000; Tsuda, 2001).

Possible Functions of Spatiotemporal Chaos in Neural Networks

Several aspects of useful computational dynamics with spatiotemporal chaos in neural systems are reviewed with the model of chaotic neural networks below. Since each constituent element of the chaotic neural network model has its own chaotic dynamics, the spatiotemporal dynamics of the whole network is fundamentally very rich, complicated, flexible, and dependent on the coupling of chaotic neuronal elements. The adjustment of chaotic dynamics in elemental neurons by constraining architecture, interactions among neurons, and other parameters in the model leads to functional harnessing of the spatiotemporal dynamics of the chaotic neural networks (Aihara et al., 1990; Rabinovich and Abarbanel, 1998; Kaneko and Tsuda, 2000).

Dynamical Association

Associative memory is one of the most popular applications of neural networks (see ASSOCIATIVE NETWORKS; COMPUTING WITH ATTRACTORS; DYNAMICS OF ASSOCIATION AND RECALL; STATISTICAL MECHANICS OF NEURAL NETWORKS). As an example, we can consider the nonlinear dynamics of autoassociative memory networks composed of 100 chaotic neurons, where the synaptic weights are determined by the correlation matrix of four stored patterns: "cross," "triangle," "wave," and "star" (Adachi and Aihara, 1997).

Figure 2A shows spatiotemporal patterns generated by the associative dynamics of the chaotic neural network. In a recalling process in the conventional autoassociative memory network, the network state quickly converges from a perturbed stored pattern to the exact stored pattern, which is an asymptotically stable fixed point of the network dynamics. The network state of the chaotic neural network, on the other hand, displays itinerant behavior among stored patterns, as shown in Figure 2A, if the network parameters are appropriately adjusted so that the refractoriness is strong enough to escape from any fixed points except the totally

quiescent state at which all neurons are resting. The pattern at $t = 1$ in Figure 2A is very close to the stored pattern of "wave." If the network were the conventional autoassociative memory, it would keep this pattern of "wave" forever after $t = 1$. However, the chaotic network continues recalling different stored patterns and their reversed patterns successively, intermittently, nonperiodically, and unpredictably; e.g., the patterns at $t = 23$ and $t = 80$ in Figure 2A are very close to the pattern "cross" and the reversed pattern of "triangle," respectively. It should be noted that this dynamical association process is different from cyclic memory, in which the order of recalled patterns is predictably periodic, as designed by synaptic connections.

Moreover, external inputs to neurons corresponding to a stored pattern can attract the spatiotemporally chaotic dynamics near the stored pattern for a while but not eternally, as shown in Figure 2B, where external inputs corresponding to the stored pattern "triangle" are applied to the chaotic neural network of Figure 2A. Actually, chaotic dynamics is effective for achieving a quick response to changing external inputs (van Vreeswijk and Sompolinsky, 1996; Rabinovich and Abarbanel, 1998).

The spatiotemporal dynamics of Figure 2A can be interpreted as either searching stored patterns or generating spatiotemporal patterns composed of static patterns memorized by synaptic weights. This dynamic may be related to a nonlinear phenomenon called *chaotic itinerancy*, widely observed in different nonlinear models of cortical neural networks, globally coupled maps, and optical turbulence (Kaneko and Tsuda, 2000; Tsuda, 2001). Recurrent neural networks composed of McCulloch-Pitts types of neurons display similar dynamics with chaotic association when reducing asymmetric connectivity storing cyclic memories (Nara et al., 1995).

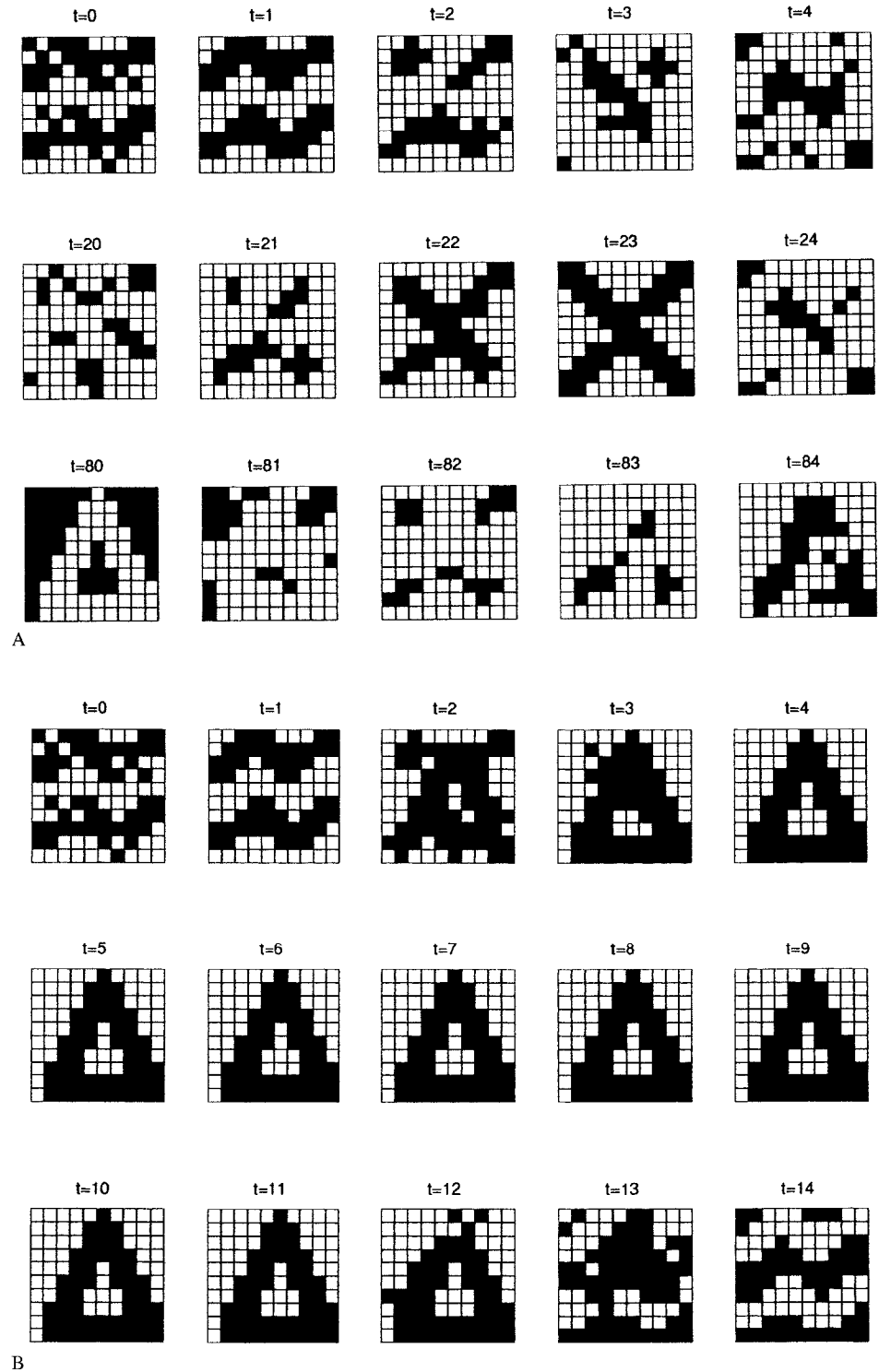
Related to this kind of memory dynamics, Freeman and colleagues proposed, on the basis of their physiological experiments on the olfactory neural system, an intriguing hypothesis: Chaotic neural activities play functional roles in a global and unstructured basal state with the capability of rapid and unbiased access to learned patterns evoked by trained odors and a catchbasin state classifying and learning unknown patterns evoked by novel or unfamiliar odors (Freeman, 2000).

Dynamical Optimization

The spatiotemporal searching dynamics demonstrated in Figure 2A also can be applied as a new kind of heuristic method to combinatorial optimization problems, especially the *NP*-hard problems that are quite important in many industrial and engineering applications. In the field of neural computation, gradient descent neurodynamics with a computational energy function has been applied to combinatorial optimization problems like the traveling salesman problems (TSP) (see OPTIMIZATION, NEURAL). Although this dynamical method for combinatorial optimization is an interesting heuristic technique, its naive applications would usually suffer from the so-called local minimum problem.

Since the spatiotemporal dynamics of the chaotic neural networks can easily escape from local minimum states, the performance of the chaotic neural networks for *NP*-hard problems is expected to be better than that of the conventional neural networks with simple gradient descent dynamics. This method utilizes chaotic fluctuation rather than stochastic fluctuation for combinatorial optimization. We can theoretically guarantee the global searching capability of the chaotic neural networks under some sufficient conditions for fully developed chaotic dynamics (Aihara, 2002). However, it is speculated that efficient searching is achieved when the chaotic dynamics generates an appropriate strange attractor with fractal structure, which may be a zero measure with respect to the Lebesgue measure of the whole state space but includes globally

Figure 2. Associative dynamics of a chaotic neural network with four stored patterns. The network is composed of 100 chaotic neurons with refractoriness (Adachi and Aihara, 1997). *A*, Spatiotemporally chaotic patterns of the chaotic neural network. The output pattern of 100 constituent neurons is displayed in the form of a 10×10 matrix with black and white squares showing firing and quiescent neurons, respectively. *B*, Response of the chaotic neural network to the stimulation corresponding to the stored pattern "triangle."



optimal or near-optimal solutions. The optimization efficiency can be further improved by introducing chaotic simulated annealing with changing network bifurcation parameter values ranging from fully chaotic states to fixed points and by combining chaotic neurodynamics with effective heuristic methods like the 2-opt algorithm and the tabu search algorithm (Aihara, 2002).

Cantor Coding

Tsuda and colleagues (Tsuda, 2001) proposed a chaos-driven contracting system, a simple example of which is a chaotic neural network with three neurons in which an unstable chaotic neuron drives a contracting system composed of a stable excitatory neuron and a stable inhibitory neuron. In this network, information of a

symbolic sequence generated by the forcing chaotic neuron is encoded on a Cantor-like set in the state space of the contracting subsystem. Because the system can be extended to a more general architecture composed of a recurrent neural system with chaotic itinerant dynamics and a contracting neural network stimulated by the former network, similar phenomena are expected to be experimentally observed at CA3 and CA1 in the hippocampus and at the olfactory bulb and the prepyriform cortex in the olfactory neural system. These neural networks may ultimately be involved in the formation of episodic memory (Tsuda, 2001).

Discussion

The chaotic neural networks and the possible computational dynamics with dynamical association, combinatorial optimization, and Cantor coding have been reviewed as simple examples of neural systems with chaotic dynamics. Although theoretical and experimental studies on chaos in neural systems further suggest other possible roles of chaotic neurodynamics in relation to higher functions of the brain, such as adaptation, perception, episodic memory, learning, awareness, intentionality, and thought (Freeman, 2000; Tsuda, 2001), it is still an important open question to experimentally prove or disprove the existence of these functions of chaos in real biological systems (Glass, 2001; see CHAOS IN BIOLOGICAL SYSTEMS). It is certain, however, that real nerve membranes have chaotic dynamics, while many artificial neuronal models are based upon oversimplified threshold dynamics. The chaotic dynamics in the level of single neurons may be the key to emergence of organized functions in biological neural networks (Rabinovich and Abarbanel, 1998). Thus, it is an important future problem to examine in detail dynamical cell assemblies and possible spatiotemporal coding (Fujii et al., 1996) in neural networks composed of chaotic neurons.

Neural systems with chaotic spatiotemporal dynamics, on the other hand, can be attractive models for analog neurocomputing (see also ANALOG NEURAL NETS: COMPUTATIONAL POWER; NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY; SPIKING NEURONS, COMPUTATION WITH) because chaotic dynamics can read out the complexity of a real number with time due to instability and nonlinearity. In fact, analog IC chips of chaotic neural networks have already been implemented as a new kind of an-

alog computing circuits for practical applications of spatiotemporal chaos in the fields of engineering (Aihara, 2002).

Road Map: Dynamic Systems

Background: Computing with Attractors

Related Reading: Chaos in Biological Systems; Synaptic Noise and Chaos in Vertebrate Neurons

References

- Adachi, M., and Aihara, K., 1997, Associative dynamics in a chaotic neural network, *Neural Networks*, 10:83–98.
- Aihara, K., 2002, Chaos engineering and its application to parallel distributed processing with chaotic neural networks, *Proc. IEEE*, 90(5): 919–930. ♦
- Aihara, K., Takabe, T., and Toyoda, M., 1990, Chaotic neural networks, *Phys. Lett. A*, 144:333–340.
- Elbert, T., Ray, W. J., Kowalik, Z. J., Skinner, J. E., Graf, K. E., and Birbaumer, N., 1994, Chaos and physiology: Deterministic chaos in excitable cell assemblies, *Physiol. Rev.*, 74:1–47. ♦
- Freeman, W. J., 2000, *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*, London: Springer-Verlag. ♦
- Fujii, H., Ito, H., Aihara, K., Ichinose, N., and Tsukada, M., 1996, Dynamical cell assembly hypothesis: Theoretical possibility of spatio-temporal coding in the cortex, *Neural Networks*, 9:1303–1350. ♦
- Glass, L., 2001, Synchronization and rhythmic processes in physiology, *Nature*, 410:277–284. ♦
- Kaneko, K., and Tsuda, I., 2000, *Complex Systems: Chaos and Beyond*, Berlin: Springer-Verlag. ♦
- Lewis, J. E., and Glass, L., 1992, Nonlinear dynamics and symbolic dynamics of neural networks, *Neural Computat.*, 4:621–642.
- Nara, S., Davis, P., Kawachi, M., and Totsuji, H., 1995, Chaotic memory dynamics in a recurrent neural network with cycle memories embedded by pseudo-inverse method, *Int. J. Bifurc. Chaos*, 5:1205–1212.
- Ott, E., Sauer, T., and Yorke, J. A. (Eds.), 1994, *Coping with Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, New York: Wiley. ♦
- Rabinovich, M. I., and Abarbanel, H. D. I., 1998, The role of chaos in neural systems, *Neuroscience*, 87:5–14.
- Segundo, J. P., Sugihara, G., Dixon, P., Stüber, M., and Bersier, L. F., 1998, The spike trains of inhibited pacemaker neurons seen through the magnifying glass of nonlinear analyses, *Neuroscience*, 87:741–766.
- Tsuda, I., 2001, Towards an interpretation of dynamic neural activity in terms of chaotic dynamical systems, *Behav. Brain Sci.*, 24:575–628. ♦
- van Vreeswijk, C., and Sompolinsky, H., 1996, Chaos in neuronal networks with balanced excitatory and inhibitory activity, *Science*, 274:1724–1726.

Cognitive Development

Yuko Munakata

Introduction

Two classic questions in the study of cognitive development are:

1. *Where does our knowledge come from?* Questions about origins (whether of knowledge, life, the universe, etc.) form the basis for some of the most interesting, challenging, and hotly debated issues in science and popular culture. In regard to the origins of knowledge, the debate has taken the form of nature versus nurture, and more recently of specifying the interactions between them.
2. *How does change occur?* This question builds on the question of origins. A complete developmental account must specify not only the beginnings (of knowledge, life, the universe, etc.), but also the mechanisms that govern change in the system. The

question of change is one of the most important yet relatively unanswered questions in the study of cognitive development. How do we progress, for example, from not speaking a word to communicating with short phrases to participating in lengthy, complex conversations?

Answering these two classic questions requires the ability to assess what is known at different points in development. For example, what do newborns, older infants, and toddlers know about the permanence of objects in the world? Specifying the time course of development of such knowledge could shed light on questions about the origins of this knowledge and how change occurs. Toward this end, researchers have developed a variety of clever techniques to assess knowledge in children of all ages, even preverbal

infants. However, children sometimes provide conflicting impressions of what they know across this variety of techniques, raising another important question:

3. *Why does what children know depend so much on how they are tested?* That is, why is children's knowledge so task dependent? Why do children pass one measure of knowledge with flying colors, while simultaneously failing another measure meant to tap the same knowledge? For example, infants as young as 3.5 months seem to understand the permanence of objects, as assessed by their looking times at events with occluded objects (Baillargeon, 1995). However, infants fail to understand object permanence for many more months, as assessed by whether or not they search for toys that are presented and then hidden.

Resolving these discrepancies in what children seem to know is critical for understanding the origins of knowledge and how change occurs.

Connectionist models have provided a useful tool for exploring these three fundamental questions, as well as many other facets of cognitive development (McClelland and Plunkett, 1995; Elman et al., 1996). Such models provide a useful complement to behavioral studies in general (O'Reilly and Munakata, 2000), and particularly in the study of development. This article describes contributions from connectionist models in exploring the origins of knowledge, mechanisms of change, and the task-dependent nature of children's knowledge. These models span the domains of perception, memory, language, problem solving, and rule use.

Origins of Knowledge

The debate about origins has played out in the domain of infants' perception of the world—whether they experience a “blooming, buzzing confusion,” as William James claimed, or come equipped to make sense of objects moving around them. Young infants appear sensitive to the continuity of object motion, the fact that objects move only on connected paths, never jumping from one place to another without traveling a path in between. For example, infants as young as 2.5 months look longer at events in which objects appear to move discontinuously than at events in which the objects move continuously (Spelke et al., 1992). Such longer looking times suggest that infants find the discontinuous events unnatural, and so possess some understanding of object continuity. What are the origins of such knowledge?

An understanding of object continuity may be part of our innate core knowledge (Spelke et al., 1992). However, the label of “innate” leaves many questions unanswered about the origins of knowledge. For example, how do infants come to understand object continuity, and what are the mechanisms underlying such understanding?

These questions were explored in a connectionist model initially developed to explore imprinting behavior in chicks, and object recognition more generally (O'Reilly and Johnson, 2002). The model viewed a simplified environment in which objects moved continuously. Based on this experience, the model developed receptive field representations of objects that encoded continuous locations in space, thereby demonstrating a sensitivity to object continuity.

What were the origins of the model's sensitivity to object continuity? First, the network had recurrent excitatory connections and lateral inhibitory connections in the hidden layers that allowed active units to remain active; specifically, active units continued to send activation to themselves via the recurrent excitatory connections, and they prevented other competing units from becoming active via the lateral inhibitory connections. Thus, when the network viewed an object, certain hidden units became active, and they tended to stay active even as the object moved around in the

input. As a result, the network represented the different input patterns associated with a moving object as the same object on the hidden layer. Second, the network learned according to a Hebbian learning rule, which led the model to associate this hidden unit pattern of activity with the object in different locations in the input. As a result, whenever the object appeared in any of these locations, the network came to activate the same hidden units, or the same object representation. In this way, with exposure to events in the world that conformed to the principle of continuity, the model developed receptive field representations of objects that encoded continuous locations in space, and so learned to “recognize” objects that moved continuously in its environment. Thus, the model provided a precise specification of the potential origins of knowledge about object continuity: a system with recurrent excitatory and lateral inhibitory connections and a Hebbian learning rule, and an environment with objects that move continuously.

One might describe this model in terms of an innate predisposition to understand the continuity of objects, given that the network was structured “from birth” with recurrent excitatory and lateral inhibitory connections and a Hebbian learning rule; all it needed was the typical experience of viewing objects moving continuously in its environment. However, again, it is not clear what benefits would be conferred by calling the developmental time-course of the model innate. In contrast, the benefits of the model should be clear in providing an explicit, mechanistic account of the potential origins of sensitivity to object continuity.

Mechanisms of Change

Most connectionist models, including the object continuity model, address some aspect of change. Networks change as activations are propagated through them, connection weights adjust during learning, and new forms emerge from the complex interactions of elements in the networks (as with the development of representations of object continuity). Here, we focus on models exploring two important issues regarding developmental changes: critical periods and stages.

Critical Periods

Humans mature at a slower rate than any other species. Counter-intuitively, this slow rate of development may provide the perfect biological environment (and a critical period) for mastering complex domains such as language. A connectionist model was developed to explore such critical periods in language learning (Elman, 1993). The model “heard” sentences of varying degrees of complexity, with the goal of predicting what word would come next at each point in a sentence. From this experience, the model learned to hold on to words it had heard recently in order to predict what might come next, allowing it to abstract the grammatical structure of the language. The model could thus correctly predict grammatically appropriate words within a sentence, even in complex sentences such as “Dogs see boys who cats who Mary feeds chase.” The model displayed a critical period, in that “young” networks were able to master the language input, while “older” networks were unable to.

Why did this model display a critical period? The young and older networks differed only in their working memory spans, or how many words they could keep in mind. Although children of different ages differ in many more ways, this single manipulation in the model allowed the potential contribution of working memory to be isolated and evaluated. The young network could keep only three or four words in mind, because its working memory was cleared after every three or four words that were heard; this span was increased gradually as the network aged. In contrast, for the older network, no such clearing of working memory ever occurred,

so it could keep more words in mind from the start of learning. Intuitively, one might have expected the older network to be more capable of learning. Instead, by being able to keep more in mind, the older network got bogged down in all of the information it was processing, and was unable to abstract the key kernels (of grammatical category, number, and verb argument type) for understanding the structure of its language. (Imagine trying to learn grammar by studying complex sentences like the “dogs” example above—there is quite a bit to get lost in!) In contrast, by being able to keep only a few words in mind, the young network was better able to focus on these key kernels. (Imagine trying to learn grammar by studying simple portions of complex sentences like the “dogs” example above, e.g., “Dogs see boys”—much less to get lost in.) After abstracting these key kernels, the young network was then able to build on this grammatical knowledge to process more complex sentences as its working memory span increased. Elman (1993) referred to this advantage of limited capacity as “the importance of starting small.”

The model thus instantiated an explicit, mechanistic account for the potential causes of critical periods in language learning, leading to the conclusion of the importance of starting small. The conclusions from the simulations are bolstered by similar conclusions from the detailed analyses of behavioral results, and the conclusions from the behavioral results are bolstered by the working implementation in a connectionist model of the somewhat counterintuitive theory.

Stages

Children appear to pass through qualitatively different stages in solving certain tasks. For example, in the balance-scale task, children view a scale with weights on each side at particular distances from the fulcrum, and they must decide which arm of the scale will fall when supports underneath the scale are released. Children initially answer these problems randomly, using no apparent rule about the physical properties of weight and distance to guide their decisions. They next attend only to the amount of weight on each side of the fulcrum (rule I), then include the distance of weights from the fulcrum if weights are equal on each side of the fulcrum (rule II), and eventually attend to both weight and distance information regardless of whether weights are equal on each side (rule III) (Siegler, 1976).

A connectionist model of the balance-scale task (McClelland, 1995) demonstrated how such stage-like progressions can result from small, successive adjustments to connection weights. The model received balance scale problems as patterns of activity corresponding to weight and distance information, and simulated children’s performance of progressing from no-rule behavior to rule I, rule II, and rule III behavior.

Why did the model display stage-like behavior? Using error-driven learning, the network’s initially random weights were slowly modified with each experience to reduce the discrepancies between the network’s predictions about balance scale problems and the actual outcomes. For some time the network’s answers were random, because the connections from input to hidden layer and from hidden layer to output layer were not yet meaningful. The network then began to develop representations of weight in its hidden layer. In one simulation, the earlier sensitivity to weight over distance arose because the network received greater exposure to problems where weight predicted the balance-scale outcome (reflecting the possibility that children have more experience with the effects of variations in weight than with the effects of variation in distance). In subsequent simulations, the earlier sensitivity to weight arose with equal exposure to the two cues, owing to the weight cue (a single piece of information) being more simple than the distance cue, which requires computing the relation between

two pieces of information. As the hidden layer representations of weight became more fully formed, such that units were becoming more distinct in their activation patterns, units could be more readily credited for their contribution to the network’s correct performance, or blamed for their contribution to the network’s errors. The connections from the hidden units to the output units (and from the input units to the hidden units) could then be adjusted appropriately to improve the network’s performance, producing a stage-like transition from random answers to rule I answers based on weight.

Sensitivity to the distance cue followed this same pattern of gradual development of hidden unit representations, facilitating credit and blame assignment and the adjustment of connections, leading to a stage-like transition to rules II and III. In this way, incremental weight adjustments in neural networks can result in small representational changes that then support relatively fast learning, producing stage-like behavior.

Task-Dependent Behavior

Children’s task-dependent behavior provides a challenge to stage theories, and to developmental theories more generally as discussed earlier. If children are in a given stage or have mastered a particular construct, why do they simultaneously pass and fail different measures of this knowledge? This section focuses on models that explore task-dependent behavior in two domains: memory for hidden objects (often studied under the rubric of a concept of object permanence) and perseveration (the repetition of behaviors that no longer make sense).

Object Permanence

Infants show an incredible range of task-dependent behaviors in their memory for hidden objects. One of the most salient is their failure to search for hidden objects for months after demonstrating sensitivity to hidden objects in their looking times (specifically, looking longer at impossible than at possible events with objects that become occluded; Baillargeon, 1995). Competing interpretations of such task-dependent behaviors have fueled controversies surrounding the origins of knowledge and the mechanisms of change in this domain.

A connectionist model demonstrated how object permanence knowledge could develop without being prespecified, and how such knowledge could lead to success in one task but not another (Munakata et al., 1997). The model viewed a simplified environment in which objects conformed to the principle of object permanence (e.g., disappearing from view behind occluders and reappearing after the occluders were removed). Based on this experience, in the absence of any explicit signal that hidden objects continued to exist, the model became sensitive to the permanence of objects, continuing to represent objects even after they were hidden.

What were the origins of the model’s knowledge of object permanence? As in the object recognition model described above, the object permanence model had recurrent excitatory connections that allowed active units to remain active. Like the language learning and balance scale models, the object permanence model also had a goal of predicting what would happen in its environment. Through error-driven learning, the network adjusted its weights if its predictions were incorrect, for example, if the network predicted that an occluded object would not reappear when the occluder was removed, and then the object did in fact reappear. So, when an object moved out of view, the network gradually learned to use its recurrent connections to maintain a representation of the object, allowing the network to accurately predict its environment (and the reappearance of such hidden objects).

Why was the model's knowledge of object permanence task dependent? The network's memory for hidden objects was graded in nature, gradually becoming more like the network's representations of visible objects. The network's perceptual prediction system could use weak memories of occluded objects to expect their reappearance, whereas the network's reaching system, with a delayed and slowed time course of development, could not use these weak memories of occluded objects to reach. This early inability to reach to occluded objects was not due simply to deficits in the reaching system, because this system was able to reach for visible objects (for which the network quickly developed strong representations). Thus, the strength of the graded representations was critical to the task-dependent behavior of the network. Moreover, the strengthening of the networks' memories alone was sufficient to allow the system to progress from initially reaching only for visible objects to then reaching for occluded objects as well. In this way, memory development may be critical to infants' increasing abilities to demonstrate sensitivity to hidden objects across a range of tasks.

Thus, with exposure to events that conformed to the principle of object permanence, the model provided an explicit, mechanistic account of the potential origins of our sensitivity to the permanence of objects, and of the task-dependent nature of infant knowledge in this domain.

Perseveration

Task-dependent behaviors may be most compelling when infants and children *perseverate*, repeating prepotent or habitual behaviors when they no longer make sense. In these cases, participants pass and fail different measures of the same knowledge in a single testing paradigm, at the same moment or very close in time. For example, as soon as infants will search for a toy that is presented and then hidden, they search perseveratively, continuing to reach back to old hiding locations after watching as the toy is hidden in a new location (Diamond, 1985). However, even as infants reach perseveratively, they occasionally gaze at the correct hiding location. Similarly, after 3-year-olds correctly sort cards according to one set of rules (e.g., with blue cards going into one pile and red cards going into another pile), most perseverate in sorting by this rule even after the rules have changed (e.g., to sort the cards by their shape rather than their color) (Zelazo, Frye, and Rapus, 1996). However, the children can correctly answer questions about the new rule they should be using, such as where trucks should go in the shape game. Thus, even as infants and children perseverate with their previous responses, they sometimes seem to indicate through other measures that they have some awareness of the correct response.

Connectionist models have been used to explore the mechanisms that lead to perseveration, support improvements with development, and yield task-dependent behaviors (Munakata, Morton, and Stedron, in press). These models simulate all of these aspects of performance based on a competition between "active" and "latent" memory traces. In the connectionist framework, active traces take the form of sustained activations of network processing units, and latent traces take the form of changes to connection weights between units. Latent traces build as a network repeats a behavior (e.g., searching in one hiding location, or sorting according to one rule), biasing the network to repeat that behavior. These latent traces are thought to be subserved by posterior cortical areas, and to develop quite early in life. The ability to maintain active traces of currently relevant information (e.g., a new hiding location or rule) appears to depend on gradual developments in the prefrontal cortex. The strength of these active traces is manipulated in the models in terms of the strength of recurrent connections within a layer, which allow active units to maintain their activity. Networks perseverate when latent traces for previously relevant information

are stronger than active memory traces for current information. Improvements in active memory abilities allow networks to overcome their prior biases when the task changes. Thus, perseveration may be understood in terms of neural specializations for different types of representations, which develop at different rates and compete with one another (see O'Reilly and Munakata, 2000, for further discussion of neural/computational trade-offs and their behavioral consequences).

Why do the models show task-dependent perseveration? As in the object permanence model, the strengthening of active representations is not an all-or-nothing process; these representations are graded in nature. The models can use weak active representations for certain tasks but not for others. For example, a network can use a weak active representation of a new card sorting rule to answer questions about the rule ("Where do trucks go in the shape game?"), because there is no competition from latent representations (e.g., to sort by color) in this task. No conflict needs to be resolved to answer this question, so weak representations suffice. In contrast, the network cannot use such weak representations to sort cards correctly. The inherent conflict in the sorting task (e.g., a card is both red and a truck) requires the active representation of the new rule to be strong enough to overcome the latent bias toward the old rule.

Similarly, a network can use weak memories for a toy's current hiding location for some tasks (gazing at the correct location) but not for others (reaching to the correct location). In this case, the gazing system updates more frequently than the reaching system (reflecting the general difference between these systems in infants, as well as in how often infants are allowed to use them in the typical toy-hiding task). Networks can thus gaze successfully with relatively weak memories, because they can update their gaze to the proper location frequently, thereby countering perseverative tendencies. In contrast, networks require stronger memories to reach to the correct location, because longer delays pass before they update their reaching, by which time they have become more susceptible to perseverative biases.

Discussion

Connectionist models have thus addressed several key issues in the study of cognitive development—the origins of knowledge, the mechanisms of change, and the task-dependent nature of developing knowledge—across a variety of domains (see also LANGUAGE ACQUISITION; PAST TENSE LEARNING; and DEVELOPMENTAL DISORDERS). In each case, the models provided explicit instantiations and controlled tests of specific theories of development, and allowed the exploration of complex, emergent phenomena. As a result, these models have provided insight into how sophisticated object-processing mechanisms might develop in the first months of life, why we might have critical periods for learning language, how some problem-solving skills can progress in stages, and how we can simultaneously pass and fail different measures of memory and rule use.

Many challenges remain, however, for connectionist models to provide more complete accounts of cognitive development. Two important areas for advancement are the following:

1. *Models that shape their environments.* Most connectionist models are "fed" their inputs, activity pattern after activity pattern, regardless of what they output. For example, language learning models hear sentence after sentence, no matter how poorly the models do in their comprehension. Similarly, models see the same objects in their environments regardless of how they behave toward the objects. Quite in contrast, even very young children shape their environments based on how they behave. For example, infants who reach for objects receive different

sensory information about the objects than infants who simply gaze at the objects. Children who show more advanced language skills may shape the ways in which caregivers speak to them. Capturing these important aspects of development requires models that shape their environments, such that their behaviors influence their subsequent inputs (e.g., Schlesinger and Parisi, 2001).

2. *Models that are more all-purpose.* Most connectionist models are designed for and tested on a single task within a single domain, such as the balance-scale task or the task of searching for hidden objects. A single model sees only this single task during the course of its development. Again, quite in contrast, children face a multitude of tasks across a range of domains each day. Capturing this important aspect of processing requires models that take in a variety of types of information and determine how to appropriately process them to perform successfully across a number of tasks (Karmiloff-Smith, 1992).

These kinds of developments in modeling, together with further explorations of the origins of knowledge, mechanisms of change, and task-dependent behaviors, should help to advance our understanding of fundamental issues in cognitive development.

Road Map: Psychology

Related Reading: Concept Learning; Developmental Disorders; Language Acquisition

References

- Baillargeon, R., 1995, Physical reasoning in infancy, in *The Cognitive Neurosciences* (M. Gazzaniga, Ed.), Cambridge, MA: MIT Press.
- Diamond, A., 1985, Development of the ability to use recall to guide action, as indicated by infants' performance on *AB*, *Child Dev.*, 56:868–883.
- Elman, J. L., 1993, Learning and development in neural networks: The importance of starting small, *Cognition*, 48:71–99.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., 1996, *Rethinking Innateness: A Connectionist Perspective on Development*, Cambridge, MA: MIT Press. ♦
- Karmiloff-Smith, A., 1992, *Beyond Modularity: A Developmental Perspective on Cognitive Science*, Cambridge, MA: MIT Press. ♦
- McClelland, J. L., 1995, A connectionist perspective on knowledge and development, *Developing Cognitive Competence: New Approaches to Process Modeling* (T. J. Simon and G. S. Halford, Eds.), Hillsdale, NJ: Erlbaum, pp. 157–204. ♦
- McClelland, J. L., and Plunkett, K., 1995, Cognitive development, in *The Handbook of Brain Theory and Neural Networks*, 1st ed. (M. A. Arbib, Ed.), Cambridge, MA: MIT Press, pp. 193–197.
- Munakata, Y., McClelland, J. L., Johnson, M. H., and Siegler, R., 1997, Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks, *Psychol. Rev.*, 104:686–713.
- Munakata, Y., Morton, J. B., and Stedron, J. M., in press, The role of prefrontal cortex in perseveration: Developmental and computational explorations, in *Connectionist Models of Development* (P. Quinlan, Ed.), East Sussex: Psychology Press. ♦
- O'Reilly, R. C., and Johnson, M. H., 2002, Object recognition and sensitive periods: A computational analysis of visual imprinting, in *Brain Development and Cognition: A Reader*, 2nd ed. (M. H. Johnson, Y. Munakata, and R. O. Gilmore, Eds.), Oxford, Engl.: Blackwell, pp. 392–413.
- O'Reilly, R. C., and Munakata, Y., 2000, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, Cambridge, MA: MIT Press. ♦
- Schlesinger, M., and Parisi, D., 2001, The agent-based approach: A new direction for computational models of development, *Dev. Rev.*, 21:121–146.
- Siegler, R., 1976, Three aspects of cognitive development, *Cognit. Psychol.*, 8:481–520.
- Spelke, E., Breinlinger, K., Macomber, J., and Jacobson, K., 1992, Origins of knowledge, *Psychol. Rev.*, 99:605–632.
- Zelazo, P. D., Frye, D., and Rapus, T., 1996, An age-related dissociation between knowing rules and using them, *Cognit. Dev.*, 11:37–63.

Cognitive Maps

Nestor A. Schmajuk and Horatiu Voicu

Introduction

According to Tolman (1932), animals acquire the *expectancy* that the performance of response R1 in a situation S1 will be followed by a change to situation S2 (S1-R1-S2 expectancy). Tolman hypothesized that a large number of local expectancies can be combined, through inferences, into a *cognitive map*. Tolman proposed that place learning, latent learning, detours, and shortcuts illustrate animals' capacity for reasoning by generating inferences. In place learning, animals learn to approach a given spatial location from multiple initial positions, independently of any specific set of responses. In latent learning, animals are exposed to a maze without being rewarded at the goal box (Figure 1). When a reward is later presented, animals demonstrate knowledge of the spatial arrangement of the maze, which remains "latent" until reward is introduced. Detour problems are those in which animals have to take an alternative, longer path to the goal. Shortcuts are those problems in which animals can take an alternative, shorter path to the goal.

When seeking reward in a maze, organisms compare the expectancies evoked by alternative paths. For Tolman, *vicarious trial-and-error behavior*, i.e., the active scanning of alternative pathways at choice points, reflects the animal's generation and comparison of different expectancies. At choice points, animals

sample different stimuli before making a decision. For example, a rat often looks back and forth between alternative stimuli before approaching one or the other. According to Tolman's *stimulus-approach* view, organisms learn that a particular stimulus situation is appetitive, and therefore it is approached. It has been suggested that, in the presence of numerous intra- and extramaze cues, animals typically learn to approach a set of stimuli associated with reward and to avoid a set of stimuli associated with punishment. However, in a totally uniform environment, animals learn to make the correct responses that lead to the goal.

Maze Learning

Deutsch (1960) presented a formal description of cognitive mapping that incorporates many of Tolman's cognitive concepts. Deutsch assumed that when an animal explores a given environment, it learns that stimuli follow one another in a given sequence. Internal representations of the stimuli are linked together in the order the stimuli are encountered by the animal. Deutsch suggested that a given drive activates its goal representation, which in turn activates the linked representations of stimuli connected to it. When the animal is placed in the maze, it searches for stimuli stimulated by the goal representation. When a stimulus activated by the goal

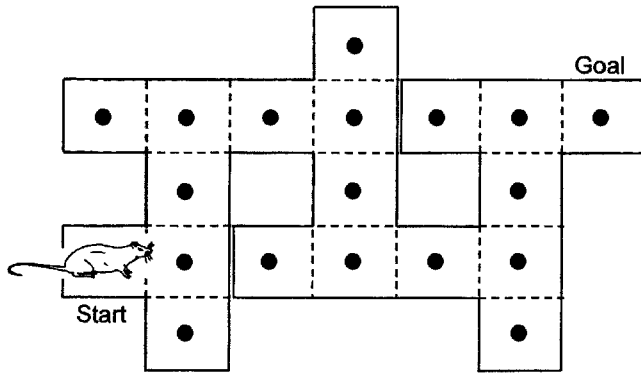


Figure 1. Latent learning. Schematic top view of a multiple T-maze. Places used in Figure 3 are represented by squares in broken lines, with the centers indicated by solid circles.

is perceived, the activation of lower stimuli in the chain is cut off, and behavior is controlled by the stimulus that is closer to the goal. Deutsch's theory can account for latent learning in the following terms. When animals are exposed to the maze without being rewarded at the goal box, they learn about the connections between different places in the maze. When a reward is subsequently presented, activation of the goal representation activates the representations of the stimuli connected to it.

Milner (1960) proposed a system capable of building a Tolmanian spatial map and of using it to control the animal's movements in a spatial environment. The model has nodes that are active only when a particular response (R_i) has been made in a particular location (S_j). The output of these nodes can be associated with nodes representing the location (S_k) that results from making response R_i at location S_j . When the organism is placed in location S_j , random responses are generated. When response R_i appears, the node with inputs S_j and R_i is active and, in turn, activates the node representing S_k . If S_k is associated to an appetitive stimulus, it activates a mechanism that holds R_i in the response generator, and location S_k can be reached. Some views of hippocampal function (e.g., McNaughton, 1989) suggest that S_j - R_i - S_k associations would be stored in the hippocampus (see HIPPOCAMPUS: SPATIAL MODELS).

Lieblich and Arbib (1982) addressed the question of how animals build a cognitive model of the world. Lieblich and Arbib posited that spatial representations take the form of a directed graph in which nodes represent *recognizable situations* in the world. In Lieblich and Arbib's scheme, a node represents not only a place, but also the motivational state of the animal. Consequently, a place in the world might be represented by more than one node if the animal has been there under different motivational states. Lieblich and Arbib postulated that each node in the world graph is labeled with *learned vectors*, \mathbf{R} , that reflect the drive-reduction properties for multiple motivations of the place represented by that node. Based on the value of \mathbf{R} , the animal moves to the node most likely to reduce its present drive. More recently, Guazelli et al., (1998), following O'Keefe and Nadel's (1978) notions, integrated the world graph model, used for map-based navigation, with a "behavioral orientation" model, used for route navigation, in order to explain experimental data in normal and fornix-lesioned rats.

Hampson (1990) analyzed maze navigation in terms of a model that combines stimulus-response and stimulus-stimulus mechanisms into a system capable of assembling action sequences in order to reach a final goal.

Schmajuk and Thieme (1992) presented a real-time, biologically plausible neural network approach to purposive behavior and cog-

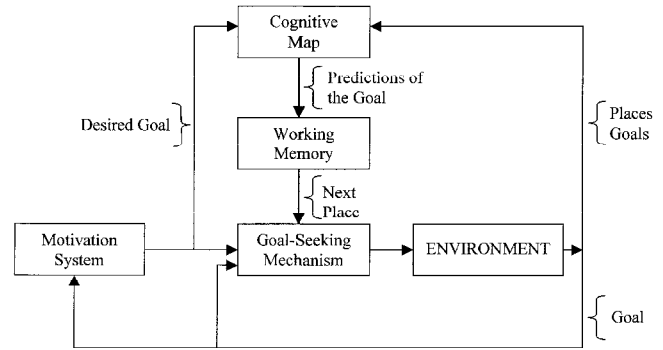


Figure 2. The model. Interactions between the motivation system, cognitive map, working memory, goal-seeking mechanism, and environment. (After Voicu and Schmajuk, 2001a.)

nitive mapping. This biologically plausible theory includes (1) an action system, consisting of a goal-seeking mechanism with goals set by (2) a motivation system; (3) a cognitive system, in which a neural network implements a cognitive map; and (4) a working memory, where the readings of the cognitive map are temporarily stored (Figure 2).

The goal-seeking mechanism in the action system changes from exploratory behavior to approach behavior when either (1) the goal is found or (2) one place in the cognitive map generates a prediction of the goal that is stronger than the predictions generated by all other alternative places.

The cognitive map is a *topological map* that represents the adjacency, but not distances or directions, between places, as well as the associations between places and goals. Figure 3 shows a heteroassociative network (Kohonen, 1977) capable, through recurrent connections, of cognitive mapping. The network includes three types of inputs: current places, neighboring places, and goals.

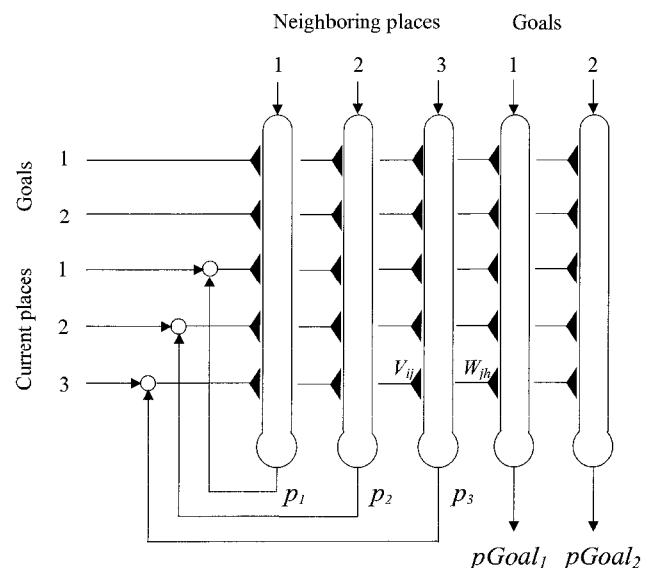


Figure 3. Cognitive Map. The prediction of neighboring place j , p_j , is fed back into the neuron representing place j as a current place. V_{ij} : association between place i and place j . W_{ih} : association between goal h and place i . W_{ih} : association between place i and goal h . g_h : activation of goal h . Arrows denote fixed excitatory connections; triangles denote variable excitatory connections. (After Voicu and Schmajuk, 2001a.)

rent places and neighboring places are defined by a system that determines the position of the animal in space. Goals represent a feature (e.g., food, unexamined places) that, under the appropriate motivation (e.g., hunger, exploration), the agent wants to approach.

Associations between place i and place j , V_{ij} , are the elementary internal learned representations of the links in the external world. These associations are stored in modifiable synapses, indicated by triangles in Figure 3. Whereas a positive V_{ij} association means that place j can be accessed from place i , a positive V_{ji} association means that place i can be accessed from place j . In both cases, $V_{ij} = V_{ji} = 0$ mean that each place cannot be accessed from the other. At the beginning of the exploration, all adjacent places are assumed to be linked in the cognitive map (all $V_{ij} = 1$). As the agent explores the environment, connections V_{ij} are modified in order to reflect the real environment. "Trodden" links between places are represented by stronger connections ($V_{ij} = V_{ji} > 1$) than unexplored links.

At the beginning of the exploration, each place is designated as "unexamined," $W_{j,GOALh} = W_{GOALh,j} = 1$. When place j can be accessed and occupied by the agent, it changes its status to "examined," $W_{j,GOALh} = W_{GOALh,j} = 0$, meaning that it is no longer a goal for exploration. When place j , adjacent to place i currently occupied by the agent, can be examined, place j also changes its status to examined, $W_{j,GOALh} = W_{GOALh,j} = 0$, even if it cannot be entered.

The cognitive map combines multiple associations V_{ij} to infer spatially remote locations. This is achieved by recurrently reinjecting the signal representing neighboring place j (as predicted by place i according to V_{ij}) into the representation of current place j . Current place j now predicts neighboring place k according to V_{jk} , and the signal representing neighboring place k is reinjected into the representation of current place k . At each reinjection, the signal representing a neighboring place is attenuated, for instance, in half. The process is halted when the representation of the remote position is eventually activated.

The network successfully describes how rats demonstrate latent learning, detour behavior, and place learning. In order to ascertain the power of the network in problem solving, Schmajuk and Thieme (1992) applied the network to the Tower of Hanoi task. Simulations show that the network takes a few trials to solve the problem in the minimum number of movements.

Schmajuk, Thieme, and Blair (1993) incorporated a route system to the network described by Schmajuk and Thieme (1992). Whereas the cognitive map stores associations between places and reward, the route system establishes associations between cues and reward. Both systems compete with each other to establish associations with the reward, with the cognitive system generally overshadowing the route system. In agreement with O'Keefe and Nadel (1978), after hippocampal lesions (see HIPPOCAMPUS: SPATIAL MODELS), animals navigate through mazes by making use of the route system.

Recently, Voicu and Schmajuk (2001a, 2001b) introduced some modifications to the Schmajuk and Thieme (1992) model. First, whereas the early model assumed no a priori knowledge of the space to be explored, the modified model assumes a representation of the environment as a set of potentially connected and unexamined locations, each approximately the size of the footprint of the agent. Second, instead of random exploratory behavior, the model assumes that a curiosity drive guides the animal to examine all unvisited places. Third, whereas in the original model the decision of what place to move to next was based on a comparison of the predictions of the goal when each of the alternative places is briefly entered, in the new model this decision is based on a comparison of the activation of each of the alternative places when the goal is activated. This approach is similar to that described by Deutsch (1960) and subsequently used by Mataric (1991) and Reid and

Staddon (1998). Fourth, the model differentiates between links that connect places but have not been traversed by the agent, and those that have. These links are part of what is referred to as "trodden path."

In addition to latent learning, detour behavior, and place learning, the new model (1) performs a thorough and efficient exploration of the environment, (2) describes shortcuts, and (3) generates novel predictions for a case in which animals have to circumvent an obstacle placed on their path to the goal and then either retake their usual route (detour) or advance directly to the goal (shortcut).

As mentioned above, in order to plan the shortest path between any two places, the model presented by Voicu and Schmajuk (2001a, 2001b) spreads activation between the representations of the goal and the current location of the animal. The resulting activities are stored in working memory. However, two problems might arise when navigating in a large environment. One, the capacity of working memory might be too small to store the activity of all intermediate places. Two, the attenuation suffered by the spreading activation is too large to reach the representation of the location of the animal. These potential problems call for the use of a hierarchical cognitive map (Voicu and Schmajuk, 2002).

In the hierarchical map, the environment is represented at multiple levels. At the highest level, the environment is divided in a number of regions equal to the size of working memory. The size of the working memory is such that the activation spreading through it can reach and activate the most remote spatial representations. At the lowest level, the environment is divided in parts equal in size to that of the footprint of the organism. Between the highest and the lowest level, each part of the previous level is divided in a number of parts equal to the size of the working memory. Path planning starts at the level that contains the points between which navigation is desired and ends at the lowest level, at which motion is produced.

According to the hierarchical cognitive model, and in agreement with experimental data, when information about the relationship between two places that belong to two different regions (e.g., Seattle and Montreal) is lacking, humans wrongly rely on the relation between the regions where those places are located (e.g., United States and Canada). Also in agreement with experimental results, the hierarchical cognitive model suggests that spatial memory is organized in a hierarchical fashion even when objects are uniformly distributed in space. Finally, experimental results suggest that hierarchical spatial representations support semantic priming, a result that the hierarchical model can explain in terms of one item activating the representation of a second one in the cognitive map, and therefore decreasing the response time to the second item.

Discussion

In the last decades, Tolman's (1932) ideas about cognitive maps have been refined and given mechanistic embodiments. In recent models (e.g., Voicu and Schmajuk, 2001a), Tolman's vicarious trial-and-error behavior has been regarded as reflecting the animal's comparison, but not the generation, of different expectancies. Expectancies are generated by activating the representation of the location of the goal with the motivation system. That activation is subsequently spread over the cognitive map until the representation of the location of the agent becomes active. At choice points, animals make a decision after sampling the intensity of the activation elicited by the different alternative paths. Several models (Mataric, 1991; Reid and Staddon, 1998; Voicu and Schmajuk, 2001a) still use Tolman's stimulus-approach view and assume that animals approach the place with the strongest appetitive activation, thereby performing a gradient ascent toward the goal.

Interestingly, Tolman (1932, p. 177) suggested that the relations between initial and goal positions can be represented by a directed

graph, and called this graph a means-end field. Current artificial intelligence theories describe problem solving as the process of finding a path from an initial to a desired state through a directed graph (see COMPETITIVE QUEUING FOR PLANNING AND SERIAL PERFORMANCE).

In addition to storing the representation of the environment in terms of the contiguity between places (Schmajuk and Thieme, 1992), cognitive maps can store information about differences in height and in the type of terrain between adjacent places, contain a priori knowledge of the space to be explored (Reid and Staddon, 1998; Voicu and Schmajuk, 2001a), distinguish between roads taken and those not taken (Voicu and Schmajuk, 2002), and keep track of which places have been examined.

Schmajuk et al. (1993) used Kohonen's (1977) two-layer associative network to mechanistically implement the cognitive map. Networks with more than two layers can also be used to represent not only the contiguity between places, but also the relative position of those places.

Recently, Voicu and Schmajuk (2001b) introduced the idea of hierarchical cognitive maps in which the environment is represented at multiple levels. At each level, the environment is divided into a number of parts equal to the size of working memory. In contrast to their nonhierarchical counterparts, hierarchical maps can plan navigation in large environments, use a smaller number of connections in their networks, and have shorter decision times.

Road Map: Psychology

Related Reading: Hippocampus: Spatial Models; Potential Fields and Neural Networks

References

- Deutsch, J. A., 1960, *The Structural Basis of Behavior*, Cambridge, Engl.: Cambridge University Press.
- Guazelli, A., Corbacho, F. J., Bota, M., and Arbib, M. A., 1998, Affordances, motivations, and the world graph theory, *Adapt. Behav.*, 6:435–471.
- Hampson, S. E., 1990, *Connectionistic Problem Solving*, Boston: Birkhauser.
- Kohonen, T., 1977, *Associative Memory*, Berlin: Springer-Verlag.
- Lieblich, I., and Arbib, M. A., 1982, Multiple representations of space underlying behavior, *Behav. Brain Sci.*, 5:627–659. ♦
- Mataric, M. J., 1991, Navigating with a rat brain: A neurobiologically-inspired model of robot spatial representation, in *From Animals to Animals I*, Cambridge, MA: MIT Press, pp. 169–175.
- McNaughton, B. L., 1989, Neuronal mechanisms for spatial computation and information storage, *Neural Connections and Mental Computations* in (L. Nadel, L. Cooper, P. Culicover, and R. Harnish, Eds.), New York: Academic Press.
- Milner, P. M., 1960, *Physiological Psychology*, New York: Holt, Rinehart, and Winston.
- O'Keefe, J., and Nadel, L., 1978, *The Hippocampus as a Cognitive Map*, Oxford, Engl.: Clarendon Press.
- Reid, A. K., and Staddon, J. E. R., 1998, A dynamic route finder for the cognitive map, *Psychol. Rev.*, 105:585–601.
- Schmajuk, N. A., and Thieme, A. D., 1992, Purposive behavior and cognitive mapping: An adaptive neural network, *Biol. Cybern.*, 67:165–174. ♦
- Schmajuk, N. A., Thieme, A. D., and Blair, H. T., 1993, Maps, routes, and the hippocampus: A neural network approach, *Hippocampus*, 3:387–400.
- Tolman, E. C., 1932, Cognitive maps in rats and men, *Psychol. Rev.*, 55:189–208. ♦
- Voicu, H., and Schmajuk, N. A., 2001a, Three-dimensional cognitive mapping with a neural network, *Robot. Auton. Syst.*, 35:21–35.
- Voicu, H., and Schmajuk, N. A., 2001b, Hierarchical cognitive maps, presented at the Fifth International Conference on Cognitive and Neural Systems, Boston, MA, May 29.
- Voicu H., and Schmajuk, N. A., 2002, Exploration, navigation and cognitive mapping, *Adapt. Behav.*, 8:207–223.

Cognitive Modeling: Psychology and Connectionism

Amanda J. C. Sharkey and Noel E. Sharkey

Introduction

Connectionism has had a major impact on psychology and cognitive modeling. In the next section we consider this influence, identifying four typical features of connectionist models of cognition:

1. They can be used both to model cognitive processes and to simulate the performance of tasks.
2. Data analogues can be derived from the models in such a way as to provide a good fit to data from cognitive psychology experiments.
3. Unlike traditional computational models, connectionist models are not explicitly programmed by the investigator.
4. They encourage the development of new accounts of empirical data.

Having identified these features, we point out that although the connectionist stance has some real benefits, it would be unwise to assume that it has been fully developed. There are aspects of it that still need to be understood and clarified. For instance, establishing equivalences between the performance of a net and the tasks it simulates involves a number of assumptions. At the same time, important aspects of the performance of a connectionist net are

controlled by the researcher, by means of *extensional programming*. In particular, decisions are made about the selection and presentation of training data. This means that the achievement of a good fit to the psychological data depends both on the way in which analogues to the data are derived and on the results of extensional programming. Nonetheless, a major advantage of a connectionist approach to the modeling of psychological processes is that it has encouraged the development of new accounts of old data in a number of domains.

We subsequently consider the potential of connectionist computation for the development of a new theory of cognition. Rather than modeling the details of psychology experiments, the research efforts of a new breed of “cognitive connectionists” have been directed toward computational experiments. These experiments have demonstrated the ability of connectionist representations to provide a viable and different account of important characteristics of cognition (compositionality and systematicity), previously assumed to be the exclusive province of the classical symbolic tradition. Although connectionism has only an abstract relationship to neural processing, the promise here is that the use of brain-style computation will enable steps to be taken toward a unified account of how a mind emerges from a brain. Of course, it is not yet known whether connectionism will ultimately be able to account for all aspects of

cognition. We discuss the likelihood that a more complete connectionist account of cognition will almost certainly involve the coordination of connectionist modules, and we point to the beginning of a debate over the extent to which such coordination requires a classical architecture or could be accomplished in a purely connectionist manner.

Models of Cognition

In 1986, a two-volume edited work by McClelland and Rumelhart presented a number of connectionist models of different aspects of cognition that had been *trained* by exposure to samples of the required tasks. This work was indebted to earlier pioneering neural network research related to cognitive processing and memory (e.g., Anderson, 1972), but it was these two volumes that set the agenda for connectionist cognitive modelers and offered a methodology that has become the standard. Connectionist cognitive models are now legion. The domains that have been simulated include, in memory, retrieval and category formation; in language, phoneme recognition, word recognition, speech perception, acquired dyslexia, and language acquisition; and in vision, edge detection and object and shape recognition.

In this article we have chosen to discuss only supervised learning techniques, as these techniques have been most commonly used in connectionist cognitive modeling. In the simplest case of supervised learning, a net consists of a set of input units, a layer of hidden units, and a set of output units, each layer being connected to the next via modifiable weights. This is a feedforward net. When the net is trained on a set of input-output pairs, the weights are adjusted via a learning algorithm (e.g., backpropagation) until the required output is produced in response to each input in the training set. When tested on a set of previously unseen inputs, the net will, to a greater or lesser extent, display an ability to generalize, that is, to go beyond the data it was trained on, and to produce an appropriate response to some of the test inputs. The ability of the net to generalize depends on the similarity between the function extracted as a result of the original training and the function that underlies the test set. If the training set was sufficiently representative of the required function, generalization results are likely to be good. Where the inputs and outputs of such a net are given an interpretation relevant to the performance of a cognitive task, the net may be seen as a model of that task.

In addition to their basic architecture and mode of operation, it is also possible to identify four typical features of connectionist models of cognition that, in combination, account for much of the popularity of the approach they exemplify: (1) They can be used both to model mental processes and to *simulate* the actual behavior involved. (2) They can provide a “good fit” to the data from psychology experiments. (3) The model, and its fit to the data, is achieved without explicit programming, or “handwiring.” (4) They often provide new accounts of the data. We discuss each of these features in turn.

The first two features, namely, the way in which connectionist nets can both provide a model of a cognitive process and simulate a related task, and their ability to provide a good fit to the empirical data, combine some of the characteristics of two earlier routes to modeling. One of these routes, taken by the cognitive psychology community, involved building models that could account for the results from psychology experiments with human subjects but did not incorporate simulations of experimental tasks. The second route, followed by the artificial intelligence (AI) community, was to build computer models that actually performed the task in ways that resembled human performance, without regard to detailed psychological evidence. The connectionist approach, as described here, provides the benefits both of simulating the performance of

human tasks and, at the same time of fitting the data from psychological investigations.

As an example of the latter, consider Seidenberg and McClelland’s (1989) model of word pronunciation. This model simulates a number of experimental behaviors, including pronunciation of novel items, differences in performance on lexical decision and naming tasks, and ease of pronunciation in relation to variables such as frequency of occurrence, orthographic redundancy, and orthographic-phonological regularity. At the same time, the model is generally accepted as providing a good fit to the experimental data. This is achieved by deriving an analogue to the data from the performance of the model. In the case of Seidenberg and McClelland’s account of lexical decisions, the comparison with lexical decision data is accomplished by computing an “orthographic error score,” based on the sum of squares of the differences between the feedback pattern computed by the network and the actual input to the orthographic units. Here a low error score is equated with a fast reaction time. Others have assumed different data analogues. For example, Plaut et al., (1996) present an attractor network that accounts for latency data in time to settle on a response.

The third typical feature of connectionist models of cognition is that the model and its fit to the data are achieved without explicit handwiring. This feature can be favorably contrasted to the symbolic programming methodology employed in AI, where the model must be programmed step by step, leaving room for ad hoc modifications and kludges.

The fourth feature, perhaps the most scientifically exciting, is the possibility of providing a novel explanation of the data. In their model of word pronunciation, Seidenberg and McClelland (1989) showed that their network provided an integrated (single mechanism) account of data on both regular and exception words; by contrast, the old cognitive modeling conventions had forced an explanation in terms of a dual route. A criticism of the Seidenberg-McClelland model was that it failed to provide an account of the reading of nonwords. However, a more recent formulation of the model (Plaut et al., 1996) shows that a single mechanism is able to provide an account of the basic patterns of word and nonword reading within a single mechanism. (See Christiansen, Conway, and Curtin, 2000, for a more recent example of a connectionist model that exhibits rule-like behavior from a single mechanism.)

On first consideration, the four features discussed above seem to provide support in favor of a connectionist approach to cognitive modeling, as opposed to the approaches taken in the two preceding routes to cognition, by AI and cognitive modelers. When the result has been the development of new, and simpler, accounts of the data, there are obvious benefits. The connectionist approach would also seem to be preferable to more symbolic approaches inasmuch as it makes it possible both to simulate the tasks and to provide a good fit to the data, and to do this without having to program these abilities in hand. However, we must tread warily here in claiming a clear distinction between the connectionist approach and those that preceded it.

Let us first reexamine the idea that connectionist network models are not explicitly programmed by hand. Although this statement is strictly true, a number of decisions that affect the subsequent performance of the model still have to be made by the researcher. The following factors, for example, both govern the creation of a model and are in the hands of the researcher: (1) the architecture of the net and its initial structure, (2) the learning technique, (3) the learning parameters (e.g., the learning rate), (4) the input and output representations, and (5) the training sample. The term *extensional programming* can be used to refer to the manipulation of these factors. By means of extensional programming, the clever experimenter can determine the ultimate form of a model and hence its performance in terms of any data set.

Control of the content and presentation of the training sample is an important aspect of extensional programming. Its potential influence on the performance of a connectionist model can be illustrated by considering some of the criticisms of McClelland and Rumelhart's past tense model (see PAST TENSE LEARNING). Their model is said to mirror several aspects of human learning of verb endings. However, Pinker and Prince (1988) pointed out that the experimenters had unrealistically tailored the environment to produce the required results and that the results were an artifact of the training data. More specifically, the results indicated a U-shaped curve in the rate of acquisition, as occurs with children, but, Pinker and Prince argued, this curve occurred only because the net was exposed to the verbs in an unrealistically structured order. These criticisms have largely been answered by further research, but the point remains. Selection of the input, and control of the way that it is presented to the net, affect what the net learns. A similar argument can be made about the selection of input representations. Plaut et al. (1996), in their model of normal and impaired word reading, explicitly discuss the role of their chosen input representation in achieving their results, claiming that is the change in input representation from the earlier Seidenberg and McClelland model that accounts for an improved ability to handle nonword data.

If the performance of a connectionist net is determined by its extensional programming, this means that there are a number of parameters that can be altered until a good data fit is achieved. Moreover, accepting that a good fit to the data has been achieved requires accepting a number of other assumptions about the way that the relevant tasks have been simulated, the consequent way that analogues to the data have been derived from the performance of the model, and so on. Thus, when a model is said to explain the empirical data from lexical decision tasks, the model does not actually perform a lexical decision task as humans do, outputting a yes when the input is a word and a no when it is not. An equivalence must be drawn between the task as it is performed by humans and the actual input-output relationships encoded by the net. In fact, the actual task in question is rarely performed by the model. What is performed is often something that approximates the task and is assumed to capture its essential characteristics. Thus, in a model of word pronunciation, the output of the model does not take the form of spoken words but rather a set of phonological features. Here the approximation is fairly straightforward, but in other cases the relationship is less obvious. Ratcliff (1990), for instance, equates the phenomenon of "recognition memory" with that of auto-association, assuming that the ability of a net to reproduce its inputs as outputs corresponds to the human ability to recognize inputs that have been seen before.

From the foregoing discussion it is apparent that some of the advantages of connectionist models of cognition are not entirely straightforward. Nonetheless, their potential for developing new accounts that go beyond the data is an important justification for their employment. Much of the excitement in the psychology community has been about their ability to handle apparently rule-governed phenomena without any explicit rules, and in particular to account for the processing of regular and exceptional material. For example, learning to produce the appropriate past tense form of verbs was always considered a rule-governed phenomenon until McClelland and Rumelhart (1986) showed that the behaviors could be trained in a model that did not contain any explicit rules. The different style of computation afforded by connectionism has also been the focus of much debate. The issue being debated is the extent to which connectionism is able to support the operations required of a theory of mind, namely, structure-sensitive operations. This debate is examined in the next section.

An Emerging Theory of Cognition

The past two decades saw considerable discussion about whether or not the new connectionism actually constituted a paradigm shift.

Whatever the eventual outcome of this discussion may be, it is clear that the adoption of connectionist modeling techniques has led to changes. One such change has been the emergence of a new breed of cognitive modelers who are concerned to explore the implications of connectionist computation for a new theory of cognition, with the ultimate aim of providing an account of the way in which a mind emerges from a nervous system. The methodology employed by these "cognitive connectionists" represents a move away from the detailed modeling of cognitive psychology experiments toward more abstract computational experimentation. This should broaden the scope of their investigations, for although there are benefits to models that provide a new explanation of a wide spectrum of data, the maintenance of a close relationship between model and data can limit investigations to previously identified questions from an older paradigm.

The promise of a new connectionist theory of cognition is that it will further our understanding of the relationship between brain and mind. However, this does not necessarily mean that such a theory will incorporate details of neural processing. It is important to make a clear distinction between the use of *brain-style* computation and neuropsychological modeling. Although connectionist models are sometimes described as being "neurally inspired," their relationship to neural processing is delicate and tenuous. In a model of cognitive processes, it is unlikely that the computation performed corresponds to what goes on at the neural level. Indeed, it has been suggested (Edelman, 1987) that it may take units on the order of several thousand neurons to encode stimulus categories of significance to animals. Clearly, when the inputs to a net are entities like noun phrases or disease symptoms or even the phonological representations of words, the inputs cannot be equated with neural inputs but must represent substantially preprocessed stimuli. In fact, there are few cases in which actual facts about the nervous system are used to constrain the architecture and design of a model. It is, of course, important to build computational models of real brain circuits in all of their glorious detail. But if one is concerned with cognition rather than with the details of neural processes, an appropriate research strategy is to use broader brush strokes, relying on computational abstractions.

In support of a movement away from the details of neurophysiology and away from psychology experiments, we can cite an example of progress from the history of connectionism itself. Arguably, it was two major simplifying assumptions made by McCulloch and Pitts that enabled them to develop their theoretical analyses without getting bogged down in the physical and chemical complexity of the nervous system. Their first simplification was based on the observation that neural communication is thresholded, and thus the spike activation potential is all or none: it either fires or does not fire. Thus, the neuron could be conceived of as a binary computing device. The second simplification was to view synapses as numerical weightings between simple binary computing elements. This meant that computation proceeded by summing the weighted inputs to an element and using the binary threshold as an output function (see Sharkey and Sharkey, 1994, for historical details). The position of the cognitive connectionist described here is therefore that we can best proceed by being constrained by simplifying but principled assumptions about neural computation and cognition.

Nonetheless, if the use of connectionist computation, as opposed to symbolic computation, is to lead to the development of a new theory of cognition, connectionist computation must first be shown to be capable, in principle, of supporting higher mental processes, and to do so in a novel manner; mere implementation of symbolic architectures will not do. The question of whether or not connectionism is capable of supporting a cognitive architecture has mainly been addressed in the context of discussions about the novelty and value of connectionist representation. This question is considered

below. (See also STRUCTURED CONNECTIONIST MODELS; CONNECTIONIST AND SYMBOLIC REPRESENTATIONS; COMPOSITIONALITY IN NEURAL SYSTEMS.)

One of the properties of uniquely connectionist representations, as identified by N. E. Sharkey (1997), is that they are distributed and nonsymbolic. Proponents of the classical symbolic tradition have claimed that such representations are in principle incapable of supporting a cognitive architecture, because in order to account for the systematic nature of human thought, representations must be able to support structure-sensitive processes, and this requires compositional representations. The assumption is that there is only one kind of compositionality, namely, the concatenative compositionality of symbolic strings. This permits structure-sensitive operations, because in their mode of combination, the constituents of complex expressions are tokened whenever the complex expression is tokened. For example, in order to develop an expression from a sentence like "John kissed Mary," arbitrary symbols representing the constituents JOHN, KISSED, and MARY are combined in a contextually independent concatenation to produce the propositional representation KISS (JOHN, MARY). Whenever this latter complex expression is tokened, its constituents, KISS, MARY, and JOHN, are also tokened. This makes the manipulation of the representations by a mechanism sensitive to the syntactic structure resulting from concatenative compositionality relatively easy.

Distributed representations, on the other hand, do not exhibit this kind of compositionality. Instead, cognitive connectionists have identified an alternative form of compositionality, one that has been described as merely functional, nonconcatenative compositionality (see COMPOSITIONALITY IN NEURAL SYSTEMS). Distributed representations combine tokens without those tokens appearing in the complex expression, since the tokens of the input constituents are destroyed in their combination. The point is that such representations can still be shown to be *functionally* compositional, because there exist general and reliable procedures for combining constituents into complex expressions and for decomposing those expressions back into the constituents. It is possible, for example, to encode simple syntactic trees in terms of connectionist distributed representations, and to decode them back into the same syntactic trees (Pollack, 1990). Thus the constituents of the tree have been combined into a form of representation that is nonconcatenative but that preserves the necessary information.

A considerable and growing body of research has shown that not only are distributed representations compositional, they can also enable *systematic* structure-sensitive operations. Chalmers (1990), for example, found that it was possible to use connectionist nets to transform distributed representations for active sentences into distributed representations for passive sentences. Thus, distributed representations allow at least a limited form of systematicity without emergence onto the symbol surface. Moreover, this is not just an example of old wine in new bottles, a mere implementation of the classical account. These uniquely connectionist representations operate in a different manner.

N. E. Sharkey (1997) describes seven major properties of uniquely connectionist representations. In addition to properties that overlap with those we identified earlier in this article, he identifies a further property of uniquely connectionist representations: namely, they are reusable, or portable, for other tasks. This property has been demonstrated in a number of cognitive tasks, such as Chalmers' (1990) active-passive transformations, and more recently in a control task. N. E. Sharkey (1997) reports the development of a disembodied arm control system in which the connectionist representations that had been developed as a result of training a net to output transformationally invariant position classes were reused as input to a net trained to direct a mobile robotic arm to pick up objects.

There is ongoing debate over the extent to which connectionism is capable of supporting a new style of cognitive architecture. The arguments in favor stress the capability of connectionist representations for both compositionality and systematicity, that the representations are reusable for other tasks, and that they accomplish these properties by different means than those of classical symbolic representations. Opponents stress the problems connectionists have in dealing with structured representations and expressing relationships between variables, and our ability to immediately follow a linguistically expressed rule (see Marcus, 2001). However, it should be noted that although it can be argued that connectionist representations can support structure-sensitive operations, or that connectionist nets can process regular and exception material within a single mechanism, it does not necessarily follow that the connectionists assume that the brain should be modeled as a unitary mechanism. There are a number of reasons for assuming that connectionist modeling of more complex cognitive functions will require some degree of modularization. For instance, there is the observed difficulty of training a net to perform two unrelated tasks at the same time, and concomitantly the advantages often observed in terms of performance when a task is modularized (A. J. C. Sharkey, 1999).

It seems likely that future connectionist research will have to attend to issues of modularity and the control and coordination of distinct connectionist modules. This research area is receiving increasing attention in the field of behavior-based robotics (see REACTIVE ROBOTIC SYSTEMS; BIOLOGICALLY INSPIRED ROBOTICS), where the challenge is not only to coordinate modules but also to provide an account of the way in which distinct modules, and the representations they employ, emerge in the first place (N. E. Sharkey, 1998). A related classical versus connectionist debate may also surface here, regarding the extent to which the coordination of modules requires a classical architecture (Hadley, 1999) or can be achieved without recourse to a central executive (N. E. Sharkey, 1998).

Discussion

This article began by outlining the main characteristics of connectionist models of cognition. Two of the main apparent advantages of connectionist models are (1) their ability to both model and simulate empirical data, and (2) their ability to do so without being explicitly programmed. These features segregate connectionist models from the earlier models of cognitive psychologists, who modeled the data without simulating the tasks involved, and from AI models, which are explicitly programmed and which simulate behavior without regard to the details of psychological evidence. On further reflection, however, the distinction between connectionist models and earlier approaches becomes less clear. Obtaining a good fit to the data relies on a number of assumptions and is not, as is sometimes supposed, achieved without experimental intervention. The researcher must make a number of decisions that determine the performance of the model. We introduced the term *extensional programming* to refer to these decisions, which include the selection of the training set and its manner of presentation. Nonetheless, even if there are some aspects of connectionist models that bear further consideration, it is still the case that the novel style of computation they embody often encourages the development of new accounts of previously investigated data.

Subsequently we considered the ability of connectionism to support a new theory of cognition. Cognitive connectionists, are more concerned to exploit the implications of connectionist computation than to provide detailed models of psychological data. An important issue for the cognitive connectionist is the notion of connectionist representation. It has been claimed that connectionist representations are in principle incapable of supporting an account of

higher mental processes. To counter this claim, computational experiments have been conducted with the aim of demonstrating that connectionist representations have a complex internal structure, capable of fulfilling the requirements for a cognitive architecture.

Connectionist representations have been shown to be capable of much of the systematicity and compositionality required of them by their critics. They do not merely implement symbolic computation but operate in a novel manner. Whether it will be possible to provide a connectionist account of all aspects of cognition remains to be seen. An important issue to be addressed is that of modularity, in particular the extent to which the coordination of distinct modules implies a classical architecture or can be accomplished in a uniquely connectionist manner.

Road Map: Psychology

Related Reading: Artificial Intelligence and Neural Networks; Philosophical Issues in Brain Theory and Connectionism; Systematicity of Generalizations in Connectionist Networks

References

- Anderson, J. A., 1972, A simple neural network generating an interactive memory, *Math. Biosci.*, 8:137–160.
- Chalmers, D. J., 1990, Syntactic transformations on distributed representations, *Connect. Sci.*, 2:53–62.
- Christiansen, M., Conway, C. M., and Curtin, S., 2000, A connectionist single mechanism account of rule-like behaviour in infancy, in *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum, pp. 83–88.
- Edelman, G. M., 1987, *Neural Darwinism*, New York: Basic Books.
- Hadley, R. F., 1999, Connectionism and novel combinations of skills: Implications for cognitive architecture, *Minds Machines*, 9:197–221.
- Marcus, G. F., 2001, *The Algebraic Mind: Integrating Connectionism and Cognitive Science*, Cambridge, MA: MIT Press. ♦
- McClelland, J. L., Rumelhart, D. E., and PDP Research Group, Eds., 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (2 vols.), Cambridge, MA: MIT Press.
- Pinker, S., and Prince, A., 1988, On language and connectionism: Analysis of a parallel distributed processing model of language acquisition, in *Connections and Symbols* (S. Pinker and J. Mehler, Eds.), Cambridge, MA: Bradford/MIT Press, pp. 73–194.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K., 1996, Understanding normal and impaired word reading: Computational principles in quasi-regular domains, *Psychol. Rev.*, 103:56–115.
- Pollack, J., 1990, Recursive distributed representations, *Artif. Intell.*, 46:77–105.
- Ratcliff, R., 1990, Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions, *Psychol. Rev.*, 96:523–568.
- Seidenberg, M. S., and McClelland, J. L., 1989, A distributed, developmental model of visual word recognition and naming, *Psychol. Rev.*, 96:523–568.
- Sharkey, A. J. C., 1999, Multi-net systems, in *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems* (A. J. C. Sharkey, Ed.), New York: Springer-Verlag, pp. 1–30. ♦
- Sharkey, N. E., 1997, Artificial neural networks for coordination and control: The portability of experiential representations, *Robot. Auton. Syst.*, 22:345–359.
- Sharkey, N. E., 1998, Learning from innate behaviours: A quantitative evaluation of neural network controllers, *Machine Learn.*, 31:115–139.
- Sharkey, N. E., and Sharkey, A. J. C., 1994, Emergent cognition, in *Handbook of Neuropsychology*, vol. 9, *Computational Modeling of Cognition* (J. Hendlar, Ed.), Amsterdam: Elsevier. ♦

Collective Behavior of Coupled Phase Oscillators

Yoshiki Kuramoto

Introduction

Certain aspects of brain functions seem largely independent of the neurophysiological details of the individual neurons and their interconnections. The phase oscillator model for neural populations, like other neural network models, relies on this anticipation, still trying to recover the *phase information* that has been ignored by conventional models using all-or-none threshold elements. Here, phase information means the kind of information encoded in the form of specific temporal structures of the sequence of neuronal spikings that the real brain should make full use of. In this article, we present a brief survey of the collective behavior of coupled oscillators using the phase model and assuming all-to-all type interconnections. With these mathematical simplifications, a number of definite results can be obtained. Because such a model could be too idealistic for practical purposes, we will not try to interpret the types of behavior obtained as relevant to specific brain activities.

Within the above restrictions on the model, the variety in collective behavior exhibited is still great. We will particularly be concerned in this article with the onset and persistence of *collective oscillation* in frequency-distributed systems; splitting of the population into a few subgroups, called *clustering*; and the more complex collective behavior called *slow switching*. It turns out that knowing the type of phase coupling between an interacting pair of oscillators is helpful for interpreting the resulting collective behavior. In the final section, a few remarks are provided on the limitations of the present model.

Phase Model

By suitably defining the phase ϕ (mod 2π) on the limit-cycle orbit of a given oscillator, its free motion can always be described by $\dot{\phi} = \omega$. When weak coupling is introduced between a pair of identical oscillators 1 and 2, each described by $\dot{\phi}_{1,2} = \omega$, phase reduction theory (Kuramoto, 1984) tells that these equations are modified as

$$\dot{\phi}_1 = \omega + \Gamma_{12}(\phi_1 - \phi_2) \quad (1)$$

combined with the similar equation for oscillator 2, where the coupling functions $\Gamma_{ij}(x)$ are 2π -periodic. What is important here is that the coupling depends only on the phase difference. Whenever necessary, $\Gamma_{ij}(x)$ can be computed numerically from knowledge of the original dynamical system model for the coupled oscillators.

The above argument can easily be generalized to systems of slightly nonidentical oscillators by replacing ω with $\omega + \delta\omega_{1,2} = \omega_{1,2}$. Thus, a general network of N similar oscillators with pairwise coupling reduces to

$$\dot{\phi}_i = \omega_i + \sum_{j=1}^N \Gamma_{ij}(\phi_i - \phi_j) \quad (2)$$

This model, possibly with generalizations by including systematic and/or random forcing terms, provides a canonical model for oscillator networks.

Whether or not the weak coupling assumption is valid, the phase model has also been used as a convenient phenomenological model for the study of visual processing, motor control, and other life

processes (Cohen, Holmes, and Rand, 1982; Ermentrout and Kopell, 1984; Sompolinsky et al., 1991).

Types of Phase Coupling

Phase coupling functions can be classified into a few basic types, and they lead to qualitatively different dynamics of a coupled pair. Assuming that the coupling is symmetric, or $\Gamma_{12}(x) = \Gamma_{21}(x) = \Gamma(x)$, we obtain the equation for the phase difference $\psi \equiv \phi_1 - \phi_2$ in the form

$$\dot{\psi} = \Delta\omega + \Gamma_a(\psi) \quad (3)$$

where $\Delta\omega = \omega_1 - \omega_2$ and $\Gamma_a(\psi)$ is twice the antisymmetric part of $\Gamma(\psi)$, i.e., $\Gamma_a = \Gamma(\psi) - \Gamma(-\psi)$. Note that $\Gamma_a(x)$ satisfies $\Gamma_a(0) = \Gamma_a(\pm\pi) = 0$. When the oscillators are identical ($\Delta\omega = 0$), there are three typical situations (Figure 1). For type A coupling, $\Gamma'_a(0) < 0$ and $\Gamma'_a(\pm\pi) > 0$, so that the phase-synchronized state $\psi = 0$ is stable, while the antiphase state $\psi = \pm\pi$ is unstable. This form of coupling is called *in-phase type*. For type B coupling, in contrast, we have $\Gamma'_a(0) > 0$ and $\Gamma'_a(\pm\pi) < 0$, so that the coupling is called *antiphase type*. For type C coupling, both the in-phase and antiphase states are unstable, while the phase difference is locked at some intermediate value. We will call this *out-of-phase type* coupling. More complicated situations are possible where multiple values of ψ become stable.

When the oscillators are nonidentical, each $\dot{\psi}$ versus ψ curve in Figure 1 will be shifted upward or downward, so that the stable

value of ψ also changes. Clearly, outside a certain range of $\Delta\omega$, no fixed point can exist. Then the oscillators fail to synchronize, and the system as a whole exhibits quasi-periodic motion with two independent frequencies, their difference being given by the long-time average of $\dot{\psi}$. Numerical computation of Γ_a for some neural oscillator models in a self-oscillatory regime assuming diffusive or pulsatile coupling suggests that any of these three types of coupling is possible in real neural systems. It is remarkable that type C coupling can be obtained for self-oscillatory Hodgkin-Huxley neurons with excitatory synaptic coupling (Hansel, Mato, and Meunier, 1993a).

Neural oscillators are often modeled with the so-called leaky integrate-and-fire (LIF) neurons. In order to see the connection of this model to the phase model (Kuramoto, 1990), let a LIF neuron be described with variable u ($0 \leq u \leq 2\pi$), and identify the states $u = 0$ and 2π just as we did for the phase variable. The intrinsic dynamics of this neuron is such that u is monotone increasing with t , so that when u reaches the level $u = 2\pi$, it is immediately reset to the zero value. This instant is interpreted as the time of firing. Specifically, u is supposed to obey the equation

$$\dot{u} = a - u \quad (4)$$

Obviously, the neuron repeats firing if $a > 2\pi$, while it is non-oscillatory but only excitable when $a < 2\pi$. As for coupling with another LIF neuron, the usual assumption is that u changes by a small amount ε each time t_n ($n = 1, 2, \dots$) the second neuron fires. The coupling is excitatory if $\varepsilon > 0$ and inhibitory otherwise. This dynamical rule is conveniently represented by the term $\varepsilon \sum_n \delta(t - t_n)$ added to the right-hand side of Equation 4.

In the self-oscillatory regime ($a > 2\pi$), the natural frequency is given by $\omega = 2\pi / \ln(1 - 2\pi/a)$. It can be shown that the network of self-oscillatory LIF neurons with weak coupling is equivalent to that of phase oscillators given by the form of Equation 2 with the coupling function $\Gamma(\psi) = \varepsilon \omega a^{-1} \exp(\psi/\omega)$ ($0 \leq \psi < 2\pi$). Note that $\Gamma(\psi)$ has a discontinuity at $\psi = 0$. It is easy to check that for positive (negative) ε , the in-phase (antiphase) state gives a unique and strongly stable state.

Throughout the subsequent sections, the coupling will be assumed to be pairwise and symmetric, belonging to either A, B, or C type.

Collective Oscillation

The collective behavior of coupled phase oscillators differs drastically for different coupling types. If the coupling is of the in-phase type and the oscillators are identical, then the whole assembly is in perfect phase synchrony, behaving as a single giant oscillator.

Many examples of collective oscillations in living and nonliving systems are known (Pikovsky, Rosenblum, and Kurths, 2001). With respect to sensory processing in the brain, in particular, collective synchronization may play a crucial role in the linking of sensory inputs across multiple receptive fields (Gray & Singer, 1998; Malsburg and Schneider, 1986). There are some studies on the last topic using the phase oscillator model (Sompolinsky et al., 1991).

Collective oscillations, if functionally relevant at all, should persist robustly against random perturbations from various sources, as would be unavoidable in the real world. How such robustness is guaranteed can theoretically be explained by using our phase model with all-to-all coupling. For a short history of this small branch of nonlinear science, see the review article by Strogatz (2000). The simplest model for this problem is given by

$$\dot{\phi}_i = \omega_i + \frac{1}{N} \sum_{j=1}^N \Gamma(\phi_i - \phi_j) \quad (5)$$

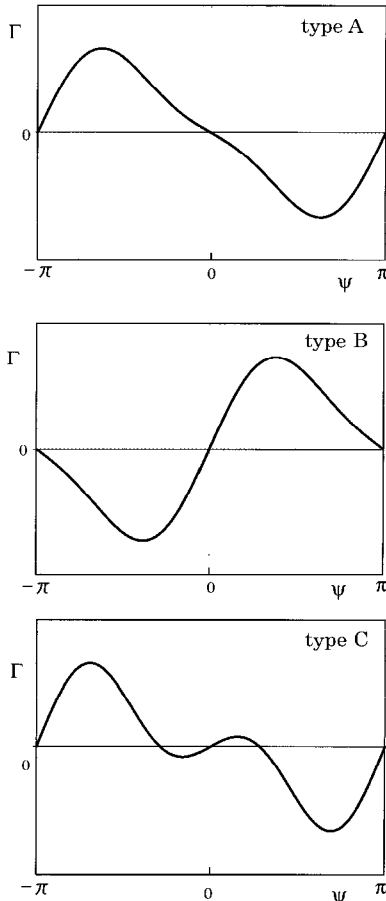


Figure 1. Three types of phase coupling functions.

Here the randomness is assumed in the natural frequencies ω_i with distribution $g(\omega)$, which is unimodal and symmetric about $\omega = \Omega$. For the simplest coupling function $\Gamma(x) = -K \sin x$, and in the limit of the population size N going to infinity, an exact solution with phase-transition-like behavior is available in a rather simple form. Macroscopic oscillation amplitude can be measured by the complex *order parameter* $w = N^{-1} \sum_{j=1}^N e^{i\phi_j}$. Analysis shows that the order parameter behaves like $w = \sigma \exp(i\Omega t)$, where σ satisfies the transcendental equation

$$\sigma = K\sigma \int_{-\pi/2}^{\pi/2} dx g(\Omega + K\sigma \sin x) \cos x e^{i\lambda} \equiv S(\sigma) \quad (6)$$

$S(\sigma)$ is an S-shaped odd function saturating to 1 as $\sigma \rightarrow \infty$. Therefore, besides the trivial solution $\sigma = 0$, which always exists, a nontrivial solution appears when K exceeds a critical value given by $K_c = 2/(\pi g(\Omega))$ at which $S'(0) = 1$. It is remarkable that the period Ω of the macroscopic oscillation becomes infinitely precise as $N \rightarrow \infty$.

There are also studies for more general coupling functions. A noticeable result of such studies is that the order parameter for K slightly larger than K_c behaves like $\sigma \propto (K - K_c)^\beta$ with the exponent $\beta = 1$ rather than the classical value $1/2$, insofar as the coupling function is not completely free from the second harmonic component.

As the origin of randomness, the distribution in natural frequency may be replaced with external noise. When the noise is additive, white Gaussian, and drives the oscillators individually, the collective dynamics can also be studied analytically by means of a Fokker-Planck equation for the number density $\rho(\phi, t)$. Similar phase-transition-like behavior with $\beta = 1/2$ is then obtained. Such a stochastic population model with the extension of including external stimuli, and with particular emphasis on its phase-resetting characteristics, was extensively studied by Tass (1999) with a view to medical applications such as the analysis of magnetoencephalography/electroencephalography data and deep brain stimulation techniques used in Parkinsonian patients.

Clustering and Complex Collective Behavior

We now consider the population dynamics of Equation 5 for the out-of-phase type coupling. The oscillators are assumed to be identical. Actually, owing to the desynchronizing nature of the coupling, the effects of randomness would be of secondary importance in causing complex collective behavior. Since perfect coherence is impossible, we want to find out what types of collective behavior can arise. The extreme opposite of perfect coherence is perfect *incoherence*, for which the phase distribution $\rho(\phi)$ is uniform. Such ρ certainly exists as a particular solution of the problem. However, analysis shows that its stability is guaranteed only for a special form of the coupling function. Thus, what occurs generically seems to be something intermediate between perfect coherence and perfect incoherence. Numerical study shows that most commonly, the whole population splits into two-point clusters, each in perfect phase synchrony, their phase difference preserving a constant value. Such collective behavior is obtained for the coupling function

$$\Gamma(\psi) = -\sin \psi + a \sin 2\psi \quad (7)$$

and illustrated in Figure 2 in comparison with the case of perfect coherence. The size proportion of the clusters, $p : 1 - p$, is not unique and depends on initial conditions; there is a certain range of p in which such two-cluster states remain stable. On a macroscopic level, the transition from one-cluster state to two-cluster state may appear as *rhythm splitting*, such as is often reported in the literature on circadian oscillations.

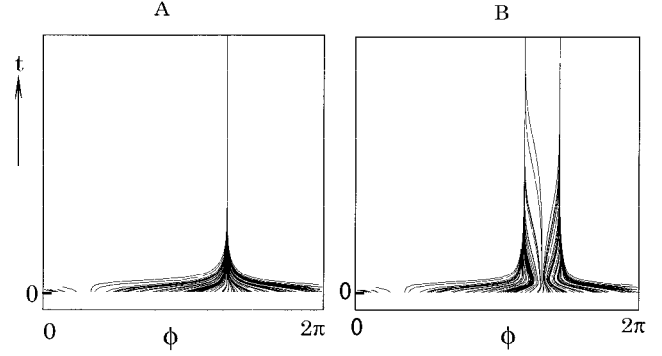


Figure 2. Formation of phase-synchronized clusters in a co-moving frame of reference for which ϕ remains constant for steady oscillation. A coupling function (Equation 7) with $\alpha = 0$ and random initial conditions are assumed. For $a = 0.4$ (A), the whole assembly converges to a single point cluster, while for $a = 0.7$ (B) it splits into two subpopulations, each in perfect phase synchrony.

The two clusters rigidly rotating in the phase space may become unstable for any p . When this happens, one or the other cluster may further split into multiple clusters, or otherwise a peculiar behavior called *slow switching* may occur (Hansel, Mato, and Meunier, 1993b). The latter phenomenon can be described as follows. For most periods, the system stays practically in a two-cluster state even if it is unstable. However, one of the clusters eventually becomes unable to maintain its internal phase synchrony, and starts to dissolve. This is followed by a short period of strong disorder. The entire system then relaxes to another two-cluster state, i.e., the state obtained by a constant phase shift of the original two-cluster state. Since the last state is also unstable, the same process repeats indefinitely, but with longer and longer time scales. The peculiar dynamics here is interpreted as resulting from the formation of an attracting heteroclinic loop connecting the first and second two-cluster states mentioned above. In the presence of weak randomness, in the form of either external noise or inhomogeneity, the switchings occur nearly periodically, with the period prolonged as the randomness is decreased.

Discussion

A few remarks should be about the two simplifying assumptions we have worked with. They are first, the weak coupling assumption, which enabled the phase description, and second, the all-to-all coupling. The first assumption has two implications. First, it implies that the oscillators practically keep staying on their intrinsic limit-cycle orbit when perturbed by the other oscillators. This allows us to ignore all degrees of freedoms other than the phases. Second, the time scale associated with the coupling is so long (much longer than the period of oscillation) that the effect of coupling can be time-averaged over one cycle of oscillation. This results in the coupling function depending only on the phase difference, thus facilitating the mathematical analysis greatly. Assuming the weakness in coupling in the first sense seems neurobiologically quite reasonable (Hoppensteadt and Izhikevich, 1997). However, the second condition seems more restrictive. For instance, whenever neural oscillators establish mutual synchronization only within a few interspike intervals, this condition must certainly be violated.

Regarding the connectivity of neural oscillators, all-to-all coupling is apparently an extreme idealization. Local coupling represents another idealization. Actual coupling must be something in

between, but the studies on coupled oscillator networks with general nonlocal coupling are still few. A neurobiologically plausible form of coupling often employed is characterized by short-range excitation and long-range inhibition. For this specific coupling type, and using relaxation oscillators, Terman and Wang (1995) studied the collective dynamics of the network in an attempt to lay a physical foundation for the oscillatory correlation theory of feature binding.

Road Map: Dynamic Systems

Related Reading: Chains of Oscillators in Motor and Sensory Systems; Synchronization, Binding and Expectancy; Visual Scene Segmentation

References

- Cohen, A. H., Holmes, P. J., and Rand, R. H., 1982, The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model, *J. Math. Biol.*, 13:345–369.
- Ermentrout, B., and Kopell, N., 1984, Frequency plateaus in a chain of weakly coupled oscillators, *SIAM J. Math. Anal.*, 15:215–237.
- Gray, C. M., and Singer, W., 1989, Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex, *Proc. Natl. Acad. Sci. USA*, 86:1698–1702.
- Hansel, D., Mato, G., and Meunier, C., 1993a, Phase dynamics for weakly coupled Hodgkin-Huxley neurons, *Europhys. Lett.*, 23:367–372.
- Hansel, D., Mato, G., and Meunier, C., 1993b, Clustering and slow switching in globally coupled phase oscillators, *Phys. Rev. E*, 48:3470–3477.
- Hoppensteadt, F. C., and Izhikevich, E. M., 1997, *Weakly Connected Neural Networks*, New York: Springer-Verlag.
- Kuramoto, Y., 1984, *Chemical Oscillations, Waves, and Turbulence*, Berlin: Springer-Verlag. ♦
- Kuramoto, Y., 1990, Collective synchronization of pulse-coupled oscillators and excitable units, *Physica D*, 50:15–30.
- Malsburg, C. von der, and Schneider, W., 1986, A neural cocktail-party processor, *Biol. Cybern.*, 54:29–40.
- Pikovsky, A., Rosenblum, M., and Kurths, J., 2001, *Synchronization: A Universal Concept in Nonlinear Science*, Cambridge, Engl.: Cambridge University Press. ♦
- Sompolinsky, H., et al., 1991, Cooperative dynamics in visual processing, *Phys. Rev. A*, 43:6990–7011.
- Strogatz, S. H., 2000, From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators, *Physica D*, 143:1–20.
- Tass, P. A., 1999, *Phase Resetting in Medicine and Biology*, Berlin: Springer-Verlag.
- Terman, D., and Wang, D., 1995, Global competition and local cooperation in a network of neural oscillators, *Physica D*, 81:148–176.

Collicular Visuomotor Transformations for Gaze Control

J. A. M. Van Gisbergen and A. J. Van Opstal

Introduction

Neurophysiological studies on oculomotor control started in the early 1970s with descriptions of firing patterns at the motoneuron and the premotor level. Meanwhile, a wealth of information has become available on signal processing in the midbrain superior colliculus (SC) and at cortical levels. This article examines recent developments concerning the role of the SC in the control of gaze shifts (combined eye-head movements) and its possible involvement in the control of eye movements in 3D space (direction and depth). New experimental paradigms, putting fewer behavioral constraints on the eye and head motor systems, have led to novel views on the nature of collicular signals.

In classifying modes of oculomotor behavior, it appears useful to make a distinction between two behavioral states that alternate continuously in daily life. During attentive fixation the vestibulo-ocular reflex (VOR) and slow vergence maintain binocular foveal fixation to correct for body movements. When the task requires inspection of an eccentric stimulus, a complex synergy of coordinated movements comes into play. Such refixations typically involve a rapid combined eye-head movement (saccadic gaze shift) and often require binocular adjustment in depth (vergence). It is thought that the system responsible for rapid eye-head movements is actively suppressed during fixation, so that a gating signal (WHEN signal) is required which enables it to respond to the signals that specify the amplitude and the direction of movement (WHERE signals). There is evidence that large gaze shifts involve concurrent suppression of the VOR, which would otherwise counteract the system's effectiveness. As discussed later in this article, there is good reason to think that the SC is involved in this cyclic alternation between gaze holding and gaze shifting. This article will review related theoretical concepts, together with the underlying experimental data. For a broader orientation, see Scudder, Kaneko, and Fuchs (2002), and Wurtz (1996).

Fixation cells (FIX) in the rostral pole of the SC are active during attentive fixation and pause during large saccades (see Wurtz, 1996,

for review). In the caudal zone, the pattern of activity is just the opposite. A group of saccade-related cells (SAC) becomes active prior and during the saccade and has low levels of activity during active fixation. A similar antagonistic pattern of activity, functionally reminiscent of the rostral-caudal distinction in the SC, characterizes omnipause neurons and saccadic burst cells in the brainstem. These cells pause and burst during saccades; it is thought that they are driven by FIX and SAC cells, respectively. The brainstem pause cells, and burst cells have mutual inhibition and it has recently been suggested that this is also the case for FIX and SAC cells in the SC (Munoz and Istvan, 1998). In the visual system, target position is topographically coded. At the level of burst cells and motoneurons, saccade size is coded temporally by the number of spikes in the burst or the instantaneous firing rate. This transition from spatial to temporal coding is a classical problem in the study of the saccadic system in which the SC is thought to play a key role. Early studies of SAC cells suggested that the SC contains a topographical map specifying desired saccade displacement. These recordings showed that each SAC neuron is recruited only for a limited range of saccade amplitudes and directions, denoted as its movement field. As one moves from the fixation zone to the caudal border, the movement field of local SAC neurons becomes larger and more eccentric. Since saccade direction is coded along the perpendicular dimension, the colliculus can be considered as a map of saccade amplitude and direction that is organized in polar coordinates. By virtue of its topographical organization, the SC has become a key area for experimental and modeling approaches to the question of how sensory signals can be transformed into goal-directed movements. It has become clear that the SC activity during saccades to auditory and somatosensory targets conforms to the same motor map, suggesting that considerable sensorimotor remapping must take place (for references and further explanation see DYNAMIC REMAPPING).

An early electrical stimulation study by Robinson showed for the first time that the SC motor map is highly nonhomogeneous:

relatively more space is devoted to the representation of small saccades. Based on these results, the collicular motor representation in the monkey has been modeled by a complex-logarithmic mapping function (Van Gisbergen, Van Opstal, and Tax, 1987). This description characterizes the relation between retinal target location and the site of maximum SAC cell activity (afferent mapping) as well as the inverse relation between the locus of movement cell activity and the resulting saccade vector (efferent mapping). In the model, the population activity resembles a Gaussian function centered on the point defined by the afferent mapping. This activity profile is translation-invariant so that the number of active cells is assumed constant, independent of saccade size. Such a topographical representation may generate goal-directed saccades if each cell has fixed connections with the horizontal and vertical burst cells downstream, allowing it to generate a small movement vector, proportional to its firing rate, into the direction of the retinal location to which it is connected by the afferent mapping.

Several recent developments have called for extensions of the model. For example, microstimulation in the SC (Van Opstal, Van Gisbergen, and Smit, 1990) has revealed that the eye displacement, encoded by a given site, can be systematically modified by the stimulation parameters. Furthermore, it now seems clear that the SC is involved with rapid gaze movements rather than being a purely saccadic eye-movement control center. Finally, there is increasing evidence that the SC is also involved in directing the eyes in depth. These latter two developments will now be discussed.

Role of the Superior Colliculus in the Control of Gaze Shifts

That activity in the colliculus can generate gaze movements has become clear from electrical stimulation studies in a number of species. The possibility that the monkey SC may code desired gaze displacement has been investigated at the single unit level by Freedman and Sparks (1997). Their approach was to compare the activity of movement cells in trials in which either the gaze displacement vector, the eye displacement vector, or the head displacement vector was constant while the other two vectors varied widely. The conclusion from this work is that a gaze displacement signal is derived from the locus of activity in the SC motor map, which is subsequently decomposed into separate eye and head displacement signals downstream from the colliculus (see Figure 1). The work of Cullen and Guitton (1997) suggests that this separation is not yet evident at the level of premotor burst cells (see discussion later in this article).

Less is known about the extent to which the WHEN systems for eye and head are coupled. If omnipause cells would govern both subsystems, one would expect the onset latencies of eye and head movements to be strongly coupled. Although microstimulation in cat SC seems to support this notion, experiments in primates reveal a considerable degree of independence, arguing against a single shared WHEN system.

Struck by the similarities between eye and head contributions to natural gaze shifts in the cat, Guitton, Munoz, and Galiana (1990) have suggested the possibility that eye and head are driven by a common signal. However, experiments in humans where eye and head had nonaligned starting positions before the gaze movement started (Goossens and Van Opstal, 1997) revealed that their movements may have different directions, incompatible with a common input signal. This led them to propose a model in which eye and head are driven by signals expressed in oculocentric and craniocentric coordinates, respectively, which are nevertheless derived from a common collicular gaze displacement command. One feature of the model is that the driving signals for both the eyes and the head depend on absolute eye position (see Figure 1). The need for such an arrangement has also become apparent from electrical

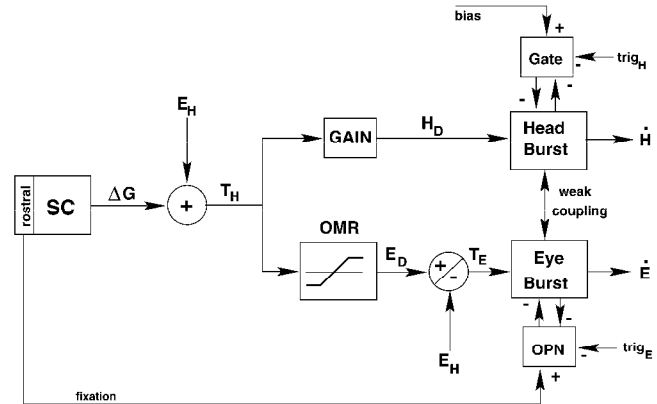


Figure 1. Primate gaze control model. As explained in the text, the collicular gaze displacement command, ΔG , cannot be used to drive eyes and head directly. The scheme proposes that it is first transformed into a desired craniocentric position, T_H , by adding current eye position, E_H . A scaled version of this signal (H_D , desired head position) controls the head movement generator and accounts for the finding that the head-movement gain is typically less than one. The craniocentric signal is also passed through a neural saturation element (OMR) to prevent the oculomotor system from running against the limits of its motor range. Dynamic eye motor error, T_E , which drives the oculomotor burst generator, is then obtained by subtracting current eye position from the clipped T_H signal, E_D (dynamic desired eye position). The model also proposes a weak excitatory coupling between the two motor systems (Galiana and Guitton, 1992; Goossens and Van Opstal, 1997). This coupling accounts for the differences in eye and head contributions to the gaze shift, as well as for the modulation of head movement trajectories by eye position. The gating mechanisms for the eye (OPN) and head burst generators (Gate) each have their own trigger and bias signals. Internal feedback loops that are assumed to control the eye and head burst generators to ensure that the eye reaches the target, as well as the role of the vestibular system, have been omitted for clarity (after Goossens and Van Opstal, 1997).

stimulation studies in the superior colliculus of the head-free monkey (Freedman, Stanford, and Sparks, 1996).

Activity of Burst Cells During Gaze Shifts

Earlier discussions of the role to be assigned to the SC in models of the oculomotor system have generally taken for granted that the role of short-lead burst neurons (BN) was already well established. These cells were thought to carry a saccadic velocity signal. However, since the underlying evidence was based on studies in head-fixed animals, it cannot be excluded that the activity of these cells may in fact be related to gaze velocity. The main reason for considering this possibility is that the colliculus, their major source of input, is now thought to specify a desired gaze shift (see previous discussion).

Cullen and Guitton (1997) have made a thorough study of saccadic burst cells to find out whether their activity during head-free gaze shifts might be gaze rather than eye-related. To investigate this question, they fitted the discharge patterns of BNs during head-free gaze shifts, $B(t)$, with a descriptive model of the form: $B(t) = a + b\dot{E} + c\dot{H}$ where a is a bias term, \dot{E} represents instantaneous eye velocity, \dot{H} is current head velocity, and b and c are coefficients. This formulation provides a framework for unambiguous definitions of gaze-related versus eye-related coding and yields numbers allowing each cell to be placed along the continuum between these two extremes.

If a cell is to code gaze velocity, it should not matter by which combination of eye and head velocity the gaze-velocity signal was

created. In such an idealized cell, coefficients b and c would have identical values. In a pure eye velocity cell, coefficient c would be zero. Remarkably, the tacit assumption in the literature that burst cells are pure eye velocity cells, which seemed to make sense because these neurons have direct connections with oculomotor neurons, was not confirmed. Most cells had a head-velocity contribution in addition to the expected eye-velocity term. On the other hand, the cells were not straightforward gaze cells either because the weights are generally not equal ($c < b$) so that, in fact, their behavior is a compromise between eye related and gaze related.

Control of Refixations in Direction and Depth

Delineation of the Problem

Most refixations made in daily life involve both a change in direction and in depth. It has long been held that these movements are made by largely independent subsystems, the saccadic and the vergence system, which were conceived of as binocular controllers. The striking contrast between the quite slow movements in pure vergence and the rapidity of saccades tended to support the idea that these systems are quite different in nature. Early modeling studies have further emphasized the notion of distinct systems with unique properties. It was recognized early on that saccades are much too fast to allow the benefit of direct sensory feedback to guide the movement on a moment to moment basis. By contrast, it has often been proposed that pure vergence movements are slow enough to be under direct visual feedback. These distinctions, based on the extreme cases of pure vergence and pure saccades, become much less clear in the more common situation where both systems act together as occurs during refixations in 3D visual space. The finding that, in such cases, vergence shows a very clear velocity enhancement during the saccade has led to the proposal that perhaps saccades may be disconjugate (unequal in the two eyes). It should be noted that, for the portion of the vergence response that is executed as part of a fast movement, the role of direct sensory feedback must be extremely limited.

Since the saccadic and the vergence subsystems typically operate in joint fashion, we must ask how this coordination comes about. Some provision is necessary to ensure that the saccadic system and the vergence system will move to the same target when there are several alternatives. Chaturvedi and Van Gisbergen (1998) performed behavioral experiments in humans instructed to make binocular refixations to a green target in 3D space and to ignore a red distracter at another location. In short-latency responses, errors (to the red stimulus) and compromising responses (in between) were not uncommon but it appeared that the two systems always worked in unison, strongly suggesting that there must be a common central target selection system operating on 3D stimulus location information.

Possible Role of SC in Coding Refixations in 3D Space

Recent evidence suggests that the SC, which receives depth information from parietal cortex, may be directly involved in saccade-vergence coordination. Chaturvedi and Van Gisbergen (1999) applied electrical stimulation in the caudal SC when the monkey was just preparing or executing the 3D refixation to a visual target. Electrical stimulation alone produced a saccade without an overt vergence component. Stimulation in midflight, just after the monkey had started his visually guided movement, caused a compromise saccade and a marked reduction of the fast vergence component. The vergence perturbation was not simply an epiphenomenon of a change in saccade duration and was comparable in strength to the saccadic disturbance. The vergence effect cannot be understood from classical models proposing that the SC

specifies only the movement in the frontal plane, leaving the depth movement coding to some other area (see above). Instead, Chaturvedi and Van Gisbergen (1999) proposed that the population of cells at any SC locus shares the same direction movement vector, like in the old 2D-coding schemes, but suggest that there is also a depth movement component. The latter is different from cell to cell so that the entire depth dimension along the locally represented direction in space is covered (see Figure 2). With such an extended scheme it is understandable, in principle, why electrical stimulation in isolation may cause only a pure saccadic movement, without any overt sign of vergence. In this situation, the local cells with different depth tuning will be excited indiscriminately, thereby causing a zero vergence displacement command which can counteract the visually induced vergence signal. By contrast, the saccadic signal will still emerge during such local artificial intervention by virtue of the topographical organization of the motor map, which represents the desired horizontal and vertical components spatially on a much coarser scale.

Coding of Eye Movements by Burst Cells

At this point it is interesting to consider how fast eye movements in 3D are coded at the premotor level. It was thought until very recently that burst cells in the reticular formation coded saccades binocularly by providing a conjugate velocity signal. From a systematic study based on eye movements in 3D space, Zhou and King (1998) concluded that most cells code eye velocity for either one or the other eye, irrespective of the depth component involved. This means that the total population contains a 3D code, including both directional and depth components, and that the old idea that they

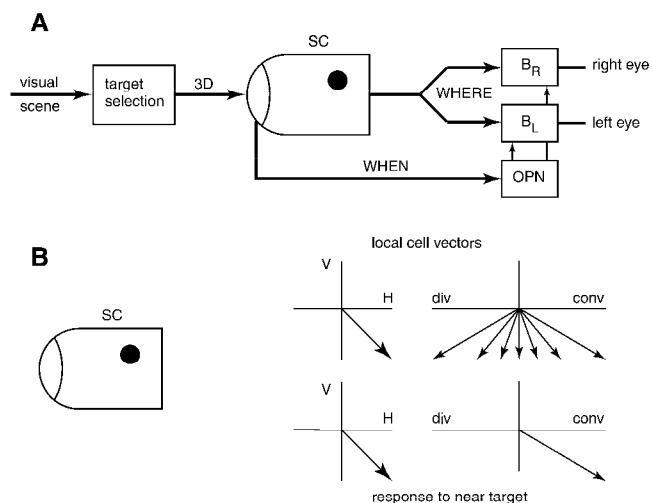


Figure 2. Possible role of SC in control of rapid eye movements in direction and depth. *A*, Based on evidence discussed in the text, the scheme proposes that the SC receives 3D information about the location of the selected target and emits a 3D WHERE signal to activate the populations of right eye and left eye burst cells in the brainstem (Zhou and King, 1998). The initiation system (WHEN) is embodied by the rostral zone in the SC and omnipause cells in the brainstem (OPN). The locus of the active zone in the SC determines the change in fixation in the frontal plane but has no topographic organization for depth coding. *B*, Proposal on how activity in the SC may code direction and depth. All the cells in the active zone code for an oblique downward movement but the cells differ in their depth coding directions as shown in the upper-right panel. For a near target in the locally represented direction of the visual field, the population will yield the direction vector shown at bottom left and the net convergence movement shown in the bottom-right panel.

code only conjugate saccades must be abandoned. It is tempting to suggest that the 3D coding at the level of the burst cells may be a further elaboration of the proposed 3D coding in the SC, which is known to be a major source of input signals for these neurons.

Discussion

Decomposition of SC Signals for Eye and Head

As explained in the previous section, it is now thought that the deep layers of the SC specify a desired gaze displacement, which typically involves movement of both the eyes and the head (Guitton et al., 1990; Freedman et al., 1996; Freedman and Sparks, 1997). However, accumulating evidence indicates that this signal cannot be used directly to drive eyes and head. First, the desired gaze shift may encode a movement well beyond the oculomotor range. For example, suppose a SC site that specifies a 40° rightward gaze displacement is active. When the eyes are looking 20° to the left, the entire movement could be covered by the oculomotor system alone. However, when the eyes start from a position 20° to the right, part of the gaze shift has to be made by the head so that the relative contributions of the eyes and the head to the total gaze shift will (in part) be determined by the initial position of the eyes in the orbit. Second, experiments have shown that the movement directions of eyes and head can be quite different when their initial starting positions are unaligned. Both motor systems appear to move in the direction of the spatial target position. Multiple regression of human data has shown that the direction of head movement is best described by the head motor-error vector (rather than by gaze error). In contrast, the eye-in-head displacement is best described in terms of an eye motor-error signal (Goossens and Van Opstal, 1997). Third, in primates the difference in movement onset of eyes and head varies considerably from trial to trial (see previous discussion). This variability is largely explained by the relative contributions of the two motor systems to the gaze shift: the larger the planned head movement, the earlier it starts (Freedman et al., 1996; Goossens and Van Opstal, 1997). These relative contributions are not fully determined by initial eye position, since stimulus modality is also an important factor: auditory gaze shifts tend to have larger (and thus earlier) head movements (Goossens and Van Opstal, 1997). Finally, the movement trajectories of the head are influenced by eye position, which suggests a subtle coupling between the two motor systems. The simplified scheme in Figure 1 incorporates these different aspects, and highlights the transformations that are required between the SC and the two motor systems.

Binocular Versus Monocular Control of Eye Movements

The question of how the brain coordinates binocular eye movements has been debated already in the previous century, long before any neurophysiological evidence was available. One view, known as the principle of equal innervation, advocates that there are two binocular control systems that affect both eyes. The conjugate system moves both eyes equally and in the same direction. The disconjugate system in this scheme moves both eyes equally in opposite directions, to generate vergence movements. In this way, all binocular eye movements can be described mathematically as the linear sum of a conjugate and a disconjugate movement command, but it is not known whether this is also a valid description of how the real system operates.

The alternative idea is that each eye is controlled separately (monocular control systems). Quite unexpectedly, a recent study by Zhou and King (1998) has yielded evidence for the latter scheme at the level of burst cells. That the situation is not simple is indicated by the even more surprising finding by Zhou and King that

motoneurons in the abducens nucleus often carry binocular signals. This is counterintuitive since, at first glance, it would seem hard to think of a more likely candidate for monocular eye movement coding than the oculomotorneuron level. The finding of monocular coding in premotor burst cells raises interesting questions for the organization at more central levels. If the SC indeed contains a 3D code for rapid eye movements, as discussed above, the question to be faced is how the functional projections of its cells may be organized. Sylvestre, Galiana, and Cullen (2002) have recently proposed a model featuring a shared vergence/saccade controller in the SC, which can yield monocular burst cells without requiring an exceedingly complex wiring diagram.

Is the SC an Open-Loop Controller?

The two schemes in this chapter suggest that the SC provides a motor command to downstream platforms in open-loop fashion. There is general agreement that vision is much too slow to provide feedback during saccadic refixations. In other words, if ongoing platform movements are to affect current SC activity, the only possibility is to use internal (or local) feedback. Currently, there is no clear evidence for the idea that the SC has access to fast moment-to-moment local feedback about saccade execution. There are signs of a limited degree of local feedback but its precise role remains to be determined. For an extensive review on experimental and theoretical work in this vast and highly controversial field we refer to Scudder et al. (2002).

Road Maps: Mammalian Brain Regions; Mammalian Motor Control; Vision

Related Reading: Dynamic Remapping; Pursuit Eye Movements; Vestibulo-Ocular Reflex

References

- Chaturvedi, V., and Van Gisbergen, J. A. M., 1998, Shared target selection for combined version-vergence eye movements, *J. Neurophysiol.*, 80:849–862.
- Chaturvedi, V., and Van Gisbergen, J. A. M., 1999, Perturbation of combined saccade-vergence movements by microstimulation in monkey superior colliculus, *J. Neurophysiol.*, 81:2279–2296.
- Cullen, K. E., and Guitton, D., 1997, Analysis of primate IBN spike trains using system identification techniques. II. Relationship to gaze, eye, and head movement dynamics during head-free gaze shifts, *J. Neurophysiol.*, 78:3283–3306.
- Freedman, E. G., Stanford, T. R., and Sparks, D. L., 1996, Combined eye-head gaze shifts produced by electrical stimulation of the superior colliculus in rhesus monkeys, *J. Neurophysiol.*, 76:927–952.
- Freedman, E. G., and Sparks, D. L., 1997, Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: Evidence for a gaze displacement command, *J. Neurophysiol.*, 78:1669–1690.
- Galiana, H. L., and Guitton, D., 1992, Central organization and modeling of eye-head coordination during orienting gaze shifts, *Ann. NY Acad. Sci.*, 656:452–471.
- Goossens, H. H. L. M., and Van Opstal, A. J., 1997, Human eye-head coordination in two dimensions under different sensorimotor conditions, *Exp. Brain Res.*, 114:542–560.
- Guitton, D., Munoz, D. P., and Galiana, H. L., 1990, Gaze control in the cat: Studies and modeling of the coupling between orienting eye and head movements in different behavioral tasks, *J. Neurophysiol.*, 64:509–531.
- Munoz, D. P., and Istvan, P. J., 1998, Lateral inhibitory interactions in the intermediate layers of the monkey superior colliculus, *J. Neurophysiol.*, 79:1193–1209.
- Scudder, C. A., Kaneko, C. R. S., and Fuchs, A. F., 2002, The brainstem burst generator for saccadic eye movements, *Exp. Brain Res.*, 142:439–462. ♦
- Sylvestre, P. A., Galiana, H. L., and Cullen, K. E., 2002, Conjugate and

vergence oscillations during saccades and gaze shifts: Implications for integrated control of binocular movement, *J. Neurophysiol.*, 87:257–272.

Van Gisbergen, J. A. M., Van Opstal, A. J., and Tax, A. A. M., 1987, Collicular ensemble coding of saccades based on vector summation, *Neuroscience*, 21:541–555.

Van Opstal, A. J., Van Gisbergen, J. A. M., and Smit, A. C., 1990, Com-

parison of saccades evoked by visual stimulation and collicular electrical stimulation in the alert monkey, *Exp. Brain Res.*, 79:299–312.

Wurtz, R. H., 1996, Vision for the control of movement, *Invest. Ophthalmol. Vis. Sci.*, 37:2131–2145.

Zhou, W., and King, W. M., 1998, Premotor commands encode monocular eye movements, *Nature*, 393:692–695.

Color Perception

Robert Kentridge, Charles Heywood, and Jules Davidoff

Introduction

Color is the name we assign to the experience elicited by an attribute of a surface, namely, its spectral reflectance. Color sensations have a reliable, though complex, relationship to the spectral composition of light received by the eyes. The visual system tackles a series of computational problems in the course of processing wavelength. Variation in the wavelength of light is isolated from variation in its intensity. The spectral reflectance properties of surfaces are isolated from the effects of the spectral composition of light illuminating them (matching surfaces with the same reflectance properties in different parts of the visual scene or under different illuminants are the two problems of color constancy). Finally, the resulting continuous color space is partitioned into discrete color categories. In addition, it will become clear that wavelength signals can be used in the course of perceiving form or motion, independent of their role in the subjective experience of color.

Wavelength-Dependent Differences Within the Visual System

Color percepts derive from light that varies in both wavelength and intensity. A single type of photoreceptor in the eye responds with differing efficiency to light over a wide range of wavelengths. Consequently, a visual system in which there is only a single type of photoreceptor inevitably confounds wavelength and intensity. A visual system containing photoreceptors that differ in their spectral response can, in principle, disambiguate wavelength and intensity by comparing the responses of different types of receptors.

Receptors

Wavelength-selective processing can be traced from differentially wavelength-sensitive cone types in the retina to the lateral geniculate nucleus (LGN) and then on to striate cortex and extrastriate areas beyond it. There are three cone types in the human retina, with peak sensitivities at 560 nm, 530 nm, and 430 nm and referred to as L, M, and S (long-, medium-, and short-wavelength-sensitive) cones, respectively. In some people one or more of these cone types are missing; hence, color sensations that would normally be perceived as distinct are confused, and the individuals are “color-blind.” The functions relating the sensitivities of these cone types to the wavelength of stimulating light can be inferred by comparing the wavelength sensitivities of color-blind and normal observers or by examining the effects of adaptation to light of one wavelength on sensitivity to light of other wavelengths. Figure 1 shows the relative absorption efficiencies of the three cone types and the typical pattern of behavioral sensitivity to intensity modulation.

The output of cones provides information about an object’s state, for example, allowing ripe and unripe fruit to be discriminated. The peak sensitivities of photoreceptors appear exquisitely matched to maximize the discriminability of the foliage or fruits that form the

diets of a number of species of primates (Sumner and Mollon, 2000). Studies of the genetic coding of cone pigments indicate that human trichromacy evolved from ancestral dichromacy through the division of a single long-wavelength-sensitive pigment into distinct L and M pigments (Bowmaker, 1998).

The Combination of Receptor Signals in the M, P, and K Channels

Three anatomically distinct cell types in the retina combine cone signals in distinct ways (Dacey, 2000). In all cases, the response has a “center-surround” organization. A set of cones from one part of the visual field influences the cell in one way, while a set of cones from the surrounding area influences it in a different way. Parasol cells receive input from L and M, but not S, cones. Inputs from L and M cones are summed in both the center and surround fields of parasol cells (Figure 2A). Parasol cells cannot convey information about wavelength independent of intensity. They project to the magnocellular layer of the LGN, which in turn projects to layers 4C α and 4B of primary visual cortex (V1). This pathway, and its onward projections, is known as the M-channel. The M-channel contributes to the perception of luminance and motion but does not convey wavelength-coded signals.

Midget ganglion cells have color-opponent receptive fields. This center-surround organization sharpens the effective wavelength selectivity of the ganglion cell, helping to unconfound wavelength and intensity variation. Consider first a nonopponent cell, sensitive to medium wavelength light. This cell will produce the same re-

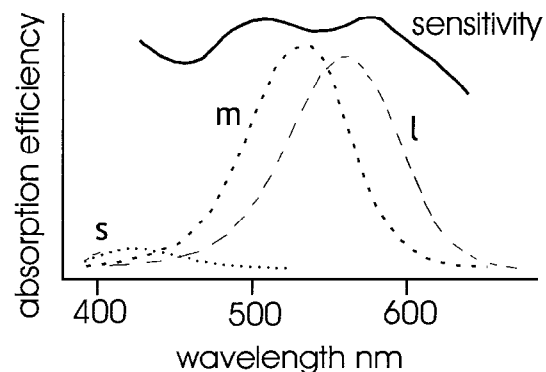


Figure 1. The relative absorption efficiencies of short-, medium-, and long-wavelength cone types, labeled s, m, and l, are shown as dotted, short-dashed, and long-dashed lines, respectively. The solid line shows sensitivity to increments in luminance for lights of different wavelengths. The sensitivity decreases falling between the peaks of the cone absorption spectra are known as Sloan-Crawford notches.

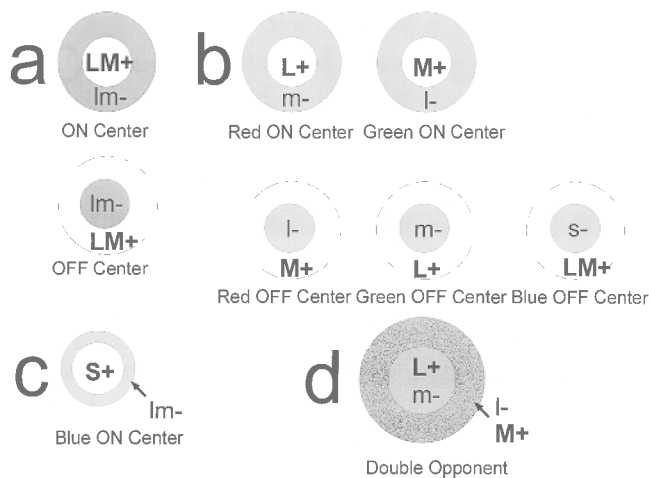


Figure 2. Schematic representations of receptive field organization of cells in (A) the M-channel, (B) the P-channel, and (C) the K-channel. (D) An example of receptive field organization of a cortical double-opponent cell.

sponse to a given intensity of medium wavelength light, or a stronger intensity of longer wavelength light. Although its peak sensitivity is to medium rather than longer wavelength light, because sensitivity only reduces gradually as wavelength deviates from the peak, the longer wavelength light still produces a response. Now consider the responses of an opponent cell excited by medium wavelength light in the center of its field and inhibited by long wavelength light in the surround to different intensities and wavelengths of light falling on its entire receptive field. Medium wavelength light produces excitation in the center and no inhibition in the surround; there is a net increase in the cell's firing rate. Higher intensities of medium wavelength light elicit stronger net responses. A slightly longer wavelength produces some excitation in the center field of the ganglion cell but also a small inhibitory response in the surround. These roughly balance, and so the firing rate of the cell is largely unaffected by the stimulus. The same situation applies to a high-intensity stimulus; again, central excitation is balanced by surround inhibition. This ganglion cell is therefore capable of conveying information solely about the intensity of medium wavelength light.

The vast majority of foveal midget ganglion cells are driven by L or by M cones in the center of their receptive field; these centers can be either excitatory or inhibitory. Away from the fovea, midget ganglion cells lose their spectral opponency, as more than one cone type drives both the center and surround. There appears to be little input from S cones to midget ganglion cells, just as there are very few S cones in the retina. About 2%–3% of parafoveal midget ganglion cells have S-OFF central receptive fields with an ON surround driven by both L and M inputs (Figure 2B). There are no S-ON center midget ganglion cells.

Small bistratified ganglion cells receive inputs from all three cone types; however, their central field always appears to be driven by an excitatory input from S cones, while their surround combines inhibitory L and M inputs. The bipolar cells that convey signals from cones to the central field receive inputs only from S cones and are driven by multiple cells, unlike the bipolar cells that drive the central fields of midget ganglion cells, which receive inputs from single cones. The result is that, although these cells do show clear spatial and spectral opponency, the size of the central field (100 μm standard deviation) is much larger than that found in midget ganglion cells (25 μm) (Figure 2C). The surround fields of small bistratified cells are smaller than those of midget ganglion

cells (140 μm and 205 μm , respectively), so these cells show relatively weak spatial opponency. One additional consequence of the S-cone specificity of the small bistratified cells is that the S-ON center, LM-OFF surround organization extends into the peripheral visual field, whereas midget ganglion cells lose spectral opponency beyond the parafovea. There are no S cones in the central 0.3 degrees of the visual field, so foveal vision is effectively color-blind to color variation mediated by S cones.

Midget ganglion cells project to the parvocellular layer of the LGN, and thence to layer 4C β of V1. This pathway, and its onward projections, is known as the P-channel. The P-channel conveys information about long and medium wavelengths and fine detail. It has been suggested that small bistratified ganglion cells conveying short wavelength information also contribute to the P-channel. However, it is now widely believed that small bistratified cells drive a distinct class of geniculate cells. The P-channel does contribute to motion perception; however, its contribution is weaker than that of the M-channel and nonveridical—the speed of perceived motion depends on the chromatic contrast of the stimulus.

Small bistratified ganglion cells form the start of the K-channel (Hendry and Reid, 2000). They project to koniocellular neurons in the LGN, distinguished from magno and parvo cells on the basis of their cell membrane chemistry. These cells mainly form layers intercalated between the parvo- and magnocellular layers, but some K-cells are also found in the parvocellular layer, with a smaller number being found in the magnocellular layer. K-cells project not only to layer 1 of V1, but also directly to V2. There is a particularly rich innervation of V2 by K-cells with foveal receptive fields. K-cells' receptive fields are large (at least as large as those of cells in the magnocellular layer) and often have irregular shapes. K-cells convey information contributing to color sensations, depending on contrasts of the output of S cones to combinations of M and L cone outputs; they may also contribute to motion perception.

The position summarized above remains controversial and has been challenged on a number of counts. In particular, it has been argued that the K-channel alone conveys chromatic signals (including L versus M information), while the P-channel is dedicated to fine spatial vision (Calkins and Sterling, 1999).

Primary Visual Cortex

The M, P, and K pathways project to groups of cells within V1 that can be distinguished on the basis of cytochrome oxidase reactivity (Livingstone and Hubel, 1984). K and P, but not M, pathways innervate cytochrome oxidase-stained regions known as blobs. P and M, but not K, pathways innervate the remaining regions, known as interblobs. There is recent evidence that cells show different specificities for wavelength processing in V1 (Conway, 2001; Johnson, Hawken, and Shapley, 2001). The cells discussed earlier in this article had a "single-opponent" organization. They can convey information about the intensities of light of particular wavelengths while being relatively uninfluenced by other wavelengths. They cannot, however, convey information about wavelength contrast. This requires "double-opponent" cells in which a central receptive field excited by one wavelength and inhibited by another is surrounded by a field in which the same two wavelengths have the opposite actions (Figure 1D). Double-opponent organization allows a cell to convey a consistent response to the boundary between two surfaces, regardless of the light illuminating them. If the illuminant changes, for example lengthening in wavelength, then longer wavelength light will be reflected from both sides of the boundary. Consider a double-opponent cell whose central receptive field is excited by long wavelengths and inhibited by medium wavelengths and whose surrounding field is inhibited by long wavelengths and excited by medium wavelengths. Imagine that the cell's receptive fields fall on a boundary between a pair of surfaces,

one of which is good and one poor at reflecting long wavelength light, so that the good reflector falls in the cell's central field. The net result will be excitation—that ratio of long to medium wavelength light is high in the central field and low in the surround. When the light illuminating both sides of the boundary lengthens in wavelength, the L/M ratios in both the excitatory center and the inhibitory surround will increase. The response of the cell is therefore largely unaffected by a change in illuminant. Obviously, such cells perform the preliminary computation necessary for color constancy. Of course, their responses only indicate spatially local changes in surface reflectance. To recover absolute reflectances throughout a scene, then, one also needs to estimate the response likely to be elicited by some fixed “anchoring” color in that scene and then to integrate local border contrasts from that anchoring point (see, e.g., Gilchrist et al., 1999, for similar arguments with respect to lightness perception). Until recently, evidence for the existence of double-opponent cells was controversial; however, recent findings indicate that such cells occur in V1 and, moreover, are sensitive to the orientation of chromatic (wavelength-dependent) borders as well as to the contrast of cone ratios across them.

Extrastriate Cortex

The clinical condition of cerebral achromatopsia, in which patients lose the ability to perceive color not as a result of retinal abnormalities but rather as a consequence of brain damage, provides strong evidence that brain areas specialized for color perception exist beyond striate cortex. The identification of these areas is, however, wreathed in controversy. The damaged areas include extrastriate cortex in the vicinity of the fusiform and lingual gyri. Neuroimaging studies have also shown increases in cerebral blood flow (implying increased brain activity) in these areas when normal subjects observed colored scenes. Zeki et al. (1991) therefore suggested that there was a specific color center in human extrastriate cortex. Early studies in which the responses from single neurons in monkeys were recorded in response to visual stimuli suggested that the color center might correspond to cortical area V4. A number of problems arose with this interpretation. The selectivity of the response of neurons to particular characteristics of stimuli differs only in degree between brain areas. Some neurons in nearly all visual areas respond selectively to wavelength; the proportion in V4 is not comparatively large. In addition, damage to area V4 in monkeys did not cause deficits in discriminations based on wavelength, although deficits were induced by damage to areas anterior to V4 (Heywood and Cowey, 1998). These findings appeared consistent with neuroimaging studies in humans identifying a color-selective area anterior to V4, christened V8 (Hadjikhani et al., 1998). Whether V8 really corresponds to the anterior areas that, when damaged, caused deficits for wavelength discrimination in monkeys remains controversial.

Wavelength Information Contributes to More Than Color Perception

The fact that our visual system can disambiguate wavelength and intensity makes it possible to ignore variations or sharp changes in intensity caused by shadows. One role of a wavelength-selective visual system is therefore segmentation of the visual scene on the basis of chromatic boundaries. Often chromatic boundaries will provide better cues for segmenting objects from their backgrounds than brightness boundaries—for example, in the dappled sunlight of a forest floor.

The residual abilities found in cerebral achromatopsia indicate that wavelength is exploited in more than one way. Although cerebral achromatopsics deny a phenomenal experience of color and

cannot discriminate between stimuli differing only in wavelength, they can effortlessly perceive boundaries between areas differing only in wavelength (Heywood, Kentridge, and Cowey, 1998). Their ability to use wavelength information to perceive form or motion, but not to perceive color, suggests that these functions may have distinct anatomical bases. Destruction of the putative color center, be it V4 or V8, disrupts the perception and experience of color, but not other functional uses of wavelength.

Discussion

Some of the earliest insights into the coding of color derived from work on color mixing. Following his discovery of the composition of white light, Newton developed the concept of the color circle, an arrangement of light sources around the periphery of a circle in which the mixture of any pair of diametrically opposite lights would produce white. Despite the color circle showing a continuum of light sources, Newton identified five primary colors (red, yellow, green, blue, and a violet-purple). However, attempts were soon made to discover how few colors were required in order to produce all other colors by mixing. Although there was some disagreement about which colors were primary, it was apparent to most investigators that three were sufficient. This culminated in the Young-Helmholtz trichromatic theory of color vision. Young believed that the primaries were red, green, and violet.

The fact that we require three primary colors in order to produce the full range of colored sensations reflects the fact that we have photoreceptors sensitive to three distinct wavelength distributions. The consequence is that any combination of lights that produces the same amount of activation in the three receptor types will produce the same response in the visual system and the same perception of color. There are, therefore, a large number of colors that are potential primaries.

Other features of color perception suggest an alternative to the trichromatic theory. In particular, there are limits to our abilities to see pairs of colors tinting one another. People perceive bluish reds and yellowish reds, but never greenish reds; they perceive reddish yellows and greenish yellows, but never bluish yellows. These opponent color pairings, red-green and blue-yellow, are also apparent in afterimages, color shadows, and color contrast. Observations such as these led Hering in 1905 to suggest a four-color opponent-process theory of color vision.

Both these theories assumed that the similarities between color sensations are completely determined by the outputs of the wavelength-dependent neurons in the visual system. For example, it is tempting to believe that opponent processes operating in V1 are the direct precursors of our space of colors. They have been taken as the sources of four primary colors (red, green, yellow, and blue) that are irreducible to other colors, and each contains one sensation that is pure (unique) in that it contains no trace of any other primary. However, the outputs of the cells in V1 would not produce these unique colors even if there were agreement as to what they might be (Saunders and van Brakel, 1997; Webster et al., 2000). Nor would the categories of color arise from variation in discrimination across the visible spectrum. The wavelengths at which there are minima in threshold do not correspond to the boundaries between primary colors. Controversially, it has been argued from cross-lingual evidence that color categories are determined by the speaker's color terms. The neurophysiology produces a given percept, but the assignment of that percept to a color category is a matter of agreement among observers.

Road Map: Vision

Related Reading: Contour and Surface Perception; Retina

References

- Bowmaker, J. K., 1998, Evolution of color vision in vertebrates, *Eye*, 12:541–547.
- Calkins, D. J., and Sterling, P., 1999, Evidence that circuits for spatial and color vision segregate at the first retinal synapse, *Neuron*, 24:313–321.
- Conway, B. R., 2001, Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1), *J. Neurosci.*, 21:2768–2783.
- Dacey, D. M., 2000, Parallel pathways for spectral coding in primate retina, *Annu. Rev. Neurosci.*, 23:743–775. ♦
- Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X. J., Spehar, B., Annan, V., and Economou, E., 1999, An anchoring theory of lightness perception, *Psychol. Rev.*, 106:795–834.
- Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., and Tootell, R. B. H., 1998, Retinotopy and color sensitivity in human visual cortical area V8, *Nature Neurosci.*, 1:235–241.
- Hendry, S. H. C., and Reid, R. C., 2000, The koniocellular pathway in primate vision, *Annu. Rev. Neurosci.*, 23:127–153. ♦
- Heywood, C. A., and Cowey, A., 1998, With color in mind, *Nature Neurosci.*, 1:171–173.
- Heywood, C. A., Kentridge, R. W., and Cowey, A., 1998, Cortical color blindness is not “blindsight for color,” *Consciousness Cognit.*, 7:410–423.
- Johnson, E. N., Hawken, M. J., and Shapley, R., 2001, The spatial transformation of color in the primary visual cortex of the macaque monkey, *Nature Neurosci.*, 4:409–416.
- Livingstone, M. S., and Hubel, D. H., 1984, Anatomy and physiology of a color system in the primate visual cortex, *J. Neurosci.*, 4:309–356.
- Saunders, B. A. C., and Van Brakel, J., 1997, Are there non-trivial constraints on colour categorization? *Behav. Brain Sci.*, 20:167–228. ♦
- Sumner, P., and Mollon, J. D., 2000, Catarrhine photopigments are optimized for detecting targets against a foliage background, *J. Exp. Biol.*, 203:1963–1986.
- Webster, M. A., Miyahara, E., Malkoc, G., and Raker, V. E., 2000, Variations in normal color vision: II. Unique hues, *J. Opt. Soc. Am.*, 17:1545–1555.
- Zeki, S., Watson, J. D. G., Lueck, C. J., Friston, K. J., Kennard, C., and Frackowiak, R. S. J., 1991, A direct demonstration of functional specialization in human visual-cortex, *J. Neurosci.*, 11:641–649.

Command Neurons and Command Systems

Jorg-Peter Ewert

Introduction

If a certain interneuron is stimulated electrically in the brain of a marine slug, the animal then displays a species-specific escape swimming behavior, although no predator is present. If a certain brain area of the optic tectum of a toad is stimulated in this manner, snapping behavior is triggered, although no prey is present. In both cases, a commanding trigger, which in the natural situation is associated with adequate predator or prey signals, activates an appropriate motor program. We address the notion “command” to a stimulus which elicits a rapid ballistic response. The transformation involves sensory pattern recognition and localization on the one side and motor pattern generation on the other; command functions provide the sensorimotor interface (Kupfermann and Weiss, 1978). This interface translates a specific pattern of sensory input mediated by sensory neurons (SN) into an appropriate spatiotemporal pattern of activity in premotor and motor neurons. The latter coordinate the corresponding action pattern, which, in various animal groups and depending on the task, may be rather stereotyped (fixed action pattern) or may leave room for variability (modal action pattern). This correspondence can be innate, modified innate, or acquired, strategy related. Commands are motivated. Once an action pattern is commanded, it tends to carry on to completion, although there may be cases in which a command is “countermanded.” How does the sensorimotor interface operate?

Before we tackle this question in depth, a few comments on pattern generation are in order (see MOTOR PATTERN GENERATION). In the present context, a motor pattern generator (MPG) is defined as an internuncial network that, in response to a commanding input, coordinates appropriate muscle contractions. The network is activated if and only if a specific (combination of) input occurs. An intrinsic pattern of neuronal connectivity in the network ensures the generation of a consistent spatiotemporal distribution of excitation and inhibition, often involving oscillatory circuits. The output of the network, mediated by premotor neurons, has privileged access to the requisite motor neuronal pools. Proprioceptive and internal feedback—or feedback from components of the motor network to the command neuron(s)—can play a role in the coordination and maintenance of a motor pattern (e.g., Jing and

Gillette, 2000). Hence, command neurons can be involved to carry timing information for the MPG. Recurrent feedback from the motor system to the command and from the command to the sensory afferents may terminate the command and may also prevent sensory input during the animal’s movement.

There are cases in which a short-term command *triggers* the corresponding MPG whose activity outlasts the command activity. In other cases, a sustained (tonic) command is necessary to drive the MPG. The efficacy of a triggering or a driving command depends on the requirements under which an action pattern is released: (1) locus of the stimulus, (2) presence of adequate stimulus features, (3) behavioral state (dominance), (4) motivation (variable on a relatively long time scale), attention, and arousal (variable in the short-term), (5) gating input, and (6) evaluation of the stimulus in connection with prior experience. The first two listed concern aspects of the releasing stimulus; the latter four refer to modulatory functions.

Sensorimotor interfaces with command functions occupy an important topic in neuroethology, the science of the neurobiological fundamentals of behavior (e.g., see Carew, 2000). Neuroethology contributes a valuable perspective to neuroscience of which the present article presents one component.

The Command Neuron Concept

The question as to whether a sensorimotor command can be mediated by a command neuron (CN),

sign stimulus → SN → CN → MPG → behavior pattern

was and occasionally still is controversial (Kupfermann and Weiss, 1978; Eaton, Lee, and Foreman, 2001). After a definition by Kupfermann and Weiss (1978), a command neuron (CN) is an interneuron whose excitation is both necessary (n) and sufficient (s) to activate the corresponding MPG. Test criteria of this n&s condition include (1) the recording of the activity from the putative CN during the presentation of a stimulus signal in registration with the corresponding action pattern (link between stimulus and motor activity), (2) electrical stimulation of the putative CN with demonstration that this action pattern is executed (s-criterion), and (3)

removal of the putative CN and demonstration that this action pattern is no longer elicited by the stimulus signal (n-criterion).

The best candidates for this approach are among the identified CNs of various invertebrates, such as the pair of lateral giant fibers in crayfish that, in response to mechanical stimulation, trigger the fast tail flip escape reaction. For the function of the lateral giant, the term “command” was first introduced by C.A.G. Wiersma and K. Ikeda (for a review, see Edwards, Heitler, and Krasne, 1999). Among vertebrates, there is the reticulospinal Mauthner cell of teleost fish, which, in response to certain acoustic-vibratory stimulation, commands the fast C-shaped body bend escape reaction, called “C-start” (Eaton et al., 2001). The C-start is triggered by a spike in one of the bilateral pair of Mauthner cells.

However, in such cases in which identified putative CNs activate corresponding MPGs, it was not trivial to evaluate the n&s condition. Eaton and co-workers performed a scholarly structured examination on the Mauthner cell (see Eaton et al., 2001). Surprisingly, it turned out that the n-criterion was not fulfilled clearly; after both Mauthner cells in the goldfish were lesioned, the C-start escape reaction could still be elicited by acoustic sign stimuli—however, at a longer latency. This suggests that appropriate auditory receptive “backup” systems, which are normally inhibited by Mauthner cell activation, commanded the C-start. Interestingly, also the s-criterion was not fulfilled unequivocally, because electrical stimulation of the Mauthner cell did not trigger the complete C-start. This suggests that for the full C-start pattern, the activity in other, parallel descending reticulospinal cells is required, too.

Is the CN concept restricted to very few cases, or is it even obsolete? Regarding unknown network interactions (serial, parallel, convergence, divergence, feedforward, and lateral inhibition), the n&s condition may be too rigorous in many cases. Depending on a sensorimotor function, we therefore envision a spectrum of possibilities by which command functions can be executed. The idea behind the revisited command concept (cf. Eaton et al., 2001) is furthermore fruitful both in the neurophysiological analysis of behavior (e.g., Ewert, 1997; Edwards et al., 1999) and in perceptual robotics (e.g., Lukashin, Amirikian, and Georgopolous, 1996; REACTIVE ROBOTIC SYSTEMS).

Experiments to test the n&s condition of particular neurons for triggering behavioral acts are a historical fact that is part of the research record on these types of systems. However, this should not distract us from the complex results that emerge from the experiments done to test this paradigm or from the complex ways in which these cells really function.

Command (Releasing) Systems

Plastic ballistic behavior elicited by distributed networks is triggered by populations of command-like interneurons that form a command system (CS) (Kupfermann and Weiss, 1978). Such a CS consists of collectively operating neurons, each of which is called a command element (CE) or command-like neuron. For just one CE of a CS, the n&s condition cannot be fulfilled. For example, if the CEs of a CS are connected to the MPG like a logical AND-gate, in this case a CE fulfills the n-criterion but not the s-criterion; if the CEs of a CS are connected to the MPG like a logical OR-gate, in that case a CE fulfills the s-criterion but not the n-criterion. However, when the CEs of a CS are treated as a unit, the CS will meet the n&s condition in both cases. We learn that the CS notion has a much broader applicability than the CN notion. In fact, the pair of identified bilateral cerebral interneurons of the type *cc5* in the marine snail *Aplysia* works conjointly as a two-neuron CS that fulfills the n&s condition for the bilateral pedal arterial-shortening component of defensive head withdrawal behavior (Xin, Weiss, and Kupfermann, 1996).

In Mauthner cell-initiated escape behavior, the current concept suggests that the C-start is elicited by a CS consisting of at least two groups of command-like reticulospinal neurons (CEs). One group, in which the Mauthner cell participates, controls the agonist muscular contraction; another group controls the antagonist contraction that, depending on the direction of the acoustic stimulus, precisely shapes the trajectory for certain C-start escape angles (Eaton et al., 2001). If we ask how the two parts of the CS control trajectory, it appears, for example, that the agonist and antagonist CEs can vary their respective output magnitudes and timing patterns to regulate the escape trajectory. This example introduces more sophisticated CSs and relates to the question whether a “backup” system really exists. Since it is difficult to explain how a backup system could evolve, it is a simpler interpretation to recognize the parallel nature of the CS. The role of the parallel output may be to produce variable trajectory angles. A consequence of its organization is that it is “robust” such that if one element (e.g., a Mauthner cell) does not participate, the behavior still can occur. But it could well be that an organism that is missing one Mauthner cell may not be able to accurately control its trajectory.

The concept of a command-releasing system (CRS) is suggested for functions requiring a more complex perception of the sensory input (Ewert, 1997). Different types of electrically uncoupled CEs (a population of each) cooperatively trigger an MPG. Each type of CE evaluates a certain aspect of the releasing stimulus signal, such as features or feature combinations and their location in space, respectively; we call this a *sensorimotor code*. The concept of coded commands stresses that firing of various types of efferent neurons (CEs) in a certain combination (CRS) characterizes an object in space and selects the appropriate goal-directed action pattern (MPG). In amphibians, such cells with different visual response properties project their axons from the optic tectum (T-type neurons) or thalamic/pretectal nuclei (TH-type neurons) to the bulbar/spinal motor systems via discrete tecto-bulbarspinal and pretecto-bulbar/spinal pathways (Figure 1). The hypothesis illustrated in Figure 1 suggests the following:

- The code {T4, X, T5.2} says: “an object is moving anywhere in the visual field [T4]” and “the object occurs at a certain position in the visual field x-y coordinates [X]” and “the object has prey features [T5.2]” → *orient!*
- The code {T1.3, T3, T5.2} says: “an object is moving in the frontal binocular field of vision at a narrow distance [T1.3]” and “the small object is approaching [T3]” and “the object has prey features [T5.2]” → *snap!*
- The code {TH6, T6} says: “a large object is moving in the dorsal visual field [T6]” and “the large object is approaching [TH6]” → *duck!*

After the hypothesis, in such a multifunctional network, on one hand, the same goal can be reached by differently combined CRS whose CEs may be distributed in different brain structures; on the other hand, certain CEs can be shared by different CRSs for different goals. Evidence of such multifunctional properties, for example, is provided for *Aplysia* by Xin et al. (1996): The type *cc5* neuron contributes to different aspects of behavior, such as defensive head withdrawal, tentacular withdrawal, feeding, and locomotion. For some (aspects of) behaviors, the neuron probably acts as CE; for other aspects or components of behavior, however, the same interneuron may participate in a distributed circuit where it is neither necessary nor sufficient.

Also, Mauthner cells participate in behaviors other than escape, such as prey capture. The story of Mauthner cell function may be even more complex. In the electric fish *Gymnotus carapo*, auditory activation of Mauthner cell leads, by its connections to medullary pacemaker neurons of the electric organ, to an abrupt and pro-

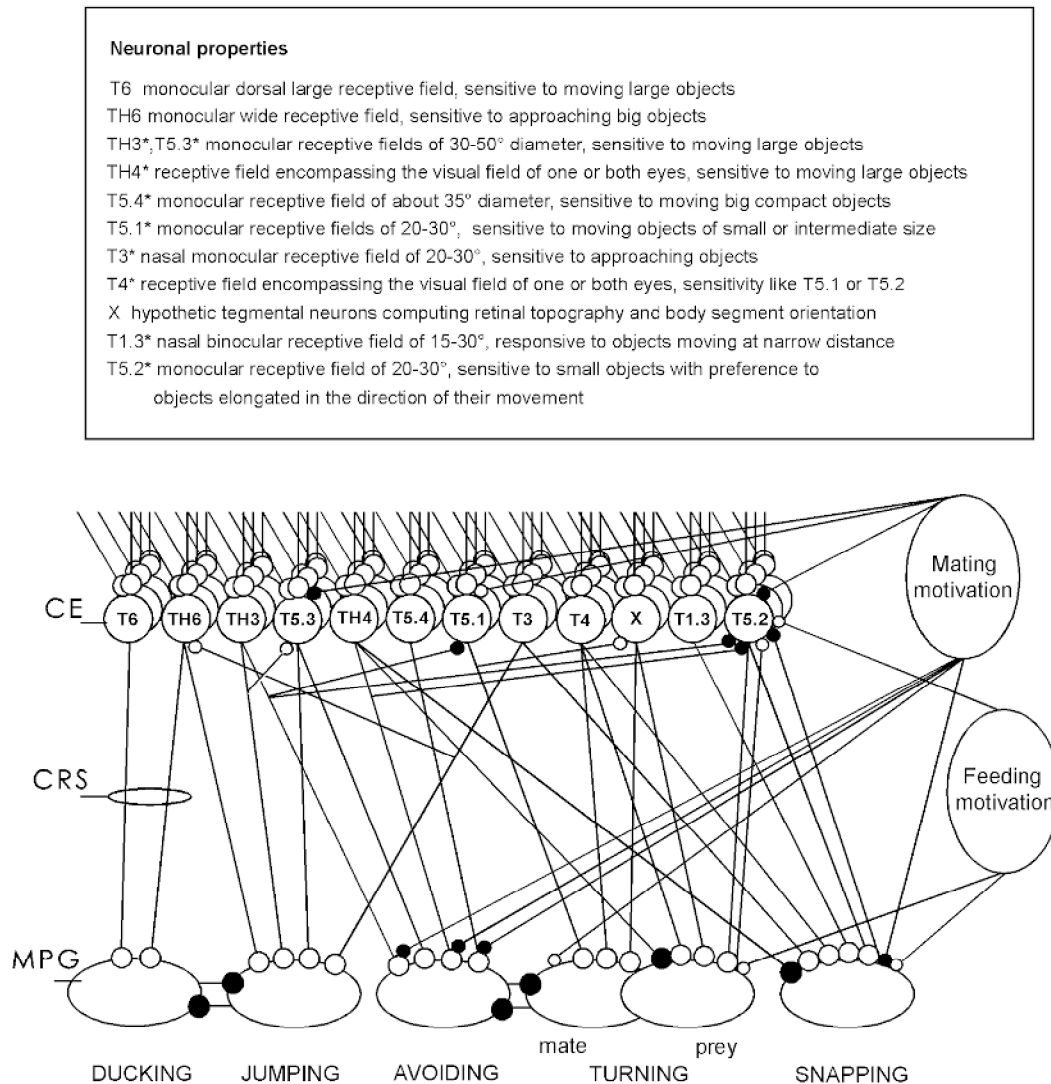


Figure 1. Concept of sensorimotor codes implemented by command-releasing systems (CRSs) in common toads. Different subsets of functionally identified types of visual thalamic/pretectal (TH) and tectal (T) neurons as command elements (CE) contribute to commanding and parameterizing motor patterns of feeding (prey orienting, snapping), escape (ducking, jumping, avoidance turning), and orienting by the male toad toward the female in the mating season. For example, *T5.1* neurons respond sensitively to prey objects, whereas *T5.2* neurons display prey-selective properties with respect to configurational visual cues; *T6*, *TH6*, *TH4*, and *T5.4* neurons

respond best to various aspects of predator stimuli. Breeding motivation and hunger state deliver different modulatory influences. Each circle represents a set of neurons, each oval stands for a motor-pattern-generating circuit (MPG) with access to motor neuronal pools. The open dots refer to excitatory influences, and the solid dots to inhibitory influences; the CE level obtains sensory inputs, for example, from retinal ganglion cells and interneurons. The asterisk indicates that the projective character of the neuron (command element) is evidenced by means of the antidromic stimulation/recording method. (Modified from Ewert, 1997.)

longed increase in the rate of discharges of the electric organ. This probably enhances the electrolocative sampling of the environment during Mauthner cell mediated escape or prey-catching behavior (Curti et al., 1999).

Properties of Commands

With respect to requirements 1 through 6 listed in the introduction, the properties of a command can be manifold. We select three aspects.

Localization

Certain CEs of a CRS monitor the stimulus in space to select and direct the behavior in relation to the target. Visual space, the x - y - z

position of an object, are translated into appropriate motor space. In mammals, the superior colliculus is involved in visual orientation of the eye, head, and trunk. Information about the position of these movable segments must be integrated in the topographic correspondence between sensory input and motor output. For example, the sensory map in the collicular superficial layers is transformed into a motor map in the deep layers, in which a vector from an initial eye position to a goal eye position is represented. Hypotheses concerning gaze, eye, and head displacement commands are discussed by Freedman and Sparks (1997).

Information on visual depth can be obtained in various ways, such as by binocular vision or motion parallax. Such information is necessary not only to select appropriate behavior, but also to estimate the real size of the target.

Feature Analysis

Certain CEs of a CRS are involved in the feature analysis of the sensory signal. For example, the toad's visual system selects between prey and nonprey through an analysis of the moving object in terms of configurational features, namely, the dimensions of an object parallel versus perpendicular to its direction of movement. A bug or a millipede is thus not recognized explicitly; rather, they are implicit in the prey schema determined by an object-features-relating algorithm (Ewert, 1997). The discharge rate of toad's prey-selective T5.2 neurons in response to a moving object is correlated with the probability that the configuration of this object fits the prey schema. These neurons, as CEs, obviously display predictive responses, since they show a strong premotor buildup of activity before the animal orients toward prey, but they are silent before or during a spontaneous head movement. In the response properties of T5.2 neurons, certain kinds of interactions of distributed tectal (T) and thalamic/pretectal (TH) neuronal populations are expressed. This feature analyzing T/TH filter system is integrated in a macronetwork that allows various kinds of modulation (Figure 2), for example, with respect to response gating (involving basal/ganglionic nuclei) or to raising or lowering the classification threshold depending on motivation (involving preoptic/hypothalamic structures) or to generalization or specification by learning (involving the hippocampal ventral medial pallium).

Motivation and Attention

A command cannot operate effectively if state-dependent inputs are not appropriate. If a toad is inattentive or not motivated to catch prey, the prey-selective T5.2 neurons will continue to discharge, at a lower rate, in response to configurational visual prey features; however, they do not show the strong buildup of activity that in prey-responsive animals precedes and predicts grasping the prey with the tongue. Comparably, neurons of area 5 of the posterior parietal association cortex in monkeys discharge strongly before the animal performs a ballistic grasping movement with its arm toward a rewarding banana. In keeping with the proposal by Mountcastle and co-workers, these "arm projection neurons" function in a broad sense as a command system for the motivated exploration of extrapersonal space (for a complementary point of view, see GRASPING MOVEMENTS: VISUOMOTOR TRANSFORMATIONS). However, if the monkey was satiated, the banana did not elicit the predictive buildup of activity in these neurons required to elicit grasping. So whether or not the s-criterion is met depends on the behavioral state. This illustrates again one of the problems with rigidly designating neuronal function according to the n&s paradigm: Whether the cell is either n or s for a behavior depends on the conditions.

In crickets, the efficacy of the calling-song commanding interneuron, too, depends on the behavioral state (Hedwig, 2000). In silent animals that were previously singing, experimentally evoked commands by this interneuron triggered the song according to the s-criterion. In the resting state, however, experimentally evoked commands by the interneuron elicited an incomplete calling song, so the interneuron failed to meet the s-criterion.

Behavioral Choice

Behavioral dominances exist, and these may depend, for example, on the season due to hormonal influences. In summertime, the toad's predator-avoiding behavior overrides prey-catching behavior if appropriate visual releasers are present simultaneously, whereas the male toad's mate-approaching behavior to an adequate stimulus is absent. In spring during the mating season, mate-approaching behavior dominates predator-avoiding behavior,

whereas prey-catching behavior fails to occur. The behavioral choice may result from dual excitatory/inhibitory influences to the toad's tectal or thalamic/pretectal CEs and/or MPGs (cf. Figure 1).

The rat's colliculus superior (SC), too, in response to certain unexpected visual stimuli can initiate rapid ballistic opposite behaviors such as orienting/approaching versus freezing/avoiding. Each type of behavior is commanded by a different set of neurons located in the intermediate or deep SC layers, respectively, that project their axons in discrete phylogenetically old SC-bulbar/spinal pathways (homologous to the tecto-bulbar/spinal pathways in amphibians). Krout et al. (2001) provide anatomic evidence for the hypothesis that the decision in the SC to select one of these behaviors—approach versus avoidance—in an appropriate stimulus situation involves visual projections from SC to (1) thalamo-basal/ganglionic, (2) thalamo-amygdaloid, and (3) thalamocortical circuits. It is suggested that loop 1 is involved in behavioral choice via disinhibitory (gating) and inhibitory (selecting) pathways, and the connections with loop 2 add information related to emotional learning and memory, whereas the connections with loop 3 allow changes in strategy.

Studies in invertebrates show that there are several ways of interaction to explain the suppression of one of two mutually exclusive behaviors, each one organized by a network N1 and a network N2, respectively.

1. An influence of network N1 cancels the command in N2: for example, dominance of escape swimming over feeding in the snail *Pleurobranchaea*; the behavioral switch is caused by swim-induced inhibition of feeding command neurons, that is, activation of feeding interneurons that inhibit the feeding command neurons (Jing and Gillette, 2000).
2. An influence of network N1 suppresses both the command in N2 and the MPG in N2: for example, choice of the body shortening withdrawal behavior in favor of swimming in the leech; in response to mechanical stimuli, body shortening is commanded by a set of interneuronal parallel pathways that inhibit one of the swim command interneurons and an interneuron of the swim MPG (Shaw and Kristan, 1999).
3. An influence of network N1 inhibits the MPG in N2 but not the command in N2: for example, in stridulating crickets, wind-evoked signals that are silencing the calling song do not inhibit the ongoing activity of the interneuron that commands this song but do inhibit the song MPG. This allows immediate singing when the wind wanes (Hedwig, 2000).

CRS and Schema Theory

A command-releasing system (CRS) can be regarded as the neurobiological correlate of Nico Tinbergen's concept of (innate) releasing mechanism (see Ewert, 1997), originally called by Konrad Lorenz "the (innate) releasing schema."

SCHEMA THEORY (q.v.) offers an interdisciplinary science that allows one to treat principles of neuroethology or neural engineering in the same language. In the language of schema theory, the sensorimotor code of a CRS embodies a *perceptual schema* that exists for only one purpose: to determine the conditions for the activation of a specific MPG embodying a *motor schema*. The CRS must also ensure that the resultant movement is directed in relation to the target. A schema and its instantiation usually are coextensive; that is, instantiation of a schema appears to be identifiable with appropriate activity in certain populations of neurons of the brain, whereby each schema may involve several cell types or brain regions, while a given cell type or brain region may be involved in several schemas. The motor schemas of directed appetitive behaviors and consummatory behavior need not occur in a fixed order; rather, each may proceed to completion, followed by perceptual

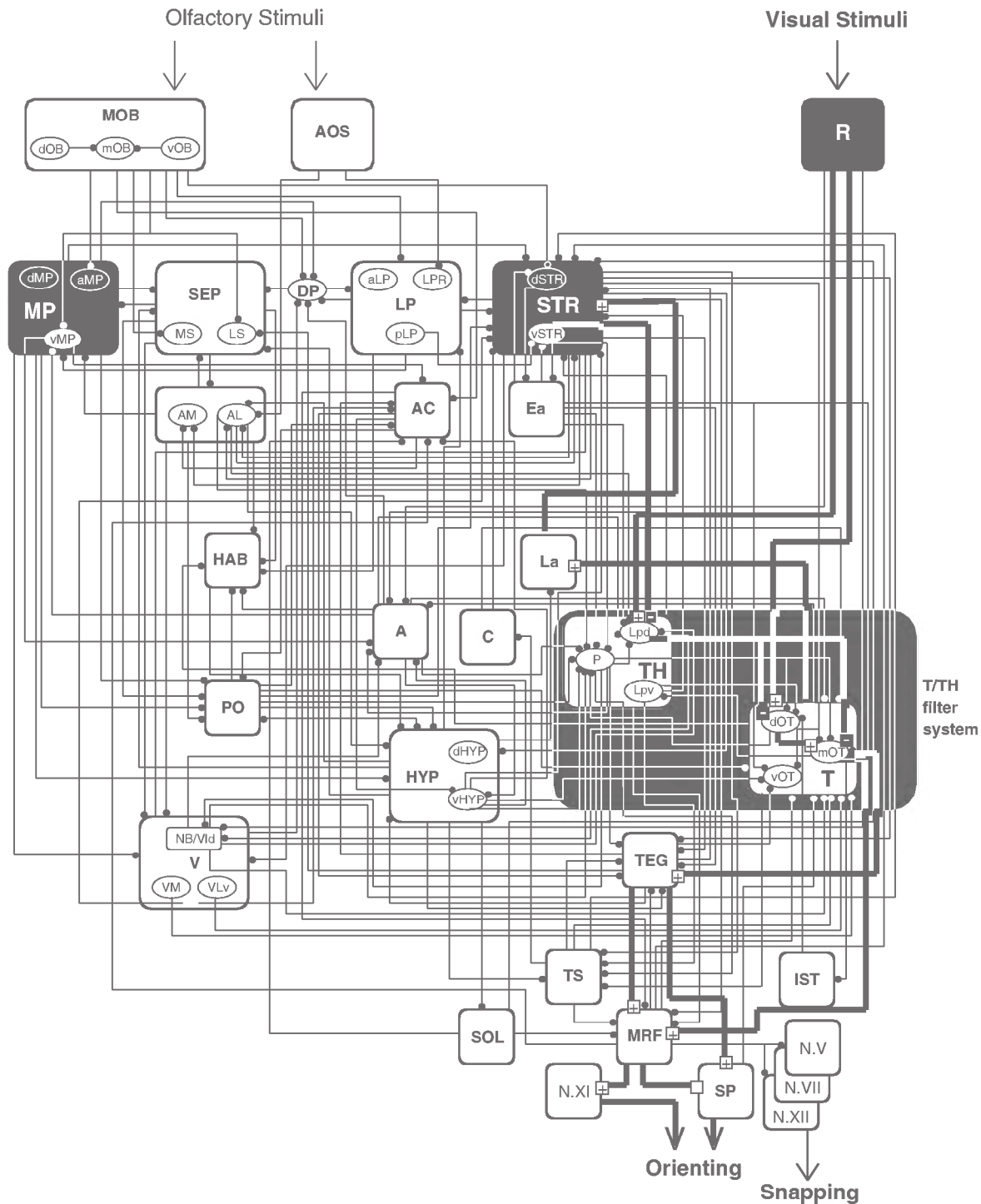


Figure 2. The visual configurational prey schema of common toads results from interactions between retina (R)-fed optic tectal (T) networks and retina-fed thalamic/pretectal (TH) networks. The former network evaluates the dimension of an object parallel to the direction of its movement; the latter evaluates its dimension perpendicular to the direction of movement. Pretectotectal inhibitory influences determine the prey-selective properties of tecto-bulbar/spinal projecting neurons of the type T5.2 (cf. Figure 1). The T/TH filter system (see the black labeled large area) is integrated in a macronetwork, whose components may influence this system in different ways. For example, the basal ganglionic ventral striatum (vSTR) is involved in gating the orienting response toward prey (cf. thick labeled lines): retinal (R) output is fed both to T and TH (Lpd nucleus); vSTR obtains T-infor-

mation via the lateral anterior thalamic nucleus (La); a descending striato(vSTR)-pretecto (Lpd)-tectal (T) disinhibitory connection gates the tecto (T)/tegmento (TEG)-bulbar (MRF)/spinal (SP) processing stream. As a result, visual perception (prey recognition) is translated into action (prey-catching orienting). Suggested excitatory and inhibitory influences are indicated by plus and minus signs, respectively. Other, not labeled, loops involving the hippocampal ventral medial pallidum (vMP) via anterior thalamic (A), preoptic (PO), and hypothalamic nuclei (HYP) modify the selectivity of the prey schema in the course of nonassociative or associative learning, such as by combining visual and olfactory cues. (For details, see Ewert et al., 2001.)

schemas that will determine which motor schema is to be executed next. Schemas may be linked by so-called *coordinated control programs*. Motor schemas, for example, can take the form of compound motor coordinations (such as a frog's programmed jump-snap-gulp sequence), which make up a set that will proceed to completion without intervening perceptual tests, for example, in such a manner that schema A proceeds to completion, and completion of schema A triggers the initiation of schema B, or that schema A passes a parameter X to schema B. It is also possible that two or more motor schemas are co-activated simultaneously and interact through competition and cooperation to yield a more complicated motor pattern.

Perspectives

Operations between sensory analysis and motor response represented by sensorimotor codes (CRS) are of general importance. Of particular interest are the alternative ways in which different organisms solve similar problems. For example, it was shown that the same algorithm that underlies configurational prey selection is implemented by such different neuronal networks as those of a toad, which is a vertebrate, and a praying mantis, which is an insect (Prete, 1992). This biologically justifies the artificial neural network approach. Cervantes-Pérez (see VISUOMOTOR COORDINATION IN FROG AND TOAD) demonstrates that such an algorithm can be modeled by neural networks applying principles of parallel-distributed processing and convergence. Pavlásek (1997) uses artificial neuronal networks to investigate how the precise timing and structuring of neural commands controlling goal-directed movements can be realized in general. Lukashin et al. (1996) apply a computational paradigm that utilizes the actual impulse population activity of directionally tuned cells recorded from the motor cortex of monkey during the performance of a motor task as command signals to drive an artificial mechanical device: an artificial neuronal network recodes the brain signals into the motor schema of a simulated actuator, a method that is suitable for electronically driven prostheses such as a multijoint artificial limb. This is just one example of the applications of research on command functions in sensorimotor systems for perceptual and reactive robotics (cf. also REACHING MOVEMENTS: IMPLICATIONS FOR COMPUTATIONAL MODELS; REACTIVE ROBOTIC SYSTEMS; and ROBOT ARM CONTROL).

Road Map: Motor Pattern Generators; Neuroethology and Evolution

Related Reading: Motor Pattern Generation; Neuroethology, Computational; Visuomotor Coordination in Frog and Toad

References

- Carew, T. J., 2000, *Behavioral Neurobiology: The Cellular Organization of Natural Behavior*, Sunderland, MA: Sinauer. ♦
- Curti, S., Falconi, A., Morales, F. R., and Borde, M., 1999, Mauthner cell-initiated electromotor behavior is mediated via NMDA and metabotropic glutaminergic receptors on medullary pacemaker neurons in a Gymnotid fish, *J. Neurosci.*, 19(20):9133–9140.
- Eaton, R. C., Lee, R. K. K., and Foreman, M. B., 2001, The Mauthner cell and other identified neurons of the brainstem escape network of fish, *Prog. Neurobiol.*, 63:467–485. ♦
- Edwards, D. H., Heitler, W. J., and Krasne, F. B., 1999, Fifty years of a command neuron: The neurobiology of escape behavior in the crayfish, *Trends Neurosci.*, 22(4):153–161. ♦
- Ewert, J.-P., 1997, Neural correlates of key stimulus and releasing mechanism: A case study and two concepts, *Trends Neurosci.*, 20:332–339. ♦
- Ewert J.-P., Buxbaum-Conradi, H., Dreisvogl, F., Glasgow, M., Merkel-Harff, C., Röttgen, A., Schürg-Pfeiffer, E., and Schwippert, W. W., 2001, Neural modulation of visuomotor functions underlying prey-catching behaviour in anurans: Perception, attention, motor performance, learning, *Comp. Biochem. Physiol. A*, 128:417–461.
- Freedman, E. G., and Sparks, D. L., 1997, Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: Evidence for a gaze displacement command, *J. Neurophysiol.*, 78:1669–1690.
- Hedwig, B., 2000, Control of cricket stridulation by a command neuron: Efficacy depends on the behavioral state, *J. Neurophysiol.*, 83(2):712–722.
- Jing, J., and Gillette, R., 2000, Escape swim network interneurons have diverse roles in behavioral switching and putative arousal in *Pleurobranchaea*, *J. Neurophysiol.*, 83(3):1346–1355.
- Krout, K. E., Loewy, A. D., Westby, G. W. M., and Redgrave, P., 2001, Superior colliculus projections to midline and intralaminar thalamic nuclei of the rat, *J. Comp. Neurol.*, 431:198–216.
- Kupfermann, I., and Weiss, K. R., 1978, The command neuron concept, *Behav. Brain Sci.*, 1:3–39. ♦
- Lukashin, A. V., Amirikian, B. R., and Georgopoulos, A. P., 1996, A simulated actuator driven by motor cortical signals, *Neuroreport*, 7(15–17):2597–2601.
- Pavlásek, J., 1997, Timing of neural commands: A model study with neuronal networks, *Biol. Cybern.*, 77(5):359–365.
- Prete, F. R., 1992, Discrimination of visual stimuli representing prey versus non-prey by the praying mantis *Sphodromantis lineola* (Burr.), *Brain Behav. Evol.*, 39:285–288.
- Shaw, B. K., and Kristan, W. B., Jr., 1999, Relative roles of the S cell network and parallel interneuronal pathways in the whole-body shortening reflex of the medicinal leech, *J. Neurophysiol.*, 82(3):1114–1123.
- Xin, Y., Weiss, K. R., and Kupfermann, I., 1996, An identified interneuron contributes to aspects of six different behaviors in *Aplysia*, *J. Neurophysiol.*, 16(16):5266–5279.

Competitive Learning

Nathan Intrator

Introduction

Competitive learning is described by a family of algorithms that use some sort of competition between lateral neurons during learning and normal activity. In a typical competitive network architecture, neurons in each layer are connected to the next layer, but, in addition, there are lateral connections between neurons in the same layer that cause the competition. Competitive learning includes a wide variety of algorithms performing different tasks, such as encoding, clustering, and classification.

The notion of a limited pool of individually adaptable tuned units is more than a functional abstraction useful for psychological mod-

eling. The reality of a distributed representational substrate whose members compete for a share in the representation of the stimulus has been demonstrated in a variety of electrophysiological studies (Gilbert, 1994). These studies can be divided into two major groups. In the first group, competition among members of a pool of tuned units has been enhanced by the withdrawal of stimulation from some of the units. For instance, the induction of a retinal scotoma leads to invasion of the visual space of affected cells by the receptive fields of other, neighboring cells. An analogous effect, albeit on a much longer time scale, is observed in the somatosensory modality (Merzenich et al., 1988). In comparison, in the sec-

ond group, the increased prominence of a subpopulation of functional units (whose emergence may in itself be a product of learning) is precipitated merely by the prevalence of the preferred stimuli of those units in the sensory repertoire of the system. For example, in the monkey, this phenomenon has been glimpsed in relation to face representation (Rolls et al., 1989), as well as the representation of general natural and artificial (Logothetis and Scheinberg, 1996) objects.

Overview

In recent years, competitive learning has received considerable attention because of its demonstrated applicability and its biological plausibility. The idea of competition between neurons leads to sparse representations of data that are easy to decode and conserve energy. Current competitive learning algorithms can be distinguished by their learning rule (which is driven by their desired objective function) or by the form and role of the competition during learning. One form of competitive learning algorithms can be described as an application of a successful single neuron learning algorithm in a network with lateral connections between adjacent neurons. The lateral connections are needed so that each neuron can extract a different feature from the data. For example, if one has an algorithm for finding the principal component in a data set, i.e., the first eigenvector of the correlation matrix of the data, then a careful application of this learning rule in a lateral inhibition (competitive) network gives an algorithm that finds the first few principal components of the data (Sanger, 1989). This set of algorithms can be characterized by the fact that if the lateral connections are turned off, then neurons are likely to learn the most easily extracted feature in the data (based on the objective function), or the feature that corresponds to the closest local minimum of the objective function. For example, in a principal components network with no lateral inhibition, all neurons will converge to the first principal component of the data. In such algorithms, it is often sufficient to study the learning rule of a single neuron and deduce from it the performance of the network as a whole.

A second family of algorithms is characterized by the fact that turning off the lateral connections will result in a great loss of ability to extract useful features. Thus, the competition between neurons has an important role in improving, sharpening, or even forming the features extracted from the data by each single neuron. This is an important distinction that suggests that competitive learning may be useful in cases where other forms of learning may have difficulties. One such example is a network that searches for clusters. Clustering is the process of grouping together data points based on some measure of distance. When no inhibition exists between neurons, it is likely that neurons will converge to the mean of the data distribution. This follows from the fact that based on a Hebbian rule, neuronal weights (of radial basis neurons) tend to converge to the center of the distribution in order to maximize correlation of input activity with output activity. An algorithm that looks for tight clusters will probably find those that are occurring with highest probability, or those cluster centers that are closest to the initial weight values. When inhibition is turned on, the cluster centers will move to new locations and will become sharper and more distinct. There is evidence that inhibition plays a similar role in various brain functions such as the creation of orientation columns in visual cortex (Ramoa, Paradiso, and Freeman, 1988) or sharpening receptive fields in sensory cortex (Merzenich et al., 1988).

There are other points of view on competitive learning, in particular, competition over presynaptic versus postsynaptic resources (Willshaw et al., 1997). A more detailed review of these issues can be found in Intrator and Edelman (1997).

Local Competition over Space

Frequently in a classification problem, there are regions in pattern space in which classification is easier, while there are other regions in which classification is not that simple and thus requires a larger (more complex) network. For example, there may be regions in which the different classes are linearly separable and other regions in which the boundary is highly nonlinear. Therefore, one of the key ideas in machine learning and statistical parameter estimation is recursive partitioning of the observation space in order to separate such regions. Motivation comes from the desire to study structure of high dimensional space by searching for homogeneous subregions that present a simpler structure than the original whole input space. Statistical theory suggests that the variance of the estimator, which directly influences generalization performance, is affected by the complexity of the estimator (polynomial degree, number of hidden units, etc.). Thus, performing the estimation with a simpler architecture is likely to improve generalization performance. However, if the network architecture is "too simple," in the sense that the architecture is not rich enough to allow a flexible enough function to fit the data, a large training error is unavoidable.

One solution to this trade-off is to recursively partition the data into several homogeneous regions so that a smaller network will suffice for each region. In this context we shall discuss in the next section the network of competing experts.

In a lateral competition network, a cluster center (the mean of a Gaussian distribution) is associated with a neuronal weight vector, and neurons compete between themselves to add each of the input patterns to their cluster. For a given pattern, the winner, in hard competition, is the neuron with highest probability for that pattern to belong to the Gaussian distribution represented by its center (see WINNER-TAKE-ALL NETWORKS). In soft competition, the cluster centers are organized in such a way that the distance of each data point from each of the cluster centers corresponds to the probability of this point under each of the Gaussian distribution (Jacobs et al., 1991).

If an input $x \in R^n$ is drawn with probability π_i from one of J independent Gaussians (for simplicity, assume that the Gaussians are radially symmetric), its a posteriori probability of belonging to the j th Gaussian is $\pi_j p_j(x)$, for

$$p_j(x) = \frac{1}{M\sigma_j} \exp\left(-\frac{\|x - \mu_j\|^2}{2\sigma_j^2}\right)$$

where $\sigma_j^2 I$ is the covariance matrix of Gaussian j , and M is a normalizing constant. The partial likelihood that measures the probability of the data under the model of input x_k in this model is $L(x_k) = \sum_j \pi_j p_j(x_k)$. The partial likelihood of a pattern set $\{x_k\}_{k=1}^K$ is

$$\mathcal{L} = \prod_k L(x_k)$$

assuming K independent patterns. This could be turned into a likelihood if the marginal probabilities are taken into account to convert the partial likelihood into a probability function, but since the marginals do not participate in the optimization, there is no need to do that. Parameter estimation of such a model involves adjusting the centers $\{\mu_j\}$, the covariances $\{\sigma_j^2\}$, and the prior probabilities $\{\pi_j\}$ so as to maximize the likelihood of the model for a given training set. It is mathematically equivalent but more convenient to maximize the log likelihood, since the maximization is then on summation and not multiplication. In a radial basis function network realization of this model, there are J radial basis hidden units and a linear output unit that gives the likelihood function under the conditions of the model.

Assuming for simplicity that σ_j are equal, the hard competition approach for solving this problem would assume that each input can belong to only one cluster center. (Thus, the competition be-

tween neurons will lead to only one active neuron at a time), and, therefore, the above definition of $P(x)$ is replaced by

$$P(x) = \max_i \pi_i p_i(x)$$

The analytic solution for cluster centers (under assumptions about radial symmetry and σ_i and π_i being equal) is

$$\mu_i = \frac{\sum_{k \in C_i} x_k}{N_i}$$

where C_i is the set of training patterns that were assigned to belong to the cluster center i and N_i is the number of these patterns. Note that this is not an explicit solution, since C_i depend on all of the μ_k s. A learning rule that converges to this solution is given by

$$\Delta \mu_{ij} = \eta(x_{ki} - \mu_{ij})p_i(x)$$

for the winning cluster i , and no change for the other clusters. This learning rule moves the closest cluster center in the direction of the data points that belong to that cluster, finally forming stable cluster boundaries and converging to the mean of the cluster. During this process, input patterns may "cross" from one cluster to another as a result of the movement of the cluster center. However, the process will converge to a stable solution simply because the mean squared distance of the patterns from their corresponding cluster centers goes down.

The soft competition paradigm assumes that each pattern can belong to any of the clusters. In this case, the solution that maximizes the likelihood of the model is given by

$$\mu_i = \frac{\sum_k p(i|x_k)x_k}{\sum_k p(i|x_k)}$$

where

$$p(i|x_k) = \frac{p_i(x_k)}{\sum_j p_j(x_k)}$$

A learning rule for the soft competition is similar to that for the hard competition. The difference is that each of the cluster centers is modified in the direction of the line connecting the cluster center and pattern x_k in proportion to the probability that it belongs to that cluster. This follows from the formulation of the data as a mixture-of-Gaussians model. The soft competition approach is closely related to k -means clustering (Duda and Hart, 1973).

In practice, this algorithm does not perform well when the number of cluster centers is not known a priori or when the clusters are not radially symmetric. A lateral inhibition network in which each unit is locally pushing other units not to become selective to the region it is normally selective to can alleviate those problems. Kohonen's self-organizing map (SOM) is one of the earlier implementations of such networks (Kohonen, 1984). Its 2D lateral inhibition structure, which is also called learning vector quantization, implements clustering and has been shown to be useful in numerous applications, most recently in the WEBSOM project for text clustering.

The idea of having experts (entire modules), rather than units, compete globally over the whole space is a direct extension of the more local competitive learning approach. It leads to a competitive network and hierarchical mixture of experts (see MODULAR AND HIERARCHICAL LEARNING SYSTEMS).

Competitive Learning and Resource Allocation

Competition over limited resources and the allocation of such resources are important characteristics of learning in biological sys-

tems. How do such systems manage and distribute their resources? It seems plausible to assume that a sophisticated system would choose an allocation strategy in response to the characteristics of the task. Consider, for instance, a situation that involves two consecutive learning tasks. In such a case, one may distinguish between two possibilities: (1) the second task involves the same data as the first one; (2) the second task involves new data (albeit presented in the same sensory modality).

We hypothesize the existence of a low-dimensional internal representation of the data, whose computation incurs a cost in terms of time and effort. It is then clear that in case 1, the system will be better off if it retains the same internal representation in both tasks. In comparison, in case 2, the existing representation may be modified, or a new one may be created. An intriguing glimpse into the approach taken by the brain in such a situation is provided by recent experimental results concerning the representation of extracorporeal space in rat hippocampus (Wilson and McNaughton, 1993; see HIPPOCAMPUS: SPATIAL MODELS). In that study of place cells, Wilson and McNaughton used chronic multielectrode implantation techniques that allowed long-term recording of cell activities. They recorded more than a hundred cells from free-moving rats that had been trained so they were familiar with part of a simple maze. In these rats, certain hippocampal cells fired in relation to their place in the maze; when the rats were suddenly exposed to a new place, they underwent a short (about 5 minutes) phase of learning in which cells that were silent in the previous location were recruited to encode the new location. This process is consistent with a possible effective reduction in inhibition of cells that have not been recently active. Such a reduction in inhibition may enable them to modify their response patterns rapidly and to become coherently active in the new location.

Discussion

There are various ways in which one can construct controller networks and they resemble various statistical approaches. One can construct either a wide shallow network, in which input patterns are split only once into one of several experts performing the task, or a deep tree with many splits, so that the regression or classification task is performed only by a bottom (terminating) node of the tree. This approach, known as the CART method, has gained considerable attention with the appearance of the classification and regression trees (Breiman et al., 1984). This method constructs a tree leading to different decisions in different regions of pattern space. A direct extension of competitive learning ideas motivated by the success of the above recursive partitioning methods is the hierarchy-of-experts approach (see MODULAR AND HIERARCHICAL LEARNING SYSTEMS).

There is no doubt about the importance of competition in the formation of neural and artificial networks. Experimental evidence from olfactory cortex suggests the performance of hierarchical clustering by a biological network (Lynch, 1986) and has led to the proposal of a model for hierarchical clustering as well. It uncovered the potential of recursive partitioning in complex pattern encoding and classification. Work by Merznick on motor cortex demonstrates that when inhibition is reduced (through amputation), a cortical region that has been responsive to input from one sensory region very quickly adapts to become sensitive to adjacent sensory regions. Competition between units and between collections of units is crucial for such a process. The exact nature of such competition, as well as the optimal architectures that exploit competition for self-organizing, have yet to be uncovered and are likely to benefit from recent advances in multiple cell recording.

Road Map: Learning in Artificial Networks

Related Reading: Data Clustering and Learning; Self-Organizing Feature Maps; Winner-Take-All Networks

References

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., 1984, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Belmont, CA: Wadsworth.
- Duda, R. O., and Hart, P. E., 1973, *Pattern Classification and Scene Analysis*, New York: Wiley.
- Gilbert, C. D., 1994, Neuronal dynamics and perceptual learning, *Curr. Biol.*, 4:627–629.
- Intrator, N., and Edelman, S., 1997, Competitive learning in biological and artificial neural computation, *Trends Cognit. Sci.*, 7:268–272. ♦
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E., 1991, Adaptive mixtures of local experts, *Neural Computat.*, 3:79–87.
- Kohonen, T., 1984, *Self-Organization and Associative Memory*, Berlin: Springer-Verlag.
- Logothetis, N. K., and Scheinberg, D. L., 1996, Visual object recognition, *Annu. Rev. Neurosci.*, 19:577–621.
- Lynch, G., 1986, *Synapses, Circuits and the Beginnings of Memory*, Boston: MIT Press. ♦
- Merzenich, M. M., Recanzone, G., Jenkins, W. M., Allard, T. T., and Nudo, R. J., 1988, Cortical representation plasticity, in *Neurobiology of Neocortex* (P. Rakic and W. Singer, Eds.), New York: Wiley, pp. 41–68. ♦
- Ramo, A. S., Paradiso, M. A., and Freeman, R. D., 1988, Blockade of intracortical inhibition in kitten striate cortex: Effects on receptive field properties and associated loss of ocular dominance plasticity, *Exp. Brain Res.*, 73:285–296.
- Rolls, E. T., Baylis, G. C., Hasselmo, M. E., and Nalwa, V., 1989, The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey, *Exp. Brain Res.*, 76:153–164.
- Sanger, T. D., 1989, Optimal unsupervised learning in a single-layer linear feedforward neural network, *Neural Netw.*, 2:459–473.
- Willshaw, D., Hallam, J., Gingell, S., and Lau, S. L., 1997, Marr's theory of the neocortex as a self-organizing neural network, *Neural Computat.*, 9:911–936. ♦
- Wilson, M. A., and McNaughton, B. L., 1993, Dynamics of hippocampal ensemble code for space, *Science*, 261:1055–1058. ♦

Competitive Queuing for Planning and Serial Performance

Daniel Bullock and Bradley J. Rhodes

Introduction

In neural network studies of planning and serial performance, there is a long history of what Hartley and Houghton (1996) called *competitive queuing* (CQ) models (Figure 1). Such models follow naturally from two assumptions: (1) More than one plan representation can be simultaneously active in a planning layer; and (2) The most active plan representation is chosen, in a second neural layer, by a competition run to decide which plan to enact next. In CQ models, activation is the “common currency” used to compare alternative plans, and simple maximum-finding or winner-take-all (WTA) dynamics can be used as the choice mechanism in the choice layer. Once a plan wins the competition and is used to initiate a response, its representation is deleted from the field of competitors in the planning layer, and the competition is run again. This iteration allows the two-layer network to transform an initial activity distribution across plan representations, often called a *primacy gradient*, into a serial performance (Grossberg, 1978).

As a representation of serial order, the primacy gradient across plan representations in a CQ model is a fundamentally parallel representation. For this reason, CQ models provide a much different basis for control of serial behavior than what have come to be called recurrent neural networks (RNNs). An RNN, in this restrictive usage, is a network in which each output is fed back to the input (or other pre-output) stage as a way of helping to create a distinctive context for eliciting the correct next output. In such an RNN, the representation of the learned sequence is itself fundamentally serial, in the sense that the information that specifies the sequence only becomes available as the serial performance unfolds. In contrast, all the information needed to specify a forthcoming sequence is present in the current state of the planning level of a CQ system, although this current state may itself be dynamically evolving. Having such an explicit, parallel, activation-based representation of sequential plans becomes a substantial advantage for many purposes. For example, such representations can be learned and recalled via compressive and expansive coding operations, as noted later in this article.

Although CQ networks are a radically different basis for sequence control than RNNs, CQ networks also use neural signal recurrence, or internal feedback, for various purposes. The deletion signal sent to the planning layer once a plan is chosen for enactment is one example. Moreover, Grossberg (1978) showed how to implement both the planning layer and the choice layer of a CQ model as *recurrent competitive fields* (RCFs), which exhibit approximate normalization of the total activity level distributed across competing sites in a neural layer. These RCFs were interpreted as parts of a working memory system, in which activity distributions are maintained through significant intervals by recurrent self-excitation combined with recurrent inhibition of competitors. Yet simple changes of signal functions can transform a pattern-holding RCF into a WTA RCF, or even an RCF that quickly forgets initial activity differences. Beyond this ready tunability to realize both the pattern-holding and WTA properties of a CQ system, RCFs have two further properties that enhance their suitability as core models of the planning layers within biological CQ systems. If one plan representation is deleted (by zeroing its activity) from a planning layer RCF, then activity automatically redistributes among the remaining plan representations in a way that preserves the rank ordering of preexisting activation levels. This third property is crucial if iterated deletion is not to disrupt the planning layer's parallel representation of the serial order of subsequent plans.

A fourth property is another consequence of RCF normalization in the planning layer. Because total activity is conserved, *more* simultaneously active plans imply *less* activation per plan, including the plan scheduled to be performed first. If the initial activation level of this plan predicts the reaction time (RT) to perform the first planned action (as is the case if the choice layer is a WTA network with a high threshold for generating an output to the response execution system, and the latter system is set to perform as soon as it receives an input) then this variant of a CQ model makes a surprising prediction. The RT to *initiate* performance of a prepared sequence should increase with the number of plans in the prepared sequence. Such a “sequence length effect on RT” is true for humans performing lightly practiced sequences from working

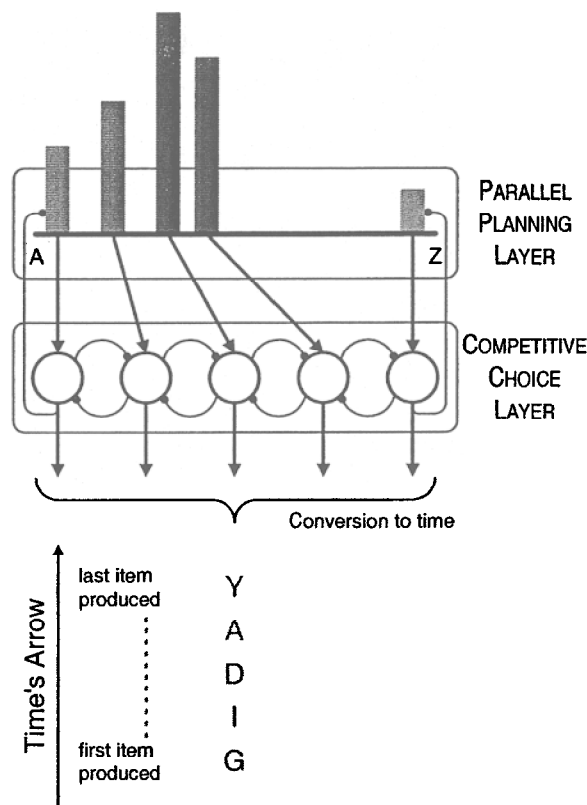


Figure 1. Initial state of a two-layer competitive queuing (CQ) system, prior to production of a five-letter sequence. The sequence that will emerge is shown in the lower part of the figure. Excitatory connections terminate with arrowheads, inhibitory connections with filled circles. The most active plan is selected for execution in the lower, competitive choice, layer by a winner-take-all dynamic whose outcome is wholly determined (in the absence of noise) by the activation gradient (representing the to-be-performed sequence) present in the parallel planning layer. Once a plan representation wins at the competitive layer, a large output signal is sent to initiate execution of the corresponding response (descending arrow) and to delete the plan's representation in the parallel activation layer (ascending path to parallel planning layer). This process iterates until all plans have been enacted and all planning layer activities deleted. The result is sequential plan execution that corresponds to the initial rank ordering (primacy gradient) of plan activation levels in the upper field of the CQ network. Although each competitive layer node would send an inhibitory connection to its correspondent in the parallel planning layer, only two such connections are shown here, to avoid clutter. In this example, which uses recurrent inhibition in the choice layer, each competitive layer node would inhibit all others, but only nearest-neighbor inhibition is actually depicted. (Adapted from Rhodes, 1999, with permission.)

memory (Sternberg et al., 1978; review in Rhodes, 1999). Simulations by Boardman and Bullock (1991) verified these RT properties for a two-layer CQ system coupled to a response generation network. They also showed that the model correctly predicted patterns of inter-item intervals observed in the Sternberg et al. RT task. Bradski et al. (1994, cited in Rhodes, 1999) developed a neural network that could serve as a perceptual preprocessor for a CQ model. In this network, a sequence of perceptual inputs induced an appropriate primacy gradient across plan representations in the planning layer of the CQ circuit.

Applying the CQ Model to Immediate Serial Recall

In the Sternberg et al. (1978) task, subjects were told to repeat short prepared lists as fast as possible following an external signal. This

qualified it as an RT task. A related list-recall task is the immediate serial recall (ISR) task, in which subjects also recall a short list from working memory, but without explicit instruction to initiate or perform recall as fast as possible. This non-RT sequence production task has also been modeled within the CQ framework. To the three core assumptions noted previously (primacy gradient, deletion on enactment, and iterated competitive choice of most-active remaining plan), Page and Norris (1998) added two auxiliary assumptions: that the choice is noisy, and that decay of activity in the planning layer occurs during input to the planning layer and during intervals spent performing items from the list. Error data favor both assumptions, and this extended model was able to address data on errors of serial recall. One kind of error, simple failure to recall, is most probable for list-final plans in long sequences. The extended model explains this as a consequence of their low initial activation level (due to being last in the primacy-gradient-coded sequence), which in turn makes them more susceptible to falling into inactivity due to the decay that can occur during enactment of the prepared sequence. Another feature of error data from ISR studies is that the vast majority of transposition errors (items are recalled, but in incorrect order) are simple exchanges with immediately adjacent items in the planned sequence. Given moderately noisy choice, this likewise follows from the gradient representation, because noise in the choice layer is less likely to illicitly promote a plan by two positions in the activity gradient than by one position.

Brain Substrates of Competitive Queuing

Data that strikingly confirm one of the main physiological predictions of CQ models was recently reported in Chafee et al. (2001). Since the CQ model depends on a working memory to hold a gradient of activations across plan representations prior to and during performance, the model predicts that such a gradient will be observable in the part of the brain responsible for working memory, namely the prefrontal cortex (PFC). Chafee et al. observed three ensembles of prefrontal activities corresponding to three segments of a forthcoming line drawing that a monkey had been trained to produce. The relative strength of activation of the three cellular ensembles predicted the order of the forthcoming segments: the higher the premovement activation, the earlier in the sequence the corresponding segment was produced. Error data were also in accord with predictions of the CQ model.

Another property of some versions of the CQ model is the normalization of total activity in the planning layer. This property predicts that peak frontal cortical activities associated with plan representations will decline as the number of coactive plans increases. Data that confirm this prediction for plan representations' activities were recently reported by Cisek and Kalaska (2002). They found a class of cells in the rostral part of the dorsal premotor cortex (rPMd) that they called "potential response cells." Different subsets of such cells, representing different potential responses, were coactive during a delay period when there was uncertainty regarding which, if any, of one or two alternative responses should be performed. This delay period activity was much more vigorous in each plan representation (cell subset) when there was only one potential response to hold in working memory than when there were two potential responses to hold in working memory.

These data on rPMd and PFC are consistent with representations in a normalized working memory capable of storing a primacy gradient—the planning layer of a CQ system. What part of the brain serves as the choice layer for frontocortical plan representations? One idea that has grown in popularity in recent years is that the striatum, a subcortical structure that encompasses the caudate, putamen, and accumbens nuclei of the basal ganglia, may provide competitive arenas in which cortically represented plans vie for

execution. According to the recent model of Brown, Bullock, and Grossberg (2000), the winning plan selectively activates a small subset of spiny projection neurons (SPNs) in the striatum, which receives a massive excitatory input from the cerebral cortex. Via a return pathway that traverses the pallidum and thalamus, this focal striatal activation enables selective activation of output cells in a cortical layer (layer Vb) that is below, and excited by, the layer in which the cortical planning cells are found. One point of current debate regards how the striatum runs its competition to choose the next plan to execute. This debate arises because the striatum contains both feedback inhibition, mediated by SPNs, and feedforward inhibition, mediated by inhibitory interneurons which, like the SPNs, are direct recipients of excitatory input from the cortex. The simulations reported in Brown et al. (2000) were based on recent physiological data indicating that the dominant factor in striatal choice making is feedforward inhibition, not recurrent inhibition. In a situation with many strongly activated competing plans, reliance on feedback inhibition to resolve the competition would require that many cells in the choice field would be transiently active. Therefore a high threshold downstream from the choice field would be necessary to prevent multiple premature response activations. With feedforward competition, none of the choice field's output cells—in the striatal case, the SPNs—need activate until the moment that the competition yields a winner.

Plan Layer Loading by Rapid Parallel Recall from Procedural Long-Term Memory

One of the advantages of CQ models' explicit parallel representation of sequential plans—an advantage unavailable to RNNs as such—is that these distributed representations can be learned and recalled via compressive and expansive coding operations. In the Sternberg task and the ISR task mentioned previously, novel sequence information was provided to the performer. According to the CQ interpretation, performers hold a corresponding parallel representation for a few seconds in working memory (WM) before generating the sequence under the guidance of WM. However, Verwey (1996), among others, showed that a very high number of practice trials with short fixed sequences leads to *disappearance* of the sequence length effect on RT originally discovered by Sternberg et al. (1978). Is this result explicable in terms of a CQ model that uses activity normalization in the planning layer? Rhodes and Bullock (2002) have recently reported successful simulations of several sets of list learning and performance data, using a neural network in which the cerebellum, modeled as a substrate for procedural long-term memory (LTM), learns activation gradients over item nodes and rapidly recalls them into a normalized motor buffer (planning layer), which is a WM for action plans. The recall process is rapid because it entails *parallel loading of sequence chunks* into a WM from LTM. When the procedural LTM of a fixed sequence representation becomes strong enough (due to extensive practice), it causes preselection of the first list item within the CQ subsystem. Such preselection explains the disappearance of the sequence length effect on RT, which Verwey (1996) showed to be a reliable effect if subjects were given high levels of practice. This hybrid cerebellar-CQ model's assumption that the cerebellum can load parallel sequence representations into a frontocortical motor buffer is supported by recent neuroanatomical tracing studies, which have discovered pathways that run from the dentate nuclei of the cerebellum, via the thalamus, to several frontocortical zones, including premotor cortex and the PFC. More generally, the hybrid model illustrates how the CQ model, which focuses on WM dynamics that support sequential performance, can be interfaced with an LTM system that compressively learns and stores, and expansively recalls, oft-used sequences. Such a system is critical for functions

that require frequent re-use of subsequences, such as musical performance or language production.

CQ as a General Basis for Serial Behavior

For CQ to qualify as a core model for all sequence planning and control, it must be shown that it is extensible to a full range of human serial performance domains. Given the pervasive involvement of the brain substrate outlined previously, it is reasonable to look for CQ signatures far beyond the types of skill tasks considered thus far. The most highly developed example of human serial behavior, syntactic language production, is a critical test case for the thesis that CQ is general. Another instructive case is sequential control of attentional focus during information pickup from complex scenes.

Is CQ Extensible to Language Production?

It might be thought that the CQ model cannot apply to syntactic language production, because sequencing errors in language production do not usually follow the “immediately adjacent items exchange” pattern found in ISR studies (which typically use non-grammatical item sequences). In most sequencing errors in language production, exchanges respect grammatical constraints, as when a sequencing error transforms the intended “flying saucers” into the spoonerism “sighing flossers.” But note: the same example supports the CQ postulate that the initial segments of both words were already coactive in a planning field prior to production of either word. Moreover, it is plausible that the exchange error occurred because noise transiently rendered the plan for “fi” less active than the plan for “s” at the instant that “flying” should have been spoken. In fact, several neural network theorists have used CQ as a core of extended models that have offered explanations of many of the grammar-respecting patterns of sequencing-errors observed in language production (e.g., Dell, Burger, and Svec, 1997; Hartley and Houghton, 1996).

The most sustained treatment of CQ in language generation is that in Ward (1994). Far from simply explaining how the “emergent choice” that operates in CQ models is compatible with grammar-respecting sequencing errors in language production, Ward argues that *only* emergent choice offers a basis for overcoming more traditional language generators' failures to mimic the “flexible incremental generation” (FIG) exhibited in the real-time behavior of human speakers as they compose sentences “on the fly.” Ward's FIG model combines CQ principles with principles inspired by *construction grammar* (e.g., Goldberg, 1995) to build a comprehensive connectionist model of grammatical sentence generation. The FIG algorithm is an iterated cycle: (1) Each node of an input conceptualization is a source of activation to “construction” nodes of various types, including words; (2) Activation is allowed to flow freely through the structured network of nodes; (3) When the network settles (or is forced to make an output) the most highly activated word representation is selected and enacted; (4) Any node or nodes of the input conceptualization that are expressed by the enacted word are inhibited, and activation levels are updated to represent the new current state; and (5) Steps 2–4 iterate until the conceptual content of the input has been expressed by the enacted word sequence. For the system to work well, the word plan that has the highest activation must be for a word that will be both syntactically and semantically correct if spoken as the next word in the utterance. This requirement is met, in part, by having the activation level of a word be determined by the product of its semantic and syntactic inputs, not by their sum.

Although much remains to test such models of CQ in the service of language production, initial simulation successes indicate that CQ may provide an ideal foundation for speech and language pro-

duction models: one that is well grounded in neurobiology, and one that overcomes the inflexibility and other limitations inherent in more traditional, less parallel, theories.

Can CQ Be Considered an Attention Control Mechanism?

To gather information from the visible world around us, we typically, and rapidly, move our attention sequentially from one part of the scene to another. Often such search is purposeful, as when we are attempting to find a particular item or object, such as a familiar face in a crowd. In other cases, the scan sequence can be driven by salient features present in the visible scene. Either way, there must exist some mechanism for determining where to focus attention first, where to move the focus next, and so on. As defined previously, CQ provides a very simple and elegant candidate mechanism. The combination of a “saliency map” with an “inhibition of return” mechanism forms the basic mechanism for controlling attention deployment in contemporary computational models of focal visual attention (e.g., Itti and Koch, 2001). The most salient location in the map has the highest activation and draws the focus of attention to that location, whereupon map activity at that location is inhibited. Consequently, some other location in the map becomes most active—i.e., most salient—and now attracts the attentional focus. This process is another example of CQ. That it applies to attention shifts is no surprise because of the close link between such shifts and saccadic eye movements, sequential control of which has been attributed to interactions between frontal cortex and the basal ganglia (e.g., Brown et al., 2000).

Discussion

The fundamental scenario represented by the CQ model is almost palpable. We seem to be able to feel that two considered plans, or candidate words, begin with nearly equal potency, but that upon further deliberation, one waxes as the other wanes, and the waxing plan is first used to guide performance. Moreover, our capacity to simultaneously consider multiple plans in WM is limited to a small number. This limitation is predicted by the activity normalization property, which itself is needed to preserve the computational function of differences in activation levels. If forced to decide quickly, we feel that we choose the more vivid plan even if the rejected or deferred plan is also vivid and attractive. In the brain’s parallel planning system, some common currency must be available for incrementing and decrementing the value of response alternatives, and in biological neural networks the use of excitation and inhibition to achieve these effects leaves activation level as the natural measure of relative priority. This, in combination with a maximum-finding network, creates a CQ system, which is probably an ancient invention in the evolution of animals. Recent computational studies, cited previously, have begun to explore how this ancient system

may still serve as a viable core for the highest levels of planning and skilled sequencing exhibited by humans.

Road Map: Artificial Intelligence

Related Reading: Artificial Intelligence and Neural Networks; Decision Support Systems and Expert Systems; Multiagent Systems; Recurrent Networks: Learning Algorithms; Winner-Take-All Networks

References

- Boardman, I., and Bullock, D., 1991, A neural network model of serial order recall from short-term memory, in *Proceedings of the International Joint Conference on Neural Networks* (Seattle) vol. II, Piscataway, NJ: IEEE Service Center, pp. 879–884.
- Brown, J., Bullock, D., and Grossberg, S., 2000, How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades, *Boston University Technical Report CAS/CNS-2000-023*.
- Chafee, M. V., Averbeck, B. B., Crowe, D. A., and Georgopoulos, A. P., 2001, Motor sequence representation of prefrontal neurons predicts error patterns in drawing, *Abstr. Soc. Neurosci.*, 533.3.
- Cisek, P., and Kalaska, J. F., 2002, Simultaneous encoding of multiple potential reach directions in dorsal premotor cortex, *J. Neurophysiol.*, 87:1149–1154.
- Dell, G. S., Burger, L. K., and Svec, W. R., 1997, Language production and serial order: A functional analysis and a model, *Psychol. Rev.*, 104:123–147. ♦
- Goldberg, A. E., 1995, *Constructions: A Construction Grammar Approach to Argument Structure*, Chicago, IL: U. of Chicago Press.
- Grossberg, S., 1978, A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans, in *Progress in Theoretical Biology*, vol. 5 (R. Rosen and F. Snell, Eds.), New York, NY: Academic Press, pp. 233–374. ♦
- Hartley, T. A., and Houghton, G., 1996, A linguistically constrained model of short-term memory for nonwords, *J. Mem. Lang.*, 35:1–31.
- Itti, L., and Koch, C., 2001, Computational modelling of visual attention, *Nature Rev. Neurosci.*, 2:194–203.
- Page, M. P. A., and Norris, D., 1998, The primacy model: A new model of immediate serial recall, *Psychol. Rev.*, 105:761–781. ♦
- Rhodes, B., 1999, *Learning-Driven Changes in the Temporal Characteristics of Serial Movement Performance: A Model Based on Cortico-Cerebellar Cooperation*, Doctoral Dissertation, Cognitive and Neural Systems Department, Boston University. ♦
- Rhodes, B., and Bullock, D., 2002, Neural dynamics of learning and performance of fixed sequences. Latency pattern reorganizations and the N-STREAMS model. *Boston University Technical Report CAS/CNS-02-005*.
- Sternberg, S., Monsell, S., Knoll, R. L., and Wright, C. E., 1978, The latency and duration of rapid movement sequences: Comparisons of speech and typewriting, (Reprinted) in *Perception and Production of Fluent Speech* (R.A. Cole, Ed.), 1980, Hillsdale, NJ: Erlbaum, pp. 469–505.
- Verwey, W. B., 1996, Buffer loading and chunking in sequential key pressing, *JEP: Human Perception and Performance*, 22:544–562.
- Ward, N., 1994, *A Connectionist Language Generator*, Norwood, NJ: Ablex Publishing. ♦

Compositionality in Neural Systems

Barbara Hammer

Introduction

In real life, people deal with composite structures: Written English language is built out of 26 characters (and a few additional symbols) that form syllables, words, sentences, articles, road maps, and finally the *Handbook of Brain Theory*. Spoken language consists of raw acoustic waves at a basic level; at a higher level, it can be

decomposed into phonemes that form words, sentences, a poem or a speech. Visual data can be decomposed into pixels with various colors and intensities; alternatively, the raw image data may be represented by features like edges or texture, which are grouped to complex contours, objects, and, finally, a complete scene, such as the image of a grandmother sitting in a chair and knitting. Moreover, not only real-life data are processed as composite objects:

artificial data created by humans or virtual objects also have a composite structure. As examples, web sites consist of single pages with a head and body, links, tabulars, figures, enumerations, and so on. Computer programs decompose into procedures and functions, or objects and methods. Logical formulas and terms are recursive objects built out of symbols for constants, variables, and functions, logical connectives, and quantors.

In respect to neural systems, two questions arise:

1. Artificial neural networks are developed in order to model important aspects of the human brain and to explain how biological neural networks process information. A common characteristic of the way in which data are created, processed, and stored by humans is compositionality. *Which neural dynamics allow composite structures, grouping, and binding to emerge* in such a way that the single parts can be restored rapidly and, at the same time, the whole composite structure can be identified with a single object?

2. Artificial neural networks are a powerful and universal machine learning tool that is used in scientific and industrial applications for data contained in a finite dimensional vector space. Everyday data are composite. *How can we adapt standard neural techniques to composite structures* such that connectionistic methods can be used in everyday applications and combined with other compositional machine learning tools?

These two questions constitute extreme positions in a single problem spectrum: How is information processed in the brain? Since it is unlikely that models that are satisfactory with regard to both efficiency and biological plausibility will become available in the near future, the focus can be on one of two aspects: either one can develop practically applicable and efficiently trainable approaches or one can design biologically plausible and universal systems. Most existing approaches lie somewhere in between, partly because biological systems are indeed very effective, suggesting the universal principle of compositionality and dynamic binding, e.g., in the visual system (Bienenstock, 1996) (see DYNAMIC LINK ARCHITECTURE).

Properties of Compositionality

What are the general properties of composite objects like the sentences “John loves Mary” or “Mary loves that John loves Mary”? The structures are composed of *basic primitives*—here, the words *John*, *Mary*, *loves*, and *that*. The primitives are instances of a certain type, e.g., a noun or verb. Composite expressions arise if the primitives are combined in specified *relations*. The sentence “John loves Mary” is an instance of a relation of the form subject–predicate–object, where subject and object can be instantiated with a specific noun, for example. Alternatively, these positions may be filled with composite objects, such as “John loves Mary” in the example “Mary loves that John loves Mary.” The complexity of recursively generated structures is usually not limited a priori. As an example, we could build the sentences “John loves that Mary loves that John loves Mary,” or “Mary loves that John loves that Mary loves that John loves Mary,” and so on. Hence, as pointed out in van Gelder (1990), the combination of primitives of certain types within constituency relations yields composite structures of *a priori unlimited complexity and an infinite number of possible combinations*.

The *semantics* of composite objects is determined by the semantics of the simple primitives and their relation in the structure. The primitives alone do not determine the semantics. The sentences “John loves Mary” and “Mary loves John,” for example, have identical constituents but different meanings. Moreover, the decomposition of complex objects into simpler parts and their interpretation may depend on the whole structure: the meaning of *her* in the sentence “Mary loves that John loves her” depends on the context of the part “John loves her.” Conversely, the same situation

may be described with different composite objects, such as “Mary loves that John loves Mary” or “Mary loves that John loves her.” In other words, parts of a structure may be substituted by entirely different representations without affecting the semantics. Hence, the whole structure as well as the involved primitives and their relation should be available for referring the semantics.

Although an unlimited number of combinations of the primitives in constituency relations is possible in principle, commonly only a small number of all possible combinations are used in practice and make sense. Either basic syntactic rules or semantic limitations restrict the variety of possible compositions. “That John loves Mary loves Mary” does not make sense, for example, if the composite structure “that John loves Mary” is used as subject. Restrictions of possible combinations are often context dependent. For example, the sentence “John loves himself” would be preferred to “John loves John” unless the word “John” refers to two different persons.

Humans are capable of understanding and producing composite objects. Moreover, they can deal with new complex structures although they have never seen the specific combination of primitives before. As an example, having read the above sentences concerning John and Mary and the sentence “Mary loves Peter,” people would be capable of understanding the sentence “John loves that Mary loves Peter,” and they would possibly infer that this sentence is false. People can infer the meaning of partially new structures. They have knowledge about primitives and relations, and additional information such as rules for composition or experience with similar structures. Moreover, compositional structures play an important role if the capability of humans for analogy-based reasoning is investigated (see ANALOGY-BASED REASONING AND METAPHOR).

Appropriate artificial neural systems for use in processing compositional data should take these properties into account. Although they have to deal with an unrestricted amount of partially new and ambiguous data, they can use the sparseness and hierarchical structure of the data for efficient processing.

Neural Systems and Compositionality

Popular neural methods perform pattern recognition, for which classical statistics provides a well-founded theory (see PATTERN RECOGNITION). Standard neural networks are adapted to real vectors contained in a finite dimensional Euclidean real-vector space. Concerning compositional structures, the question arises how to encode and process an arbitrary amount of information with this finite dimensional machinery.

Compared to their biological counterparts, artificial neural networks often neglect the fine temporal structure of spike trains (see OSCILLATORY AND BURSTING PROPERTIES OF NEURONS AND ADAPTIVE SPIKE CODING): standard artificial networks process real values or binary values in a well-defined topological order. Real values correlate to the mean spiking frequency of the biological neurons; hence, the local temporal structure and respective correlation of spikes are not taken into account. Binary values might encode single spikes, although the topology allows processing on a discrete or fixed time scale only. Experiments provide evidence that the fine temporal structure may indeed carry important information for encoding complex scenes in biological networks (Engel et al., 1992). However, the precise way in which information is encoded in the respective parts of the brain is not yet understood. Moreover, some effects can possibly already be explained at the abstract level of rate coding (see RATE CODING AND SIGNAL PROCESSING). Hence it is worth considering the whole spectrum of neural architectures capable of dealing with compositionality. Our taxonomy for characterizing the various approaches is based on the connection structure of the neural architectures and on their fundamental dynamical behavior.

A key property of appropriate systems is their capability of processing a priori unlimited information. *Static solutions* with feedforward networks have the advantage that efficient training algorithms are readily available (see BACKPROPAGATION: GENERAL PRINCIPLES). Unfortunately, either their capacity is limited or they need an a priori unlimited amount of neural resources.

Recurrent neural networks use the additional degree of freedom provided by a priori unlimited processing time in order to map the information appropriately. Recurrent networks may be *partially recurrent*, where the processing dynamics are determined by the respective data structure. The restriction to limited recursive data structures allows the immediate generalization of efficient standard learning tools (see RECURRENT NETWORKS: LEARNING ALGORITHMS). Alternatively, the networks may be *fully recurrent*, in which case the dynamics are determined by the respective process. This allows more flexibility, but the processing time cannot be limited a priori, and alternative learning algorithms are necessary (see TEMPORAL SEQUENCES: LEARNING AND GLOBAL ANALYSIS). In particular, continuous-time fully recurrent neural networks can use the fine temporal structure for encoding complex information, such as in networks of spiking neurons (see INTEGRATE-AND-FIRE NEURONS AND NETWORKS). These approaches are biologically plausible and flexible. Unfortunately, no efficient learning algorithm comparable to standard backpropagation for feedforward networks has been established for networks of spiking neurons until today.

Static Approaches

The *grandmother neuron doctrine* assumes that each simple or composite object is represented by the activity of a specific neuron for this object (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS). Neurons for complex objects are hierarchically connected to neurons representing their simple parts, and they only become active if all neurons for their parts become active. Putative “grandmother cells” have been found in the visual system of the macaque monkey, for example (see SPARSE CODING IN THE PRIMATE CORTEX) but the data do not preclude the firing of such cells for a variety of patterns. Some invariance, for example with respect to the specific lighting, may be involved in the hierarchical computation, such that the neuron that encodes our grandmother fires independent of the specific context. Naturally, this approach requires additional neural resources for each new primitive, relation, or composite object, and therefore suffers from combinatorial limitations. Moreover, it is not obvious how entirely new combinations of well-known ingredients could be processed properly, i.e., in such a way that the possibility of analogies and transfer is taken into account. However, the hierarchical feedforward computation makes it possible to identify each neuron with a specific meaning and to use standard neural techniques.

Alternatively, composite objects may be represented in a *distributed manner* via the activation of a group of neurons for the single features (see POPULATION CODES). This reduces the number of necessary neurons, but it is not obvious how invariance and not instantiated or new features are to be integrated. The necessary amount of resources depends on the respective task and cannot be uniformly limited. Moreover, the composition of two objects into a composite object is to be distinguished from the simple superposition of the activation of each; otherwise, the respective decomposition would become ambiguous.

Naturally, both encodings may be combined in *localized distributed representations* of objects that correspond to the firing of neurons at a certain area of the pool. This encoding is suggested by the presence of topology-preserving maps in biological systems where similar stimuli cause neural activities in similar neural regions (see OCULAR DOMINANCE AND ORIENTATION COLUMNS).

Either localized or distributed encoding of composite objects in a vector space of fixed dimension has so far been the standard encoding for practical applications (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS). Hierarchical extraction of relevant features and finally of the respective class is a promising technique that has been successfully applied in various applications, such as the recognition of visual objects or text processing (Riesenhuber and Poggio, 1999; Mel and Fiser, 2000). Given the limits with respect to the number of recognized objects, static approaches suffer from a priori limitations and can only explain parts of biological information processing.

Partially Recurrent Systems

Discrete time partially recurrent neural networks are widely used in time series prediction, speech recognition, and the processing of sequences of real vectors in general (Kremer, 2001) (see CONSTITUENCY AND RECURSION IN LANGUAGE; IDENTIFICATION AND CONTROL). They deal with sequences of real vectors as opposed to simple real vectors and hence are capable of processing very simple composite objects with a priori unlimited informational content. Their dynamics directly mirror the recursive structure of the data: a standard feedforward network encodes in its internal activations the context of the computation, i.e., the first part of a sequence, and recursively maps one entry of the sequence after another to new contexts, depending on its internal state. Since the dynamic is fixed according to the data structure, standard gradient descent techniques can be used for supervised learning of mappings with sequences as inputs or outputs (see RECURRENT NETWORKS: LEARNING ALGORITHMS).

The a priori unlimited information in the data structure changes the learning theoretical properties compared to standard feedforward networks: On the one hand, the power of recurrent networks can be demonstrated by relating them to classical symbolic computing mechanisms, such as Turing machines (Hammer, 2002) (see NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY). On the other hand, valid generalization can no longer be guaranteed for training set sizes that are both independent of the underlying input distribution and independent of the specific output of the training algorithm (Hammer, 2002) (see PAC LEARNING AND NEURAL NETWORKS; VAPNIK-CHEVONENKIS DIMENSION OF NEURAL NETWORKS).

One can think of the recursive dynamics of simple recurrent networks as a recursive coding of sequences to distributed representations in a finite dimensional vector space. This idea can immediately be generalized to more complex composite objects, provided they have a recursive structure: trees consist of a label and a fixed number of subtrees. Hence, a network that is to encode trees instead of simple sequences can recursively map the root's label and the already computed codes for the subtrees to a code for the entire tree. It can also decode a distributed representation of a tree by computing the root's label and codes for the subtrees, which can be recursively processed further (Frasconi, Gori, and Sperduti, 1997; Hammer, 2002). This mechanism can be found in various STRUCTURED CONNECTIONIST MODELS (q.v.). Note that terms and formulas have a natural representation as tree structures: the single symbols are contained in the respective nodes, and the subformulas or subterms respectively correspond to subtrees. Thus, networks capable of encoding or decoding tree structures constitute a universal mechanism that enables the use of neural techniques in symbolic domains.

Concrete implementations of this basic idea differ in how the respective networks are trained. With the dynamics being determined by the data structure, gradient descent techniques can be used for supervised learning of general mappings, with trees as input or output. Recursive networks are trained directly for the

specific learning problem (Frasconi et al., 1997; Hammer, 2002). The recursive autoassociative memory trains only encoding and decoding, such that their composition yields the identity on the data, leading to universal encoding (Frasconi et al., 1997; Sperduti, 1994). Holographic reduced representation uses a fixed transition function that is not trained at all (Plate, 1995). These approaches have been successfully applied in such different areas as chemistry, theorem proving, and language processing (Frasconi et al., 1997; Hammer, 2002). Learning is quite similar to the training of simple recurrent networks as far as practical algorithms as well as theoretical properties like approximation and generalization capability are concerned (Hammer, 2002).

This approach, however, is limited to recursive compositions with a well-defined recursive structure; cyclic structures cannot be processed in this way. Moreover, gaining access to single parts of recursive structures may be time-consuming and subject to noise, depending on the level of recurrence. There exist fundamental mathematical limitations to the possibility of coding infinite tree structures in a finite dimensional vector space: encoding has to be nested or fractal, which means that the Euclidean metric is no longer appropriate for the resulting distributed representations (Hammer, 2002). Thus, various neural methods that are based on the Euclidean metric cannot be used for further processing. Reliable decoding is a difficult task with a lower-bounded complexity (Hammer, 2002) (see VAPNIK-CHEVONENKIS DIMENSION OF NEURAL NETWORKS). Therefore, neural networks with the above dynamics are not appropriate for efficiently decoding a large amount of distributed data.

Fully Recurrent Systems

Fully recurrent neural systems are networks where the neuron's activations evolve in time in a continuous or discrete manner. Commonly, the dynamics can be described by nonlinear difference equations in discrete time or differential equations in continuous time. The main difference from partially recurrent networks is that the processing dynamics and consequently the required computation time are not directly determined by the data structures. Time and complexity of computation and representation of information are a priori unlimited. The systems may use the temporal structure of the neuron's activations for storing important information in specific spatiotemporal activation patterns such as synfire chains, i.e., successive spiking of specified neurons in precise time intervals (see SYNFIRES CHAINS). As proposed by Bienenstock (1996), for example, the synchronous oscillation of different neurons or assemblies may indicate that they represent objects that are bound together (see SYNCHRONIZATION, BINDING AND EXPECTANCY AND ADAPTIVE SPIKE CODING). This type of binding could easily be further processed with coincidence detectors. Alternatively, information may be stored in simple localized or distributed patterns of the neuron's activities, as in the static and partially recurrent approaches. Various systems obey a gradient dynamics as an example and converge to a fixed point that contains the important information (see TEMPORAL SEQUENCES: LEARNING AND GLOBAL ANALYSIS).

Concrete implementations differ considerably in their complexity and intention. Several approaches merely demonstrate that important effects such as *oscillation*, *synchronization*, and *coincidence detection* can be found in experiments on biological neural activities and can be simulated in an artificial, though biologically plausible, environment (see CHAOS IN NEURAL SYSTEMS; COLLECTIVE BEHAVIOR OF COUPLED OSCILLATORS, and CORTICAL POPULATION DYNAMICS AND PSYCHOPHYSICS). Particularly in the context of networks of spiking neurons, methods that allow a mathematical investigation of complex systems have been developed over the last years (see INTEGRATE-AND-FIRE NEURONS AND

NETWORKS). First steps in possible training mechanisms use the principles of self-organization and Hebbian learning (see POST-HEBBIAN LEARNING ALGORITHMS). Binding via the temporal structure and synchronous oscillation is a biologically plausible mechanism that is supported by experimental results (Engel et al., 1992). It is not yet obvious how complex recursive and hierarchical structures can be represented in this way, since methods such as iterated period doubling or a superposition of the oscillation, for example, are restricted by the computation accuracy of neurons. Moreover, efficient learning algorithms for practical applications are not yet available.

Other approaches *develop practical tools for concrete tasks* that involve binding mechanisms, such as feature grouping or edge detection in images (see VISUAL SCENE SEGMENTATION). Grouping may be encoded in synchronously oscillating neurons or in the localized activation of specific neurons in a limiting state that minimizes an energy function, such as in the competitive layer model (Wersing, Steil, and Ritter, 2001). Most approaches are based on appropriate excitation of similar neurons and an inhibition of cells with dissimilar activation. Again, only limited possibilities of automatic learning of the connections are available so far.

Finally, recurrent systems for *complex analogical reasoning and symbolic processing* have been proposed (see STRUCTURED CONNECTIONIST MODELS). Popular approaches are LISA, SHRUTI, and INFERNET, which have in common that binding is realized via synchronous oscillation of neurons or pools of neurons (Hummel and Holyoak, 1997; Sougné, 1999; Shastri, 1999). Structures are represented through localized or distributed cell activations. Rules correspond to specific neural connections that allow human-like analogical reasoning and are mostly hand encoded. They are capable of simulating various effects and limitations of human-like reasoning. However, like most fully recurrent approaches, the systems suffer from the lack of universal and efficient training algorithms.

Discussion

Compositionality as a common principle of information processing should find its counterpart in artificial neural systems. Solutions may either enhance static feedforward systems and represent the objects by static activation patterns, or may rely on recursive and adaptive encoding mechanisms in partially recurrent networks, or may use complex dynamics and the fine temporal structure of the neuron's activation for reliable representation. Of course, it is possible to transfer more practical tools and theoretical guarantees from classical network techniques to these systems if they are similar to classical systems. It should be pointed out that it is still disputed whether artificial neural networks are capable of adequately handling compositional data, and if so, which type of network is the most suitable one. Remarkable results have been obtained with simple recurrent networks, although some researchers argue that more complicated dynamics or dynamics similar to classical symbolic processing mechanisms are necessary for successful modeling within the context of compositionality (see CONNECTIONIST AND SYMBOLIC REPRESENTATIONS and Elman (1998) and references therein for a discussion of this subject).

Important directions for future research include the exploration of real neural codes in biological systems. Of particular interest is whether effects such as synchronous activation indeed contain information that is necessary for the representation of relations. Oscillations could constitute a byproduct of information processing that merely enables efficient adaptation in biological systems.

Restrictions on compositionality are another avenue of exploration. In practice, only a small subset of all possible combinations of primitives and relations occurs. Mathematical analysis of structure-processing systems implies usually a worst-case analysis

and might indicate, for example, that static approaches are not appropriate for this field. However, neural mechanisms that are not capable of representing arbitrary composite objects in principle might be well suited for restricted, though important, domains.

Road Map: Artificial Intelligence

Related Reading: Analogy-Based Reasoning and Metaphor; Dynamic Link Architecture; Structured Connectionist Models; Systematicity of Generalizations in Connectionist Networks

References

- Bienenstock, E., 1996, Composition, in *Brain Theory: Biological Basis and Computational Theory of Vision* (A. Aertsen and V. Braitenberg, Eds.), New York: Elsevier, pp. 269–300. ♦
- Elman, J., 1998, Generalization, simple recurrent networks, and the emergence of structure, in *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (M. A. Gernsbacher and S. J. Derry, Eds.), Mahwah, NJ: Erlbaum.
- Engel, A. K., König, P., Kreiter, A. K., Chialen, T. B., and Singer, W., 1992, Temporal coding in the visual cortex: New vista on integration in the nervous system, *Trends Neurosci.*, 15:218–225.
- Frasconi, P., Gori, M., and Sperduti, A., 1997, A general framework for adaptive processing of data sequences, *IEEE Trans. Neural Netw.*, 9:768–786. ♦
- Hammer, B., 2002, Recurrent networks for structured data: A unifying approach and its properties, *Cogn. Systems Res.* (in press). ♦
- Hummel, J. E., and Holyoak, K. J., 1997, Distributed representation of structure: A theory of analogical access and mapping, *Psychol. Rev.*, 104:427–466.
- Kremer, S. C., 2001, Spatio-temporal connectionist networks: A taxonomy and review, *Neural Computat.*, 13:249–306. ♦
- Mel, B., and Fiser, J., 2000, Minimizing binding errors using learned conjunctive features, *Neural Computat.*, 12:247–278.
- Plate, T., 1995, Holographic reduced representations, *IEEE Trans. Neural Netw.*, 6:623–641.
- Riesenhuber, M., and Poggio, T., 1999, Are cortical models really bound by the “binding problem”? *Neuron*, 24:87–93.
- Shastri, L., 1999, Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony, *Appl. Intell.*, 11(1):79–108.
- Sougné, J. P., 1999, *INFERNET: A neurocomputational model of binding and inference*, PhD thesis, Université de Liège.
- Sperduti, A., 1994, Labeling RAAM, *Connect. Sci.*, 6:429–459.
- van Gelder, T., 1990, Compositionality: A connectionist variation on a classical theme, *Cognit. Sci.*, 14:355–384. ♦
- Wersing, H., Steil, J. J., and Ritter, H., 2001, A competitive layer model for feature binding and sensory segmentation of features, *Neural Computat.*, 13:357–387.

Computing with Attractors

John Hertz

Introduction

This article describes how to compute with networks with feedback that exhibit complex dynamical behavior. In order to compute with any machine, we need to know how data are to be fed into it and how the result of a computation is to be read out. These questions are trivial for layered feedforward networks, but not for networks with feedback. A natural proposal is to wait until the network has “settled down” and then read the answer off a suitably chosen set of units. The state a dynamical system settles into is called an *attractor*, so this paradigm is called computing with attractors.

The term “settling down” is not meant to restrict this picture to cases in which the dynamical state of the network stops changing. This is one kind of attractor, but it is also possible to settle down into periodic or even chaotic patterns of activity, as described below.

It is possible in principle to perform computations based on the transient approach to the attractor, in addition to or instead of on the basis of the attractor alone. However, we will not consider this alternative in this article.

Computing with attractors is appealing because it does not require the person reading the result to observe the entire evolution of the network. Neither need she know when the computation was started or how long it took. All that is required is the identification of the attractor that a given initial condition evolves toward (and a way of recognizing that the network has reached the attractor). Therefore, this survey will begin with a brief description of the kinds of attractors one can meet: fixed points, limit cycles, and strange attractors. We will illustrate the paradigm with a simple example, the Hopfield model for associative memory. We then indicate briefly both how the connections necessary to embed desired patterns can be learned and how the paradigm can be extended to time-dependent attractors. Finally, we examine the possible relevance of attractor computation to the functioning of the brain.

Networks, Attractors, and Stability

We will focus our attention on networks described by systems of differential equations like

$$\tau_i \frac{du_i}{dt} + u_i(t) = \sum_{j \neq i} w_{ij} g[u_j(t)] + h_i(t) \quad (1)$$

Here $u_i(t)$ is the net input to unit i at time t and $g(\)$ is a sigmoidal activation function ($g' > 0$), so that $V_i = g(u_i)$ is the activation of unit i . The connection weight to unit i from unit j is denoted w_{ij} , $h_i(t)$ is an external input, and τ_i is a relaxation time.

We can also consider discrete-time systems governed by

$$V_i(t + 1) = g \left[\sum_j w_{ij} V_j(t) + h_i(t) \right] \quad (2)$$

Here it is understood that all units are updated simultaneously.

These models are deterministic. Noisy networks can be analyzed by the methods of statistical mechanics and are treated extensively elsewhere in this *Handbook* (see, e.g., STATISTICAL MECHANICS OF NEURAL NETWORKS and STATISTICAL MECHANICS OF GENERALIZATION).

Viewing such a network as a computer, data can be read into it in two ways. One is as the $h_i(t)$ on a subset of the units, which we can call input units. The $h_i(t)$ values might be held fixed or varied in time, depending on the problem. This way of loading data is, of course, just carried over directly from the conventional input-output paradigm as we normally apply it to feedforward networks. Alternatively, we can load the data by setting the values of the initial activations $V_i(0)$. For layered feedforward networks this procedure is equivalent to the previous one, but for recurrent nets it is fundamentally different. We can also use both these input mechanisms simultaneously. The first is appropriate when we want to have the output vary with the input (e.g., continuous mapping), while we use the second when we want an output to be insensitive

to small changes in the input (e.g., error correction). As the second one is intrinsic to recurrent networks, we will focus most of our attention on it.

The program of such a computer is its connection weights w_{ij} . Some of them may be zero, but the questions we address here are rather trivial unless there is some feedback, i.e., unless our networks are *recurrent*. We will not restrict our attention to symmetric connections; w_{ij} need not equal w_{ji} .

Finding the correct weights to implement a particular computation is a highly complex problem. However, for recurrent networks, we must first understand something about the attractors that represent the results of computations, so we now turn our attention to this problem.

To describe the dynamics of our networks, we make use of a picture in which the activation of each unit in the network is associated with a direction in a multidimensional space, called the *configuration space*. Every point in this space represents a possible state of the network, called the *state vector*, and the motion of this vector represents its evolution in time. For all recurrent networks of interest, there are just three possibilities for the asymptotic state:

1. The state vector comes to rest, i.e., the unit activations stop changing. This is the simplest case to analyze and is called a *fixed point*. Different results of a computation (owing to different input data) are characterized by settling into different fixed points. The region of initial states that settles into a single fixed point is called its *basin of attraction*. Most of the examples of recurrent networks in the literature, such as the Hopfield model and many related ones, compute with fixed points.
2. The state vector settles into a periodic motion, called a *limit cycle*.
3. The state vector moves chaotically, in the sense that two copies of the system that initially have nearly identical states will grow more and more dissimilar as they evolve: the two state vectors diverge from each other. However, the way they diverge is restricted. At any time the two state vectors are actually growing closer together in many directions in the configuration space; the divergence occurs only in some (typically a few) directions. A Poincaré map showing, e.g., the states of some of the units every time the state vector passes through some hyperplane in configuration space will be a fractal object with a dimensionality greater than zero (Schuster, 1989). This kind of attractor is called *strange* (see CHAOS IN BIOLOGICAL SYSTEMS and CHAOS IN NEURAL SYSTEMS).

Which kind of attractor we obtain will depend on the connections in the network and, possibly, the input data. Suitable learning algorithms make it possible to design the desired type. In simple applications, fixed points are naturally easiest to deal with. However, it may sometimes be advantageous to exploit the richer dynamical possibilities available in non-fixed-point attractors. For example, limit cycles allow the timing of the network's response to be controlled.

For all three kinds of attractors, the computation performed is a mapping from an initial condition to a particular attractor. It is evident that the dynamics partitions the configuration space into basins of attraction around the attractors. All initial conditions within a given basin map to the same attractor; that is, they are classified in the same way by the computation.

There are conditions under which the attractors will always be fixed points. For nets described by the continuous dynamics of Equation 1, a sufficient (but not necessary) condition is that the connection weights be symmetric: $w_{ij} = w_{ji}$.

General results about the stability of recurrent nets were proved by Cohen and Grossberg (1983). They showed that for static external input h_i , if the connection matrix w_{ij} is symmetric, the at-

tractors of Equation 1 are always fixed points, even if the activation function is allowed to differ from unit to unit, the $u_i(t)$ on the left-hand side is replaced by a general monotonic function $b_i(u_i)$, and τ_i is a (positive) function of u_i .

The proof illustrates the basic mathematical strategy for proving stability. Suppose we can find some quantity, a nontrivial function of the state variables u_i , which always decreases under the dynamics described in Equation 1 except for special values of u_i at which it does not change. These values are fixed points. For values of u_i near such a point, the system will evolve either toward it (an attractor) or away from it (a repeller). For almost all starting states, the dynamics will end at one of the attractor fixed points. Furthermore, these are the only attractors. If there were a limit cycle, for example, our function would decrease everywhere on a closed curve, which is impossible. Thus, whether such a quantity exists for a given network is very important. A function with this property is called a *Lyapunov function*.

There is indeed such a function for the Cohen-Grossberg extension of the dynamics described in Equation 1:

$$L(\mathbf{u}) = \sum_i \int^{u_i} [h_i(u) - h_i] g'_i(u) du - \frac{1}{2} \sum_{ij} w_{ij} g_i(u_i) g_j(u_j) \quad (3)$$

We can show directly that it is a Lyapunov function by computing its time derivative. Making use of the equations of motion and the symmetry of w_{ij} , we find

$$\dot{L} = - \sum_i g'_i(u_i) \tau_i(u_i) \dot{u}_i^2 \leq 0 \quad (4)$$

Thus L is always decreasing, except at the fixed points $\dot{u}_i = 0$.

When a Lyapunov function exists, we may think about the dynamics in terms of sliding downhill on a surface in configuration space, the height of which is given by $L(\mathbf{u})$. The motion is not simple gradient descent, since \dot{u}_i is not exactly proportional to $-\partial L/\partial u_i$ (there is an extra factor $g'_i(u_i) \tau_i(u_i)$). Nevertheless, the motion is always downhill, and the bottoms of the valleys correspond to the fixed points.

If we know the form of the Lyapunov function for a particular kind of network, this picture gives us a clue about how to program desired fixed-point attractors: we try to choose the connection weights w_{ij} and biases h_i so that L has minima at these points in configuration space.

To gain a little more insight, we restrict ourselves to the case $b_i(u) = u$ and an activation function $g(u) = \tanh(\beta u)$. Using the activation variables $V_i = g(u_i)$ instead of u_i , we find that we can write L in the form

$$L(\mathbf{V}) = \frac{1}{\beta} \sum_i \int^{V_i} dy \tanh^{-1} y - \sum_i h_i V_i - \frac{1}{2} \sum_{ij} w_{ij} V_i V_j \quad (5)$$

For large gain β , the first term is small. It is natural to think of the other two simply as a "potential energy" that the system tries to minimize. The w_{ij} and h_i should thus be chosen so that their sum has minima at or near the desired fixed points. The main effect of the first term is just to prevent the activations from reaching 1 or -1, since its derivatives diverge there.

Sometimes it is simple to construct a potential energy with the desired minima, at least approximately. In other problems, this strategy may be inadequate, and we have to resort to iterative learning algorithms to determine the network parameters.

Limit cycles and strange attractors are harder to handle mathematically. Often, however, it is possible to proceed in some kind of analogy with the fixed-point case.

Associative Memory

The most celebrated application of computing with fixed points is ASSOCIATIVE NETWORKS (q.v.). Here we follow the treatment due

to Hopfield (1984) (see also Hertz, Krogh, and Palmer, 1991, chaps. 2 and 3). There is a set of patterns to be stored somehow by the computer. Given as input a pattern that is a corrupted version of one of these, the result of the computation should be the corresponding uncorrupted one.

The strategy for solving this problem is to try to guess a form for the potential energy part of Equation 5 that will have minima at the configurations corresponding to the patterns to be stored. We take patterns $\xi_i^\mu = \pm 1$. The subscript i labels the N elements of the pattern (e.g., pixel values), and the superscript μ labels the p different patterns in the set. The patterns are assumed random and independent for both different i and different μ . (This assumption is artificial for most potential applications, but studying this simple case affords us some insight into how the network works, and we will see how to relax it in the next section.) If we wanted to store just one such pattern ξ_i , a natural choice is just to take the potential energy function proportional to $-(\sum_i \xi_i V_i)^2$. The quantity $\sum_i \xi_i V_i$ measures the similarity between the pattern and the state of the network. It achieves its maximal value at $+N$, and therefore $-(\sum_i \xi_i V_i)^2$ is minimal if and only if every V_i coincides with ξ_i . For more than one pattern we try one such term for each pattern, yielding a total potential energy

$$H = -\frac{1}{2N} \sum_{\mu=1}^p \left(\sum_i \xi_i^\mu V_i \right)^2 \quad (6)$$

Multiplying out the square of the sum, we can identify the connection weights in Equation 5 as

$$w_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu \quad (7)$$

The form of this equation suggests a Hebbian interpretation (see HEBBIAN SYNAPTIC PLASTICITY). For each pattern, there is a contribution to the connection weight proportional to the product of sending (ξ_j^μ) and receiving (ξ_i^μ) unit activities when the network is in the state $V_i = \xi_i^\mu$. This is just the form of synaptic strength proposed by Hebb (1949) as the basis of animal memory, so this is sometimes called a Hebbian storage prescription. This matrix is symmetric and has positive definite eigenvalues, so our earlier results guarantee that the attractors of our network dynamics are fixed points for both continuous and discrete-time dynamics.

The hope is that this will produce a fixed point of the network dynamics at or near each ξ_i^μ . (Because our H is purely quadratic in the V_i , we also expect fixed points near $-\xi_i^\mu$.) We can see how well this works by examining the stationary points of the Lyapunov function described by Equation 5, which are

$$V_i = \tanh \left(\beta \sum_j w_{ij} V_j \right) \quad (8)$$

We would first like to know whether there are solutions of Equation 8 that vary across the units like the individual patterns ξ_i^μ .

The quality of retrieval of a particular stored pattern ξ_i^μ is measured by the quantity $m_\mu = N^{-1} \sum_i \xi_i^\mu V_i$. Using Equation 8, with the weight formula of Equation 7, we obtain

$$m_\mu = \frac{1}{N} \sum_i \xi_i^\mu \tanh \left(\beta \sum_v \xi_i^\mu m_v \right) \quad (9)$$

The kind of solution we are looking for should describe a state of the network that is correlated with only one of the stored patterns, i.e., just one of the $m_\mu \neq 0$. With this restriction, Equation 9 reduces to

$$m = \tanh(\beta m) \quad (10)$$

where m is the value of the one non-zero m_μ . This equation has nontrivial solutions whenever $\beta > 1$. Next we have to inquire

whether these solutions are truly attractors, i.e., whether they are stable. At this point the story gets mathematically involved, so we simply survey the results (Kühn, Bös, and van Hemmen, 1991). Everything here is derived in the limit of a large network, i.e., $N \rightarrow \infty$; thus, the tools of statistical mechanics can be brought to bear.

The story is quite simple when p (the number of stored patterns) is a negligible fraction of N , the number of units in the network. Then the nontrivial solutions are globally stable, while the solution $m = 0$ is unstable, whenever $\beta > 1$.

If the gain is high enough, there are other attractors in addition to the ones we have tried to program into the network with the choice of Equation 7. In the simplest of these, the state of the network is equally correlated with three of the ξ_i^μ , say, $\xi_i^{\mu_1}$, $\xi_i^{\mu_2}$, and $\xi_i^{\mu_3}$. These other attractors are thus mixtures of three of our desired attractors. Such solutions exist whenever $\beta > 1$, but they are locally stable only when $\beta > 2.17$. Turning the gain up higher still, combinations of greater numbers of the desired memories also become stable. Thus, by keeping the gain between 1 and 2.17, we can limit the attractor set to the desired states.

When p is of the same order as N , the analysis is more involved. The root of the problem is that the different terms in the weight formula in Equation 7 interfere with each other, even for independent random patterns. (The overlap between two patterns is of order $N^{-1/2}$, but as there are of order N such overlaps, the net effect is of order 1.) This cross-talk has three important effects. First, it induces small mismatches, which grow with increasing $\alpha = p/N$, between the original patterns ξ_i^μ and the attractors. Second, and more dramatically, it destroys the pattern-correlated attractors completely above a critical value of α , α_c . This critical value depends on the gain β , and in the limit $\beta \rightarrow \infty$, α_c approaches 0.14. Finally, one finds that whenever the gain exceeds $\beta_s = (1 + 2\sqrt{\alpha})^{-1} \leq 1$ there are infinitely many fixed points, all completely uncorrelated with the patterns ξ_i^μ .

Nevertheless, as long as we are not trying to store too many patterns ($\alpha < \alpha_c$), there will be attractors that are strongly correlated with the patterns. The unwanted other attractors can no longer be completely eliminated by suitable tuning of the gain, but as they are uncorrelated with the patterns, they do not have much effect on the retrieval of a pattern, starting from an initial configuration not too far from the attractor.

Thus, attractor computation works in this system over a wide range of model parameters. It can be shown to be robust with respect to many other variations as well. These include dilution (random removal of connections), asymmetry (making some of the $w_{ij} \neq w_{ji}$), and quantization or clipping of the weight values.

Obviously, both this network and the kind of computation it performs are very simple. It is not directly applicable to problems like scene analysis, where several objects, as well as relationships between them, have to be identified and characterized. It is possible to construct more complex networks, with modular and hierarchical structure, to perform such computations, but space does not permit us to treat them here.

A number of other problems, in particular in optimization theory, have been treated using the same strategy of choosing the connections so that the potential energy has minima in the appropriate places (see OPTIMIZATION, NEURAL). The features we have noted in the associative memory problem appear to be universal. It is possible to obtain the desired attractors (at least approximately), but other, undesired attractors are generally also created. These can be controlled in some degree by suitable adjustment of the gain or other parameters.

Learning

The weight formula in Equation 7 was only an educated guess. It is possible to obtain better weights that reduce the cross-talk and increase α_c by employing systematic learning algorithms.

One of the simplest of these is *Boltzmann learning* (see SIMULATED ANNEALING AND BOLTZMANN MACHINES and Hertz et al., 1991, chap. 7). Originally, Boltzmann learning was formulated for stochastic binary units. Here we use a formulation for continuous-valued deterministic units. Suppose, as above, that we want to make an attractor of the configuration in which the unit activations are proportional to pattern ξ_i^μ . Now if we start the network in the configuration $V_i = \xi_i^\mu$, it will settle into some fixed point V_i^μ . The algorithm is to change the weights according to

$$\Delta w_{ij} = \eta(\xi_i^\mu \xi_j^\mu - V_i^\mu V_j^\mu) \quad (11)$$

The first term is a Hebb-like learning term, like Equation 7, and the second term ensures that learning stops when the fixed point V_i^μ coincides with ξ_i^μ . This is then performed for every pattern and repeated until the attractors converge to the desired locations. It is evident from Equation 11 that the resulting connections will be symmetric if the initial ones are. Boltzmann learning can also be used when there are hidden units. In that case, when i or j is a hidden unit, the patterns ξ_i^μ or ξ_j^μ are simply replaced by the stationary values those units take when the nonhidden units are clamped at the pattern values.

In the stochastic Boltzmann machine algorithm, the degree of stochasticity is controlled by a “temperature” parameter T . It corresponds in our deterministic network of continuous-valued units to the degree of softness in the sigmoidal activation function $g(u)$. (For $g(u) = \tanh \beta u$, the parameter β can be identified with $1/T$.) For both kinds of networks, convergence to the fixed point V_i^μ is aided by starting at high T and gradually lowering it. This procedure goes by the name *simulated annealing*.

The other simple learning rule that can be used to learn particular attractors is the delta rule. When there are no hidden units, its weight updating rule is

$$\Delta w_{ij} = \eta(\xi_i^\mu - V_i^\mu)\xi_j^\mu \quad (12)$$

As in Equation 11, the role of the second or “unlearning” term in the parentheses is to turn learning off when the desired fixed points are achieved. With hidden units, the delta rule becomes what is known as *backpropagation* (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION). There is insufficient space here to go into the mathematical description of backpropagation in recurrent networks (but see RECURRENT NETWORKS: LEARNING ALGORITHMS). We only remark that it is describable as propagating weight adjustments through the original network with all the directions of the connections reversed (see Hertz et al., 1991, chap. 7).

Nonstationary Attractors

So far, we have worked with networks with first-order dynamics (1 or 2) and a symmetric weight matrix. If we relax either of these conditions, non-fixed-point attractors are possible.

It is possible to extend the Hopfield model described above to store pattern sequences, by including suitable delays into the discrete-time dynamics (2). This problem is treated in detail in TEMPORAL SEQUENCES: LEARNING AND GLOBAL ANALYSIS (q.v.). It can be mapped onto the one with static patterns, and much of the analysis for that model can be carried over to the dynamic case. Recent experimental investigations (Markram et al., 1997; Bi and Poo, 1998) have found synaptic changes that depend in sign on the relative timing of pre- and postsynaptic spikes. Such synaptic dynamics could provide a physiological substrate for sequence learning.

In a purely computational context, iterative learning algorithms can also be brought to bear to stabilize specific desired limit cycles. In particular, the recurrent backpropagation algorithm mentioned above for learning fixed points can be extended rather straightforwardly

wardly to learning arbitrary time-dependent patterns (see Hertz et al., 1991, chap. 7).

If networks can learn arbitrary periodic attractors, the obvious next question is whether they can learn strange attractors. The answer to this is also affirmative. The initial work on this problem was done by Lapedes and Farber (1987), who succeeded in teaching a network a strange attractor generated by a nonlinear differential-delay equation known as the Mackey-Glass equation, which was originally introduced in a model of blood production.

Discussion: Attractors in the Brain?

Both local (see MOTOR PATTERN GENERATION) and macroscopic neural activity (as in epilepsy; see EEG and MEG ANALYSIS) can be described in terms of non-fixed-point attractors. However, the most interesting questions about attractors in the brain have to do with their functional roles in processes such as perception, recognition, and memory. For example, are particular attractors associated with the recognition of particular objects? Can the settling of the brain's activity into such an attractor be identified with the recognition process?

It has also been suggested (Skarda and Freeman, 1987) that a strange attractor in which the system hops irregularly between two or more regions of configuration space can provide a model for understanding why the brain does not get trapped forever in a fixed-point or limit-cycle attractor.

There is progress toward more biologically realistic modeling of the neural structures that could carry out the kinds of computations we have been discussing (see CORTICAL HEBBIAN MODULES).

Some interesting evidence in favor of attractor computation comes from modeling by Grinasty, Tsodyks, and Amit (1993) of some experiments in which monkeys learn to identify a set of visual patterns. During training, the patterns are presented in a particular order. After learning, the patterns are presented in random order and multicellular recordings are made in a small region of anterior ventral temporal cortex during the period between stimuli. The spatial pattern of the mean firing rates across the electrode array is found to be stimulus specific. The interesting finding is that although the stimuli are not correlated with each other, the resulting firing patterns are. Strong correlations are found only between pairs of activity patterns evoked by stimuli that were close together in the training sequence.

In the theoretical analysis, the firing rate patterns are identified with attractors of the cortical dynamics. They suppose there are pre- and postsynaptic delays in the learning of the patterns, which leads to new terms proportional to $N^{-1} \sum_{\mu} \xi_i^\mu \xi_j^{\mu \pm 1}$ added to the weight formula of Equation 7. This has the consequence that the attractor corresponding to a particular pattern is correlated with those of nearby patterns in the sequence. The form of this correlation can be calculated and is quite similar to that observed in Miyashita's experiments. This result lends credence to both the idea of attractor computation and the Hebbian learning picture.

Another phenomenon in which attractor networks appear to play an important role is working memory. The basic experimental finding is the following. Macaque monkeys perform a match-to-sample task: in each trial the animal sees a sequence of visual patterns, and it must press a bar when the one shown first is repeated. It thus has to keep a memory of the first stimulus during the intervening ones in the sequence. Some neurons in prefrontal (PF) cortex that are selectively sensitive to the first (sample) pattern exhibit continuing activity, above background level, in the entire delay period up to the re-presentation of that pattern, despite the intervening stimuli. Some neurons in inferior temporal (IT) cortex, which provides input to PF, exhibit similarly persistent “delay activity,” but this activity is terminated by the presentation of different, intervening

stimuli. Thus, these areas seem to be involved in the temporary storage of memories for visual patterns.

Associative memory networks incorporating important features of cortical circuitry and firing dynamics and exhibiting delay activity for learned attractors were developed by Amit and Brunel (1997). Their approach was developed further and applied to the phenomenology described above by Renart, Parga, and Rolls (2000) and Renart et al. (2001), who modeled IT and PF cortices as a pair of coupled associative memory modules. For appropriate coupling between the modules, the initial stimulus leads to an attractor state, associated with the first pattern, that persists in the PF module, even after the stimulus is changed and the state of the IT module is pushed into another attractor by the new (different) stimulus.

Related work by Compte et al. (2000), focusing on PF cortex alone, models experiments on an oculomotor delayed-response task in which a monkey has to remember the location of a visual cue during a delay period of a few seconds before it makes a behavioral response. Although the details of the two groups' models differ, both reproduce the phenomenology of the experiments quite well, lending support to an interpretation in terms of attractor computation.

It is thus evident that the attractor hypothesis at least provides a useful framework within which to address a number of problems in cortical computation, both theoretically and experimentally. It is too soon to know how widely it will be applicable, but it plays a key role in the current interplay of theory and experiment, as we try to build a coherent mathematical framework for cognitive neuroscience.

Road Maps: Dynamic Systems; Grounding Models of Networks

Related Reading: Dynamics and Bifurcation in Neural Nets; Energy Functionals for Neural Networks

References

- Amit, D. J., and Brunel, N., 1997, Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex, *Cerebral Cortex*, 7:237–252.
- Bi, G. Q., and Poo, M. M., 1998, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type, *J. Neurosci.*, 18:10464–10472.
- Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J., 2000, Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model, *Cerebral Cortex*, 10:910–923.
- Cohen, M., and Grossberg, S., 1983, Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Trans. Syst. Man Cybern.*, 13:815–826.
- Griniasty, M., Tsodyks, M. V., and Amit, D. J., 1993, Conversion of temporal correlations between stimuli to spatial correlations between attractors, *Neural Computat.*, 5:1–17.
- Hebb, D. O., 1949, *The Organization of Behavior*, New York: Wiley. ♦
- Hertz, J. A., Krogh, A. S., and Palmer, R. G., 1991, *Introduction to the Theory of Neural Computation*, Redwood City, CA: Addison-Wesley. ♦
- Hopfield, J. J., 1984, Neurons with graded responses have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA*, 79:3088–3092.
- Kühn, R., Bös, S., and van Hemmen, L., 1991, Statistical mechanics of graded-response neurons, *Phys. Rev. A*, 43:2084–2087.
- Lapedes, A., and Farber, R., 1987, Nonlinear Signal Processing Using Neural Networks: Prediction and Signal Modelling, Technical Report LA-UR-87-2662, Los Alamos, NM: Los Alamos National Laboratory.
- Markram, H., Lubke, J., Frotscher, M., and Sakmann, B., 1997, Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs, *Science*, 275:213–215.
- Renart, A., Moreno, R., de la Rocha, J., Parga, N., and Rolls, E. T., 2001, A model of the IT-PF network in object working memory which includes balanced persistent activity and tuned inhibition, *Neurocomputing*, 38–40:1525–1531.
- Renart, A., Parga, N., and Rolls, E. T., 2000, A recurrent model of the interaction between prefrontal and inferotemporal cortex in delay tasks, *Adv. Neural Inform. Proc. Syst.*, 12:171–177.
- Skarda, C. A., and Freeman, W. J., 1987, How brains make chaos in order to make sense of the world, *Behav. Brain Sci.*, 10:161–195.
- Schuster, H. G., 1989, *Deterministic Chaos*, 2nd ed., Weinheim, Germany: VCH Verlagsgesellschaft. ♦

Concept Learning

Thomas J. Palmeri and David C. Noelle

Introduction

Concepts are mental representations of kinds of objects, events, or ideas. We have concepts because they allow us to see something as a kind of thing rather than just as a unique individual. Concepts allow generalization from past experiences: By treating something as a kind—as an instantiation of some concept—as a member of some category—we can use what we have learned from other examples of that kind. Concepts permit inferences: Deciding predator from prey, edible from inedible, friend from enemy involves concepts. Concepts facilitate communication: Describing something as a kind of thing may obviate the need to provide details of the thing itself. Concepts permit different levels of abstraction: We know the difference between objects and ideas, animals and plants, dogs and cats, and terriers and collies. Concepts bring cognitive economy: What we learn about animals generally can be applied to specific animals without needless replication of that knowledge throughout our conceptual hierarchy. Concepts are fundamental building blocks of human knowledge (see Margolis and Laurence, 1999).

The focus of this article is on learning mental representations of new concepts from experience. We will also address how we use

mental representations of concepts to make categorization decisions and other kinds of judgments.

Overview of Concept Learning Models

One goal of model development is to test specific hypotheses regarding the kinds of representations created during learning and the kinds of processes used to act upon those representations to make decisions. In order to develop a formal model of concept learning, a modeler must specify what perceptual information is provided by the sensory system, how that information is represented, how that information is compared with what has been learned about a concept, how this previously learned knowledge is represented in memory, and how decisions are made based on comparing perceptual information with stored conceptual representations (see PATTERN RECOGNITION).

A tacit assumption of many models of concept learning is that the perceptual system extracts information from the environment, passing a perceptual representation on to a conceptual stage of processing, which in turn generates an action. In the parlance of neural

networks, the perceptual system provides the inputs to the network, the association weights and activation functions encode the conceptual representations, and decision processes use the outputs to generate a response. Most concept learning models characterize the perceptual system as a dimensionality reduction device (see OBJECT RECOGNITION and OBJECT STRUCTURE, VISUAL PROCESSING). For example, in the case of vision, the input on the retina is an extremely high-dimensional representation, with every photoreceptor effectively encoding an independent dimension of sensation. The perceptual system creates a relatively low-dimensional representation by recoding the retinal input in terms of a smaller number of features or dimensions. These vectors of features or dimensions serve as the inputs to the concept learning network.

The distinction between features and dimensions is fully discussed elsewhere (e.g., Tversky, 1977). The members of a category—the extension of a concept—may be seen as clusters of perceptual vectors in a psychological space. An important type of concept learning entails associating regions in that space with particular category labels.

A *featural representation* in a neural network essentially consists of a vector of input nodes encoding the presence or absence of primitive elements. Similar stimuli share many common features. Dissimilar stimuli correspond to uncorrelated vectors. Often in simulation modeling, feature representations have no direct relationship to the actual stimuli of an experiment, but are instead designed to capture the statistical relationships among stimuli.

A *dimensional representation* is not discrete, but represents information in terms of values along continuously varying psychological dimensions. Similar stimuli have similar values along the dimensions, occupying adjacent locations in psychological space. Often in simulation modeling, dimensional representations may be derived from physical properties of actual stimuli or may be derived from similarity ratings (or other measures of psychological proximity) made by subjects using techniques such as multidimensional scaling (although feature representations can be derived as well).

The core of this article reviews models of how concepts are learned and represented in neural networks. One important distinction between models is whether concepts have LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS (q.v.). Another is whether conceptual knowledge relies on abstractions, such as rules or prototypes, or relies on specific exemplar knowledge. Most models focus on supervised learning, where trial-by-trial feedback is supplied, but some models address unsupervised learning. Most models focus on how concepts are learned within a category learning paradigm, whereby subjects learn to produce the correct label for each stimulus, but some models address how subjects can learn to infer properties other than the category label. Most models focus on

learning from induction over examples, but some address learning from instruction or from explanation as well.

Concept Learning Models

Rule Models

Early philosophers conjectured that all concepts were decomposable into necessary and sufficient conditions for membership (e.g., Plato in Margolis and Laurence [1999]). A “triangle” is a closed form with three sides, a “bachelor” is an unmarried adult male, and so forth. Conceptual rules are like carefully worded definitions provided to a student. Indeed, a strength of this hypothesis is the apparent alignment of mental representations with self-reports of conceptual knowledge. Generally, to be considered a rule, a concise definition is required—if arbitrarily complex rules are allowed, the rule hypothesis becomes vacuous, because virtually any representation can be characterized by complex rules. Therefore, rule representations typically include just a small set of dimensions—sometimes only a single dimension. Although this may seem overly restrictive, humans frequently exhibit reliance on individual dimensions during concept learning (Nosofsky, Palmeri, and McKimley, 1994).

Simple neural network units may be connected to compute logical functions (see NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY). A rule involving a threshold along a single dimension is the simplest example, and serves as the basis for rules in some concept learning models (e.g., Ashby et al., 1998; Erickson and Kruschke, 1998). Assuming a dimensional input representation, a single-dimension rule simply involves a learned weighted connection w_{ij} from input node a_i to output unit o_j with learned bias θ_j

$$o_j = \frac{1}{[1 + \exp(-w_{ij}a_i - \theta_j)]} \quad (1)$$

The sigmoidal activation of this unit is proportional to the likelihood of category membership. A network of this kind essentially implements a linear decision boundary in psychological space, with the boundary orthogonal to one of the psychological dimensions, and with the position of the boundary specified by the learned bias θ_j (see Figure 1A).

Prototype Models

Although people often report conceptual knowledge in terms of rules, there are reasons to question rules as the universal basis for conceptual knowledge. Many common concepts are difficult to de-

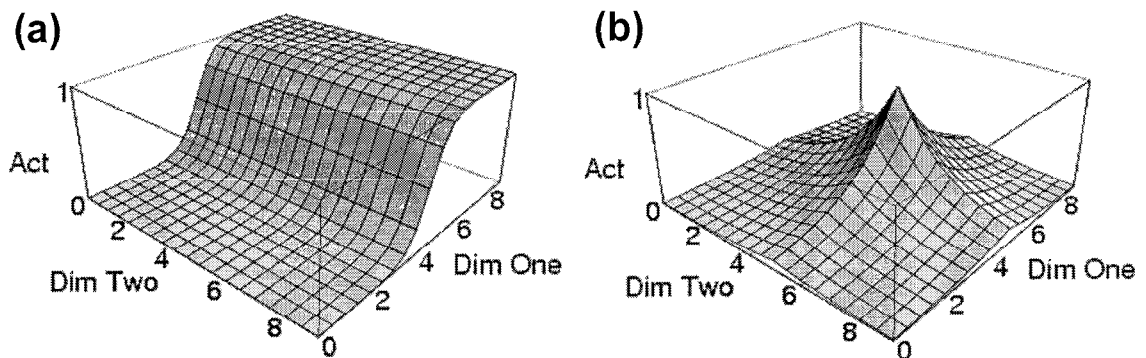


Figure 1. Examples of activation (Act) as a function of location in a two-dimensional psychological space (Dim One \times Dim Two) for (a), the logistic sigmoidal function in Equations 1–3 and for (b), the radial-basis function in Equation 4.

fine using rules—Wittgenstein suggested the example of “games” as appearing to defy definition (see Margolis and Laurence, 1999). Instead, concepts seem to possess “family resemblances,” with instances bearing many similarities but no characteristics common to all members. Human performance often belies the existence of rules in that some items appear to be “better” members than others in terms of typicality ratings, speed of processing, and inductive power.

Such findings led to an alternative view of conceptual knowledge based on abstract prototypes (see Rosch and also Lakoff in Margolis and Laurence, 1999). A prototype need never be directly experienced, but can be formed by averaging across observed instances. In psychological space, the prototype is the centroid of a cloud of instances. New items are classified according to their relative similarity to learned prototypes, with typicality effects emerging from this process. Learning just two prototypes effectively partitions psychological space into two regions separated by a linear boundary, but this boundary is a fuzzy (probabilistic) one and is unconstrained in terms of its orientation within psychological space.

A simple two-layer prototype model assumes an input layer of features (or dimensions) with learned associations to category output nodes. Each output unit o_j corresponds to a single concept prototype, and the weights w_{ij} from inputs a_i to each o_j reflects the strength of association of particular stimulus features with that concept:

$$o_j = \frac{1}{1 + \exp\left(-\sum_i w_{ij}a_i - \theta_j\right)} \quad (2)$$

The probability $P(j)$ of categorizing a stimulus as a member of category j can simply be given by the relative activation of node j compared to all other nodes. Other more dynamic mechanisms, such as lateral inhibition, can also introduce competition between concepts in WINNER-TAKE-ALL NETWORKS (q.v.).

This simple network learns to associate input stimuli with their corresponding categories. A related approach is to train ASSOCIATIVE NETWORKS (q.v.) to reproduce the features of each training instance as well as the correct category label (e.g., McClelland and Rumelhart, 1985). This pattern completion approach permits the network to not only categorize stimuli, but also to infer other missing features as well. More powerful pattern completion arises when COMPUTING WITH ATTRACTORS (q.v.) such as in the Brain State in a Box model (see ASSOCIATIVE NETWORKS). Recurrent connections between all category label units and all feature units permit the network to encode soft constraints guiding how activation settles over time. In addition to their pattern completion properties, learned basins of attraction in such networks can instantiate nonlinear decision boundaries between categories, potentially creating

a kind of prototype model that incorporates information about both the mean and the variability of a distribution of category exemplars.

Exemplar Models

Simple concept learning networks of the sort just described can learn category structures (a) and (b) depicted in Figure 2—linearly separable categories—but cannot learn category structure (c)—nonlinearly separable categories. This is the classic XOR problem that stymied early developments in neural networks (see PERCEPTONS, ADALINES, AND BACKPROPAGATION). In contrast, multi-layered networks with an input layer, a hidden layer, and an output layer can learn these category structures. Activation of hidden and output nodes is determined by a nonlinear sigmoid function like that shown in Equation 2. Learning takes place via gradient descent on error, with knowledge fully distributed throughout the network connections (see BACKPROPAGATION: GENERAL PRINCIPLES). These “backpropagation networks” are powerful learning devices and, with sufficient hidden nodes, they can acquire concepts of nearly unlimited complexity.

Psychological models of concept learning attempt to model human behavior, with all its errors and apparent inefficiencies. Although backpropagation networks are powerful machine learning devices, they make relatively poor models of human concept learning. For one, backpropagation networks are insensitive to the psychological dimensional structure within the categories. From a statistical standpoint, structure (a) and structure (b) in Figure 2 have equivalent complexity, so backpropagation networks learn both types equally quickly. But people find structure (a) far easier to learn than structure (b) because structure (a) permits attending to a single dimension. Backpropagation networks generally learn linearly separable categories, such as structure (b), far more quickly than nonlinearly separable categories, such as structure (c). By contrast, people find structure (c) easier to learn. More generally, people are not constrained by linear separability, and often exhibit more rapid learning of nonlinearly separable than linearly separable categories. Finally, backpropagation networks suffer from catastrophic forgetting in which new concept learning overwrites previous concept learning. There have been many modifications to simple backpropagation networks to combat this problem. Common to many approaches is the use of semilocalized representations instead of fully distributed representations (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS).

Exemplar-based models assume local representations (see Smith and Medin in Margolis and Laurence, 1999). In contrast to rule and prototype models, concepts are represented extensionally, in terms of specific category instances. A number of variations of exemplar models have been proposed, and a vast set of empirical phenomena is consistent with them (Nosofsky and Kruschke, 1992). Perhaps the best-known neural network exemplar model is ALCOVE (Kruschke, 1992), which is largely derived from the gen-

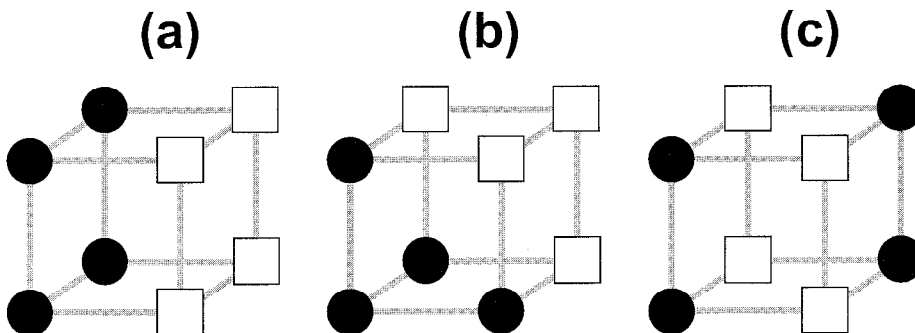


Figure 2. Depictions of three category structures. Individual stimuli are composed of three binary-valued dimensions. The dimensional values of a stimulus are specified by its location in the three-dimensional psychological space. For each structure, black circles represent stimuli in one category and white squares represent stimuli in another category.

eralized context model of categorization (Nosofsky, 1986), but incorporates error-driven learning.

ALCOVE is a three-layered feedforward network in which activation passes from a dimensional input layer, with each dimension scaled by a learned selective attention weight, to an exemplar-based hidden layer, to a category output layer via learned association weights. In contrast to backpropagation networks (compare Figures 1(a) and 1(b)), activation h_j of hidden exemplar node j is based on its similarity to the input stimulus

$$h_j = \exp \left[-c \left(\sum_i \alpha_i e_{ji} - a_i \right)^{1/r} \right] \quad (3)$$

where a_i is the value of the input stimulus along dimension i , e_{ji} is the value of exemplar j along dimension i , α_i is the learned attention to dimension i , c is a similarity scaling parameter, and r determines the psychological distance metric (see also RADIAL BASIS FUNCTION NETWORKS). When optimally allocated, selective attention weights emphasize differences along diagnostic dimensions and deemphasize differences along nondiagnostic dimensions (Nosofsky, 1986).

ALCOVE also learns to associate exemplars with category outputs. Activation of category output node k is given by

$$o_k = \sum_j w_{kj} h_j \quad (4)$$

where w_{kj} is the learned association weight between exemplar j and output node k . The probability of categorizing the input stimulus as a member of category K is given by

$$P(K) = \frac{\exp(\phi o_K)}{\sum_k \exp(\phi o_k)} \quad (5)$$

where ϕ is a response mapping parameter. Attention weights (α_i) and association weights (w_{kj}) are learned by gradient descent on error.

Although exemplar models like ALCOVE have accounted for a wide variety of fundamental categorization phenomena (Nosofsky and Kruschke, 1992), until recently they have ignored the time-course of making a categorization decision. One avenue of recent theoretical development has addressed the time-course of the accumulation of perceptual evidence used to categorize a stimulus (Lamberts, 2000), modeling how perceptual processes make some information available sooner than others. Another avenue of development has examined the time-course of making categorization decisions. Nosofsky and Palmeri (1997) proposed a stochastic exemplar-retrieval model with a competitive decision process to account for both categorization response probabilities and response times. Although this theoretical development was formalized using mathematical modeling tools, such as a random walk decision process, the dynamics of these stochastic, competitive categorization models could be implemented in various neural network architectures as well.

Mixed Models

Exemplar models have been shown to account for a variety of results, including some that were originally thought to unequivocally indicate rules or prototypes as concept representations. Yet, there seems to be some emerging evidence that people do use abstractions, particularly abstract rules, as concept representations. Clearly, people can be instructed to use rules before they have experienced any examples (Palmeri, 1997; Noelle, Cottrell, and McKinley, 2002). Also, it appears that people may approach the task of learning categories by testing simple categorization rules (e.g., Nosofsky et al., 1994), although people may eventually shift to using exemplars with experience (Johansen and Palmeri, in

press). One important focus of current research is developing and testing formal models with mixed representations. At one extreme are models that posit functionally independent rule-based and exemplar-based systems that race to completion (e.g., Palmeri, 1997); exemplar-based representations gain strength with repeated exemplar experience and eventually win the race. Alternatively, rule and exemplar representations may be functionally independent, but the outputs of these systems may compete based on strength of evidence rather than completion time (e.g., Ashby et al., 1998). Erickson and Kruschke (1998) proposed a neural network model (ATRIUM) with separate rule and exemplar representations that compete, with the model learning whether rule-based or exemplar-based information should be used to categorize a particular instance. Finally, other architectures have proposed combinations of rules, exemplars, and perhaps other representations within a single representational medium.

Neuroscientific Models

Neurobiological findings have begun to constrain concept learning models by ruling out mechanisms that resist reasonable neural implementation. Some recent models have included hypotheses concerning how conceptual knowledge is instantiated in the brain.

COVIS is one example of a model grounded in neurophysiology (Ashby et al., 1998). COVIS is a mixed representational model, incorporating implicit decision boundaries and explicit unidimensional rules. The implicit learning system, assumed to reside within the striatum, encodes a category decision boundary (although other representations—such as exemplar-like coarse-coded topological maps of psychological space with regions associated with category labels—also fits within this general framework). Verbal rule processing is assumed to exist in the prefrontal cortex, with the selection of rule-attended dimensions handled by a reinforcement learning process in the anterior cingulate, mediated by projections from the basal ganglia. Given a stimulus, the implicit and rule-based systems compete to provide a category response. This model has been applied to concept learning deficits seen in the very old and the very young, in patients with Parkinson's disease and Huntington's disease, in clinically depressed patients, in individuals with focal brain lesions, and in nonhuman animals.

Other neurally oriented concept learning models have focused on the role of prefrontal cortex in representing and actively maintaining rule-based information during categorization (Noelle et al., 2002; O'Reilly et al., 2002). A distinguishing feature of these models is the use of signals that encode changes in expected future reward (see REINFORCEMENT LEARNING)—emanating from the basal ganglia dopamine system (see DOPAMINE, ROLES OF)—to determine when a useful rule representation has been found and should be gated into a prefrontal working memory system. Like COVIS, these dopamine-gating models incorporate a mixed representation, integrating rules with a kind of procedural knowledge. Unlike COVIS, these models focus on procedural knowledge embedded in multilayer networks, presumably located within the cortex, rather than on a special implicit learning system within the striatum. Rule representations, stored as patterns of activity in a prefrontal attractor network (see COMPUTING WITH ATTRACTORS), do not directly compete with these procedural systems, but rather modulate them through the injection of activity. Models of this kind have provided explanations for frontal lesion data, suggested a coarse topological organization for prefrontal cortex, captured patterns of performance on dynamic classification tasks, explained interference effects in instructed category learning, and illuminated learning deficits in schizophrenia.

Discussion

In this article, we limited our discussion to the kinds of representations and processes that subserve a particular aspect of concept

learning, namely learning to categorize. Recent work has investigated other topics, as well, including how people learn to infer properties other than the category label and how learning about categories may influence perceptual processing in a top-down manner (e.g., Schyns, Goldstone, and Thibaut, 1998).

An important focus of current research was outlined in this article: Do people use different kinds of concept representations, how are those representations learned, and is the dominance of particular representations modulated by experience or other task demands? In order to help answer these questions, and in order to develop neurally plausible models of concept learning, some researchers are beginning to incorporate the constraints imposed by various neuroscientific sources of evidence, including studies of patients with focal brain damage, functional imaging and evoked potential studies, and single unit recordings in animals.

Road Map: Psychology

Related Reading: Feature Analysis; Object Recognition; Pattern Recognition

References

- Asby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M., 1998, A formal neuropsychological theory of multiple systems in category learning, *Psychol. Rev.*, 105:442–481.
- Erickson, M. A., and Kruschke, J. K., 1998, Rules and exemplars in category learning, *J. Exp. Psychol.*, 127:107–140.
- Johansen, M. K., and Palmeri, T. J., (in press), Are there representational shifts during category learning? *Cognitive Psychology*.
- Kruschke, J. K., 1992, ALCOVE: An exemplar-based connectionist model of category learning, *Psychol. Rev.*, 99:22–44.
- Lamberts, K., 2000, Information-accumulation theory of speeded categorization, *Psychol. Rev.*, 107:227–260.
- Margolis, E., and Laurence, S., 1999, *Concepts: Core Readings*, Cambridge, MA: MIT Press. ♦
- McClelland, J. L., and Rumelhart, D. E., 1985, Distributed memory and the representation of general and specific information, *J. Exp. Psychol.*, 114:159–188.
- Noelle, D. C., Cottrell, G. W., and McKinley, C. R. M., 2002, Modeling individual differences in the specialization of an explicit rule, Manuscript under review.
- Nosofsky, R. M., 1986, Attention, similarity, and the identification-categorization relationship, *J. Exp. Psychol.*, 115:39–57.
- Nosofsky, R. M., and Kruschke, J. K., 1992, Investigations of an exemplar-based connectionist model of category learning, in *The Psychology of Learning and Motivation*, vol. 28 (D. L. Medin, Ed.), San Diego, CA: Academic Press, pp. 207–250. ♦
- Nosofsky, R. M., and Palmeri, T. J., 1997, An exemplar-based random walk model of speeded classification, *Psychol. Rev.*, 104:266–300.
- Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C., 1994, Rule-plus-exception model of classification learning, *Psychol. Rev.*, 101:53–79.
- O'Reilly, R. C., Noelle, D. C., Braver, T. S., and Cohen, J. D., 2002, Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control, *Cerebral Cortex*, 12:246–257.
- Palmeri, T. J., 1997, Exemplar similarity and the development of automaticity, *J. Exp. Psychol.*, 23:324–354.
- Schyns, P. G., Goldstone, R. L., and Thibaut, J. P., 1998, The development of features in object concepts, *Behav. Brain Sci.*, 21:1–40. ♦
- Tversky, A., 1977, Features of Similarity, *Psychology Review*, 84:327–352.

Conditioning

Nestor A. Schmajuk

Introduction

During conditioning, animals modify their behavior as a consequence of their experience of the contingencies between environmental events. This article delineates several formal theories and neural network models that have been proposed to describe classical and operant conditioning. Other important theories, omitted here owing to space limitations, are described by Schmajuk (1997).

Classical Conditioning

During classical conditioning, animals change their behavior as a result of the contingencies between the conditioned stimulus (CS) and the unconditioned stimulus (US). Contingencies may vary from very simple to extremely complex ones. For example, in Pavlov's famous experiment, dogs were exposed to the sound of a bell, the CS, followed by food, the US. At the beginning of training, animals only generated unconditioned responses (UR), salivation, when the US was presented. With an increasing number of CS-US pairings, CS presentations elicited a conditioned response (CR). In general, a CR is analogous to the UR (dogs salivate in response to the bell), CR onset precedes the US onset (salivation precedes food presentation), and the peak CR amplitude tends to be located around the time of the occurrence of the US. When acquisition is followed by presentations of CS alone, the CR extinguishes.

Different CS-US interstimulus intervals (ISI) may be employed. Conditioning is negligible with short ISIs, increases dramatically at an optimal ISI that depends on the response being conditioned, and gradually decreases with increasing ISIs.

Second-order conditioning consists of a first phase in which CS₁ is paired with the US. In a second phase, CS₁ and CS₂ are paired together in the absence of the US. Finally, when CS₂ is presented alone, it generates a CR. Sensory preconditioning consists of a first phase in which two CSs, CS₁ and CS₂, are paired together in the absence of the US. In a second phase, CS₁ is paired with the US. Finally, when CS₂ is presented alone, it generates a CR.

In latent inhibition, pre-exposure to CS retards the acquisition of CS-US associations. Latent inhibition is characterized by a large number of properties. In blocking, an animal is first conditioned to CS₁, and this training is followed by conditioning to a compound consisting of CS₁ and a second stimulus, CS₂. This procedure results in a weaker conditioning to CS₂ than would be attained if paired separately with the US. In overshadowing, an animal is conditioned to a compound consisting of CS₁ and CS₂. This procedure results in weaker conditioning to each CS than it would achieve if it was independently trained.

In conditioned inhibition, CS₂ acquires inhibitory conditioning following CS₁ reinforced trials interspersed with CS₁-CS₂ nonreinforced trials. In contrast to excitatory conditioning, presentations of CS₂ alone do not extinguish inhibitory conditioning.

In compound conditioning, two or more stimuli are presented together in the presence of the US. In a feature-positive discrimination, animals receive reinforced simultaneous compound presentations (CS₁ overlapping with CS₂) alternated with nonreinforced presentations of CS₂. Animals learn to respond to CS₁ but not to CS₂. In an occasion-setting paradigm, animals receive reinforced serial compound presentations (CS₁ preceding CS₂) alternated with nonreinforced presentations of CS₂. Animals learn to respond to

the CS₁-CS₂ compound but not to CS₁ or CS₂. In negative patterning, presentations of a reinforced component (CS₁ or CS₂) are intermixed with nonreinforced compound (CS₁-CS₂) presentations. Negative patterning is attained if the response to the compound is smaller than the sum of the responses to the components. In positive patterning, reinforced compound (CS₁-CS₂) presentations are intermixed with nonreinforced component (CS₁ or CS₂) presentations. Positive patterning is attained if the response to the compound is larger than the sum of the responses to the components.

Associations, Predictions, and Connections

Modern learning theories assume that the association between events CS_i and CS_k, $V_{i,k}$, represents the *prediction* that CS_i will be followed by CS_k. Neural network or connectionist theories frequently assume that the association between CS_i and CS_k is represented by the efficacy of the synapses, $V_{i,k}$, that connect a presynaptic neural population excited by CS_i with a postsynaptic neural population that is excited by CS_k (event k might be another CS or the US). When CS_k is the US, this second population controls the generation of the CR. At the beginning of training, synaptic strength $V_{i,US}$ is small, and therefore, CS_i is incapable of exciting the second neural population and generating a CR. As training progresses, synaptic strengths gradually increase, and CS_i comes to generate a CR.

Although some models of conditioning describe changes in $V_{i,k}$ on a trial-to-trial basis, real-time networks describe the unbroken, continuous temporal dynamics of $V_{i,k}$. In general, real-time neural networks assume that CS_i gives rise to a trace, $\tau_i(t)[d(\tau_i)/dt = K_1(CS_i - \tau_i)]$, in the central nervous system that increases over time to a maximum and then gradually decays to zero. The increment in $V_{i,k}$ is a function of the intensity of the CS_i trace at the time the US is presented.

Changes in synaptic strength $V_{i,k}$ might be described by $\Delta V_{i,k} = f(CS_i)f(CS_k)$, where $f(CS_i)$ represents the presynaptic activity and $f(CS_k)$ the postsynaptic activity. Different $f(CS_i)$ and $f(CS_k)$ functions have been proposed. Learning rules for $V_{i,k}$ either assume variations in the effectiveness of CS_i, $f(CS_i)$, the US, $f(CS_k)$, or both. The following sections describe how different types of models deal with the many experimental results presented before.

Variations in the Effectiveness of the CS: Attentional Models

Attentional theories assume that the formation of CS_i-US associations depend on the magnitude of an internal representation of CS_i, $f(CS_i)$. In neural network terms, attention may be interpreted as the modulation of the CS representation that activates the presynaptic neuronal population involved in associative learning. When focused on a particular CS, selective attention enhances the internal representation of that specific CS.

Mackintosh's (1975) attentional theory suggests that CS_i associability, $f(CS_i)$, increases when CS_i is the best predictor of (the CS most strongly associated with) the US and decreases otherwise. Mackintosh's model describes latent inhibition, overshadowing, and blocking. In contrast to Mackintosh's (1975) view, Pearce and Hall (1980) suggested that $f(CS_i)$ increases when CS_i is a poor predictor of the US, $f(CS_i) = IUS - \sum_j V_{j,US} CS_j$, where $\sum_j V_{j,US} CS_j$ represents the aggregate prediction of the US computed on all CSs present at a given moment. In addition to latent inhibition, blocking, and overshadowing, the Pearce and Hall model correctly predicts that latent inhibition might be obtained after training with a weak US. According to Grossberg's (1975) neural attentional theory, pairing of CS_i with a US causes both an association of $f(CS_i)$ with the US and an association of the US with $f(CS_i)$. Sensory representations $f(CS_i)$ compete among themselves for a limited-

capacity short-term memory activation that is reflected in CS_i-US associations.

Variations in the Effectiveness of the US: Simple and Generalized Delta Rules

A popular rule, proposed independently in psychological (Rescorla and Wagner, 1972) and neural network (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION for the Widrow-Hoff rule) domains, has been termed the delta rule. The delta rule describes changes in the synaptic connections between the two neural populations by way of minimizing the squared value of the difference between the output of the population controlling the CR generation and the US. According to the "simple" delta rule, CS_i-US associations are changed until $f(US) = US - \sum_j V_{j,US} CS_j$ is zero. In neural network terms, $f(US)$ can be construed as the modulation of the US signal that activates the postsynaptic neural population involved in associative learning. Rescorla and Wagner showed that the model describes acquisition, extinction, conditioned inhibition, blocking, and overshadowing.

Sutton and Barto (1981) presented a temporally refined version of the Rescorla-Wagner model. In the model, the effectiveness of CS_i is given by the temporal trace $\tau_i = f(CS_i(t)) = Af(CS_i(t)) + BC\tau_i(t)$, which does not change over trials. The effectiveness of the US changes over trials according to $f(US(t)) = (y(t) - y'(t))$, where the output of the model is $y(t) = f[\sum_j V_{j,US} f(CS_j) + f(US)]$, $f(US)$ is the temporal trace of the US, and $y'(t) = Cy'(t) - (1 - C)y(t)$. Computer simulations show that the model correctly describes acquisition, extinction, conditioned inhibition, blocking, overshadowing, primacy effects, and second-order conditioning. In 1990 the authors proposed a new rendering of the Sutton and Barto (1981) model, designated the temporal difference model, in which $f(US) = (US + \gamma y'(t + 1) - y'(t))$. The temporal difference model correctly describes ISI effects, serial-compound conditioning, no extinction of conditioned inhibition, second-order conditioning, and primacy effects.

Kehoe (1988) presented a network that incorporates a hidden-unit layer trained according to a delta rule. In addition to the paradigms described by the Rescorla-Wagner model, the network describes stimulus configuration, learning to learn, savings effects, and positive and negative patterning.

Schmajuk and DiCarlo (1992) introduced a model that, by employing a generalized delta rule (also known as backpropagation; see PERCEPTRONS, ADALINES, AND BACKPROPAGATION) to train a layer of hidden units that configure simple CSs, is able to solve negative and positive patterning. Interestingly, this biologically plausible, real-time rendition of backpropagation differs from the original version in that the error signal that is used to train hidden units, instead of including the derivative of the activation function of the hidden units, simply contains their activation function. Figure 1 shows real-time simulations on trials 1, 4, 8, 12, 16, and 20 in a delay-conditioning paradigm with a 200-ms CS, a 50-ms US, and a 150-ms ISI. As CR amplitude increases over trials, output weights VS_i and VN_j and hidden weights VH_{ij} may increase or decrease.

The network provides correct descriptions of acquisition of delay and trace conditioning, extinction, acquisition-extinction series, blocking, overshadowing, discrimination acquisition and reversal, compound conditioning, feature-positive discrimination, conditioned inhibition, negative patterning, positive patterning, and generalization.

Gluck and Myers (1993) presented a network that also trains a hidden layer through a backpropagation procedure. The authors assume three three-layer networks that work in parallel. The output and hidden layers of one of the networks are trained to associate CS inputs with those same CS inputs and the US. The output layers

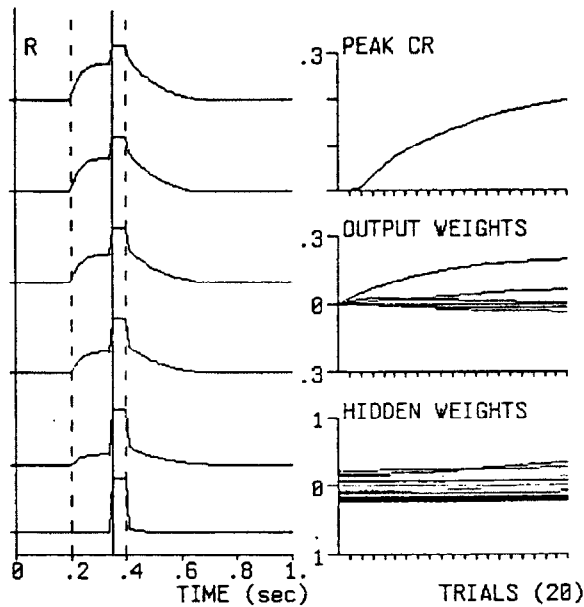


Figure 1. Acquisition of classical conditioning. *Left*, Real-time simulated conditioned and unconditioned response on trials 1, 4, 8, 12, 16, and 20. Vertical dashed lines indicate CS onset and offset. Vertical solid line indicates US onset. Trial 1 is represented at the bottom of the panel. *Right*, Peak CR: Peak CR as a function of trials. Output weights: Average $V_{i,US}$ and $V_{i,N}$ as a function of trials. Hidden weights: Average $V_{H,i}$ as a function of trials. (After Schmajuk and DiCarlo, 1992.)

of the other two networks are also trained by the US, but their hidden units are trained by the hidden units of the first network.

Variations in the Effectiveness of Both the CS and the US

To account for a wider range of classical conditioning paradigms, some theories have combined variations in the effectiveness of both the CS and the US. For example, Wagner (1978) suggested that CS_i -US associations are determined by (1) $f(US) = (US - \sum_j V_{j,US} CS_j)$ as in the Rescorla-Wagner model, and (2) $f(CS_i) = (CS_i - V_{i,CX} CX)$, where CX represents the context and $V_{i,CX}$ the strength of the CX - CS_i association.

Schmajuk, Lam, and Gray (1996) introduced a theory that assumes that $f(CS_i)$ is modulated by the association of the internal representation of CS_i with the total environmental novelty, z_i . Total environmental novelty is given by $\sum_j |\hat{\lambda}_j - \bar{B}_j|$, that is, the sum of the absolute values of the differences of the average predicted and the average observed event j . Schmajuk et al. (1996) showed that the model correctly describes most of the properties of latent inhibition.

Buhusi and Schmajuk (1996) showed that when combined with the Schmajuk and DiCarlo (1992) model (see Figure 2), the Schmajuk et al. (1996) approach can describe a wide variety of classical conditioning data; see the figure caption for details.

Multiple Representations of the CS: Timing

The fact that the peak CR amplitude tends to be located around the time of the occurrence of the US suggests that animals learn about the temporal relationship between the CS and the US. Grossberg

and Schmajuk (1989) proposed a neural network (called the spectral timing model) that is capable of learning the temporal relationships between the CS and the US. The model consists of three layers of neural elements. A step function, activated by the CS presentation, excites the first layer that contains many elements, each one having a different reaction time. The output of each element in the first layer is a sigmoid function that activates a second layer of habituating transmitter gates. In turn, the output of each transmitter gate activates a $f(CS_i)$ element. Those $f(CS_i)$ elements active at the time of the US presentation become associated with the US in proportion to their activity. All $f(CS_i)$ elements activate their corresponding $V_{i,US}$ weights and are added to generate the CR. During testing, the CR shows a peak at the time when the $f(CS_i)$ elements that have been active simultaneously with the US are active again. The model is able to describe ISI curves with single and multiple USs and a Weber's law for temporal generalization. Grossberg and Schmajuk (1989) showed that the model can explain the effects of increasing CS and US intensity, an inverted U in learning as a function of ISI, multiple timing peaks, effect of increasing US duration, and the effect of drugs on timed motor behavior.

Church and Broadbent (1991) presented a connectionist version of Church's scalar timing theory. The model consists of the following components: (1) a pacemaker that emits pulses, (2) a switch that is opened at the onset of the event to be timed and closed at its offset, (3) a counter that accumulates pulses, (4) a reference memory that accumulates pulses of reinforced times and a working memory that stores the total number of pulses accumulated in a particular trial, and (5) a comparator that compares the values stored in both memories. Values stored in working memory are compared to values stored in reference memory, and if they are similar, a response is produced. Notice that the number of stored pulses increases with the measured time. The model is able to describe timing with single but not with multiple USs and describes a Weber's law for temporal generalization.

Operant Conditioning

During operant (or instrumental) conditioning, animals change their behavior as a result of a triple contingency between its responses (R), discriminative stimuli (S_D), and the reinforcer (US). Animals are exposed to the US in a relatively close temporal relationship with the S_D and R. As they experience the S_D -R-US contingency, animals emit R when S_D is presented.

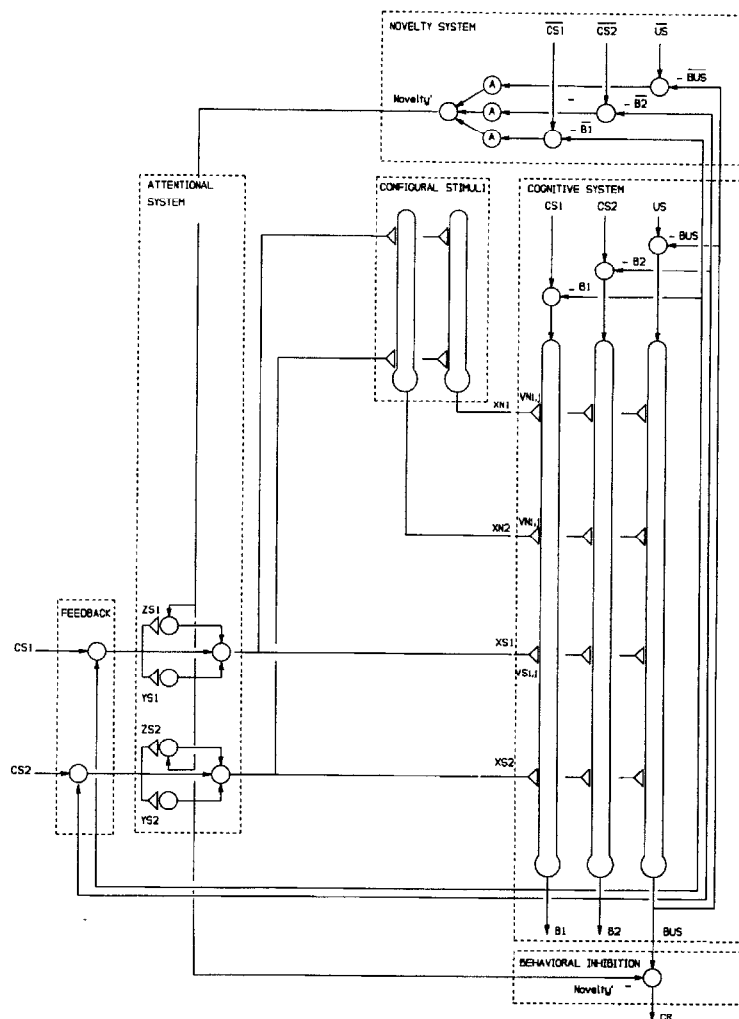
Four classes of S_D -R-US contingencies are possible: (1) positive reinforcement, in which R is followed by the presence of an appetitive US; (2) punishment, in which R is followed by the presence of an aversive US; (3) omission, in which R is followed by the absence of an appetitive US; and (4) negative reinforcement (escape and avoidance), in which R is followed by the absence of an aversive US. As in reinforcement learning (see REINFORCEMENT LEARNING), during operant conditioning, animals learn by trial and error from feedback that evaluates their behavior but does not indicate the correct behavior.

Positive Reinforcement

Operant conditioning can be obtained with free operant procedures, in which the operant response may occur repeatedly. Free operant paradigms are usually run in a Skinner box. In such a box, rats learn to press a bar (R) or pigeons learn to peck a key (R) to obtain food (US) from a dispenser when a light (S_D) is lit.

Dragoi and Staddon (1999) introduced a model that describes the major properties of free operant conditioning. According to their theory, (1) responses and stimuli become associated with re-

Figure 2. Diagram of a network that incorporates (a) an attentional system and (b) a configural system. In the attentional system, the internal representations of CS_i are modulated by the total environmental novelty, $Novelty'$, computed in the novelty system. In the configural system, the internal representations of simple stimuli XS_i become *configured* with the internal representations of other CSs in hidden units that represent configural stimuli CN_j . In the associative system, both CS_i and CN_j become associated with the other CSs and the US. The attentional system permits the description of latent inhibition, and the configural system permits the description of occasion setting. CS_i : conditioned stimulus; ZS_i , YS_i : attentional associations; XS_i : CS_i internal representation; CN_j : configural representation; VS_{ik} : XS_i - CS_k association; VS_{iUS} : XS_i -US association; VN_{jk} : CN_j - CS_k associations; US: unconditioned stimulus; B_k : CS_k aggregate prediction; B_{US} : US aggregate prediction; \overline{CS}_k : CS_k average observed value; \overline{B}_k : CS_k average predicted value; CR: conditioned response. Arrows represent fixed synapses. Solid circles represent variable synapses. (After Buhusi and Schmajuk, 1996).



inforcement, (2) response-reinforcement and stimulus-reinforcement associations are combined to generate learning expectancy, and (3) the operant response is controlled by the interaction between expected and experienced events. The model describes qualitative features of operant behavior such as discrimination learning, response selection, contingency effects, effects of reinforcement delay, matching in choice experiments, development of preference, contrast effects, resistance to extinction, spontaneous recovery, regression, serial-reversal learning, and overtraining reversal effect.

Negative Reinforcement

Operant conditioning can also be obtained with discrete trial procedures, in which the operant response occurs only once on a given trial. For instance, discrete-trial avoidance paradigms are usually run in a two-way shuttle box. The shuttle box is a chamber with two compartments separated by a barrier with a door. Each compartment has a metal grid floor that can deliver a shock (US). Lights above the chambers provide warning signals (WS) for the US. The experiment starts with both compartments being illuminated. At time zero, the light above the compartment where the animal is located turns off (WS), and the door separating both compartments opens. If the animal has not crossed to the opposite side after a given time (that may vary between 2 and 40 s), the shock US is

applied. If the animal has crossed to the opposite side before that time, it avoids the US, and the separating door closes behind it. After a constant or an average intertrial interval that varies from 15 s to 4 minutes, the whole sequence restarts.

Schmajuk, Urry, and Zanutto (1998) presented a real-time two-process theory of avoidance that combines elements of classical and operant conditioning. The network incorporates two processes: classical and operant conditioning. Whereas the classical conditioning process controls US-US, WS-US, and R-US associations, the operant conditioning process controls US- R_{escape} and WS- $R_{\text{avoidance}}$ associations. Whereas classical conditioning is regulated by a delta rule, $f(US) = US - \sum_j V_{jUS} X_j$, where X represents WS, R, or the US, operant conditioning is regulated by a novel algorithm that mirrors the classical conditioning algorithm, $f(US) = - (US - \sum_j V_{jUS} X_j)$. Schmajuk et al. (1998) applied the network to the description of escape and avoidance behavior in a shuttle box, running wheel, leg flexion, or lever-pressing paradigms as Sidman avoidance. Schmajuk et al. (1998) demonstrated through computer simulations that the model describes most of the features that characterize avoidance behavior.

Discussion

Theories and neural networks of conditioning can be evaluated at different levels. At the behavioral level, simulated behavioral re-

sults are compared with experimental data. At the computational level, simulated activity of the neural elements of the model are compared with the activity of single-neuron or neural population activity. At the anatomical level, interconnections among neural elements in the model are compared with neuroanatomical data. Finally, model and animal performances can be compared after brain lesions, induction and blockade of long-term potentiation, or administration of different psychopharmacological drugs.

Examples of these good matches include the Schmajuk et al. (1996) model, which describes a very large number of conditioning paradigms, many of the behavioral properties of latent inhibition, the activity of dopamine neurons in the nucleus accumbens as coding for the variable novelty, the effect of lesions of the hippocampus and different regions on the accumbens on latent inhibition, the effect of administration of amphetamine and haloperidol also on latent inhibition. Another interesting case is the Sutton and Barto (1990) model, which, when combined with some of the principles used in the Grossberg and Schmajuk (1998) model, describes the activity of dopamine neurons of the ventral tegmental area and substantia nigra in terms of the prediction errors for rewards (Schultz, Dayan, and Montague, 1997).

Good models are characterized by (1) a large percentage of simulation results that match experimental results and (2) a relatively small number of equation parameters. As Schmajuk (1997) explains, the ratio between (1) and (2) gives a measure of the overall quality of a model.

Road Maps: Neural Plasticity; Psychology

Related Reading: Cognitive Maps; Concept Learning; Embodied Cognition; Motivation

References

- Buhusi, C. V., and Schmajuk, N. A., 1996, Attention, configuration, and hippocampal function. *Hippocampus*, 6, 621–642. ♦
- Church, R. M., and Broadbent, H. A., 1991, A connectionist model of timing, in *Neural Network Models of Conditioning and Action* (M. Commons, S. Grossberg, and J. E. R. Staddon, Eds.), Hillsdale, NJ: Erlbaum, pp. 225–240.
- Dragoi, V., and Staddon, J. E. R., 1999, The dynamics of operant conditioning, *Psychol. Rev.*, 106:20–61.
- Gluck, M. A., and Myers, C. E., 1993, Hippocampal mediation of stimulus representation: A computational theory, *Hippocampus*, 3:491–516.
- Grossberg, S., 1975, A neural model of attention, reinforcement, and discrimination learning, *Int. Rev. Neurobiol.*, 18:263–327.
- Grossberg, S., and Schmajuk, N. A., 1989, Neural dynamics of adaptive timing and temporal discrimination during associative learning, *Neural Networks*, 2:79–102.
- Kehoe, E. J., 1988, A layered network model of associative learning: Learning to learn and configuration, *Psychol. Rev.*, 95:411–433.
- Mackintosh, N. J., 1975, A theory of attention: Variations in the associability of stimuli with reinforcement, *Psychol. Rev.*, 82:276–298.
- Pearce, J. M., and Hall, G., 1980, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli, *Psychol. Rev.*, 87:532–552.
- Rescorla, R. A., and Wagner, A. R., 1972, A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non-reinforcement, in *Classical Conditioning II: Theory and Research* (A. H. Black and W. F. Prokasy, Eds.), New York: Appleton-Century-Crofts.
- Schmajuk, N. A., 1997, *Animal Learning and Cognition: A Neural Network Approach*. New York: Cambridge University Press. ♦
- Schmajuk, N. A., and DiCarlo, J. J., 1992, Stimulus configuration, classical conditioning, and the hippocampus, *Psychol. Rev.*, 99:268–305.
- Schmajuk, N. A., Lam, Y. W., and Gray, J. A., 1996, Latent inhibition: A neural network approach, *J. Exp. Psychol. Anim. Behav. Process.*, 22: 321–349. ♦
- Schmajuk, N. A., Urry, D., and Zanutto, B. S., 1998, The frightening complexity of avoidance: An adaptive neural network, in *Models of Action: Mechanisms of Adaptive Behavior* (C. Wynne and J. E. R. Staddon, Eds.), Hillsdale, NJ: Erlbaum, pp. 201–238.
- Schultz, W., Dayan, P., and Montague, P. R., 1997, A neural substrate of prediction and reward, *Science*, 275:1593–1599.
- Sutton, R. S., and Barto, A. G., 1981, Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol. Rev.*, 88:135–170.
- Wagner, A. R., 1978, Expectancies and the priming of STM, in *Cognitive Processes in Animal Behavior* (S. H. Hulse, H. Fowler, and W. K. Honig, Eds.), Hillsdale, N.J.: Erlbaum, pp. 177–209.

Connectionist and Symbolic Representations

David S. Touretzky

Introduction

In symbolic representations, the heart of mathematics and most models of cognition, symbols are meaningless entities to which arbitrary significances may be assigned (Newell, 1980; Harnad, 1990). Composing ordered tuples from symbols and other tuples allows us to create an infinitude of complex structures from a finite set of tokens and combination rules. Inference in the symbolic framework is founded on structural comparison and rule-governed manipulation of these objects.

Many aspects of language and cognition appear rule-like but, on closer inspection, are not so amenable to axiomatization. A famous example is the production of past tense forms of English verbs (Rumelhart and McClelland, 1986; see PAST TENSE LEARNING). Regular verbs add /t/ or /d/ or /ed/ to derive their past tense phonetic form, depending on the final sound of the stem, e.g., “flipped” versus “fibbed” versus “fitted.” People generate regular past tenses of novel forms, such as “bork” to “borked,” as if they were “applying the rule.” But multiple classes of irregular verbs follow different conventions, e.g., sing/sang, bring/brought, or leave/left. People sometimes apply these patterns to novel forms, as in

“bling”/“blang,” or misapply them to stems that follow a different irregular convention (“bring” to “brang” instead of “brought”). There are stages in child language acquisition where the regular suffix is not only misapplied, as in “bring” to “bringed,” but sometimes combined with the output of an exception pattern, producing forms such as “broughted.” The interactions of regular and exception forms, and the developmental phenomena associated with past tense acquisition, are difficult to formalize. Thus, it seems unlikely that children are constructing explicit past tense *rules* in their heads.

The aim of the formal symbolic approach to understanding is to give a precise account of a domain in the language of that domain, e.g., a theory of grammar expressed in terms of morphemes, words, and phrases, or a theory of vision formulated in terms of pixels, regions, boundaries, etc. Axioms (rules) determine the structures that may be composed and the inferences that may be drawn from them. The approach has had great success in formalizing mathematics, but the hope that language, perception, or other aspects of cognition might succumb to similar treatment has faded.

It is important to note here that rule-based systems are not limited to deductively sound, consistent inferences. Artificial intelligence makes extensive use of nondeductive reasoning methods. For ex-

ample, case-based reasoning tries to match a problem description against a library of cases for which the solution is already known. Match scores are determined heuristically, in some instances simply by counting the number of shared features. The cases retrieved provide only a best guess at an answer, but this may be sufficient.

The search for sound deductive rules for cognitive domains having long been abandoned, the debate between symbolists and connectionists is over how much of the formalist enterprise should be retained. Symbolists have moved to more complex formalizations of cognitive processes, using heuristic and unsound inference rules. (No one claims humans are sound reasoners. Heuristics have proved very useful, and suitably constrained inference systems might never suffer the consequences of their unsoundness.) Connectionists explore a radical alternative: that cognitive processes are mere epiphenomena of a completely different type of underlying system, whose operations can never be adequately formalized in symbolic language (Smolensky, 1988; see PHILOSOPHICAL ISSUES IN BRAIN THEORY AND CONNECTIONISM).

That connectionist models are implemented on digital computers does not make them symbolic models. The past tense model can be described in terms of nodes and links and activation values, but there is no isomorphism between this description and one phrased in the domain language, i.e., in terms of verb stems and affixes. The relationships between elements at the two levels can only be hinted at. This is what makes the model nonsymbolic.

In summary, connectionism replaces classical discrete, set-theoretic semantics with continuous, statistical, vector-based semantics. In the following sections we examine representation and processing issues from a connectionist perspective.

Feature Vector Representations

If arbitrary symbols are replaced by feature vectors, the similarity of two concepts can be measured by their dot product. This is an attractive alternative to defining explicit axioms for determining, say, whether *chair* should be more similar to *table* than to *couch*. Similar concepts will naturally have similar vector representations, which facilitates generalization in neural networks.

Early models in this vein used hand-constructed feature vectors, often with familiar semantic features such as “human,” “animate,” “solid,” etc. But other types of representations were also used. Rumelhart and McClelland’s past tense model represented present tense verb forms with a triplet encoding. The phonemic sequence /stop/ would be encoded in triplet form as $\#s_t, s_{t_0}, \rho_p$, and $\rho_p\#$. The encoding used in Rumelhart and McClelland’s model was actually a finer-grained representation that used triplets of phonetic features instead of whole phonemes. Words were encoded as 460-element vectors, with each element corresponding to a possible triplet. A network trained on some past tense forms could thus generalize correctly to novel forms.

The advent of backpropagation learning led to the creation of feature vectors by the networks themselves. Hinton’s (1990, pp. 47–76) family tree model learned relationships about three generations of people, such as “Victoria is the mother of Colin,” “Victoria is the wife of James,” and “Penelope is the mother of Victoria.” The network took as input a pair such as “Victoria” and “mother,” and was trained to activate the output unit for “Colin.” A separate unit was used to represent each individual or relationship, but these units projected to small hidden layers that in turn projected to the central hidden layer. After training on 104 tuples defined over 24 persons and 12 relationships, feature vector representations developed in the dedicated person and relationship hidden layers that reflected the similarity structure of the domain. For example, one unit came to be strongly inhibited when coding for people in the third generation. The unit was weakly active for peo-

ple in the middle generation; and strongly active for grandparents, i.e., persons in the first generation.

Connectionist representations have been termed subsymbolic (Smolensky, 1988) because they capture graded, messy, but statistically important aspects of a domain. This idea has also been explored in Latent Semantic Analysis (Foltz, 1996), which derives feature vector representations for words in a large text corpus. First a matrix of occurrence information for each word in each of many contexts is constructed. A context can be a document, a paragraph, or even a sentence. The matrix is then reduced to a lower-dimensional form by singular value decomposition (SVD), similar to Principal Components Analysis. The result is a set of 100–300 dimensional feature vectors, one per word, which are useful for many kinds of text matching and retrieval applications. These features capture information about linguistic relations but do not look anything like the semantic features normally used in linguistic analysis. Once again, dot product can be used to compute similarity.

A problem with fixed feature vector representations is that what should count as similar is often context dependent. Another problem is that while feature vectors are a natural way to represent individual concepts, they are not so convenient for representing relationships among concepts; the latter may have to be generated on the fly during reasoning rather than constructed gradually through gradient descent learning.

Composite Structure

Symbolic models create trees of increasing size to encode structures with more components, but this route is not open to connectionist networks, which use fixed-length vectors (see COMPOSITIONALITY IN NEURAL SYSTEMS). Several solutions have been proposed. Encoder/decoder networks created using backpropagation can map composite objects into points in a vector space. Simple recurrent networks (SRNs; see Elman in Touretzky, 1991) represent sequences this way, and Recurrent Auto-Associative Memories (RAAMs; see Pollack in Hinton, 1990) use a similar approach to represent trees. The problem with this method is that the inductive bias of backpropagation learning does nothing to enforce systematicity in these representations, so novel inputs will not be treated correctly unless the model has been trained on a substantial fraction of the set of all possible structures to be encoded.

An alternative approach employs simple combinatorial operations on vectors to perform composition and extraction operations, thereby enforcing systematicity (see SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS). Plate’s (2000) holographic reduced representation is the most promising example. Holographic reduced representations encode frame-like objects where each role and each filler is a vector of several thousand bits. Roles are paired with fillers using a circular convolution operator, which produces a vector result. Multiple role/filler pairings are then combined by vector addition and normalization. *John kissed Mary* might be encoded as *agent* \otimes *john* + *patient* \otimes *mary* + *action* \otimes *kissed*. The vectors encoding composite objects can themselves serve as roles or fillers, providing a limited form of structural recursion. But the primitive vectors, e.g., for *john* and *mary*, must be nearly orthogonal to prevent interference between overly similar patterns, which precludes the use of a meaningful feature vector representation. On the other hand, dot product can still be used to do matching and retrieval, and even to measure analogical similarity (Eliasmith and Thagard, 2001).

Nonlinear Mapping

Certain types of representation support certain types of processing better than others. Most of the inference in connectionist networks takes place via nonlinear mapping of vectors, implemented by one

or more weight matrices. Properly constructed, such matrices can simulate the effects of discrete inference rules, combine rules with exceptions, express graded inferences, exploit statistical structure in the domain, and incorporate information from multiple evidence sources. Most important, these mappings can be constructed automatically using learning procedures.

A single layer of weights was sufficient to allow the Rumelhart and McClelland past tense model to capture both the regular past tense rules and various classes of exceptions. More powerful mappings can be achieved with additional layers of weights, and by allowing a network to feed back on itself. In SRNs (Elman in Touretzky, 1991), the hidden layer activity at time t is fed back as an additional input to the network at time $t + 1$ (see CONSTITUENCY AND RECURSION IN LANGUAGE). SRNs process information sequentially and can learn finite-state grammars. But Servan-Schreiber, Cleeremans, and McClelland (in Touretzky, 1991) showed that these networks do more than that: they learn and exploit the statistical structure of the training data. Servan-Schreiber et al. suggest that statistical differences between main and embedded clauses might help a language learner track long-range dependencies generated by recursive grammatical constructs.

The problems with the nonlinear mapping approach are that it requires lots of training data to adequately cover a small domain, and the “rules” are often learned imperfectly, so the network does not always generalize correctly. On the other hand, it provides a way to study complex phenomena without having to formulate rules explicitly. Plaut (1999) showed that lesioning a network that mapped orthographic to phonetic representations (and that also had attractor dynamics) could reproduce many interesting effects observed in word reading and acquired dyslexias (see LESIONED NETWORKS AS MODELS OF NEUROPSYCHOLOGICAL DEFICITS). Some of these effects, such as characteristic mixtures of error types, have proved difficult to explain with symbolic models.

Rohde’s connectionist sentence comprehension and production model (CSCP) shows that nonlinear maps can challenge symbolic language models in nontrivial domains (Rohde, 2002). CSCP is a collection of SRNs that learns a substantial fragment of English grammar, including many types of embedded and relative clauses. The model has a 300-word vocabulary and contains both parsing and sentence production components. It required 2 months on a fast workstation to train. Although its performance is not perfect, its coverage is impressive.

Parallel Constraint Satisfaction

Rule-based systems have trouble determining the best interpretation of an ambiguous stimulus because of the difficulty of formulating explicit rules that weigh multiple sources of evidence in a flexible manner. Connectionist networks approach this as a constraint satisfaction problem. Evidence sources impose weak constraints on the stimulus interpretation, and the activity of the network evolves to reflect the collective effect of all these influences.

One of the simplest constraint satisfaction architectures is the interactive activation network, in which constraints are expressed as weighted excitatory or inhibitory links between nodes standing for concepts. Semantically related units have mutually excitatory connections, while units that code for alternative hypotheses inhibit each other. Consider the treatment of *The astronomer married the star* in Waltz and Pollack’s (1985) interactive parsing model. There are two word-meaning nodes for *star*, one being “heavenly body” and the other “movie star.” There is an excitatory link from astronomer to the heavenly body node because the two are semantically related. Hence, the model predicts that this meaning of star will be primed by “astronomer,” and will initially be favored over “movie star.” But when *married* is encountered, it imposes a constraint that the subject and object must be human. This ultimately

leads to suppression of activity in the heavenly body node, causing movie star to become active.

Neural net constraint satisfaction systems, unlike their symbolic counterparts, are attractor networks (see COMPUTING WITH ATTRACTORS). We can draw on dynamical systems theory to understand their behavior. Most of these networks use “localist” or symbolic rather than “distributed” or vector representations (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS), but their inference mechanism is still nonaxiomatic.

More sophisticated constraint satisfaction can be done with Boltzmann machines, which employ hidden units to express nonlinear interactions among constraints, and simulated annealing search to find the best interpretation of the input. An example of spreading activation and simulated annealing using localist representations is Hofstadter and Mitchell’s analogy-making program, CopyCat (Mitchell, 1993). One of the few annealing reasoners developed to date that employs feature vector representations is microKLONE (Derthick in Hinton, 1990), which translated a frame-based semantic network language into a complex structure of nodes and links expressing semantic constraints among slot fillers. microKLONE was able to use these constraints not only to fill in missing information, but also to produce plausible inferences in counterfactual situations.

Discussion

Connectionist representations promise a continuous, statistical, vector-based alternative to rule-based reasoning over discrete symbol structures. Although too abstract to map directly onto neural circuitry, the new conceptual framework has strongly influenced theorizing about the brain.

As yet we have only glimmerings of how connectionist models might surpass the abilities of symbolic reasoners. One suggestion of how they might do this is Hinton’s (1990, pp. 47–76) notion of a “reduced description,” in which the encoding of a concept like “room” contains abbreviated but directly accessible information about affiliated concepts such as “window,” “door,” “doorknob,” and “keyhole.” A reasoner could thus immediately recognize how a room could have a keyhole, rather than having to follow a chain of pointers to reach that concept via “door” and “doorknob.” But we have not yet seen a successful implementation of reduced descriptions, much less a mechanism to automatically tailor such descriptions to the requirements of a domain.

The elements discussed in this article—feature vector representations, composite structure, nonlinear maps, and parallel constraint satisfaction—do not yet work well together. Multiple conceptual breakthroughs are likely required before they can be incorporated into a unified connectionist theory of symbol processing.

Road Map: Artificial Intelligence

Related Reading: Artificial Intelligence and Neural Networks; Compositionality in Neural Systems; Hybrid Connectionist/Symbolic Systems; Systematicity of Generalizations in Connectionist Networks

References

- Eliasmith, C., and Thagard, P., 2001, Integrating structure and meaning: A distributed model of analogical mapping, *Cognit. Sci.*, 25:245–286.
- Foltz, P. W., 1996, Latent Semantic Analysis for text-based research, *Behav. Res. Meth. Instr. Comput.*, 28:197–202.
- Harnad, S., 1990, The symbol grounding problem, *Physica D*, 42:335–346.
- Hinton, G. E., Ed., 1990, *Connectionist Symbol Processing*, *Artif. Intell.*, 46 (special issue). ♦
- Mitchell, M., 1993, *Analogy-Making as Perception*, Cambridge, MA: MIT Press.
- Newell, A., 1980, Physical symbol systems, *Cognit. Sci.*, 4:135–183.

- Plate, T., 2000, Analogy retrieval and processing with distributed vector representations, *Expert Syst. Int. J. Knowledge Engn. Neural Netw.*, 17:29–40. ♦
- Plaut, D. C., 1999, Computational modeling of word reading, acquired dyslexia, and remediation, in *Converging Methods in Reading and Dyslexia* (R. Klein and P. A. McMullen, Eds.), Cambridge, MA: MIT Press, pp. 339–397.
- Rohde, D. R., 2002, A connectionist model of sentence comprehension and production, Ph.D. diss., Carnegie Mellon University. Available: <http://www.cs.cmu.edu/~dr/Thesis>.
- Rumelhart, D. E., and McClelland, J. L., 1986, On learning the past tense of English verbs, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (J. L. McClelland and D. E. Rumelhart, Eds.), Cambridge, MA: MIT Press, vol. 2, pp. 216–217.
- Smolensky, P., 1988, On the proper treatment of connectionism, *Behav. Brain Sci.*, 11:1–74. ♦
- Touretzky, D., Ed., 1991, *Connectionist Approaches to Language Learning, Machine Learn.*, 7(2–3):105–252 (special issue).
- Waltz, D. L., and Pollack, J. B., 1985, Massively parallel parsing: A strongly interactive model of natural language interpretation, *Cognit. Sci.*, 9:51–74.

Consciousness, Neural Models of

John G. Taylor

Introduction

The construction of neural theories of consciousness faces the difficulty that, by its very nature, consciousness is subtle, and its defining characteristics are still poorly discerned. This leads to uncertainty about what properties create consciousness in the brain and even regarding the location of the sites of consciousness creation, the so-called neural correlates of consciousness (NCC). In order to prevent the NCC from wandering all over the brain, whether in the primary sensory or unimodal associative cortices (Pollen, referred to in Taylor, 2001), or now in the prefrontal cortex (Crick and Koch, 1998), care must be taken in assessing evidence in support of various possible sites of the NCC. New experimental results from brain imaging, as well as further insights into single cell activity and the effects of brain deficits, are now leading to a clearer picture of the NCC (Taylor, 1999, 2001a). This is an important advance, since concentration can now be turned to the nature of the neural representations crucially involved in consciousness. Neural network ideas and explicit models are helping to pin down what these representations consist of and how they function. Thus, progress is slowly being made to understand scientifically this most mysterious of human phenomena: the race for consciousness has started in earnest (Taylor, 1999).

Yet consciousness is not easy to define precisely. In order to make initial progress without spending time in logic chopping, let us follow the article on consciousness in the previous edition of the *Handbook* (Velmans, 1995). The definition given there is initially appropriate for the present purposes:

Consciousness is synonymous with awareness or conscious awareness (sometimes phenomenal awareness). The contents of consciousness encompass all that we are conscious of, aware of or experience.

This distinguishes consciousness from other mental activity; there is much in the mind of which the possessor is not conscious, so that consciousness is a special faculty of the mind. The nature of its function has been vigorously debated (Velmans, 1995), although with no justification of either extreme: consciousness has no purpose, but is solely an epiphenomenon, or alternatively it is the ultimate control system of the mind. Comments supporting the latter view will be given later in this article, as well as further discussion of the validity of the above definition.

If consciousness is difficult to define and its function is hard to discern, at least some features involved with its creation are more generally agreed upon. Characteristics needed for consciousness are as follows:

1. *Temporal duration.* Neural activity is needed to be present for at least 200 ms for awareness to arise. This is supported by the data of Libet and colleagues (1964) and by the duration of activity in buffer working memory sites, suggested by a number of workers as being the sites of creation of consciousness (Taylor, 1999).
2. *Attentional focus.* It is widely supposed that consciousness can only arise of an object at the focus of attention; there is experimental support for this position (Mack and Rock in Wright, 1997). The resulting amplification of the attended input allows it to attain consciousness and also be laid down in long-term memory, for later conscious access.
3. *Binding.* The binding of features, analyzed in separate areas of cortex, to allow recognition and experience of objects, has long been regarded as crucial for the explanation of consciousness. Such binding has been proposed to occur by several means, especially by simultaneous activations, such as by coupled oscillatory modules (Crick and Koch, 1990) or by attentive competitive/amplificatory processing (Triesman in Wright, 1997). There is evidence for both of these being involved in cortical processing.
4. *Bodily inputs.* The availability of such inputs is needed to give a perspective to inputs, following detailed investigations by psychologists (Bermudez, Marcel, and Eilan, 1995).
5. *Saliency.* Coded in the limbic system, saliency is needed to give a suitable level of importance to a given input, and can arise, for example, from the cingulate cortex, as involving motivational activation, or from other limbic sites.
6. *Past experience.* This gives content to conscious experience, by reactivating previous relevant experience (Taylor, 1999); such memories are especially involved in defining the self.
7. *Inner perspective:* Besides having awareness of an input, we each possess an inner perspective, without which we cannot experience the mental world as belonging to us. The inner perspective is thus that of ownership of our awareness. Without it we would be zombies, having content but no sense of “what it is like to be me” (Nagel, 1974).

Only by a concerted attack on the underlying brain structures involved in the creation of consciousness, and their detailed modeling, can real progress be made. Recent advances on the three counts—the “where,” the “what,” and the “how” of consciousness—will be described in this article. Several recent experimental results from brain imaging, single-cell data, and brain deficits will be described in the next section. These help in determining a site for the NCC. From these data, the inferior parietal lobes are concluded to be the most appropriate region for the site of conscious-

ness (Taylor, 2001a). In the following section the notion of the central representation will be developed; it contains the crucial contents for consciousness. A general model of attention control, the CODAM model, which incorporates the central representation and is appropriate for simulating the associated emergence of consciousness, is then presented in the next section. Relations to other models are described briefly in the penultimate section, and a short conclusion and discussion of open questions finishes the review.

Experimental Data

The data to be considered are of a variety of sorts: from single cells, lesion effects on behavior, and from brain imaging (especially that using PET and fMRI). The single-cell data involves information on the presence or absence of significant activity observed in an animal under anesthesia. This has little effect on the level of early sensory cortical responses, for example, in V1 and MT or in inferotemporal cortex. We will later consider the important effects of attention on such activity, although this does not change the overall story. There is considerable effect of anesthesia on parietal lobe single-cell responses, which led to great difficulty in measuring from such cells before the advent of the ability to record from awake, behaving monkeys. Thus, occipital and temporal cortices are not sufficient for consciousness, while parietal may be.

As a start to considering lesion effects, we note the singular lack of loss of consciousness due to frontal deficits brought about either by disease or injury, as numerous reports show. For example, there is the famous case of Phineas Gage (described in Taylor, 1999), who had a tamping iron blown through his frontal lobes with considerable loss of frontal cortex but without successive loss of consciousness as he was carried to the local doctor. There is also the case of the young man who was born bereft of most of his frontal lobes (mentioned in Taylor, 2001), but yet, apart from great social problems, lived a normal conscious existence.

One area of brain lesions of particular relevance to the NCC is that of neglect (which can be either of visual inputs or control of actions). Patients usually suffer a loss of the right parietal lobe, and subsequently lose awareness of input from their left hemifield. This loss, for example, is observed in the inability to cross out lines on the left of their field of view. It is now agreed that neglect arises specifically from damage to the inferior parietal lobe (Milner, 1997).

Much is also being discovered by brain imaging about the siting in cortex of buffer working memories. Those for spatial vision are in the right, those for language and temporal estimation in the left, inferior parietal lobes. Extinction, involving loss of awareness of the right-hand object of two similar objects, one on the left, one on the right, is sited separately in the superior parietal lobe (Milner, 1997).

Recent fMRI data indicates that the experience of the motion aftereffect (MAE) most occurs strongly in BA 40 (the supramarginal gyrus), in the inferior parietal lobe. The experimental paradigm to observe this uses motion adaptation to a set of horizontal bars moving vertically downward for 30 s, and then stopping. Subjects exposed to such a display usually experience the MAE for 9 or so s after cessation of the movement of the bars.

Whole-head fMRI measurements during this paradigm (Taylor et al., 2000) showed a network of connected areas, as in Figure 1. This network has a posterior group, involving especially the motion area MT, which was found to be responsive to all forms of motion as well as the MAE. On the other hand, there is a set of anterior modules, shown in Figure 1, which are particularly active both during the MAE period and just after the cessation of oscillatory movement of the bars both up and down (after which there is no MAE experience reported). Finally, there are inferior parietal re-

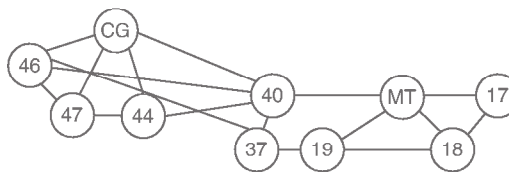


Figure 1. The network of areas active in the brain during the MAE experiment (Schmitz et al., 1998, referred to in Taylor, 2001). The lines joining the various modules denote those for which the correlation coefficient is at least 0.4. MT denotes the middle temporal area, and the other areas are numbered according to Brodmann's numeration.

gions, which demonstrate activity almost solely in response to the MAE period.

All of the inputs required to support the features described in the first section are available, it would seem almost uniquely, for the inferior parietal lobe. Thus, the buffer working memory sites for space and time (and language) have been noted previously as being there, as is a competition for consciousness associated with these sites (Taylor, 1999). Bodily inputs are also available there, as noted from effects of manipulation of the body in neglect, as well as from known neuroanatomy (connections with the vestibular apparatus and the cerebellum are well known). The limbic system is also well connected to the inferior parietal lobe and so is episodic memory.

We conclude that the inferior parietal lobe (IPL) is suitably connected and structured to satisfy all of the criteria A–F in the first section. We will discuss its relevance to criterion G when we consider the CODAM model of a Heution control.

The Central Representation

Evidence from neglect studies and brain imaging on healthy subjects has been presented in this article to implicate the IPL as playing a crucial role in controlling attention and creating awareness. Attention can occur in a range of possible frames of reference: neglect can be observed tied to an object, or to a trunk-centered frame of reference or a variety of other reference frames (Milner, 1997). This implies that the IPL is composed of a set of modules carrying information from the environment as well as modulation by possible body input. Thus, the IPL is eminently suited to carry what is termed the “Central Representation,” defined as: “The combined set of multi-modal activations involved in fusing sensory activity, body positions, salience and intentionality for future planning; it involves a competitive process between the various working memory modules it contains to single out one to be conscious and be used for report to other working memory sites for further planning or action” (Taylor, 2001a).

There are several important features of the central representation (CR) that need discussion, in relation to the criteria in the introduction to the article:

1. The CR must have access to sensory input, such as in vision, coded at a high level. Thus, it must have good access to temporal lobe representations, so as to use the categorization built there to guide action.
2. It also must have access to the bodily input needed to guide actions in terms of the intentionality coded in the superior parietal lobe. Such intentionality is coded for various sorts of actions: of the limbs, eyes, head, or fingers. This intentionality must be furnished with the parameters of the objects on which the actions must be taken; thus, cerebellar and vestibular input must also be accessible to the central representation, as it is in the parietal lobes. A neural model of this intentionality has been presented in Fagg and Arbib (1998).

3. Saliency of the inputs in the sensory field is an important attribute for the guidance of actions; that arises from limbic input already activated to provide saliencies of inputs from the orbitofrontal cortex by way of the cingulate. This is compounded by activations in the posterior cingulate gyrus, encoded as parts of episodic memory.
4. Several modules in the CR are involved in the IPL; the total activity must undergo an overall competition, possibly aided by thalamonucleus reticularis processing. A simulation of such a model has been given earlier (described in Taylor, 1999). The existence of such competition is supported by attention deficits observed in subjects with pulvinar lesions.
5. Siting the emergence of awareness in the IPL, as the result of the competition ongoing there, is supported by simulation of the data of Libet and colleagues (1964). The original experiment involved the creation of sensory experience (that of a gentle touch on the back of the patient's hand) by direct stimulation of cortex in patients being operated on for movement problems. The simulation (described in Taylor, 1999) used a simplified model of the corticothalamo-nucleus reticularis circuit, and led to the observed dependence of the delay of awareness on the strength of the threshold current for experiencing the touch on the back of the patient's hand.
6. Such a competition has also been suggested (described in Taylor, 1999) as occurring to explain experimental results of subliminal effects on lexical decision response times obtained by Marcel (1980). The experiment involved measurement of the reaction times of subjects to deciding if the first or third of three letter strings were words or not. Subliminal exposure to priming words occurred for the second letter string under one condition, with the presentation of polysemous words such as "palm," on which the lexical decision had to be made to the third word. The prior exposure caused the decision to be speeded up or delayed in characteristic ways according to the semantic relations of the three words to each other; the simulation was able to explain these results by means of a competition assumed to occur on the phonological store, aided and abetted by activations from a semantic memory store.

In conclusion, we site the CR in the IPL as the site for attention and short-term memory processing, with confluence there of information on saliency, episodic memory, high level coding of inputs, and information on body state.

An Attention Control Model (CODAM) of the Emergence of Awareness

Attention, the respectable face of consciousness for neuroscience, has received surprisingly short shrift recently from those directly

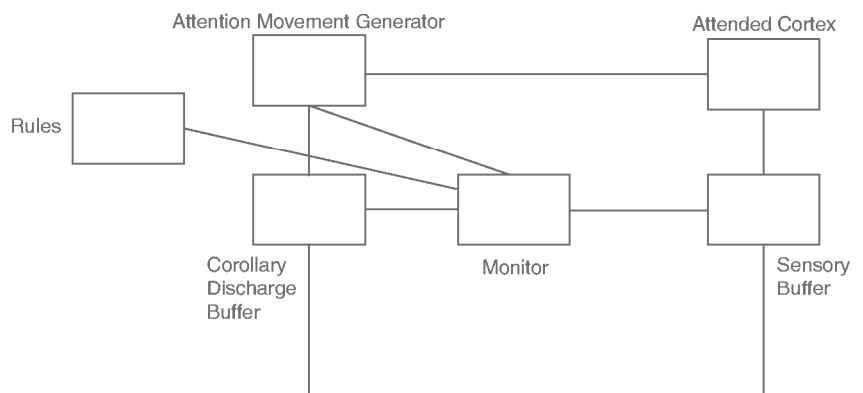
attacking the problem of consciousness. This is in spite of the phenomenon of inattention blindness, that there is no perception without attention, as noted by Mack and Rock (in Wright, 1997), and the many claims by those working on attention that there is no consciousness of an unattended input. There is now improved understanding of attention arrived at by brain imaging using a range of psychophysical paradigms and ever more careful single-cell experiments. A control view of attention is now accepted, in which signals from outside early cortex modulate inputs so as to allow selection of a desired target input from a set of distracters. We will now develop this view, describing how the Central Representation supports the creation of consciousness by means of a suitable control model.

In such a control model, there are the "plant" components of primary and secondary sensory and motor cortices containing input activations being attended to, an inverse controller in parietal/frontal as source of the attention signal, and a rules module containing the desired state into which attended input should be transformed (there being a prefrontal top-down rules module, and a bottom-up component in superior colliculus). Further components of a so-called forward model (or observer) can also be tentatively identified. This resides as buffer sites in inferior parietal and updating areas in prefrontal, as well as parts involved in monitoring effectiveness of response under the guidance of the rules modules, sited in cingulate cortex. This model (shown in Figure 2) has been simulated for a variety of paradigms (Taylor and Rogers, 2001), being an engineering control formalization of many neural models of attention.

To bring consciousness to the foreground inside the control model of attention, we turn to recent developments in phenomenology to guide us: there are two components of consciousness: "consciousness of" and the pre-reflective self. The pre-reflective self is experienced as the ownership of one's conscious experience and as the basis of all awareness; without it there would be content but no owner of that content.

To incorporate this important component in the attention control model, a very conjectural step was made in Taylor (2000): it was proposed that an observer is present that contains a buffered copy of the controller signal (more properly called a *corollary discharge*, since it is coded in the same manner as attended sensory activation on its buffer). This copy is used to achieve more rapid updating of the movement control signal. Such a copy will not be bound in any attention-based manner to the content of consciousness, since that can only be present on feedback from the plant. The corollary discharge signal will therefore not have any content. It can, however, be identified with the experience of "ownership," that of the about-to-appear amplified input that is being attended to. Such a signal can grant immunity to error through misidentification of the first-

Figure 2. A simple attention movement control model (shown in bold lines), composed of attended cortex (containing activity representing an attended input), an attention movement generator, a rules module (for either top-down or bottom-up control of attention), and a monitor (based on the error between the required attention state and that occurring as determined by a sensory buffer). The additional module and connections (in dotted lines) completes the CODAM model. It involves a buffer to hold the corollary discharge of attention movement. This corollary discharge signal is employed to speed up movement of attention as well as prevent incorrect updating of the sensory buffer until attention has been moved to the correct place (as assessed by the corollary discharge buffer acting as an observer or attention state estimator).



person pronoun. This would follow if the corollary discharge acts only to let onto the buffer what it has been told to by the inverse attention controller. As such, it inhibits all other possible entrants to contentful consciousness. This occurs for the brief period before the attentionally amplified input from sensory cortex arrives. The corollary discharge is then supposed to be inhibited in its turn. Such complex processing is supported by the siting of much of the attention control structures nearby in the parietal lobe, singled out recently as crucial for consciousness to arise. A simple model for such usage of the corollary discharge is shown in Figure 2, involving an additional observe component (shown in dotted lines), beyond the control model used in Taylor and Rogers (2001).

Following many experimental results, we site the attention controller in superior parietal lobe, the prefrontal cortex or superior colliculus, the attended plant in sensory cortices, working memory buffers in inferior parietal lobe, the monitor in the cingulate, and the goals module in the prefrontal cortex.

This approach results in the corollary discharge of attention movement (CODAM) model of consciousness (Taylor, 2000, 2001b):

Experience of the pre-reflective self is identified with the corollary discharge of the attention movement control signal residing briefly in its buffer until the arrival of the associated attended input activation at its own buffer.

The CODAM model achieves the required unpacking of the Central Representation: its essential constituents are (1) that of a signal of ownership in the corollary discharge buffer (with no content), (2) contents of external inputs represented by activity subsequently stored on the attended input buffer. There is a strict temporal order of activation, with the former briefly activated first, followed by the attended input buffer. More extended duration of the corollary discharge signal is suggested as the source of the meditative states of samadhi or nirvana: this is attention observing itself (Taylor, 2002). The presence of the pre-reflective self implies that the definition of consciousness given in the introductory section is incomplete; it needs extension by addition of the sentence "Those contents include awareness of consciousness itself, as occurs in the pre-reflective self."

The above model is only presented at the "arrows and boxes" level. Detailed neural implementations of various components are possible:

1. *Temporal duration*: achieved by the neural field recurrence—the bubble model, in which bubbles are temporally extended but spatially localised regions of neural activity with a certain degree of independence from inputs (Taylor, 1999).
 2. *Attention focus*: numerous neural models of attention processing in terms of object recognition have been created, such as the feedback amplification system of Mozer and Sitton (referred to in Taylor and Rogers, 2002), and see references therein. A general control framework encompassing such models has been suggested (Taylor, 2000, 2001b). A specific simulation, using competitive processing on the IMC is given in Taylor and Rogers (2001). Global competition across the CR is achieved by the NRT (with simulation described in Taylor, 1999).
 3. *Binding*: achieved by attention processing in the models noted above, to which synchronized activity can be added.
 4. *Bodily inputs*: the effect of these on attention processing in the parietal lobe has been modeled by modulations using various neural methods.
 5. *Salience*: amygdala encoding salience of inputs has been used in models of frontal set shifting (discussed in Taylor, 1999).
 6. *Episodic memory*: various hippocampal models have been created, although the manner in which episodic memory is built in cortex is still unclear (see discussion in Taylor, 1999).
- There are numerous important questions to be answered before the CODAM model can be accepted as a source of inner experience:
- Would it pass the Turing test if it were included in a full neural simulation?
 - How does the control aspect of attention, regarded as the highest control system in the brain, achieve learning of effective motor responses so as to achieve automaticity?
 - How is language (and thought) built from this overall framework?
- It is hoped that answers to these questions will be forthcoming by subsequent work.

Other Neural Models of Consciousness

Numerous models have been presented in the past to explain consciousness; only neural models will be considered here, and those only briefly.

1. Gray's Hippocampal Predictor model (see Gray et al. in Freeman and Taylor, 1997). This suggests that the hippocampus enables predictions to be made of future experiences, so creating consciousness. Various amnesic subjects, without hippocampus, however, still respond in a conscious manner in conversation, in spite of severe long-term memory deficits.
2. Aleksander's MAGNUS (see Browne et al. in Freeman and Taylor, 1997). This is based on a set of attractors, built by learning in a recurrent net in RAM-based hardware. Activation of an attractor is claimed to be the "artificial consciousness" of the system of the related input. This model has the defect that it involves no "internal experience" associated with attractor activity; it is a useful, if broad, model of the contents of consciousness.
3. Shallice's SAS (see reference in Taylor, 1999). This assumes total control of neural activity by the frontal "supervisory attention system." However, evidence against the NCC being sited in the frontal lobes was given in the section on experimental data. Better consistency with the CODAM model of the previous section can be achieved by extending the supervisory attention system to include parietal sites.
4. Baar's Global Workspace (see Newman et al. in Freeman and Taylor, 1997). This regards consciousness as the gaining of access to a "global workspace" (GW), considered as being on layer 1 of cortex. The GW is an incompletely defined concept, especially since it also does not necessarily have any internal experience. It can be related to the CODAM model by regarding access to the GW as the important process for determining the pre-reflective self. The CODAM model indicates how this access is controlled by suitable buffer activity, as well as determining the temporal nature of activity on ensuing access, when content enters consciousness.
5. Crick and Koch's 40 Hz (1990). That gamma-band oscillations are important for binding and segmenting is now well established (both experimentally in the brain and by computation). However, since such oscillations are observed in anesthetized animals, such activity cannot be regarded as sufficient for consciousness.
6. Pollen's Early model (see reference in Taylor, 2001). This supposes consciousness arises from feedback and relaxation to a fixed point of an attractor dynamics. Yet there is such re-entrance at many levels in the brain, such as between LGN and V1. The activity in LGN is not in consciousness, so indicating that the existence of such feedback is not sufficient for consciousness to be created.
7. Zeki's "local homunculus" (see references in Taylor, 2001a). This proposes that micro-consciousnesses arise in numerous

early cortical areas. However not only does this proliferation of homunculi add to the difficulties of consciousness but is also contradicted by the results presented in the section on experimental data.

8. Edelman's Reentrant Theory (see reference in Taylor, 1999). This is based on the special use of reentrant circuits, and so suffers from the difficulties of item 6 of this list.
9. Roll's Higher Order Theory (HOT; expanded in Roll's article in Freeman and Taylor, 1997). This uses language to enable higher order thoughts of lower order experienced inputs. The original HOT theory faces the difficulty that it rules out non-reflective consciousness in animals other than humans. Moreover the perspectival nature of experience cannot be incorporated into the HOT model in any obvious manner.
10. Harth's Inner Sketchpad Model (see Harth in Freeman and Taylor, 1997). This is based on the use of the reentry of a global scalar quantity, the degree of overlap between the input and activation, to achieve hill-climbing or attractor relaxation; it has the same defect as that of item 6 of this list, being involved in much dynamical processing in the brain but not just that specifically producing consciousness.
11. The Competitive Relational Mind model (Taylor, 1999). That a competitive process occurs in and between working memory sites for the emergence of consciousness is now becoming clear, but is still insufficient for the production of any perspectival account of awareness at the level of the pre-reflective self (Taylor, 2001b). Further circuitry is needed to build on this approach, as described in the section on the CODAM model.

In spite of the defects pointed out in the above models, they all involve important components in overall brain processing. The CODAM model introduced earlier is to be regarded as an extra component that benefits from the various computational features emphasised in the models of the previous section.

Discussion

Our main conclusions are:

1. The IPL is the essential site in the brain for consciousness.
2. The central representation, based there, gives conscious content (through activity in attended sites bound to the central representation by various mechanisms).
3. A simple control model of the movement of attention, supported by experimental data, can be extended to a mechanism for the creation of consciousness through the CODAM model. This led to a computational understanding of the minimal or pre-reflective self.

The broad range of the review touched on many areas, all of which require considerable further work. If the basic CODAM prin-

ciples are correct—as is so far supported by a range of material (Taylor, 2001b)—it provides a first viable neural approach to understanding human consciousness. It could even lead to the creation of a viable blueprint for machine consciousness. The greatest challenge facing the computational neuroscience community is to create simulations that can properly test the ideas presented here. At the same time, the challenge to neuroscience is to develop experimental proof or disproof of the CODAM model.

Road Map: Psychology

Related Reading: Action Monitoring and Forward Control of Movements; Cognitive Maps; Embodied Cognition; Emotional Circuits; Language Evolution: The Mirror System Hypothesis

References

- Bermudez, J. L., Marcel, A. J., and Eilan, N., 1995, *The Body and the Self*, Cambridge, MA: MIT Press.
- Crick, F. H. C., and Koch, C., 1990, Towards a neurobiological theory of consciousness, *Sem. Neurosci.*, 2:263–275.
- Crick, F. H. C., and Koch, C., 1998, Consciousness and neuroscience, *Cerebral Cortex*, 8:97–107. ♦
- Fagg, A. H., and Arbib, M. A., 1998, Modeling parietal-premotor interactions in primate control of grasping, *Neural Networks*, 11:1277–1304. ♦
- Freeman, W., and Taylor, J. G., 1997, Neural Networks for Consciousness, Special Issue, *Neural Networks*, 10(7).
- Libet, B., Alberts, W. W., Wright, E. W., DeLattre, L. D., Levin, G., and Feinsein, B., 1964, Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex, *J Neurophysiol.*, 27:546–578
- Marcel, A. J., 1980, Conscious and preconscious recognition on polysemous words: locating the selective effects of prior verbal contexts, in *Attention and Performance VIII* (R. S. Nickerson, Ed.), Hillsdale NJ: Lawrence Erlbaum.
- Milner, A. D., 1997, Neglect, extinction and the cortical streams of visual processing, in *Parietal Lobe: Contributions to Orientation in 3D Space* (P. Thier and H.-O. Karnath, Eds.), Heidelberg: Springer, pp. 3–22. ♦
- Nagel, T., 1974, What is it like to be a bat? *Philos. Rev.*, 83:434–450. ♦
- Taylor, J. G., 1999, *The Race for Consciousness*, Cambridge, MA: MIT Press.
- Taylor, J. G., 2000, Attentional movement: The control basis for consciousness, *Neurosci. Abst.*, 30:2231. Abstract No. 839.3.
- Taylor, J. G., 2001, The central role of the parietal lobes for consciousness, *Consc. Cognit.*, 10:379–417; 421–424.
- Taylor, J. G., 2002, From matter to mind, *J. Consc. Stud.*, 9:3–22.
- Taylor, J. G., and Rogers, M., 2002, A control model for the movement of attention, *Neural Networks*, 15:309–326.
- Taylor, J. G., Schmitz, N., Ziemons, K., Gross-Ruyken, M.-L., Mueller-Gaertner, H.-W., and Shah, N.-J., 2000, The network of areas involved in the motion after-effect, *NeuroImage*, 11:257–270. ♦
- Velmans, M., 1995, Consciousness, theories of, in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, Ed.), Cambridge, MA: MIT Press, pp. 247–250.
- Wright, R., (Ed.), 1997, Visual attention. Oxford: Oxford University Press.

Constituency and Recursion in Language

Morten H. Christiansen and Nick Chater

Introduction

Upon reflection, most people would agree that the words in a sentence are not merely arranged like beads on a string. Rather, the words group together to form coherent building blocks within a sentence. Consider the sentence, *The girl liked a boy*. Intuitively, the chunks *the girl* and *liked a boy* constitute the basic components

of this sentence (compared to a simple listing of the individual words or alternative groupings, such as *the girl liked* and *a boy*). Linguistically, these chunks comprise the two major *constituents* of a sentence: a subject noun phrase (NP), *the girl*, and a verb phrase (VP), *liked a boy*. Such *phrasal* constituents may contain two types of syntactic elements: other phrasal constituents (e.g.,

the NP *a boy* in the above VP) or *lexical* constituents (e.g., the determiner *the* and the noun *girl* in the NP *the girl*). Both types of constituent are typically defined *distributionally* using the so-called replacement test: If a novel word or phrase has the same distribution as a word or phrase of a known constituent type—that is if the former can be *replaced* by the latter—then they are the same type of constituent. Thus, the lexical constituents *the* and *a* both belong to the lexical category of determiners because they occur in similar contexts and therefore can replace each other (e.g., *A girl liked the boy*). Likewise, *the girl* and *a boy* belong to the same phrasal category, NP, because they can be swapped around, as in *A boy liked the girl* (note, however, that there may be semantic constraints on constituent replacements. For example, replacing the animate subject NP *the girl* with the inanimate NP *the chair* yields the semantically anomalous sentence, *The chair liked a boy*).

In linguistics, grammar rules and/or principles determine how constituents can be put together to form sentences. For instance, we can use the following phrase structure rules to describe the relationship between the constituents in the example sentences above:

$$\begin{aligned} S &\rightarrow \text{NP VP} \\ \text{NP} &\rightarrow (\text{det}) \text{N} \\ \text{VP} &\rightarrow \text{V (NP)} \end{aligned}$$

Using these rules we obtain the following relationships between the lexical and phrasal constituents:

$$[s_{\text{NP}[\text{det} \text{ The }]_{\text{N} \text{ girl }]}]_{\text{VP}[\text{V} \text{ liked }]_{\text{NP} [\text{det} \text{ a }]_{\text{N} \text{ boy }]}]}$$

To capture the full generativity of human language, *recursion* needs to be introduced into the grammar. We can incorporate recursion into the above rule set by introducing a new rule that adds a potential prepositional phrase (PP) to the NP:

$$\begin{aligned} \text{NP} &\rightarrow (\text{det})\text{N}(\text{PP}) \\ \text{PP} &\rightarrow \text{prep NP} \end{aligned}$$

These rules are recursive because the expansion of the right-hand sides of each can involve a call to the other. For example, the complex NP *the flowers in the vase* has the simple NP *the vase* recursively embedded within it. This process can be applied arbitrarily often, creating, for instance, the complex NP with three embedded NPs:

$$\begin{aligned} &[_{\text{NP}} \text{ the flowers } [_{\text{PP}} \text{ in } [_{\text{NP}} \text{ the vase } \\ &\quad [_{\text{PP}} \text{ on } [_{\text{NP}} \text{ the table } \\ &\quad \quad [_{\text{PP}} \text{ by } [_{\text{NP}} \text{ the window }]]]]]]]] \end{aligned}$$

Recursive rules can thus generate constructions of arbitrary complexity.

Constituency and recursion are some of the most fundamental concepts in linguistics. As we saw above, both are defined in terms of relations between symbols. Symbolic models of language processing therefore incorporate these properties by fiat. In this article, we discuss how constituency and recursion may fit into a connectionist framework and the possible implications for linguistics and psycholinguistics.

Constituency

Connectionist models of language processing can address constituency in three increasingly radical ways. First, some connectionist models are *implementations* of symbolic language processing models in “neural” hardware. Many early connectionist models of syntax used this approach; an example is Fanty’s (1986) network implementation of a context-free grammar. This kind of model contains explicit representations of the constituent structure of a sentence in just the same way as a nonconnectionist implementation

of the same model would. Connectionist implementations of this kind may be important; they have the potential to provide feasibility proofs that traditional symbolic models of language processing are compatible with a “brain-style” computational architecture. But these models add nothing new with respect to the treatment of constituency.

The remaining two classes of connectionist models *learn* to process constituent structure, rather than having this ability hardwired. One approach is to have a network learn from input “tagged” with information about constituent structure. For example, Kim, Srinivas, and Trueswell (2002) train a network to map a combination of orthographic and co-occurrence-based “semantic” information about a word onto a structured representation encoding the minimal syntactic environment for that word. With an input vocabulary consisting of 20,000 words, this model has an impressive coverage and can account for certain results from the psycholinguistic literature concerning ambiguity resolution in sentence processing. But because constituent structure has been “compiled” into the output representations that the network was trained to produce, this kind of model does not offer any fresh insight into how linguistic constituency might operate, based on connectionist principles.

The third class of connectionist models addresses the more ambitious problem of learning the constituent structure of a language from untagged linguistic input. Such models have the potential to develop a new or unexpected notion of constituency, and hence may have substantial implications for theories of constituency in linguistics and psycholinguistics.

To understand how the more radical connectionist models address constituency, we need to frame the problem more generally. We can divide the problem of finding constituent structure in linguistic input into two interrelated parts: segmenting the sentence into chunks that correspond, to some extent, to linguistic constituents, and categorizing these units appropriately. The first problem is an aspect of the general problem of *segmenting* speech into appropriate units (e.g., phonemes, words) and more generally is an aspect of perceptual grouping. The second problem is an aspect of the general problem of classifying linguistic units—for instance, recognizing different classes of phonemes or establishing the parts of speech of individual lexical items. The segmentation and classification problems need not be solved sequentially. Indeed, there may be mutual influence between the decision to segment a particular chunk of language and the decision that it can be classified in a particular way. Nonetheless, it is useful to keep the two aspects of the analysis of constituency conceptually separate.

It is also important to stress the difference between the problem of assigning constituent structure to novel sentences where the language is known and the problem of acquiring the constituent structure of an unknown language. Statistical symbolic parsers are able to make some inroads into the first problem (Charniak, 1993). For highly stylized language input, and given a prestored grammar, they can apply grammatical knowledge to establish one or more possible constituent structures for novel sentences. But symbolic methods are much less advanced in acquiring the constituent structure of language, because this requires solving the hard problem of learning a grammar from a set of sentences generated by that grammar. It is therefore in relation to the acquisition of constituency that connectionist methods, with their well-developed learning methods, have attracted the most interest.

We begin by considering models that focus on the problem of classifying, rather than segmenting, the linguistic input. One connectionist model (Finch and Chater, 1993) learns the part of speech of individual words by clustering words together on the basis of the immediate linguistic contexts in which they occur. The rationale is based on the replacement test mentioned earlier: if two words are observed to occur in highly similar immediate contexts in a corpus, they probably belong to the same syntactic category. Finch

and Chater used a single-layer network with Hebbian learning to store co-occurrences between “target” words and their near neighbors. This allowed each target word to be associated with a vector representing the contexts in which it typically occurred. A competitive learning network classified these vectors, thus grouping together words with similar syntactic categories. This method is able to operate over unrestricted natural language, in contrast to most symbolic and connectionist models. From a linguistic perspective, the model slices lexical categories too finely, producing, for example, many word classes that correspond to nouns or verbs. On the other hand, the words within a class tend to be semantically related, which is useful from a cognitive perspective. The same method can be extended to classify sequences of words as NPs, VPs, etc. An initial classification of words is used to recode the input as a sequence of lexical constituents. Then, short sequences of lexical constituents are classified by their context, as before. The resulting groups of “phrases” (e.g., determiner-adjective-noun) are readily interpretable as NPs, and so on, but again, these groupings are too linguistically restrictive (i.e., only a small number of NPs are included in any particular cluster). Moreover, this phrasal level classification has not yet been implemented in a connectionist network.

A different attack on the problem of constituency involves training simple recurrent networks (SRNs) on linguistic input (Elman, 1990). An SRN involves a crucial modification to a feedforward network: the current set of hidden unit values is “copied back” to a set of additional input units, and paired with the *next* input to the network. The current hidden unit values can thus directly affect the next hidden unit values, providing the network with a memory for past inputs. This enables it to tackle sentence processing, where the input is revealed gradually over time rather than being presented at once.

Segmentation into constituents can be achieved in two ways by an SRN trained to *predict* the next input. One way is based on the assumption that predictability is higher within a constituent than across constituent boundaries, and hence that high prediction error indicates a boundary. This method has been advocated as potentially applicable at a range of linguistic levels (Elman, 1990), but in practice it has been successfully applied only on corpora of unrestricted natural language input in finding word boundaries (Cairns et al., 1997). Even here, the prediction strategy is a very partial cue to segmentation. If the network is provided with information about naturally occurring pauses between utterances (or parts of utterances), an alternative method is to assume that constituent boundaries occur where the network has an unusually high expectation of an *utterance boundary*. The rationale is that pauses tend to occur at constituent boundaries, and hence the prediction of a possible utterance boundary suggests that a constituent boundary may have occurred. This approach seems highly applicable to segmenting sentences into phrases, but it, too, has primarily been used for finding word boundaries in real corpora of language, when combined with other cues (Christiansen, Allen, and Seidenberg, 1998).

So far we have considered how SRNs might find constituents. But how well do they classify constituents? At the word level, cluster analysis of hidden unit activations shows that, to some extent, the hidden unit patterns associated with different word classes group naturally into syntactic categories, for SRNs trained on simple artificial grammars (Elman, 1990). These results are important because they show that even though the SRN may not learn to classify constituents explicitly, it is nevertheless able to *use* this information to process constituents appropriately.

Another way of assessing how SRNs have learned constituency is to see if they can generalize to predicting novel sentences of a language. The logic is that to predict successfully, the SRN must exploit linguistic regularities that are defined across constituents, and hence develop a notion of constituency to do so. However,

Hadley (1994) points out that this type of evidence is not compelling if the novel sentences are extremely similar to the network’s training sentences. He suggests that, to show substantial evidence for generalization across constituents, the network should be able to handle novel sentences in which words appear in sentence locations where they have not previously occurred (see SYSTEMATICITY OF GENERALIZATIONS IN CONNECTIONIST NETWORKS). For example, a novel sentence might involve a particular noun in object position, where it has previously occurred only in subject position. To generalize effectively, the network must presumably develop some abstract category of nouns. Christiansen and Chater (1994) demonstrated that an SRN can show this kind of generalization.

Despite this demonstration, though, connectionist models do not mirror classical constituency precisely. That is, they do not derive rigid classes of words and phrases that are interchangeable across contexts. Rather, they divide words and phrases into clusters without precisely defined boundaries, and they treat words and phrases differently, depending on the linguistic contexts in which they occur. This *context-sensitive* constituency can be viewed either as the undoing of connectionist approaches to language or as their radical contribution.

The potential problem with context-sensitive constituency is the productivity of language. To take Chomsky’s famous example, how do we know that the statement *colorless green ideas sleep furiously* is syntactically correct, except by reference to a context-insensitive representation of the relevant word classes? This seems necessary, because each word occurs in a context in which it has rarely been encountered before. But Allen and Seidenberg (1999) argue that this problem may not be fatal for context-sensitive notions of constituency. They trained a network to mutually associate two input sequences, a sequence of word forms and a corresponding sequence of word meanings. The network was able to learn a small artificial language successfully: it was able to regenerate the word forms from the meanings, and vice versa. Allen and Seidenberg then tested whether the network could recreate a sequence of word forms presented to it, by passing information from form to meaning and back. Ungrammatical sentences were recreated less accurately than grammatical sentences, and the network was thus able to distinguish grammatical from ungrammatical sentences. Importantly, this was true for sentences in which words appeared in novel combinations, as specified by Hadley’s criterion and as exemplified by Chomsky’s famous sentence. Thus, the context sensitivity of connectionist constituency may not rule out the possibility of highly creative and novel use of language, because abstract relations may be encoded at a semantic level as well as at the level of word forms.

If the apparent linguistic limitations of context-sensitive constituency can be overcome, then the potential psychological contribution of this notion is enormous. First, context sensitivity seems to be the norm throughout human classification. Second, much data on sentence processing seem most naturally to be explained by assuming that constituents are represented in a fuzzy and context-bound manner. The resulting opportunities for connectionist modeling of language processing are extremely promising. Thus, connectionist research may provide a more psychologically adequate notion of constituency than is currently available in linguistics.

Recursion

As with constituency, connectionist models have dealt with recursion in three increasingly radical ways. The least radical approach is to hardwire recursion into the network (e.g., as in Fanty’s (1986) implementation of phrase structure rules) or to add an external symbolic (“first-in-last-out”) stack to the model (e.g., as in Kwasny and Faisal’s (1990) deterministic connectionist parser). In both cases, recursive generativity is achieved entirely through standard sym-

bolic means, and although this is a perfectly reasonable approach to recursion, it adds nothing new to symbolic accounts of natural language recursion. The more radical connectionist approaches to recursion aim for networks to *learn* to deal with recursive structure. One approach is to construct a modular system of networks, each of which is trained to acquire different aspects of syntactic processing. For example, Miikkulainen's (1996) system consists of three different networks: one trained to map words onto case-role assignments, another trained to function as a stack, and a third trained to segment the input into constituent-like units. Although the model displays complex recursive abilities, the basis for these abilities and their generalization to novel sentence structures derive from the configuration of the stack network combined with the modular architecture of the system, rather than being discovered by the model. The most radical connectionist approaches to recursion attempt to learn recursive abilities with minimal prior knowledge built into the system. In this type of model, the network is most often required to discover both the constituent structure of the input and how these constituents can be recursively assembled into sentences. As with the similar approach to constituency described in the previous section, such models may provide new insights into the notion of recursion in human language processing.

Before discussing these modeling efforts, we need to assess to what extent recursion is observed in human language behavior. It is useful to distinguish *simple* and *complex* recursion. Simple recursion consists in recursively adding new material to the left (e.g., the adjective phrases (AP) in *the gray cat* → *the fat gray cat* → *the ugly fat gray cat*) or the right (e.g., the PPs in *the flowers in the vase* → *the flowers in the vase on the table* → *the flowers in the vase on the table by the window*) of existing phrase material. In complex recursion, new material is added in more complicated ways, such as through center-embedding of sentences (*The chef admired the musicians* → *The chef who the waiter appreciated admired the musicians*). Psycholinguistic evidence shows that people find simple recursion relatively easy to process, whereas complex recursion is almost impossible to process with more than one level of recursion. For instance, the following sentence with two levels of simple (right-branching) recursion, *The busboy offended the waiter who appreciated the chef who admired the musicians*, is much easier to comprehend than the comparable sentence with two levels of complex recursion, *The chef who the waiter who the busboy offended appreciated admired the musicians*. Because recursion is built into the symbolic models, there are no *intrinsic* limitations on how many levels of recursion can be processed. Instead, such models must invoke *extrinsic* constraints to accommodate the human performance asymmetry on simple and complex constructions. The radical connectionist approach models human performance directly without the need for extrinsic performance constraints.

The SRN model developed by Elman (1991) was perhaps the first connectionist attempt to simulate human behavior on recursive constructions. This network was trained on sentences generated by a small context-free grammar incorporating center-embedding and a single kind of right-branching recursive structure. In related work, Christiansen and Chater (1994) trained SRNs on a recursive artificial language incorporating four kinds of right-branching structures, a left-branching structure, and center-embedding. The behavior of these networks was qualitatively comparable with human performance in that the SRN predictions for right-branching structures were more accurate than on sentences of the same length involving center-embedding, and performance degraded appropriately as the depth of center-embedding increased. Weckerly and Elman (1992) further corroborated these results, suggesting that semantic bias (incorporated via co-occurrence restrictions on the verbs) can facilitate network performance in center-embedded constructions, similar to the semantic facilitation effects found in human processing. Using abstract artificial languages, Christiansen

and Chater (1999) showed that the SRN's general pattern of performance is relatively invariant across network size and training corpus, and concluded that the human-like pattern of performance derived from intrinsic constraints inherent to the SRN architecture.

Connectionist models of recursive syntax typically use "toy" fragments of grammar and small vocabularies. Aside from raising concerns over scaling-up, this makes it difficult to provide detailed fits with empirical data. Nonetheless, some attempts have recently been made to fit existing data and derive new empirical predictions from the models. For example, the Christiansen and Chater (1999) SRN model fits grammaticality rating data from several behavioral experiments, including an account of the relative processing difficulty associated with the processing of center-embeddings (with the following relationship between nouns and verbs: $N_1N_2N_3V_3V_2V_1$) versus cross-dependencies (with the following relationship between nouns and verbs: $N_1N_2N_3V_1V_2V_3$). Human data have shown that sentences with two center-embeddings (in German) are significantly harder to process than comparable sentences with two cross-dependencies (in Dutch). The simulation results demonstrated that the SRNs exhibited the same kind of qualitative processing difficulties as humans on these two types of complex recursive constructions.

Just as the radical connectionist approach to constituency deviates from classical constituency, the above approach to recursion deviates from the classical notion of recursion. The radical models of recursion do not acquire "true" recursion because they are unable to process infinitely complex recursive constructions. However, the classical notion of recursion may be ill-suited for capturing human recursive abilities. Indeed, the psycholinguistic data suggest that people's performance may be better construed as being only quasi-recursive. The semantic facilitation of recursive processing, mentioned earlier, further suggests that human recursive performance may be partially context sensitive. For example, the semantically biased sentence, *The bees that the hive that the farmer built housed stung the children*, is easier to comprehend than the neutral sentence, *The chef that the waiter that the busboy offended appreciated admired the musicians*, even though both sentences contain two center-embeddings. This dovetails with the context-sensitive notion of constituency and suggests that context sensitivity may be a more pervasive feature of language processing than is typically assumed by symbolic approaches.

Discussion

This article has outlined several ways in which constituency and recursion may be accommodated within a connectionist framework, ranging from direct implementation of symbolic systems to the acquisition of constituency and recursion from untaged input. We have focused on the radical approach, because this approach has the greatest potential to affect psycholinguistics and linguistic theory. However, much of this research is still preliminary. More work is needed to decide whether the promising but limited initial results can eventually be scaled up to deal with the complexities of real language input, or whether a radical connectionist approach is beset by fundamental limitations. Another challenge is to find ways—theoretically and practically—to interface models that have been proposed at different levels of linguistic analyses, such as models of morphology with models of sentence processing.

Nevertheless, the connectionist models described in this article have already influenced the study of language processing. First, connectionism has helped promote a general change toward replacing "box-and-arrow" diagrams with explicit computational models. Second, connectionism has reinvigorated the interest in computational models of learning, including learning properties, such as recursion and constituent structure, that were previously assumed to be innate. Finally, connectionism has helped increase

interest in the statistical aspects of language learning and processing.

Connectionism has thus already had a considerable impact on the psychology of language. But the final extent of this influence depends on the degree to which practical connectionist models can be developed and extended to deal with complex aspects of language processing in a psychologically realistic way. If realistic connectionist models of language processing can be provided, then the possibility of a radical rethinking not just of the nature of language processing, but of the structure of language itself, may be required.

Road Map: Linguistics and Speech Processing

Background: Language Processing

Related Reading: Language Acquisition; Recurrent Networks; Learning Algorithms

References

- Allen, J., and Seidenberg, M. S., 1999, The emergence of grammaticality in connectionist networks, in *The Emergence of Language* (B. MacWhinney, Ed.), Mahwah, NJ: Erlbaum, pp. 115–151.
- Cairns, P., Shillcock, R. C., Chater, N., and Levy, J., 1997, Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation, *Cogn. Psychol.*, 33:111–153.
- Charniak, E., 1993, *Statistical Language Learning*, Cambridge, MA: MIT Press. ♦
- Christiansen, M. H., Allen, J., and Seidenberg, M. S., 1998, Learning to segment speech using multiple cues: A connectionist model, *Lang. Cogn. Proc.*, 13:221–268.
- Christiansen, M. H., and Chater, N., 1994, Generalization and connectionist language learning, *Mind Lang.*, 9:273–287.
- Christiansen, M. H., and Chater, N., 1999, Toward a connectionist model of recursion in human linguistic performance, *Cogn. Sci.*, 23:157–205. ♦
- Elman, J. L., 1990, Finding structure in time, *Cogn. Sci.*, 14:179–211. ♦
- Elman, J. L., 1991, Distributed representation, simple recurrent networks, and grammatical structure, *Machine Learn.*, 7:195–225.
- Fant, M. A., 1986, Context-free parsing with connectionist networks, in *Neural Networks for Computing* (J. S. Denker, Ed.), New York: American Institute of Physics, pp. 140–145.
- Finch, S., and Chater, N., 1993, Learning syntactic categories: A statistical approach, in *Neurodynamics and Psychology* (M. Oaksford and G. D. A. Brown, Eds.), New York: Academic Press, pp. 295–321.
- Hadley, R. F., 1994, Systematicity in connectionist language learning, *Mind Lang.*, 9:247–272.
- Kim, A. E., Srinivas, B., and Trueswell, J. C., 2002, The convergence of lexicalist perspectives in psycholinguistics and computational linguistics, in *Sentence Processing and the Lexicon: Formal, Computational and Experimental Perspectives* (P. Merlo and S. Stevenson, Eds.), Amsterdam: John Benjamins Publishing, pp. 109–135.
- Kwasny, S. C., and Faisal, K. A., 1990, Connectionism and determinism in a syntactic parser, *Connect. Sci.*, 2:63–82.
- Miikkulainen, R., 1996, Subsymbolic case-role analysis of sentences with embedded clauses, *Cogn. Sci.*, 20:47–73.
- Weckerly, J., and Elman, J. L., 1992, A PDP approach to processing center-embedded sentences, in *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, pp. 414–419.

Contour and Surface Perception

Heiko Neumann and Ennio Mingolla

Introduction

Accumulating evidence from psychophysics and neurophysiology indicates that the computation of visual object representations is organized into parallel interacting subsystems, or streams, that consist of mechanisms following complementary processing strategies. *Boundary formation* proceeds by spatially linking oriented contrast measures along smooth contour patterns, whereas *perceptual surface* attributes, such as lightness or texture, are derived from local ratio measures of image contrast of regions taken along contours. Mechanisms of both subsystems mutually interact to resolve initial ambiguities and to generate coherent representations of surface layout.

Even when viewing single-gray-level images, people can discern important characteristics of visual scenes, including brightness, reflectance, surface orientation, texture, transparency, and relative depth. It has been proposed that these inferences are based on neural mechanisms that compute distinct and spatially registered representations of such characteristics or features, so-called *intrinsic images*. This approach has been formalized on the basis of statistical inference theory, in which the a posteriori likelihood of estimating several scene characteristics given the available image is optimized. Common to these approaches is the requirement that representations of intrinsic scene characteristics are constrained in order to guarantee the consistency of the set of solutions, which often involve smoothness assumptions for correlated feature estimates. These consistency constraints are typically based on the laws of physical image generation, such as Lambertian surface reflectance properties, and thus the overall process comprises an ideal observer seeking an optimal solution to the problem of inferring physical scene properties from images.

After briefly summarizing fundamental approaches to the computation of intrinsic scene characteristics, we consider an alternative literature of neural models of boundary and surface computation.

Intrinsic Images and the Neural Processing of Surface Layout

Barrow and Tenenbaum (1978) proposed a framework for machine vision based on computation of view-centered families of representations of local estimates of intrinsic scene characteristics, such as illumination, surface reflectance, and orientation. The resulting images are spatially registered with the single achromatic luminance image in which all the attributes are encoded pointwise. In order to regularize the inverse problem of computing several attributes from image data, several constraints, or assumptions, need to be incorporated into the recovery mechanisms. For example, in order to achieve the continuity of homogeneous surface attributes, each intrinsic image incorporates lateral smoothing of values in a discrete grid unless an edge breaks the local continuity. Discontinuities need to be in registration with intensity edges, and the recovered values for attributes need to fulfill the image irradiance equation at each grid point. This approach has been formalized in terms of coupled Markov random field (MRF) models in which the a posteriori probability for scene interpretation given an image, $p(\text{scenelimage}) \propto \exp(-E/T)$ (E is an energy function that sums the image and prior constraints, T is a temperature term), can be maximized utilizing a stochastic sampling scheme (simulated annealing; see SIMULATED ANNEALING AND BOLTZMANN MACHINES). The MRF formulation allows us to incorporate a line pro-

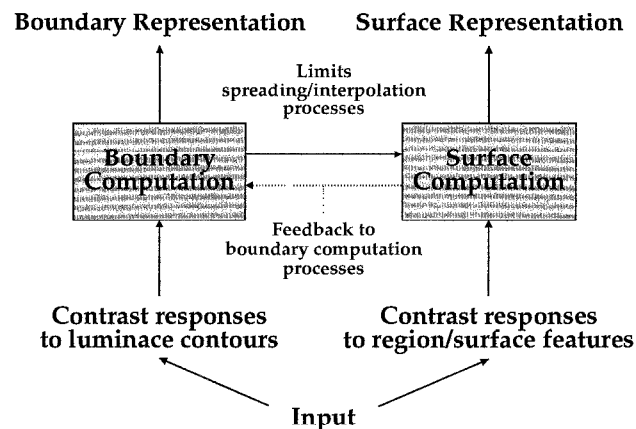


Figure 1. Schematic of the macroscopic computational elements for computing boundaries and surfaces. The flow of computation is segregated into parallel but interacting streams to determine boundaries and surface attributes. See text for details.

cess (representing discontinuities) in the form of additional associated energy terms that can break the smoothness within an intrinsic attribute representation at locations where the tension of the recovered surface exceeds certain limits. These interactions correspond to the bidirectional coupling of boundary and surface computation depicted in Figure 1. More recent developments of the MRF approach take into account the computation of perceptual surface attributes, such as transparency (Kersten, 1991).

These approaches share the same basic computational principles: representations of smooth surface characteristics are generated by interpolating sparse data from initial estimates, and an edge map is extracted to depict the discontinuities in one or several attributes. In order to generate mutually consistent maps of intrinsic image characteristics, or attributes, the energy function used in the MRF model must utilize the prior probabilities of image characteristics, which in turn are derived from physical imaging models, e.g., for reflectance or transparency. In that sense the estimation process defines an ideal observer that utilizes knowledge about image formation by trying to invert the optical image generation process. Perceptual findings, however, indicate that, for example, an attribute such as lightness (the perceived surface reflectance) depends on surface size and that the judgment of surface orientation is unreliable over variations in surface curvature or lighting conditions. These observations cannot uniquely be accounted for by the intrinsic image approach, since size effects impose constraints not considered by such local processes as are employed in implementing an inverse optics solution.

In addition, line processes representing physical discontinuities in a surface property are modeled as separate MRFs. In order to be computationally tractable their prior probability structure is formulated over a small local pixel neighborhood only. Empirical observations again suggest that contour processes act over long spatial distances in order to reliably integrate contour fragments to form surface boundaries under variable imaging conditions. In all, these observations motivate the investigation of the neural mechanisms underlying the computation of perceptual boundaries and the subsequent assignment of surface attributes.

Elements of Spatial Long-Range Integration in Boundary Finding

Formal approaches to the Gestalt concept of *good continuation* have garnered increasing attention since Field, Hayes, and Hess

(1993) popularized the notion of an “association field,” an elongated spatial zone aligned with oriented contour segments denoting facilitatory perceptual interactions with other segments. This geometry of spatial integration is summarized in the “bipole” icon of Figure 2, a figure-eight-shaped zone that was introduced as a “co-operative cell” unit in a neural network model for grouping by Grossberg and Mingolla (1985). The strength of spatial integration in the presently considered models is always some function of the distances and relative alignments of oriented units, such as contrast edges. Imagine that a contrast pattern occurs along the orientations denoted by the long axes of the dark and light ellipses at locations x and x' . The influence of the edge at the light ellipse on the representation of the contour strength in the region of the dark ellipse is calculated using the following fundamental quantities: (1) the distance, r , between the centers of the two ellipses; (2) the angle, θ , between the ray passing through the centers of the two ellipses and the principal axis of the dark ellipse; and (3) the difference in orientation, ϕ , of the principal axes of the two ellipses.

The bipole shape expresses the region of relatively high coupling strength for the influence of contour segments remote from the central ellipse on a unit whose positional and orientational preferences are denoted by the ellipse at the center. Its shape and connectivity pattern is justified by recent investigations in psychophysics, physiology, and anatomy. For example, the co-occurrence of edge segments in natural scenes have been shown to have a bipole distribution, and cortical cells of similar orientation selectivity in tree shrew are preferentially connected along the axis of their visuotopic alignment to cells of compatible orientational preference (see Neumann and Mingolla, 2001, for details). Thus, the bipole icon for grouping of contour segments, originally suggested on intuitive grounds, is validated by a growing body of empirical data.

The bipole concept can be embedded in a framework in which the pattern of connectivity among relatively tightly coupled neural units is described formally by a *spatial weighting*, or kernel, function, coding the strength of connections between units in a spatial array. Items that are closely spaced are more likely to be grouped than candidates that are located far apart. The underlying neighborhood function often selectively facilitates a sector of a spatial surround to define an anisotropic coupling that is compatible with the feature domain. Elementary features along dimensions such as (tangential) orientation, motion direction, and disparity provide the

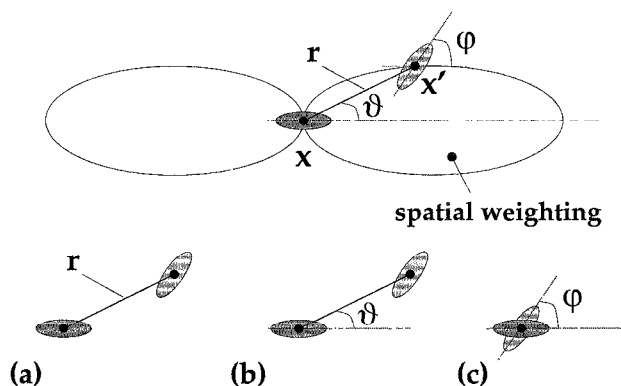


Figure 2. *Top*, The “bipole icon” for modeling feature integration in spatial grouping. The quantity to be assessed is the “contribution” of activation in an oriented unit, denoted by the light ellipse, to the activation at the center unit, denoted by the dark ellipse. *Bottom*, The figure-eight shape of a bipole expresses relations among three fundamental quantities: (a) the distance, r , between spatial locations, (b) the angle, θ , between the virtual line and the target orientation, and (c) the difference in orientation, ϕ , of the two ellipses.

dimensions of the visual representation space. The feature *relativity* (or *compatibility*) between stimulus items is defined along these dimensions. Herein, the most likely appearance of meaningful structure is encoded to represent “what feature goes with what.” In the following treatment we will focus on grouping mechanisms for static form processing and therefore consider only *orientation* as the relevant feature dimension.

The result of integration of items defines the *support* of a localized feature measurement at a given spatial location based on the configuration of other items. Based on the bipole structure of left and right subfields (Figure 2), the support of a target item can be defined as a function

$$\text{support}_{xy, \text{feature}} = \{\text{left-input}_{xy, \text{feature}}\} \circ \{\text{right-input}_{xy, \text{feature}}\} \quad (1)$$

where \circ denotes some operation to define the *combination of subfields*. Subscripts xy denote spatial locations in a two-dimensional (2D) retinotopic map, and *feature* identifies the feature dimension involved in the grouping process—in our case, *orientation*. The *activation* of the corresponding target cell results as a function $f(\cdot)$ of the support. The formal description of the mechanisms underlying the computation of activations “left-input” and “right-input,” respectively, necessitates the detailed specification of the interaction of activities in grouping.

In most models the connectivity pattern of the relatable features is prespecified, or *programmed*, referring to some measure of geometric entities. These are designed to encode static efficacies between n -tuples, e.g., pairs, of feature measurements at given locations optimizing a given functionality. To date, few approaches have investigated the possible *self-organization* of such lateral interactions in a neural architecture. Note that the spatial weighting function and the feature relativity define the components of the net *coupling strength*. This function specifies a metric for the similarity measure in the $\langle xy, \text{feature} \rangle$ -space and thus defines the distance function of *feature cooperation* for the clustering of a visual pattern in accordance with a relativity measure that underlies the visual interpolation in spatial grouping for object recognition (Kellman and Shipley, 1991). *Input activations* at the different sites are necessary to gather any support. Here, two different types of interactions can be distinguished: (1) a convergent feedforward mechanism is defined when the *bottom-up* input is integrated at the target location, whereas (2) a mechanism of (nonlinear) *lateral* interaction is defined when activity is horizontally integrated within a neural layer.

Taken together, the activation of, say, “left-input” derived from the feature integration process using the bipole mechanism is computed by

$$\text{left-input}_{x\theta} = \sum_{x'\phi} \{\text{act}_{x'\phi} \cdot \text{relate}_{xx'\theta\phi} \cdot \text{weight}_{xx'\theta}^L\} \quad (2)$$

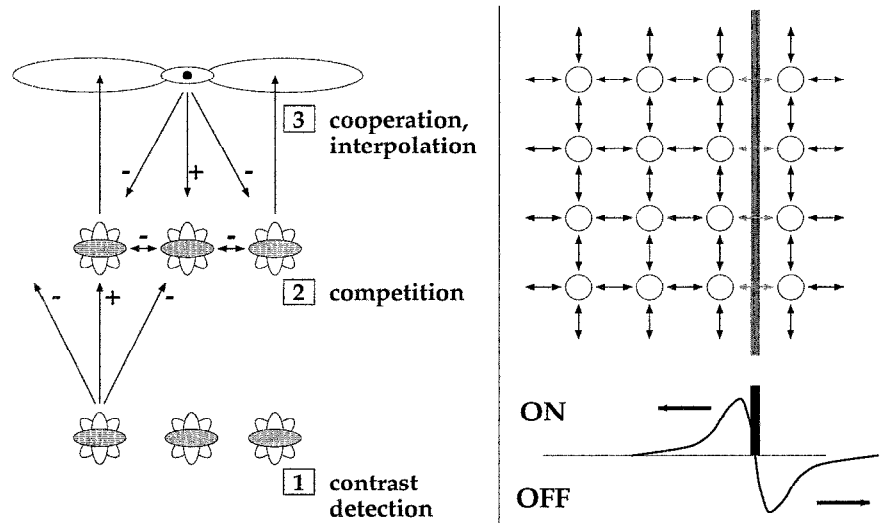
where “weight” denotes the spatial weighting kernel, “relate” the feature relativity, and “act” the input activations. The “right-input” activity is computed similarly. Here, $x = (x, y)$ and θ correspond to the location and orientation of the target feature, respectively. Other parameters refer to the specific location $\langle x', \phi \rangle$ in the space-orientation neighborhood (see Figure 2).

Grouping Models and Their Components

The initial processing stages of these models consist of some form of filtering the input luminance image. A (possibly nonlinear) center-surround mechanism (resembling segregated ON and OFF contrast channels at the retina and LGN) computes contrast ratios throughout the image. These outputs drive oriented simple and complex cells, which are often modeled as localized spatial frequency filters, such as Gabor or wavelet kernels. Their output of different orientation fields defines the interface representation for subsequent grouping processes. We focus on those models that (1) have their roots in the explanation of empirical data and (2) were most influential for subsequent scientific developments. A comprehensive treatment can be found in Neumann and Mingolla (2001).

The Boundary Contour System (BCS) has been developed as part of a unified modeling framework called FACADE (Form-And-Color-And-DEpth). As described by Ross, Gossberg, and Mingolla (2000), the BCS consists of a series of boundary detection, competition, and cooperation stages. The general layout of basic processing stages is presented in Figure 3 (left). Long-range boundary cooperation of this scheme (stage 3) accomplishes the grouping of consistent boundaries and the completion of interrupted boundaries. Spatial weighting and relativity, as elements of the support function, in this model are defined as $\text{weight}_{xx'\theta}^{L/R} = [\pm \Gamma_{xx'\theta}^{rad} \cdot \Gamma_{xx'\theta}^{ang}]^+$ (for a scheme separable in polar coordinates; $[\cdot]^+$ denoting half-wave rectification) and $\text{relate}_{xx'\theta\phi} = \cos^q(2 \tan^{-1}((y' - y)/(x' - x)) - (\theta + \phi))$ (for a co-circularity constraint; q controlling the angular width of the lobes). The support and activation function is computed employing bipole cells that fire only if both lobes of their integration fields are sufficiently activated. The left/

Figure 3. Details of the mechanisms involved in boundary and surface computations. *Left*, The three major stages for *boundary computation* involve contrast detection and competition between responses in space and orientation domain. Resulting responses are integrated over longer distances utilizing oriented bipoles, which lead to on-center/off-surround feedback interaction. *Right*, Lateral interaction for *surface computation* between sites of a regular grid. Local contrast signals measured near boundaries are laterally propagated to fill in regions void of activities. The lateral interaction is controlled by an inhibitory mechanism to reduce or switch off lateral couplings by high boundary activation, as indicated by the light shaded arrows. See text for details.



right subfield combination thus realizes a multiplicative, or AND gate, combination of input terms. The cooperative-competitive (CC) feedback loop between stages 1 and 3 (Figure 3, left) acts to complete and enhance spatially and orientationally consistent boundary groupings while inhibiting inconsistent ones, thereby also suppressing noise. This mechanism of long-range completion is also capable of signaling boundaries over gaps in the image, that is, over regions void of any contrast signals, and thus can generate illusory contours for images in which humans perceive them.

Relaxation labeling schemes have been developed as optimization procedures to find consistent interpretations for measurement problems with uncertainty. The computational goal seeks to achieve a consistent labeling assigning graded activations, or probabilities, to a limited set of labels for nodes in a graph representation. For contour integration, the labeling problem can be formulated as one of finding the most likely set of orientations for discrete grid locations corresponding to the graph nodes (a “no-line” label is taken into account to represent the nonresponsiveness of cells at locations that are not boundary elements). Parent and Zucker (1989) determine unambiguous orientation estimates along contours by evaluating consistency constraints based on a measure of compatibility between pairs of orientations. Initial responses are generated by oriented filters (Figure 3, left, stage 1) that were normalized in order to allow filter activations treated as probabilities for assigning orientation labels (stage 2). The individual strengths for orientation measures at a given image location are iteratively updated through the support that is gathered from relatable activities in a spatial neighborhood. Spatial weighting and relatability are defined as $\text{weight}_{\mathbf{x}\mathbf{x}'}^{L/R} = P_{\mathbf{x}\mathbf{x}'}^{\text{length}, L/R} \cdot d(\mathbf{x}, \mathbf{x}')$, where P denotes a predicate to compensate for path length differences on discrete grids and $d(\cdot)$ compensates for differences in interpixel distances, and $\text{relate}_{\mathbf{x}\mathbf{x}'}^{\theta\phi} = c_{\mathbf{x}\mathbf{x}'}^{\theta\phi} \cdot K_{\mathbf{x}\mathbf{x}'}^{\theta\phi} \cdot C_{\mathbf{x}\mathbf{x}'}^{\theta\phi}$. Here the co-circularity measure c is augmented by two binary predicates to exclude any candidates of incompatible local contour curvature. Input activations in the orientation field are thinned by nonmaximum suppression in order to generate localized representations of contours. The support and activation function is computed by integrating responses from the subfields of the spatial weighting functions (stage 3). Activities are iteratively updated by a nonlinear recurrent competitive/cooperative mechanism.

Contrary to the previous approaches, Heitger et al. (1998) proposed a feedforward scheme of successive filtering for contour grouping that selectively integrates activities from oriented single end-stopped (ES) filters, which respond to, e.g., line ends and corners. The result of such grouping is combined with the representation of oriented contrast responses to generate a final contour map. The core mechanism of grouping again utilizes spatial weighting functions (bipoles) virtually equivalent to those of the BCS (Figure 3, left, stage 3). Two grouping rules are distinguished for corners (para grouping) and line ends (ortho grouping). The responses of ortho and para grouping are linearly interpolated. Different curvature classes are distinguished, similar to the relaxation scheme (see above), by partitioning the bipoles into subfields such that only some of them in the left lobe can cooperate with their counterparts in the right lobe. Left/right lobes are combined in a multiplicative fashion via an AND-gating mechanism, similar to the BCS approach. This makes the grouping scheme selective to complete activations between localized ES features. The elongated bipole lobes are also capable of signaling boundaries over gaps, and thus generate illusory contours. This model does not incorporate any feedback mechanism and must, therefore, employ some stages of postprocessing in order to sharpen the final boundary response.

Based on the previous “core models,” other approaches have been developed that elaborate aspects of the general framework. For example, random walks of particles in a discrete lattice of the

sampled space-orientation domain have been investigated to determine the paths of spatially relatable items at a given location (x, y, θ) and another point (i, j, ϕ) . Particles were initiated at sparse keypoint locations corresponding to localized responses of ES cells. The probability densities in the stochastic completion field represent the strengths—and therefore likelihood—of smooth paths connecting pairs of key points. Another study investigated how V1 horizontal long-range integration could functionally account for the enhancement of texture region boundaries and pop-out effects in visual search. Here, the spatial integration field is subdivided into spatially nonoverlapping parts of excitatory and inhibitory contribution. An elongated bipole integrates activities of cells oriented such that they form smooth interpolations with the target cell at the bipole center. Activities of like orientation from a sector orthogonal to the target cell orientation generate the inhibitory signal. In a similar spirit, a scheme of excitatory long-range integration has been used that consists of two spatial regions: one coaxial bipole with cocircular relatability and one sector that extends orthogonally from the target cell’s orientation axis and integrates units oriented parallel to that of the cell. The latter component contributes to a facilitation of simple symmetric shape axes. Each cell inhibits itself based on a threshold of the average input from its immediate neighbors, so that only salient arrangements produce a net output.

Contrast integration for boundary grouping addresses the particularly important question about what the core principles are that underlie the establishing of spatial integration and grouping. Yet there still is an ongoing debate about the role of feedforward and feedback mechanisms involved in spatial grouping and the dominance of their individual contributions in visual processing. Aspects of *temporal coding* principles based on oscillator mechanisms or spiking neurons may play another important role in grouping tasks. Several neurophysiological studies indicate that distributed representations of related scene fragments are linked by temporally correlated, or synchronized, neural activation. The temporal coding hypothesis studied in isolation appears, however, to be incomplete. The temporal establishment of grouping addresses the signaling of binding, but not the “how” or “what” of its computation. The mechanisms of oriented long-range interactions for integration provide the underlying basis for grouping to establish perceptual items related to surface boundaries.

Complementary Mechanisms of Surface Perception

The computation of perceptual surface qualities complements the formation of their boundaries. Image regions that are sufficiently homogeneous in luminance or statistical distribution of contrasts can give rise to the impression of color, texture, brightness, or lightness, also known as achromatic color. How is the generation of smooth representations of surface qualities accomplished? Paradiso and Nakayama (1991) provided compelling psychophysical evidence for a long-hypothesized neural process that propagates local estimates of lightness from boundaries into region interiors. Further psychophysical as well as physiological investigations of temporal properties of brightness and texture filling-in revealed further details of the neural machinery underlying the integration of perceptual surface properties.

Computational models for the generation of perceptual surface quantities generally pursue one of three basic strategies: (1) filtering and rule-based symbolic interpretation, (2) spatial integration via inverse filtering and labeling, or (3) filling-in. In the first scheme, (nonlinear) combinations of filter responses and subsequent rule-based decision operations lead to the final prediction of lightness. The latter two approaches both begin with local luminance ratios estimated along boundaries. Computation of surface qualities proceeds by propagating local estimates into region interiors in order to generate a spatially contiguous representation of

homogeneous properties (Figure 3, right). In the spatial integration approach, the lateral propagation is the consequence of an iterative process to invert previous spatial derivative operations labeling region interiors with estimates of quantities derived at region boundaries. Filling-in approaches spread activity in a neural map such that at locations coding region interiors, activation is generated by a spatial diffusion process integrating estimates of local contrasts from remote boundaries.

Spatial filtering approaches utilize initial stages of either isotropic or oriented filters, or both, over multiple bandpass channels of spatial frequency. The filter outputs are scaled nonlinearly by a gain-control mechanism and thresholded, individually interpreted by a set of rules, and finally combined over several scales. In two dimensions this strategy leads to results that depend on the direction of the sequential application of interpretation rules. Authors pursue a simple averaging of results from forward and backward scanning over different orientations (McArthur and Moulden, 1999). Approaches in this category follow a tradition that the input luminance distribution is processed through a sequence of filtering steps during the early stages of the visual system. It is assumed that specific features in the responses directly contribute to observable brightness effects by implicitly propagating local qualities into region interiors as a consequence of applying some interpretation rules.

Spatial integration models attempt to recover object lightness of a surface by utilizing the sequence of processing stages *Filtering/Differentiation* → *Boundaries/Thresholding* → *Integration*. Differentiation and subsequent thresholding operations are intended to detect salient changes in the luminance signal, which in turn trigger the integration from local luminance ratios. Luminance ratios provide the basis to infer surface-related properties that are invariant against gradual illumination changes, thus discounting the illuminant. Furthermore, lightness can be influenced by luminances at regions remote from each other. The Retinex algorithm by Land and McCann (1971) accounts for these observations by integrating contrasts along several pathways of different lengths. The logarithms of luminance ratios at thresholded contrast locations are summed such that, as a net effect, ratios measured between the target region and distant regions are integrated and averaged. A center-surround interaction accounts for this process in an approximate form. Alternative formulations of lightness integration numerically invert the differentiated 2D luminance image. Under certain boundary conditions a unique solution exists that can be computed by numerical techniques (Hurlbert, 1986). Changes in reflectance properties occur locally, thus comprising a high spatial frequency pattern, while inhomogeneous illumination constitutes a phenomenon at low spatial frequencies. Suppression of influences in low spatial frequencies could be achieved by homomorphic filtering in which low-frequency components are reduced or even suppressed and higher frequencies are amplified. Following this idea, approximate approaches for lightness computation utilize multiple spatial frequency channels of center-surround mechanisms based on divisive, or shunting, interactions of different scales.

As a second computational strategy of spatial integration models, *filling-in* approaches proceed by taking local ratios along boundaries and subsequently propagate these measures into the void spaces of bounded regions. Generating a representation of surface quality utilizes the processing stages *Filtering* → *Boundaries* → *Filling-in*. The algorithm proposed by Grossberg and Todorović (1988) was the first implementation in two dimensions of the complementary operations of the combined BCS and Feature Contour System (FCS) model of brightness perception that was able to explain a wide variety of phenomena, such as simultaneous contrast and the Craik-O'Brien-Cornsweet illusion. The model was later demonstrated to also account for the temporal properties of brightness filling-in (Paradiso and Nakayama, 1991). Filling-in of local

contrast ratios taken along extended figural boundaries is laterally propagated in a diffusion process. Such a diffusion is controlled by a gradual permeability function Y utilizing a boundary signal such as the one generated by mechanisms of long-range integration (compare Figure 2 and Figure 3, right). This renders filling-in a spatially inhomogeneous diffusion process that generates a maximum likelihood (regularized) solution of a brightness surface given the sparse contrast estimates at the boundaries, where Y is monotonically decreasing to split apart regions of homogeneous surface properties. In the case that the function Y is solely controlled by an auxiliary boundary signal, the function solves a simple gradient descent of finding a dense activity distribution representing surface brightness. In general, Y could also depend on gradients of the filling-in signal itself, such that boundary and surface computation become mutually interdependent (dotted arrow in Figure 1) and the overall system becomes nonlinear.

Ross and Pessoa (2000) describe an important extension of previous models by suggesting a context-sensitive weighting of contrast measures prior to a final integration stage. Emphasis is put on a computational mechanism that tags boundaries from contrast detection and grouping to (partially) segment an image into different context domains. The segmentation is triggered by the presence of T-junctions, which are used as seed points to propagate the tagging signal along the roof of the Ts and the adjoining smooth boundary segments, thus generating a map of context boundaries. These boundaries are subsequently used to suppress initial contrast measures along the corresponding contours via a gating mechanism. As a result, the contribution of contrast measures that could lead to erroneous lightness estimates over segmentation boundaries is reduced or even suppressed. This computational mechanism is capable of producing region lightness estimates that account for several effects that have been previously shown to be unexplainable by simple local mechanisms of contrast integration, e.g., White's effect, the Benary cross, and Adelson's folded-card stimuli.

Although not described here, the FACADE model, with which the Ross and Pessoa model has certain parallels, proposes related explanations of the just-mentioned and related lightness effects (Grossberg, 1994). FACADE has also been developed to account for the perception of such phenomena as amodal surface completion, whereby occluded portions of objects are sensed as being present in specific locations behind foreground objects, and more generally the separation of surface regions seen as figures from backgrounds. The development of this and other models (Heitger et al., 1998) that attempt to account for more complex aspects of surface perception than are captured in planar arrays indicates that an exciting new phase of inquiry into surface perception is under way.

Stochastic Formulation of Boundary/Surface Computations

In the tradition of intrinsic image architectures, some approaches formulate the above-outlined boundary and surface computations in a Bayesian framework. For example, Lee (1995) utilized a variant of a nonlinear diffusion mechanism that incorporated piecewise smoothness of filling-in domains that were separated by contour segments denoting breaks in the smooth surface signal. Statistical signals from initial responses of oriented Gabor filters of different frequency selectivity served as filling-in signals. These inputs were subsequently diffused in the space-frequency domain incorporating a spatial modulation by the boundary signal. Mutual interaction between boundary and surface processes were modeled as MRF processes in which Bayesian priors propagated bidirectionally and interacted through local connections in each area. Since this approach focuses on texture region segregation, an explicit modeling of the optical image generation process, as in clas-

sical intrinsic image approaches, is not necessary here. This renders the model a Bayesian interpretation of above-mentioned neural processes for surface/boundary extraction.

Discussion

We have seen that the key stages of a common framework of computational models for generating perceptual surface representations utilize separate subsystems for boundary and surface computation. Depending on the modeling framework, different processes of mutual interaction are defined to achieve the goal of generating a coherent representation of object surfaces and their attributes. Modeling the *neural* processes of a perceptual surface layout has focused on the processes of contour completion based on long-range integration based on some variation of the structure of the bipole kernel and a number of additional computational principles in evaluating related activities in feature space. The bipole kernel itself graphically visualizes the oriented fan-like connectivity between sites in a space-orientation feature space.

In comparison with processes of boundary finding, the situation with respect to surface quality perception is considerably less developed, most especially with respect to attributes such as texture perception, which we have not reviewed. All models presented are based on some mechanism to laterally propagate activities to interpolate and smooth sparse estimates. Improved empirical techniques juxtaposed with hypotheses developed by computational modelers offer the hope that coming years will see a convergence in this area, as has already occurred in the field of contour completion.

Road Map: Vision

Related Reading: Global Visual Pattern Extraction; Laminar Cortical Architecture in Visual Perception; Perception of Three-Dimensional Structure; Stereo Correspondence

References

- Barrow, H. G., and Tenenbaum, J. M., 1978, Recovering intrinsic scene characteristics from images, in *Computer Vision Systems* (A. R. Hanson and E. M. Riseman, Eds.), New York: Academic Press, pp. 3–26.
- Field, D. J., Hayes, A., and Hess, R. F., 1993, Contour integration by the human visual system: Evidence for a local “association field,” *Vision Res.*, 33:173–193.
- Grossberg, S., 1994, 3-D vision and figure-ground separation by visual cortex, *Percept. Psychophys.*, 55:48–120.
- Grossberg, S., and Mingolla, E., 1985, Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentation, *Percept. Psychophys.*, 38:141–171.
- Grossberg, S., and Todorović, D., 1988, Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena, *Percept. Psychophys.*, 43:723–742.
- Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., and Kübler, O., 1998, Simulation of neural contour mechanisms: Representing anomalous contours, *Image Vision Comput.*, 16:407–421.
- Hurlbert, A., 1986, Formal connections between lightness algorithms, *J. Opt. Soc. Am. A*, 3:1684–1693. ♦
- Kellman, P. J., and Shipley, T. F., 1991, A theory of visual interpolation in object perception, *Cognit. Psychol.*, 23:141–221. ♦
- Kersten, D., 1991, Transparency and the computation of scene attributes, in *Computational Models of Visual Processing* (M. S. Landy and J. A. Movshon, Eds.), Cambridge, MA: MIT Press, pp. 209–228. ♦
- Land, E. H., and McCann, J. J., 1971, Lightness and Retinex theory, *J. Opt. Soc. Am.*, 61:1–11.
- Lee, T. S., 1995, A Bayesian framework for understanding texture segmentation in the primary visual cortex, *Vision Res.*, 35:2643–2657.
- McArthur, J. A., and Moulden, B., 1999, A two-dimensional model of brightness perception based on spatial filtering consistent with retinal processing, *Vision Res.*, 39:1199–1219.
- Neumann, H., and Mingolla, E., 2001, Computational neural models of spatial integration in perceptual grouping, in *From Fragments to Objects: Grouping and Segmentation in Vision* (T. F. Shipley and P. J. Kellman, Eds.), Amsterdam: Elsevier, pp. 354–400. ♦
- Paradiso, M. A., and Nakayama, K., 1991, Brightness perception and filling-in, *Vis. Res.*, 31:1221–1236.
- Parent, P., and Zucker, S., 1989, Trace inference, curvature consistency, and curve detection, *IEEE Trans. Pattern Anal. Machine Intell.*, 11:823–839.
- Ross, W. D., Grossberg, S., and Mingolla, E., 2000, Visual cortical mechanisms of perceptual grouping: Interacting layers, networks, columns, and maps, *Neural Netw.*, 13:571–588.
- Ross, W. D., and Pessoa, L., 2000, Lightness from contrast: A selective integration model, *Percept. Psychophys.*, 62:1160–1181.

Convolutional Networks for Images, Speech, and Time Series

Yann LeCun and Yoshua Bengio

Introduction

The ability of multilayer backpropagation networks to learn complex, high-dimensional, nonlinear mappings from large collections of examples makes them obvious candidates for image recognition or speech recognition tasks (see PATTERN RECOGNITION). In the traditional model of pattern recognition, a hand-designed feature extractor gathers relevant information from the input and eliminates irrelevant variabilities. A trainable classifier then categorizes the resulting feature vectors (or strings of symbols) into classes. In this scheme, standard, fully connected multilayer networks can be used as classifiers. A potentially more interesting scheme is to eliminate the feature extractor, feeding the network “raw” inputs (e.g., normalized images) and relying on learning algorithms to turn the first few layers into an appropriate feature extractor. Although this can be done with an ordinary fully connected feedforward network with

some success for tasks such as character recognition, there are problems.

Firstly, typical images, or spectral representations of spoken words, are large, often with several hundred variables. A fully connected first layer with, say a few hundred hidden units would already contain tens of thousands of weights. Overfitting problems may occur if training data are scarce. In addition, the memory requirement for that many weights may rule out certain hardware implementations. But the main deficiency of unstructured nets for image or speech applications is that they have no built-in invariance with respect to translations or local distortions of the inputs. Before being sent to the fixed-size input layer of a neural net, character images, spoken word spectra, or other two-dimensional (2D) or one-dimensional (1D) signals must be approximately size normalized and centered in the input field. Unfortunately, no such preprocessing can be perfect: handwriting is often normalized at the word

level, which can cause size, slant, and position variations for individual characters; and words can be spoken at varying speed, pitch, and intonation. This causes variations in the position of distinctive features in input objects. In principle, a fully connected network of sufficient size could learn to produce outputs that are invariant with respect to such variations. However, learning such a task would probably result in multiple units with identical weight patterns positioned at various locations in the input. Learning these weight configurations requires a very large number of training instances to cover the space of possible variations. Conversely, in convolutional networks, shift invariance is automatically obtained by forcing the replication of weight configurations across space.

Secondly, a deficiency of fully connected architectures is that the topology of the input is entirely ignored. The input variables can be presented in any (fixed) order without affecting the outcome of the training. But images, or spectral representations of speech, have a strong 2D local structure, and time series have a strong 1D structure: variables (or pixels) that are spatially or temporally nearby are highly correlated. Local correlations are the reasons for the well-known advantages of extracting and combining *local* features before recognizing spatial or temporal objects. Convolutional networks force the extraction of local features by restricting the receptive fields of hidden units to be local.

Convolutional Networks

Convolutional networks combine three architectural ideas to ensure some degree of shift and distortion invariance: local receptive fields, shared weights (or weight replication), and, sometimes, spatial or temporal subsampling. A typical convolutional network for recognizing characters is shown in Figure 1 (from LeCun et al., 1990). The input plane receives images of characters that are approximately size normalized and centered. Each unit of a layer receives inputs from a set of units located in a small neighborhood in the previous layer. The idea of connecting units to local receptive fields on the input goes back to the perceptron in the early 1960s, and was almost simultaneous with Hubel and Wiesel's discovery of locally sensitive, orientation-selective neurons in the cat's visual system. Local connections have been reused many times in neural models of visual learning (see Mozer, 1991; see also NEOCOGNITRON: A MODEL FOR VISUAL PATTERN RECOGNITION). With local receptive fields, neurons can extract elementary visual features such as oriented edges, end points, or corners (or similar features in speech spectrograms). These features are then combined by the higher layers. As stated earlier, distortions or shifts of the input can cause the position of salient features to vary. In addition, elementary feature detectors that are useful on one part of the image are likely to be useful across the entire image. This knowledge can be

applied by forcing a set of units whose receptive fields are located at different places on the image to have identical weight vectors (Rumelhart, Hinton, and Williams, 1986). The outputs of such a set of neurons constitute a *feature map*. At each position, different types of units in different feature maps compute different types of features. A sequential implementation of this, for each feature map, would be to scan the input image with a single neuron that has a local receptive field and to store the states of this neuron at corresponding locations in the feature map. This operation is equivalent to a convolution with a small-size kernel, followed by a squashing function. The process can be performed in parallel by implementing the feature map as a plane of neurons that *share* a single weight vector. Units in a feature map are constrained to perform the same operation on different parts of the image. A convolutional layer is usually composed of several feature maps (with different weight vectors), so that multiple features can be extracted at each location. The first hidden layer in Figure 1 has four feature maps with 5×5 receptive fields. Shifting the input of a convolutional layer will shift the output but leave it unchanged otherwise. Once a feature has been detected, its exact location becomes less important, as long as its approximate position relative to other features is preserved. Therefore, each convolutional layer is followed by an additional layer that performs a local averaging and a subsampling, reducing the resolution of the feature map, and reducing the sensitivity of the output to shifts and distortions. The second hidden layer in Figure 1 performs 2×2 averaging and subsampling, followed by a trainable coefficient, a trainable bias, and a sigmoid. The trainable coefficient and bias control the effect of the squashing nonlinearity (for example, if the coefficient is small, then the neuron operates in a quasi-linear mode). Successive layers of convolutions and subsampling are typically alternated, resulting in a *bi-pyramid* at each layer, the number of feature maps is increased as the spatial resolution is decreased. Each unit in the third hidden layer in Figure 1 may have input connections from several feature maps in the previous layer. The convolution/subsampling combination, inspired by Hubel and Wiesel's notions of "simple" and "complex" cells, was implemented in the neocognitron model, although no globally supervised learning procedure such as backpropagation was available then.

Since all the weights are learned with backpropagation, convolutional networks can be seen as synthesizing their own feature extractor. The weight-sharing technique has the interesting side effect of reducing the number of free parameters, thereby reducing the "capacity" of the machine and improving its generalization ability (see LeCun, 1989, on weight sharing, and LEARNING AND GENERALIZATION: THEORETICAL BOUNDS for an explanation of capacity and generalization). The network in Figure 1 contains about 100,000 connections, but only about 2,600 free parameters because

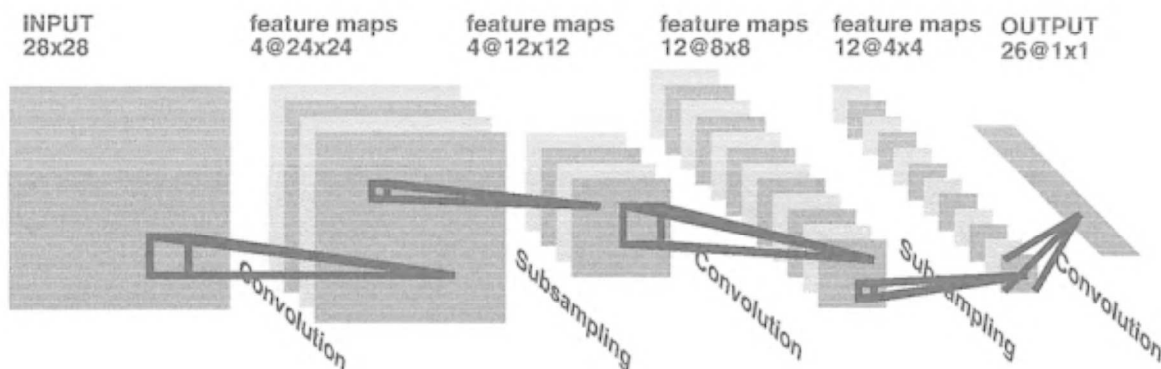


Figure 1. Convolutional neural network for image processing, e.g., handwriting recognition.

of the weight sharing. Such networks compare favorably with other methods on handwritten character recognition tasks (Bottou et al., 1994; see also PATTERN RECOGNITION), and they have been deployed in commercial applications (LeCun et al., 1998).

Fixed-size convolutional networks that share weights along a single temporal dimension are known as time-delay neural networks (TDNNs). TDNNs have been used in phoneme recognition (without subsampling) (Lang and Hinton, 1988; Waibel et al., 1989), spoken word recognition (with subsampling) (Bottou et al., 1990), and on-line handwriting recognition (Guyon et al., 1991).

Variable-Size Convolutional Networks: SDNNs

While characters or short spoken words can be size normalized and fed to a fixed-size network, more complex objects such as written or spoken words and sentences have inherently variable size. One way of handling such a composite object is to segment it heuristically into simpler objects that can be recognized individually (e.g., characters, phonemes). However, reliable segmentation heuristics do not exist for speech or cursive handwriting. A brute force solution is to scan (or replicate) a recognizer at all possible locations across the input. While this can be prohibitively expensive in general, convolutional networks can be scanned or replicated very efficiently over large, variable-size input fields. Consider one instance of a convolutional net and its alter ego at a nearby location. Because of the convolutional nature of the networks, units in the two nets that look at identical locations on the input have identical outputs; therefore their output does not need to be computed twice. In effect, replicating a convolutional network can be done simply by increasing the size of the field over which the convolutions are performed and replicating the output layer, effectively making it a convolutional layer (see Figure 2). An output whose receptive field is centered on an elementary object will produce the class of this object, while an in-between output may be empty or may contain garbage. The outputs can be interpreted as evidence for the categories of objects centered at different positions of the input field. A postprocessor is therefore required to pull out consistent interpretations of the output. Hidden Markov models (HMMs) or other graph-based methods are often used for that purpose (see LeCun et al., 1998; see also HIDDEN MARKOV MODELS, SPEECH RECOGNITION TECHNOLOGY, and PATTERN RECOGNITION). The replicated network and the HMM can be trained simultaneously by backpropagating gradients through the HMM. Globally trained, variable-size TDNN/HMM hybrids have been used for speech recognition (see PATTERN RECOGNITION for a list of references) and on-line handwriting recognition (Schenkel et al., 1993). Two-dimensional

replicated convolutional networks, called *space displacement neural networks* (SDNNs), have been used in combination with HMMs or other elastic matching methods for handwritten word recognition (Keeler, Rumelhart, and Leow, 1991; Bengio, LeCun, and Henderson, 1994). Another interesting application of SDNNs is in object spotting (Wolf and Platt, 1994).

An important advantage of convolutional neural networks is the ease with which they can be implemented in hardware. Specialized analog/digital chips have been designed and used in character recognition and in image preprocessing applications (Boser et al., 1991). Speeds of more than 1,000 characters per second were obtained with a network with around 100,000 connections (shown in Figure 1).

The idea of subsampling can be turned around to construct networks that are similar to TDNNs but can generate sequences from labels. These networks are called reverse TDNNs because they can be viewed as upside-down TDNNs: temporal resolution increases from the input to the output, through alternated oversampling and convolution layers (Simard and LeCun, 1992).

Discussion

Convolutional neural networks are a good example of an idea inspired by biology that resulted in competitive engineering solutions that compare favorably with other methods (Bottou et al., 1994; LeCun et al., 1998). Although applying convolutional nets to image recognition removes the need for a separate hand-crafted feature extractor, normalizing the images for size and orientation (if only approximately) is still required. Shared weights and subsampling bring invariance with respect to small geometric transformations or distortions, but fully invariant recognition is still beyond reach. Radically new architectural ideas, possibly suggested by biology, will be required for a fully neural image or speech recognition system.

Road Maps: Learning in Artificial Networks; Linguistics and Speech Processing

Related Reading: Feature Analysis; Hidden Markov Models; Pattern Recognition; Speech Recognition Technology

References

- Bengio, Y., LeCun, Y., and Henderson, D., 1994, Globally trained handwritten word recognizer using spatial representation, space displacement neural networks and hidden Markov models, in *Advances in Neural Information Processing Systems 6* (J. Cowan, G. Tesauro, and J. Alspector, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 937–944.
- Boser, B., Sackinger, E., Bromley, J., LeCun, Y., and Jackel, L., 1991, An analog neural network processor with programmable topology, *IEEE J. Solid-State Circuits*, 26:2017–2025.
- Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P., and Vapnik, V., 1994, Comparison of classifier methods: a case study in handwritten digit recognition, in *Proceedings of the International Conference on Pattern Recognition*, Los Alamitos, CA: IEEE Computer Society Press.
- Bottou, L., Fogelman-Soulie, F., Blanchet, P., and Lienard, J. S., 1990, Speaker independent isolated digit recognition: Multilayer perceptrons vs dynamic time warping, *Neural Netw.*, 3:453–465.
- Guyon, I., Albrecht, P., LeCun, Y., Denker, J. S., and Hubbard, W. H., 1991, Design of a neural network character recognizer for a touch terminal, *Pattern Recognit.*, 24:105–119.
- Keeler, J., Rumelhart, D., and Leow, W., 1991, Integrated segmentation and recognition of hand-printed numerals, in *Advances in Neural Information Processing Systems 3* (R. P. Lippmann, J. M. Moody, and D. S. Touretzky, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 557–563.
- Lang, K., and Hinton, G., 1988, *The Development of the Time-Delay Neural Network Architecture for Speech Recognition*, Technical Report CMU-CS-88-152, Carnegie-Mellon University, Pittsburgh, PA.

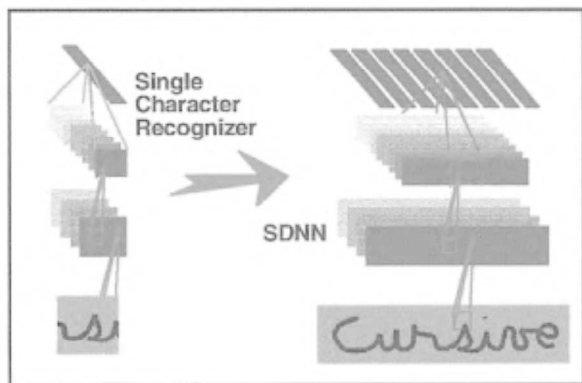


Figure 2. Variable-size replicated convolutional network called a space-displacement neural network (SDNN).

- LeCun, Y., 1989, *Generalization and Network Design Strategies*, Technical Report CRG-TR-89-4, Department of Computer Science, University of Toronto.
- LeCun, Y., Boser, B., Denker, I., Henderson, D., Howard, R., Hubbard, W., and Jackel, L., 1990, Handwritten Digit Recognition with a Back-Propagation Network, in *Advances in Neural Information Processing Systems 2* (D. Touretzky, Ed.), pages 396–404, Morgan Kaufmann, San Mateo.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998, Gradient based learning applied to document recognition, *Proc. IEEE*, 86:2278–2324.
- Moser, M., 1991, *The Perception of Multiple Objects: A Connectionist Approach*, Cambridge, MA: MIT Press.
- Rumelhart, D., Hinton, G., and Williams, R., 1986, Learning representations by back-propagating errors, *Nature*, 323:533–536.
- Schenkel, M., Weissman, H., Guyon, I., Nohl, C., and Henderson, D., 1993, Recognition-based segmentation of on-line hand-printed words, in *Advances in Neural Information Processing Systems 5* (C. Hanson and L. Giles, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 723–730.
- Simard, P., and LeCun, Y., 1992, Reverse TDNN: An architecture for trajectory generation, in Moody, J., Hanson, S., and Lipmann, R., editors, *Advances in Neural Information Processing Systems 4* (J. Moody, S. Hanson, and R. P. Lipmann, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 579–588.
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., 1989, Phoneme recognition using time-delay neural networks, *IEEE Trans. Acoustics Speech Signal Process.*, 37:328–339.
- Wolf, R., and Platt, J., 1994, Postal address block location using a convolutional locator network, in *Advances in Neural Information Processing Systems 6* (J. Cowan, G. Tesauero, and J. Alspector, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 745–752.

Cooperative Phenomena

Hermann P. J. Haken

Introduction

Most objects of scientific study in physics, chemistry, biology, neuroscience, and many other fields are composed of many individual parts that interact with each other. By their interaction, the individual parts may produce cooperative phenomena that are connected with the emergence of new qualities that are not present at the level of the individual subsystems. For instance, the human brain consists of a network of some 100 billion neurons that, through their cooperation, bring about pattern recognition, associative memory, decision making, steering of locomotion, speech, emotions, and so on. Physical systems may serve as model systems or as testing grounds for the development of new concepts and mathematical tools. One important class comprises systems in thermal equilibrium that may undergo a phase transition when the temperature changes. A typical example is water freezing to ice. At the microscopic level, where we are concerned with the motion of individual molecules, molecules above freezing temperature exhibit disordered movement. Below freezing temperature, also called the *critical temperature*, water molecules take their positions in the highly ordered ice crystal. Incidentally, at the macroscopic level, the properties of ice are different from those of water, as is quite evident in the mechanical properties of the two substances. Another example is afforded by ferromagnets. Ferromagnets are composed of many individual elementary magnets that change their orientation randomly but interact to align with each other below a critical temperature, the *Curie temperature*, and may thus produce a macroscopic magnetization typical for ferromagnets. While in metals at a somewhat elevated temperature the electrons move entirely independently of each other, in superconducting metals below a critical temperature the electrons form pairs. At the macroscopic level, the electrical resistance drops to zero, and the metals become superconducting. In both of these cases the macroscopic properties of the system change dramatically when the temperature of the system is changed from the outside and passes through a critical value. Theories that address these phase transitions were originally developed by Landau (cf. Landau and Lifshitz, 1959) and later by K. G. Wilson (1971) and others.

In contrast to biological systems, whose functioning is maintained by a continuous flux of energy or matter through them, the physical systems just described are truly dead. There are, however, physical systems whose spatial, temporal, or spatiotemporal structures are produced and maintained by a continuous influx of energy

and/or matter and which may show phase-transition-like phenomena. Thus, they seem suited to act as model systems for biological systems (including the brain), and they may also guide the development of new types of neural nets. Typical examples are provided by lasers, fluids, plasmas, and semiconductors.

Let us consider a solid state laser (Haken, 1983). In a solid matrix, laser-active atoms are embedded that are excited (“pumped”) by, for example, light from other lamps. When the excitation level is low, the individual atoms emit their light independently of each other so that microscopically chaotic—i.e., irregular—light waves emerge. If the pump power exceeds a critical value, the properties of the light change dramatically. It is now composed of a single, almost infinitely long wave track that shows only minor fluctuations in phase and amplitude. In the laser, the emission acts of the electrons of the individual atoms have become correlated to produce the collective phenomenon of coherent light. This ordering phenomenon is brought about by the system itself, not by an outside agent interfering with the system, and for this reason this ordering phenomenon is called *self-organization*. The basic mechanism for the emergence of a single coherent light wave is as follows: When a light wave has been emitted from an excited atom, it may hit another excited atom and force that atom to enhance the impinging light wave by the process of stimulated emission (as formulated by Einstein). When a number of atoms are excited, an avalanche of that light wave is generated. Again and again such avalanches are generated, and thus start to compete with each other. The one that has the largest growth rate wins the competition and establishes the laser light wave. Because the energy supply to the system is limited, the light wave eventually saturates. The light wave so established forces the individual electrons of the atoms into its rhythm (synchrony) and thus forces them to support it. In the terminology of synergetics (Haken, 1983), the light wave acts as the *order parameter*. This is a variable that describes the macroscopic order of the system and gives orders to the individual parts of it. In the laser, the order parameter *enslaves* the electrons of the atoms.

When the pump power to the laser is further increased, new effects may appear. For instance, several coherent waves may be produced simultaneously and may lead to specific spatiotemporal intensity distributions; ultrashort regular light pulses may occur; or light may show deterministic chaos. Thus, a single, rather unspecific input variable—namely, the pump power—controls the self-

organization of the system, resulting in the production of entirely new temporal or spatiotemporal structures. This input variable is called the *control parameter*. Note the difference between the concepts of order parameter and control parameter: a *control parameter* is a quantity that is imposed on the system from the outside, whereas an *order parameter* is established by the system itself via self-organization.

A variety of similar phenomena related to the formation of spatiotemporal patterns can be observed in fluid layers heated from below or from above, in fluids between two vertical, coaxial, rotating cylinders, and in semiconductors that are driven away from thermal equilibrium. Additional examples of pattern formation can be found in specific chemical reactions. All of these structure-forming processes in nonequilibrium systems are studied in the interdisciplinary field of synergetics, which affords a unified approach to such processes.

Outline of the Mathematical Approach

In order to make visible the profound analogies between the formation of patterns by quite different systems, and to prepare the ground for establishing an important analogy between pattern formation and pattern recognition, we have to adopt a rather abstract level of formulation (Haken, 1983). To describe a system at the microscopic level, we introduce the state vector

$$\mathbf{q} = (q_1, q_2, \dots, q_n) \quad (1)$$

In the example of a laser, the individual components q_j may stand for the time-dependent field amplitudes used in a decomposition of the electric field strength of the laser light into so-called modes. The modes are typically standing or running sinusoidal waves that fit in between the mirrors of the laser. Further components q_j may stand for the dipole moments of the individual atoms and for the inversion (i.e., degree of excitation) of the atoms. In fluids, q_j denotes the density, the components of the velocity field, and the temperature field. In semiconductors, q_j stands for the densities of electrons, holes, impurity centers, and the electric field. In these cases, the components are both space and time dependent. In models of chemical reactions, q_j may stand for the concentration of a chemical of kind j . In general, the state vector develops in the course of time; this time evolution is described by evolution equations of the form

$$\dot{\mathbf{q}} = \mathbf{N}(\mathbf{q}, \nabla, \alpha) + \mathbf{F}(t), \quad (2)$$

where α represents one or several control parameters. The left-hand side is the temporal derivative of the state vector \mathbf{q} , which is determined by the right-hand side of this equation. \mathbf{N} is a nonlinear function of the state vector. The state vector may be subjected to spatial differential operations $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$. Finally, \mathbf{F} represents stochastic forces that stem from internal or external fluctuations. In chemical reactions the typical reaction-diffusion equations are of the form

$$\dot{\mathbf{q}} = \mathbf{N}(\mathbf{q}, \alpha) + \mathbf{F}(t) + D\Delta\mathbf{q} \quad (3)$$

where D is a diffusion matrix and $\Delta = \nabla^2$ is the Laplace operator.

Equation 2 in this general form covers an enormous range of phenomena, and at first glance it appears impossible to devise a general method of solution. From the experimental point of view, however, we are often confronted with situations like the following. Below a certain pump power threshold, a laser acts as a lamp without the emission of coherent light. When we slowly increase the pump power, suddenly the laser forms coherent laser light. In

other words, the former state has become unstable and has been replaced by a new state. This suggests the following strategy: We assume that, for a given control parameter value α_0 , a solution \mathbf{q}_0 of Equation 2 (with $\mathbf{F} \equiv 0$) is known. The general procedure allows us to treat all kinds of \mathbf{q}_0 as such a reference state; \mathbf{q}_0 may be time-independent (representing a fixed point), time-periodic (representing a limit cycle), time-quasi-periodic (forming a torus), or time-chaotic (forming a chaotic attractor). Common features and differences with respect to bifurcation theory, an important branch of dynamic systems theory, are worth mentioning. Dynamic systems theory also considers bifurcation from a fixed point and from time-periodic solutions; however, it neglects the very important impact of fluctuations as well as dynamical behavior (i.e., relaxation processes), whereas in synergetics these phase-transition effects are included, as is the bifurcation from quasi-periodic and chaotic reference states \mathbf{q}_0 . Here we explicitly consider the case of a fixed-point \mathbf{q}_0 . To check the stability of the state \mathbf{q}_0 when we change the control parameter, we make the hypothesis that, for α close to α_0 , the state \mathbf{q}_α can be written as

$$\mathbf{q}_\alpha = \mathbf{q}_0 + \mathbf{w}(\mathbf{x}, t) \quad (4)$$

where \mathbf{w} is assumed to be a small quantity. We may thus insert Equation 4 into Equation 2. In the resulting equation for \mathbf{w} (still with $\mathbf{F} \equiv 0$), we keep only the linear terms and obtain

$$\dot{\mathbf{w}} = L(\mathbf{q}_0)\mathbf{w} \quad (5)$$

where L is a linear operator. It can be shown quite generally in all the previously mentioned cases of \mathbf{q}_0 that the solutions \mathbf{w} can be written in the form

$$\mathbf{w}(\mathbf{x}, t) = \begin{cases} e^{\lambda t} \mathbf{v}_u(\mathbf{x}, t), & \text{Re } \lambda \geq 0 \\ e^{\lambda t} \mathbf{v}_s(\mathbf{x}, t), & \text{Re } \lambda < 0 \end{cases} \quad (6)$$

where we distinguish between the two sets of modes: *unstable modes*, with $\text{Re } \lambda \geq 0$, and *stable modes*, with $\text{Re } \lambda < 0$ (Re = real part of). Here, \mathbf{v} depends on t in a way that is weaker than an exponential growth or decay. The λ 's are the eigenvalues of Equation 5. Any linear combination of Equation 6 is, of course, again a solution of Equation 5. In what follows, it, however, will be crucial to treat the solutions in Equation 6 individually and to distinguish between those eigenvalues λ whose real part is positive or zero and those whose real part is negative. It is our goal to solve the nonlinear stochastic Equation 2 exactly, i.e., not only in a linear approximation. To this end, we expand the unknown function \mathbf{q} into a superposition of the individual eigenfunctions \mathbf{v} of the linear operator L in Equation 5:

$$\mathbf{q}(\mathbf{x}, t) = \mathbf{q}_0 + \sum_u \xi_u(t) \mathbf{v}_u(\mathbf{x}) + \sum_s \xi_s(t) \mathbf{v}_s(\mathbf{x}) \quad (7)$$

In the mathematical sense, this is a complete superposition representing \mathbf{q} as a function of \mathbf{x} . The amplitudes ξ_u and ξ_s are still unknown functions of time. To obtain equations for ξ_u , ξ_s , we insert the expansion of Equation 7 into Equation 2. On the right-hand side of Equation 2 we expand the nonlinear function \mathbf{N} that has become a function of ξ_u and ξ_s into a power series of ξ_u and ξ_s . The terms *linear* in ξ_u or ξ_s read $\xi_u L(\mathbf{q}_0) \mathbf{v}_u$ or $\xi_s L(\mathbf{q}_0) \mathbf{v}_s$, respectively. Because of Equation 5, in the case of a fixed point \mathbf{q}_0 , we may replace, for instance, $L \mathbf{v}_u$ by $\lambda_u \mathbf{v}_u$.

Keeping these and all higher-order terms and projecting both sides of Equation 2 on the modes \mathbf{v} , we obtain the following two sets of equations:

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u(\xi_u, \xi_s) + \hat{F}_u(t) \quad (8)$$

and

$$\dot{\xi}_s = \lambda_s \xi_s + \hat{N}_s(\xi_u, \xi_s) + \hat{F}_u(t) \quad (9)$$

The first term on the right-hand side of Equations 8 and 9, respectively, stems from the linear part of N , where use was made of the fact that \mathbf{v}_u and \mathbf{v}_s are eigenfunctions of the linear operator L in Equation 5 with the eigenvalues λ that occur on the right-hand side of Equation 6. These equations are entirely equivalent to the former Equation 2. However, provided the inequality

$$|Re\lambda_s| \gg |Re\lambda_u| \quad (10)$$

holds, the *slaving principle* of synergetics (Haken, 1983) can be applied. According to this principle, the mode amplitudes ξ_s are uniquely determined (enslaved) by ξ_u . The possibility of expressing ξ_s by ξ_u can be made plausible in the following fashion: Let us assume that according to Equation 10, the mode amplitudes ξ_s relax much faster than the mode amplitudes ξ_u . Consider Equation 9 with slowly varying ξ_u that act as driving forces on ξ_s . When we neglect transients, ξ_s being driven by ξ_u changes much more slowly than it normally would because of the first term on the right-hand side in Equation 10. In other words, this means that ξ_s can be neglected against $\lambda_s \xi_s$, or that $\dot{\xi}_s = 0$. This turns Equation 9 into an algebraic equation that can be solved for ξ_s , expressing ξ_s by ξ_u . This approximation is called *adiabatic approximation*. The slaving principle ensures that this procedure is a first step toward a systematic procedure by which ξ_s can be expressed uniquely and exactly by ξ_u and \hat{F}_s :

$$\xi_s(t) = f_s(\xi_u(t), t) \quad (11)$$

The explicit dependence of f_s on t stems from the time dependence of the fluctuating forces, but not of that of the amplitudes ξ_u . In most cases of practical interest, f_s can be approximated by a low-order polynomial in ξ_u .

In practically all cases that have been treated in the literature, the systems are of very high dimension—i.e., they contain very many variables—but the number of unstable mode amplitudes ξ_u is very small. The amplitudes ξ_u are called *order parameters*, whereas the variables ξ_s can be called *enslaved variables*. The order parameter concept allows an enormous reduction in the degrees of freedom. Think of a single-mode laser in which we have one mode and, say, 10^{16} degrees of freedom stemming from the dipole moments and inversions of the individual laser atoms. The order parameter in the single-mode laser is identical with the lasing mode, i.e., a single degree of freedom. Once we have expressed the enslaved modes by means of the order parameters ξ_u via Equation 11, we may insert Equation 11 into Equation 8 and thus find equations for the order parameters alone:

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u(\xi_u, f_s(\xi_u, t)) + \hat{F}_u(t) \quad (12)$$

In a number of cases, these equations can be put into *universality classes* describing the similar behavior of otherwise quite different systems. The term *universality classes* is chosen in analogy to universality classes in the theory of phase transitions of systems in or close to thermal equilibrium, such as superconductors or ferromagnets, although the classes treated here are of a more general character. In the present context, the term means that Equation 12 can be put into specific general forms (see below). For instance, a single-mode laser, the formation of a roll

pattern in a fluid, or the formation of a stripe pattern in chemical reactions obey the same basic order parameter equation. Such universality classes can be established because of the following facts:

- When we are dealing with a *soft transition* of a system, its order parameters are small, close to the instability point. In analogy to conventional phase transition theory, we call a transition a soft transition if the order parameters change smoothly with the control parameter. This allows us to expand \hat{N}_u into a power series with respect to the order parameters, where we may keep only a few, leading terms.
- Furthermore, we may exploit symmetries. For instance, given a term $\beta \xi_u^2$, if there is an inversion symmetry of the system, it follows that $\beta = 0$. Symmetries lead also in a number of cases to relationships between coefficients.
- Finally, we may invoke so-called normal form theory to simplify the nonlinear functions on the right-hand side of Equation 13.

Some Examples for the Formation of Patterns

We illustrate the procedure just described with a few typical examples encountered in concrete, important cases. In the case of a *single real or complex order parameter*, a typical order parameter equation reads:

$$\dot{\xi}_u = \lambda \xi_u - \beta |\xi_u|^2 \xi_u + F(t) \quad (13)$$

where λ plays the role of a control parameter. The state vector reads

$$\mathbf{q}(\mathbf{x}, t) = \mathbf{q}_0 + \xi_u(t) \mathbf{v}_u(\mathbf{x}) + \sum_s \xi_s(t) \mathbf{v}_s(\mathbf{x}) \quad (14)$$

where the sum over the enslaved modes, s , is in general small, so that the evolving pattern is determined by the second term on the right-hand side, which is called the mode skeleton because it represents the basic features of the evolving pattern. If the system is originally spatially homogeneous, \mathbf{q}_0 does not depend on the space coordinate, and $\mathbf{v}_u(\mathbf{x})$ as a solution of Equations 5 and 6 is basically a sine function. If λ is real, a spatial stripe pattern is formed (stripes in chemical reactions, rolls in fluids, current filaments in semiconductors, etc.). If λ is complex, a coherent (and spatially modulated) wave emerges (single-mode laser). The stochastic force F leads to amplitude fluctuations and phase diffusion.

In the case of *two order parameters*, the mode skeleton is determined by

$$\mathbf{q}(\mathbf{x}, t) = \mathbf{q}_0 + \xi_1(t) \mathbf{v}_1(\mathbf{x}) + \xi_2(t) \mathbf{v}_2(\mathbf{x}) \quad (15)$$

Depending on the order parameter Equation 12, modes \mathbf{v}_u may either coexist or compete, with the result that only one remains. In the case of competition, either ξ_1 or ξ_2 vanishes, and a stripe pattern appears. In the case of coexistence, both are nonvanishing, so that the total pattern becomes a superposition of the patterns \mathbf{v}_1 and \mathbf{v}_2 (square pattern).

A further example for pattern formation in physics is provided by the computer simulation shown in Figure 1, where a liquid in a circular vessel is heated from below. Depending on a prescribed initial state, different stripe patterns may evolve, where in one stripe the fluid is up- and downwelling. In such a case the fluid has the property of an associative memory; i.e., an incomplete set of data (one stripe) is supplemented automatically by the system to a full stripe pattern.

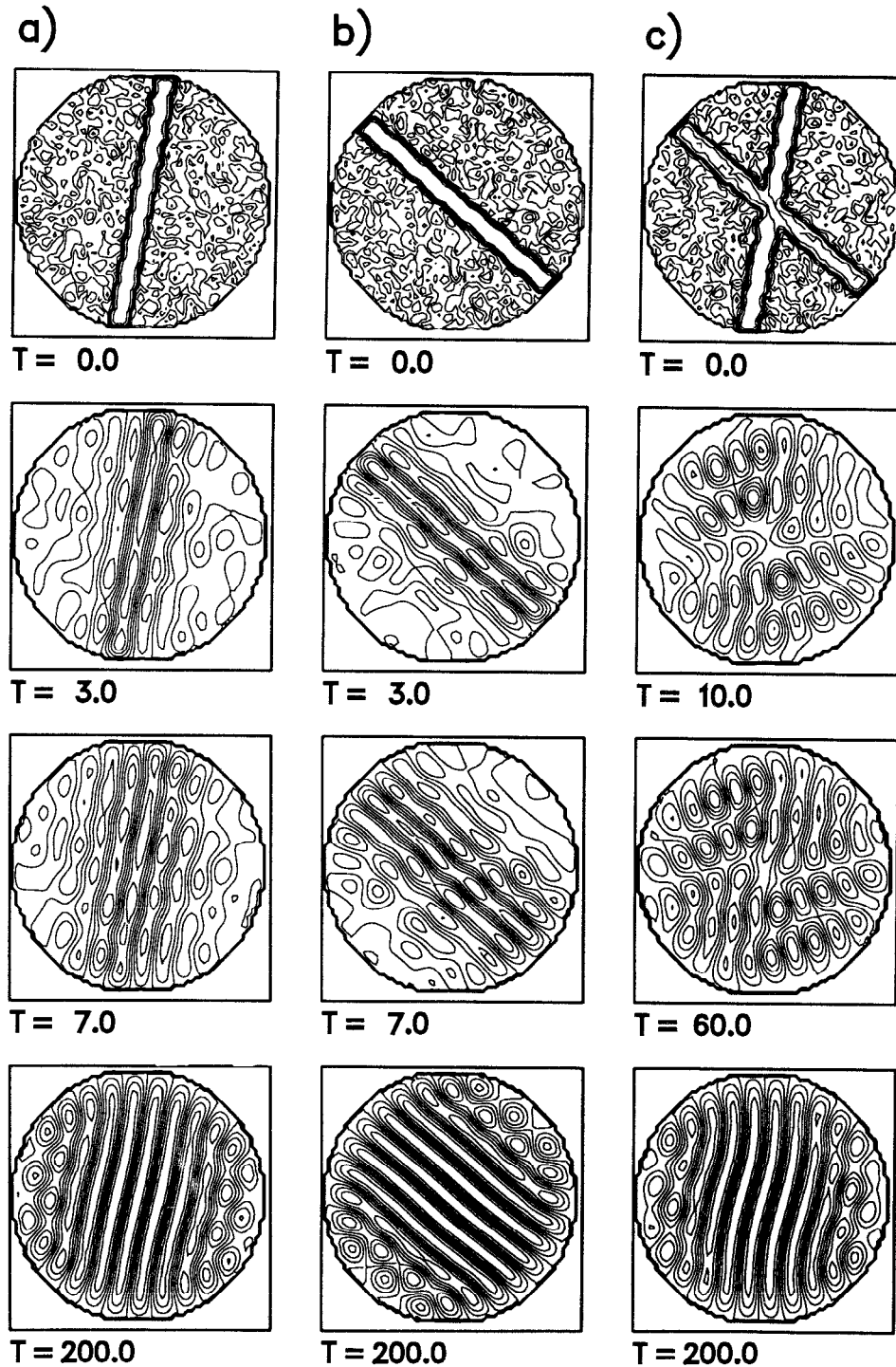


Figure 1. Computer simulation of a fluid heated from below in which initially one upwelling stripe is given. *Left column*, Completion of the single stripe to a full stripe pattern in the course of time. *Middle column*, The same as before, but with a different orientation of the initial stripe. *Right column*, Two initially given upwelling stripes, where one stripe is somewhat stronger than the other one. In the course of time, the originally stronger stripe wins the competition and determines the whole stripe pattern.

Phenomenological and Macroscopic Approaches

In many complex systems, such as biological systems, the basic microscopic evolutions in Equation 2 are not or not fully known.

Nevertheless, in this case also we may capitalize on the results shown earlier under the assumption that a nonequilibrium phase transition happens. This is the case if the macroscopic behavior of the system studied changes qualitatively. In such a case, we may

phenomenologically postulate order parameter equations. We may also analyze spatiotemporal patterns by invoking the decomposition formulated in Equation 7 and simple forms of the order parameter equations and using experimental data on evolving or changing spatiotemporal patterns.

Cooperative Phenomena in Neuroscience I

Pattern Recognition

Cooperative phenomena according with the models and interpretations discussed to this point abound in neuroscience. In all of these cases the brain is understood as a self-organizing (synergetic) system. It is possible that important aspects of brain function can be described in terms of order parameters and enslavement.

We have just seen that synergetic systems may act as an associative memory (Kohonen, 1987). This allows us to devise an algorithm (or model) for pattern recognition based on an analogy between pattern formation and pattern recognition (Haken, 1991). In both cases, an incomplete set of data (fluid: a single stripe; pattern recognition: part of a pattern) is completed to a full set of data (fluid: complete stripe pattern; pattern recognition: complete pattern) by means of order parameters and the slaving principle. The algorithm can be formulated as follows: We consider a set of prototype patterns stored in our system. These patterns are represented by vectors $\mathbf{v}^{(k)}$ of a high-dimensional space, where one component $v_i^{(k)}$ corresponds to the gray value or color value of a specific pixel, i , of a pattern labeled by an index k . In the same

way we encode a starting or test vector \mathbf{q} (a pattern to be recognized).

By means of these vectors $\mathbf{v}^{(k)}$ we construct a dynamics for the test vector \mathbf{q} in the following sense: The test vector \mathbf{q} becomes a time-dependent quantity undergoing a gradient dynamics in a potential field, which may be visualized as a mountainous landscape. This potential field possesses those and only those minima that correspond to the stored prototype pattern vectors $\mathbf{v}^{(k)}$. Note that this approach avoids the well-known difficulty of a number of neural networks, in particular of the Hopfield type (Hopfield, 1982), in which the system can be trapped in spurious minima that do not correspond to any stored patterns (compare the concept of the Boltzmann machine). Here the dynamical system leads to an identification of prototype patterns without the need to introduce simulating annealing, i.e., a statistical pushing of the test vector \mathbf{q} (see SIMULATED ANNEALING AND BOLTZMANN MACHINES).

In Figure 2, the associative property of the dynamical system is shown using three different initial conditions: part of a face, the name of a pattern, and a pattern that is disturbed by noise. In every case there is complete restoration of the original prototype. The dynamics of a "synergetic computer" can be interpreted or realized by means of a parallel network in which each component q_j of \mathbf{q} represents the activity of a model neuron j . By means of the hypothesis

$$\mathbf{q}(t) = \sum_{k=1}^M \xi_k(t) \mathbf{v}^{(k)} + \mathbf{w}(t) \quad (16)$$

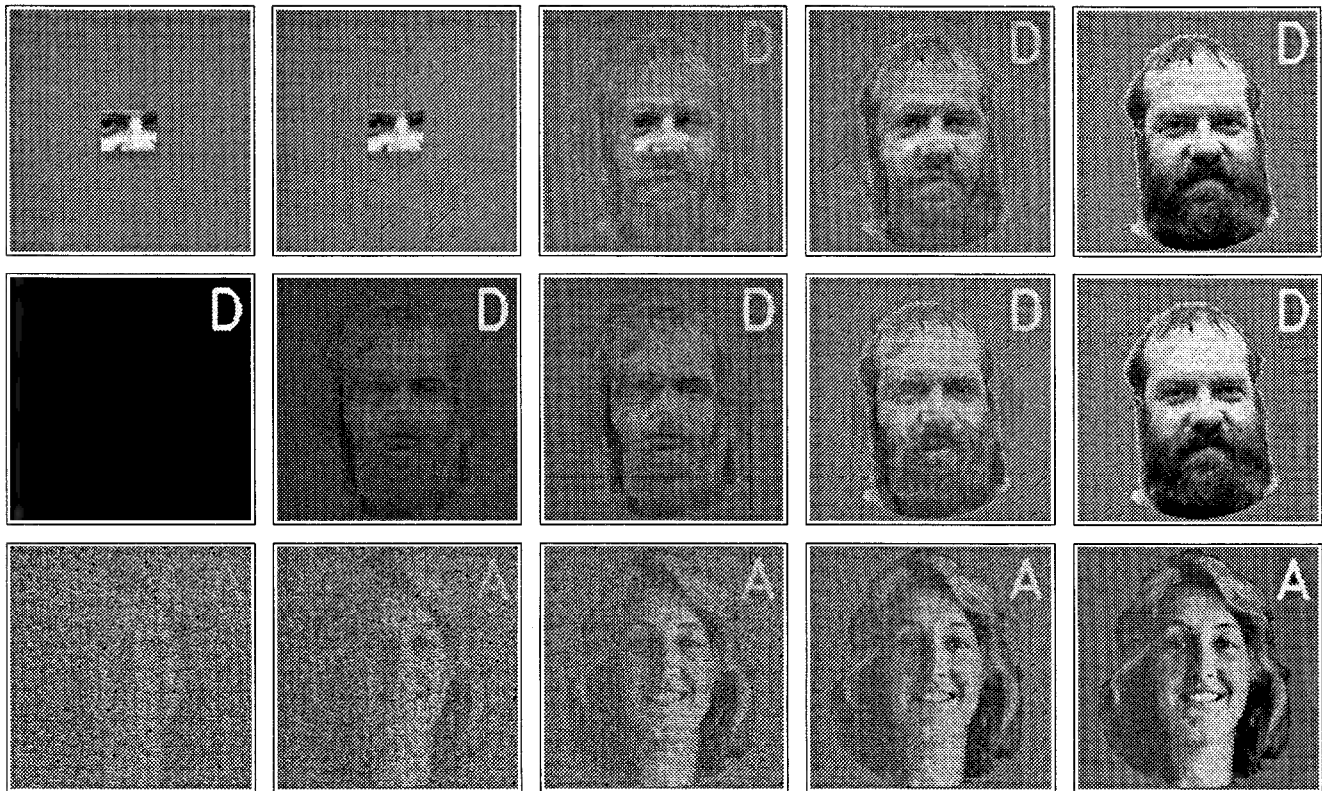


Figure 2. Upper row, Restoration of a full face from part of it. Middle row, Restoration of a full face from the single letter coding for the name. Bottom row, Restoration of a face from an originally noisy face.

where \mathbf{w} is a residual vector, the dynamics of \mathbf{q} can be transformed into the order-parameter equations

$$\dot{\xi}_k = \lambda_k \xi_k - B \sum_{k' \neq k}^M \xi_{k'}^2 \xi_k - C \sum_{k'=1}^M \xi_{k'}^2 \xi_k \quad (17)$$

The attention parameters λ_k play the role of control parameters and serve to amplify each order parameter, while the second term serves for a discrimination and the last term for a saturation of the order parameters. In this way, an individual order parameter ξ_k is attached to each perceived pattern. The formalism allows one to treat time-dependent attention parameters and the properties of the perception of ambiguous patterns (Haken, 1991, 1996). The formalism can be extended to model stereovision (cf. Haken, 1996).

Thus far we have described a formal model (algorithm) that can be implemented on a serial computer or realized by some parallel devices. But we can go a step farther. The “enslaved” parts are the neurons with their activities, while the order parameters are interpreted as the percepts. In this “synergetics” approach, a connection between the microscopic level, described by the individual parts of a system, and the macroscopic level, described by order parameters, is established. This separation is made possible by a separation of time scales for the reaction of the parts and of the order parameters. Whereas in the brain, the typical time constants of neurons are on the order of milliseconds, those of the brain’s macroscopic performance, such as recognition and speech, are on the order of hundreds of milliseconds. At the microscopic level we may mention in particular the neural network model by McCulloch and Pitts (1943), who modeled the individual neurons as two-state elements. A fruitful subsequent step was recognition of the analogy between the McCulloch-Pitts network and spin glasses, which allowed Hopfield (1982) to introduce an energy function for that network, and which gave rise to an avalanche of studies, in particular by theoretical physicists.

Cooperative Phenomena in Neuroscience II

EEG, MEG, Movement Coordination, Hallucinations

An important and very comprehensive model of brain action in terms of neurons (spike rates of action potentials) was established by Wilson and Cowan (1972), who solved their equations numerically. The spatiotemporal patterns found by the authors (for instance, to model hallucinations) can be, at least qualitatively, re-derived in the terms set out in this article. Further evidence for the occurrence of adequate order parameters in brain activities is as follows:

1. The identification of low-dimensional chaos (describable by order parameters) by Babloyantz (1985).
2. The identification of low-dimensional attractors in the olfactory bulb by Freeman (1975).
3. The analysis of petit mal epilepsy, describable by Shilnikov chaos (Friedrich and Uhl, 1992).
4. The MEG analysis of sensorimotor coordination by Kelso (Kelso, Fuchs, and Haken, 1992; see also Haken, 1996). In this work, MEG changes (in both frequency and time domains) in response to auditory and visual stimuli are revealed.
5. The analysis of movement coordination by Haken, Kelso, and Bunz (1985); see also Haken, 1996.

Because in some EEG and MEG measurements, multi-electrode or squid derivations were possible, the spatiotemporal patterns could be determined and, in particular, the basic modes in the sense of a mode decomposition could be identified. In a number of experiments, a surprisingly low number of dominating modes and thus order parameters could be found.

In conclusion, the strategy of searching for order parameters describing brain functions has found some justification, but considerable work remains to be done. Important work about connecting levels of description of cortical and behavioral dynamics in a systematic way has been done by Kelso’s group (Kelso, Fuchs, and Jirsa, 1999).

Road Map: Dynamic Systems

Background: Self-Organization and the Brain

Related Reading: EEG and MEG Analysis; Pattern Formation, Biological; Pattern Formation, Neural

References

- Babloyantz, A., 1985, Evidence of chaotic dynamics of brain activity during the sleep cycle, in *Dimensions and Entropies in Chaotic Systems* (G. Mayer-Kress, Ed.), Berlin and New York: Springer-Verlag.
- Freeman, W., 1975, *Mass Action in the Nervous System: Examination of the Neurophysiological Basis of Adaptive Behavior Through the EEG*, San Diego, CA: Academic Press.
- Friedrich, R., and Uhl, C., 1992, Synergetic analysis of human electroencephalograms: Petit-mal epilepsy, in *Evolution of Dynamical Structures in Complex Systems* (R. Friedrich and A. Wunderlin, Eds.), Berlin and New York: Springer-Verlag.
- Haken, H., 1983, *Synergetics: An Introduction*, 3rd ed., Berlin and New York: Springer-Verlag. ♦
- Haken, H., 1991, *Synergetic Computers and Cognition*, Berlin and New York: Springer-Verlag.
- Haken, H., 1996, *Principles of Brain Functioning*, Berlin and New York: Springer-Verlag. ♦
- Haken, H., Kelso, J. A. S., and Bunz, H., 1985, A theoretical model of phase transitions in human hand movements, *Biol. Cybern.*, 51:347–356.
- Hopfield, J. J., 1982, Neural networks and physical systems with emergent computational abilities, *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.
- Kelso, J. A. S., Fuchs, A., and Haken, H., 1992, Phase transitions in the human brain: Spatial mode dynamics, *Int. J. Bifurcat. Chaos*, 2:917–939.
- Kelso, J. A. S., Fuchs, A., and Jirsa, V. K., 1999, Traversing scales of brain and behavioral organization, in *Analysis of Neurophysiological Brain Functioning* (C. Uhl, Ed.), Berlin and New York: Springer-Verlag, pp. 73–125. ♦
- Kohonen, T., 1987, *Associative Memory and Self-Organization*, 2nd ed., Berlin and New York: Springer-Verlag.
- Landau, D., and Lifshitz, I. M., 1959, in *Course of Theoretical Physics*, vol. 5, *Statistical Physics*, London: Pergamon Press.
- McCulloch, W. S., and Pitts, W. H., 1943, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.*, 5:115–133.
- Wilson, H. R., and Cowan, J. D., 1972, Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.*, 12:1–24. ♦
- Wilson, K. G., 1971, Renormalization group and critical phenomena: I. Renormalization group and the Kadanoff scaling picture, *Phys. Rev. B*, 4:3174–3183; II. Phase-space cell analysis of critical behavior, *Phys. Rev. B*, 4:3184–3205.

Cortical Hebbian Modules

Daniel J. Amit

Introduction

Before describing model networks of cortical Hebbian modules, we will briefly review some experimental findings. Primates are trained to perform a delay match-to-sample (DMS) task or a delay eye-movement (DEM) task in which a relatively large set of stimuli is presented. The behaving monkey must remember sufficient information about the sample (eliciting) stimulus in order to decide on its behavioral response following the presentation of a second (test) stimulus. The test stimulus is, with equal likelihood, identical to or different from the first stimulus in the DMS task, and a “go” signal in the DEM task. Using single-unit extracellular recordings, neurophysiologists have observed elevated spike rates during the delay period, after the eliciting (sample) stimulus was removed. These elevated spike rates are reproducible and occur in areas such as inferotemporal cortex (IT) and prefrontal cortex, which have been suggested to be part of a working memory system (Fuster, 1995; Goldman-Rakic, 1987). Neurons in rather compact columns have been observed to exhibit stimulus-selective elevated rates that can persist for as long as 30 s, a very long time on the scale of neural time constants. The rates observed are in the range of about 10–20 spikes/s, against a background of spontaneous activity of a few per second. The subset of neurons that sustain elevated rates in the absence of a stimulus is selective of the preceding, first, or sample stimulus. The distribution of rates during the delay among the neurons of the cortical module, or the *delay activity distribution* (DAD), could therefore act as a neural representation and an active memory of the identity of the eliciting stimulus, the representation or memory being transmitted for processing later, when the stimulus is no longer present (Amit, 1995).

The DADs appear to be concentrated in localized columns in associative cortex. The estimate is that this column is about 1 mm² in cross-section parallel to the cortical surface. The delay activities corresponding to all the stimuli (as many as 100 in some studies) are constrained to this small module: it contains some 10⁵ cells, and 1%–2% of the cells participate in the DAD of a given stimulus (Brunel, 1996); i.e., 1,000–2,000 cells would propagate elevated rates.

The absence of the external stimulus during delay activity leads to the conclusion that the selective activity distributions must be an expression of autonomous local dynamics in the column, whose substrate either forms during the training process or is innate. Some exegesis is needed here. First, the selective delay activity may, of course, be a result of a structure in the afferents arriving at the column under observation. In other words, the particular column may be a mere readout board. However, the absence of the stimulus implies that the structured activity must be maintained somewhere in the brain, and an autonomous mechanism for sustaining a DAD must exist and be formed by learning in the alternative location. We will assume that it is sustained where it is observed. There is some experimental evidence to support such a position.

A second point concerns the assertion that the local structure underlying the delay dynamics could be formed by learning. This statement is supported by the fact that when new stimuli are added to the set after the appearance of DADs, no delay activity is observed for the new stimuli. The alternative situation may describe prefrontal cortex, where DADs appear to be built-in (Goldman-Rakic, 1987).

These observations raise two preliminary issues concerning possible models to account for the computational aspects: (1) the necessary neural elements and synaptic structures that can reproduce

the observed neuronal spike dynamics, and (2) the synaptic dynamics that can give rise, in the process of training, to a synaptic structure in the local module that is capable of sustaining selective DADs. We will concentrate on the first issue and limit ourselves to a few comments about the second.

A successful treatment of the first issue would conclude with a network of neural elements of a given internal dynamics and synaptic matrices that would lead to neurons emitting spikes in spontaneous activity, at low rates, in a very stable way, and, when stimulated by prelearned stimuli, would maintain selective DADs for a large set of eliciting stimuli. For each stimulus, a small fraction of the cells in the module would have elevated rates, and the rest of the cells would maintain spontaneous activity. The spike activities, both spontaneous and structured, would be rather noisy. In other words, there would be large variability in interspike intervals (Softky and Koch, 1992).

At this point we arrive at the second issue: the effect of neuronal activities on synaptic efficacies, which is at the basis of learning. This issue can be broken down into two major parts for further discussion. One part has to do with the feedforward synaptic dynamics, leading to the formation of the module by afferents from preceding cortical areas, the same module for an entire set of different stimuli. The other part has to do with the formation of col-lateral synapses within the module. The synaptic matrices, which should sustain the various structured activity distributions, must be generated in the process of training via the neuronal activities. If neuronal activities are expressed in terms of spikes, synaptic efficacies must be sensitive to pre- and postsynaptic spikes, in order to affect Hebbian learning, that is, to be potentiated when presynaptic and postsynaptic neurons are simultaneously active and to be depressed when the activities of the two are anticorrelated. There is a tension between the need to learn at least something from every stimulus, so as to accumulate memory, and the need for synaptic stability on long time scales (hours, days, years). Some synapses must change for every stimulus, but the learning process must be immune to spontaneous activity; synapses should not change too easily. Moreover, since neuronal spike dynamics is governed by the emerging synapses, the learning process (synaptic acquisition) may deviate while learning goes on. All of these considerations impose rather severe constraints on both neuronal and synaptic dynamics, over and above the constraints implied by the collective network dynamics (see, e.g., Fusi et al., 2000).

Minimal Elements for a Spiking Network

As a minimal model we take neurons to be of the form described in Equation 1, that is, simple point RC integrators of their afferent currents that emit a spike when the integrated level of depolarization V reaches a threshold θ , followed by an absolute refractory period τ_0 and a resetting of the depolarization to H —an integrate-and-fire (IF) device (see INTEGRATE-AND-FIRE NEURONS AND NETWORKS). Formally, the dynamics of the depolarization is given by

$$\tau \dot{V} = -V + \sum_{i=1}^C J_i \tau \sum_k \delta(t_i^k - t) \quad (1)$$

where τ is the integrator’s time constant (RC of the equivalent circuit). Effectively, the depolarization $V(t)$ is a sum of unit contributions, each with amplitude J_i , over the interval τ . This simplified form presupposes that the dynamics of the synaptic conductances is much faster than that of the depolarization (for extensions,

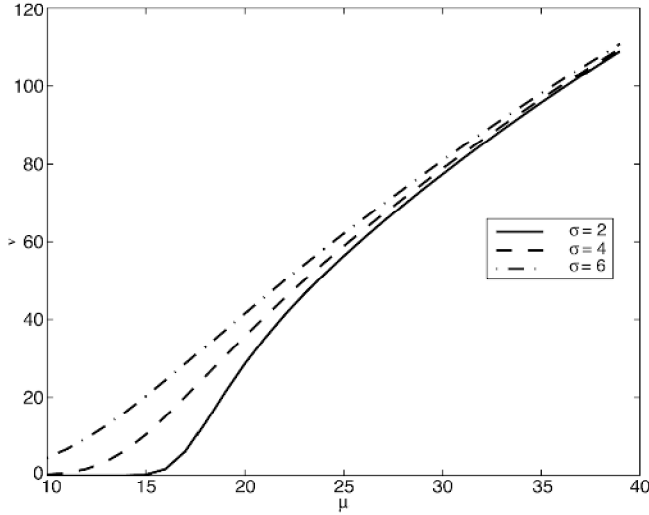


Figure 1. Gain function of integrate-and-fire neuron: rate versus mean afferent current (μ), for several values of the noise σ , for afferent Gaussian current.

see Brunel and Sergi, 1998). In that case the synaptic efficacies J_i are equivalent to the depolarization provoked by a single afferent spike.

The output spike rate of such a neuron is the inverse of the mean time between two consecutive events in which the depolarization reaches threshold. It is the mean first passage time across the threshold. If the neuron is depolarized by a current of Gaussian distribution, uncorrelated at different times, the output rate can be calculated explicitly (Tuckwell, 1988) by the following:

$$v_{\text{out}} = \left(\tau_0 + \tau \int_{(H-\mu)/\sigma}^{(\theta-\mu)/\sigma} du \sqrt{\pi} \exp(u^2) [1 + \text{erf}(u)] \right)^{-1} \equiv \Phi(\mu, \sigma) \quad (2)$$

where τ_0 is the absolute refractory period, μ is the mean of the Gaussian afferents depolarizing the cell per τ , and σ is the corresponding standard deviation (SD).

Note that the response function of this neuron depends not only on the constant part of the afferent current but also on its variance σ , and is an extension of common mean-field results (see, e.g., Amit and Brunel, 1997a). The rate versus the average part of the current (μ) is represented in Figure 1 for several levels of the noise (σ). For low noise, it is zero below a threshold and is convex throughout; with increasing noise a concave region appears, allowing for more than one stable rate. It is the concave part that makes possible the coexistence of spontaneous and selective, elevated spike rate distributions (see, e.g., Fusi and Mattia, 1999; Brunel, 2000).

The assumption of an afferent current of Gaussian distribution is quite useful in cortical conditions. If neurons emit spikes at low rates and receive them via a large number of independent channels (synapses), the total current will be a sum of Poisson processes and hence a Gaussian process. The various assumptions can be verified in microscopic simulations.

The Network

Such minimal elements are put together in a 4:1 ratio of excitatory to inhibitory neuron populations and in large numbers (if not quite as large as the 10^5 of anatomy), in a module that represents a cortical column. A putative assumption for the connectivity pattern would be total randomness, implying that (1) every neuron in the

module can synapse on any other neuron; (2) the set of neurons in the local module synapsing on any given cell is chosen at random, constrained only by a mean connectivity given by anatomy (about 10%); (3) synaptic efficacies (excitatory and inhibitory) are random, centered on some plausible mean for each population of synapses; and (4) spike transmission delays are randomly distributed among the synapses (with an average of a couple of milliseconds).

In addition to collateral connections, neurons receive afferents from outside the module (Figure 2). Those afferents may represent a general arousal level or a selective stimulus. The fraction of such synapses is estimated by Braitenberg and Schüz (1991) to be roughly equal to the collateral one. In the absence of structured stimulation, these afferents will be considered uncorrelated, non-selective, and of a fixed spike rate. Stimulation will consist of attributing to the external afferents of a given subset of neurons in the module a higher rate. This leads to higher response rates in that subset of neurons and corresponds to the “visual response” of neurons observed experimentally.

Estimates of J range from 0.05 to 0.5 mV. If one uses an average of 0.2 mV for the excitatory efficacy and a threshold of 20 mV, some 100 simultaneous excitatory synaptic events are required to bring the postsynaptic neuron to threshold.

The above description of neural elements and connectivity of the network can be studied in simulations once the numerical parameters are set. Spike activity in the simulation can be recorded under different conditions, much as would be done in a neurophysiological experiment. One can also record spikes from single or multiple neurons and observe spike rasters, peristimulus histograms (see Figure 4), and cross-correlations of spike emission times prior to learning, in spontaneous activity states. Then, one can impose a synaptic matrix, presumed to have been formed by learning, and observe the same quantities inside and outside of the subset of selective neurons. These are quantities that can also be sampled by electrodes in the cortex of performing animals.

Extended Mean-Field Theory

When the network is in asynchronous activity states, i.e., when neural spike emissions of different neurons are essentially inde-

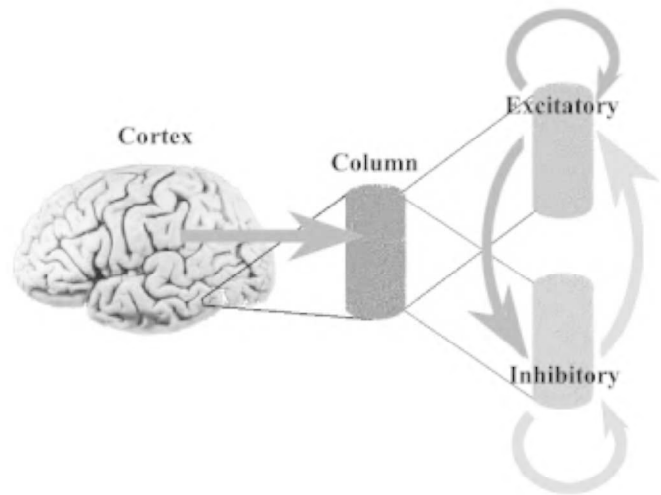


Figure 2. Connectivity scheme. The local network, embedded in the cortex and represented as a column, is divided in two populations of neurons, the excitatory population and the inhibitory population. Neurons in the local network receive collateral connections from excitatory and inhibitory neurons in the same module, as well as from excitatory neurons elsewhere in the cortex.

pendent, there are compact, analytical tools to study a wide variety of properties of large subpopulations of neurons—the collective, computational properties of the network. In other words, if the dynamics of the network leads to large subpopulations within each of which neurons act in a statistically similar way, then the dynamics of the very large network can be described at a level of computational complexity equal to the number of populations that are expected to express different statistical properties. This may appear as a severe restriction. However, one should keep in mind that observable neural phenomena must involve large numbers of neurons acting in essentially the same fashion; otherwise the probability of an electrode detecting a neuron representative of a behavioral correlate would be too low. Similarly, macroscopic probes, as in fMRI or EEG, would not have a large enough signal representing a specific phenomenon. What makes the theory particularly simple is that each neuron (in cortex) receives a very high number of afferents.

The assembly of neurons can be divided into subsets according to neuron types (e.g., excitatory, inhibitory), different sets of afferent synapses (e.g., potentiated or depressed by learning a given stimulus, or not), currently driven by a stimulus or not (e.g., receiving selective external currents), and so on. The network is divided into P subpopulations Π_δ , $\delta = 1, \dots, P$. Each neuron in population δ is assumed to have the same time constant τ_δ , the same rate v_δ , the same mean number of afferents from population γ — $C_{\delta\gamma}$, and the same mean afferent synaptic efficacy $\langle J_\delta \rangle_\gamma$. The mean afferent current to a neuron in population δ can be written as:

$$\mu_\delta = \tau_\delta \sum_\gamma C_{\delta\gamma} \langle J_\delta \rangle_\gamma v_\gamma(t) + \langle I_{\text{ext},\delta} \rangle \quad (3)$$

where the sum is over all populations γ . $I_{\text{ext},\delta}$ is the afferent current from outside the module. Here, for simplicity, we assume the rate to be constant for all neurons inside a population (for extensions, see Amit and Brunel, 1997a).

Spike trains are assumed to be Poissonian, so the variance of the input to a neuron in population δ is given as:

$$\sigma_\delta^2 = \tau_\delta \sum_\gamma C_{\delta\gamma} (\text{Var}(J_\delta)_\gamma + \langle J_\delta \rangle_\gamma^2 v_\gamma(t) + \sigma_{\text{ext}}^2) \quad (4)$$

Writing $\text{Var}(J_\delta) = \Delta \langle J_\delta \rangle^2$ (Δ independent of γ) we obtain:

$$\sigma_\delta^2 = (1 + \Delta) \tau_\delta \sum_\gamma C_{\delta\gamma} \langle J_\delta \rangle_\gamma^2 v_\gamma(t) + \sigma_{\text{ext}}^2 \quad (5)$$

The afferent currents are converted to output rates via the response function, Equation 2:

$$v_{\text{out},\delta} = \Phi(\mu_\delta, \sigma_\delta) \quad (6)$$

For a specified network, μ_δ and σ_δ are functions of the set of rates in the different populations. Hence, if we impose the condition that rates on the left-hand side of Equation 6 be equal to the input rates forming the currents, Equations 6 become a set of self-consistent equations, whose number is equal to the number of populations, for the same number of unknowns: the set of stationary rates in the network. A solution of this set of equations provides a set of rates for the populations selected, and an accompanying study of their stability will determine which of the solutions is an attractor—a DAD. The fact that the feedback depends not only on the mean of the current but also on its variance is where mean-field theory is extended.

The importance of mean-field theory goes well beyond its analytic, compact treatment of the behavior of the spiking network. The theory also allows us to identify the *collective variables* characterizing the dynamics of the system, such variables as can be observed experimentally and by other components of a multimolecular brain.

Dynamic Mean-Field Theory

In addition to the self-consistent equations given in the preceding sections, a very effective tool for testing the stability of the different stationary states is the dynamic extension of mean-field theory. From the depolarization dynamics of Equation 1 one obtains evolution equations for $\mu[V]$ and $\sigma^2[V]$, which are the asymptotic mean and variance of the depolarization in absence of a threshold (Tuckwell, 1988; Amit and Brunel, 1977b):

$$\begin{aligned} \tau_\delta \frac{d\mu_\delta[V(t)]}{dt} &= -\mu_\delta[V(t)] + \tau_\delta \sum_\gamma A_{\delta,\gamma} v_\gamma + \mu_{\delta,\text{ext}} \\ \tau_\delta \frac{d\sigma_\delta^2[V(t)]}{dt} &= -2\sigma_\delta^2[V(t)] + \tau_\delta \sum_\gamma B_{\delta,\gamma} v_\gamma + \sigma_{\delta,\text{ext}}^2 \end{aligned} \quad (7)$$

where $A_{\delta,\gamma}$ and $B_{\delta,\gamma}$ are the coefficients of the rates, as in the sums in Equations 3 and 5. But μ_δ and σ_δ^2 , required for the response function that determines the output rate of the neuron, are the means and variances of the afferent currents.

The connection between $(\mu_\delta[V], \sigma_\delta^2[V])$, computed in the dynamic equations, and $(\mu_\delta, \sigma_\delta^2)$ is made in the following manner. In a stationary state, when the left-hand sides of Equations 7 vanish, we have:

$$\begin{aligned} \mu_\delta[V(t)] &= \mu_\delta = \tau_\delta \sum_\gamma A_{\delta,\gamma} v_\delta + \mu_{\delta,\text{ext}} \\ 2\sigma_\delta^2[V(t)] &= \sigma_\delta^2 = \tau_\delta \sum_\gamma A_{\delta,\gamma} B_{\delta,\gamma} v_\delta + \sigma_{\delta,\text{ext}}^2 \end{aligned} \quad (8)$$

When the temporal variation of the input currents (and hence of the rates) is slow, Equations 8 can also be used away from the stationary state, and we can substitute $\mu_\delta[V(t)]$ and $2\sigma_\delta^2[V(t)]$ for μ_δ and σ_δ^2 , respectively, in Equation 2 and obtain a relation between $v_\delta(t)$ and $(\mu_\delta[V(t)], 2\sigma_\delta^2[V(t)])$. Substituting this relation in Equations 7 gives us a closed set of dynamical equations for $\mu_\delta[V(t)]$ and $\sigma_\delta^2[V(t)]$, which determines the dynamics of $v_\delta(t)$. A variation in $\mu_\delta[V(t)]$ and $\sigma_\delta^2[V(t)]$, as given in Equations 7, leads to a shift in the rates, which are then fed back into the dynamical equations.

A set of rates v satisfying Equations 6 and 8 is a stationary point of the dynamics (Equations 7), and vice versa. Thus, solving numerically for the stationary states of Equations 7 is one way of solving Equations 6, with the bonus that such solutions are stable, since they attract.

The Currents

Spontaneous Activity

To proceed, one expresses the mean and variance of the afferent currents in terms of the instantaneous values of the rates in different neural populations that are assumed to behave in a homogeneous way. For example, if the network is expected to sustain only spontaneous activity, one would expect two populations only, a population of excitatory neurons and a population of inhibitory neurons. They would be distinguishable either because the two types of neurons would have different physiological characteristics or because there would be different distributions of synaptic efficacies on the dendrites.

Suppose each neuron receives $C_E(C_I)$ excitatory (inhibitory) synapses on average, whose efficacy has a Gaussian distribution of mean $J_E(J_I)$ and variance $J^2\Delta$ for both. Let the time constants be $\tau_E(\tau_I)$ and their average emission rates be $v_E(v_I)$. Then,

$$\begin{aligned} \mu_E &= \tau_E(C_E J_E v_E - C_I J_I v_I) + \mu_{\text{ext}} \\ \mu_I &= \tau_I(C_E J_E v_E - C_I J_I v_I) + \mu_{\text{ext}} \end{aligned} \quad (9)$$

where μ_{ext} is the average afferent, excitatory, nonselective current from outside the module, taken equal for both types of neurons.

The variances of the external currents are σ_{ext}^2 . The variances of the total afferent currents are:

$$\begin{aligned}\sigma_E^2 &= \tau_E \left(C_E J_E^2 v_E - C_I J_I^2 v_I \right) + \sigma_{\text{ext}}^2 \\ \sigma_I^2 &= \tau_I \left(C_E J_E^2 v_E - C_I J_I^2 v_I \right) + \sigma_{\text{ext}}^2\end{aligned}\quad (10)$$

When the μ s and the σ s are introduced in the transduction functions (Equation 2) of the two types of neurons (differing by the integration constants τ_E and τ_I), one obtains two self-consistent equations (like Equation 6) for the average stationary rates in the two populations.

Selective Delay Activity

A richer example is one where learning has taken place and sets of synapses have been modified in response to stimuli presented. Each stimulus presented to the network produces an increase in the spike rates of neurons in a subpopulation of the module; those are the neurons with *visual response*. Mean-field theory is simple if one assumes that no neuron is visually selective to more than one stimulus. Hebbian learning is then expressed by the fact that synapses between neurons responding to a given stimulus have their average efficacy increased, $J_E \rightarrow J_+$, whereas synapses between a neuron responsive to a stimulus and one that is not have their average efficacy depressed, $J_E \rightarrow J_-$.

In this case, the neurons in the module can be divided into four distinct, homogeneous groups: (1) neurons selective to the current stimulus (such as those responding to an image currently presented to a monkey); (2) neurons selective to another stimulus in the set used for training but not activated in the present trial; (3) neurons not responsive to any of the stimuli used (hence with unmodified synapses); and (4) inhibitory neurons. If the number of excitatory (inhibitory) neurons in the module is N_E (N_I), if the number of neurons selective to a stimulus is fN_E , and if the number of stimuli used is p , then the number of neurons in each of the four classes is, respectively, fN_E , $(p-1)fN_E$, $(1-pf)N_E$, and N_I .

In this case one has four self-consistency equations (Equations 3 through 6), or the dynamic version (Equations 7), to find the various stationary activity states of the network following learning, at different stages before, during, and after stimulus presentation, and to study their stability. In the process one constructs the μ 's and the σ 's for each of the populations (eight expressions in all). For example, the mean and variance of the afferent current to neurons in the selective population are

$$\begin{aligned}\mu_1 &= \tau_E C_E [fJ_+ v_1 + (1-fp)J_- v_+ + f(p-1)J_- v_-] \\ &\quad - C_I J_I v_I + \mu_{\text{ext}} \\ \sigma_1^2 &= \tau_E C_E [fJ_+^2 v_1 + (1-fp)J_-^2 v_+ + f(p-1)J_-^2 v_-] \\ &\quad + C_I J_I^2 v_I + \sigma_{\text{ext}}^2\end{aligned}\quad (11)$$

in which v_1 , v_+ , v_- , and v_I are, respectively, the rates in the four different classes of neuron described above.

Stability and Other Dependence on Parameters

The example of spontaneous activity is very rudimentary, but not without significant lessons (Amit and Brunel, 1997b). The main lesson obtained from mean-field theory is about stability: in the absence of inhibition, spontaneous activity, with spike rates in the range observed, is not possible. Moreover, inhibition must be strong enough to overcome excitation for the collateral part of the afferent currents, even though there are many fewer inhibitory neurons than excitatory ones. Inhibitory neurons can win by possessing either stronger synapses ($J_I > J_E$) or faster internal dynamics ($\tau_I <$

τ_E). In either of these conditions, spikes can be emitted by neurons of the assembly only because of afferents from outside the module.

Another issue that can be studied in detail in this framework is the range of potentiation, following the appearance of selective delay activity, in which spontaneous activity can coexist with the DAD, both of which are observed experimentally and are stable. Which of them manifests itself depends on whether the stimulus presented is familiar or not. The various regimes of stability are presented in the mean-field bifurcation diagram in Figure 3. The two solid curves are the equilibrium rates of spontaneous (flat, low) and persistent delay activity. Below a potentiation of 2.05, only spontaneous activity is stable, at about 3 Hz. For potentiation between 2.05 and 2.3, both are stable, and the dotted curve delineates their corresponding basins of attraction. To the right of 2.3, the stability of spontaneous activity is destroyed, and only the selective delay activity is stable.

Simulations Versus Theory

Generic conclusions require a theory, but the theory must be confronted with simulation of the spiking system that the theory purports to describe. This comparison is quite satisfactory (Amit and Brunel, 1997a), even though the assumption about the absence of correlations in the spike processes reaching a neuron via different afferent channels is not fully justifiable, as has become evident from cross-correlations measured on simultaneously "recorded" cells in the simulation.

An example of a confrontation of theory with simulation is presented in Figure 4. A network of 16,000 excitatory and 4,000 inhibitory neurons is connected as described earlier in this article. The distribution of connectivities and of synaptic efficacies is random. Five groups of 1,600 excitatory neurons each have the average synaptic efficacies connecting them potentiated by a factor of 1.9. Each neuron receives 1,600 excitatory afferents from outside the network, at a rate of $4/\text{s}^{-1}$. The simulation is then launched with an initial random distribution of depolarizations. The neurons start in a spontaneous activity state, as is seen in the figure in the

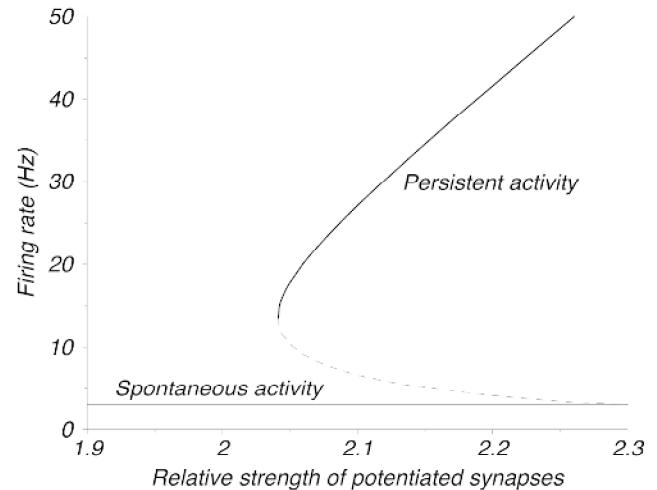
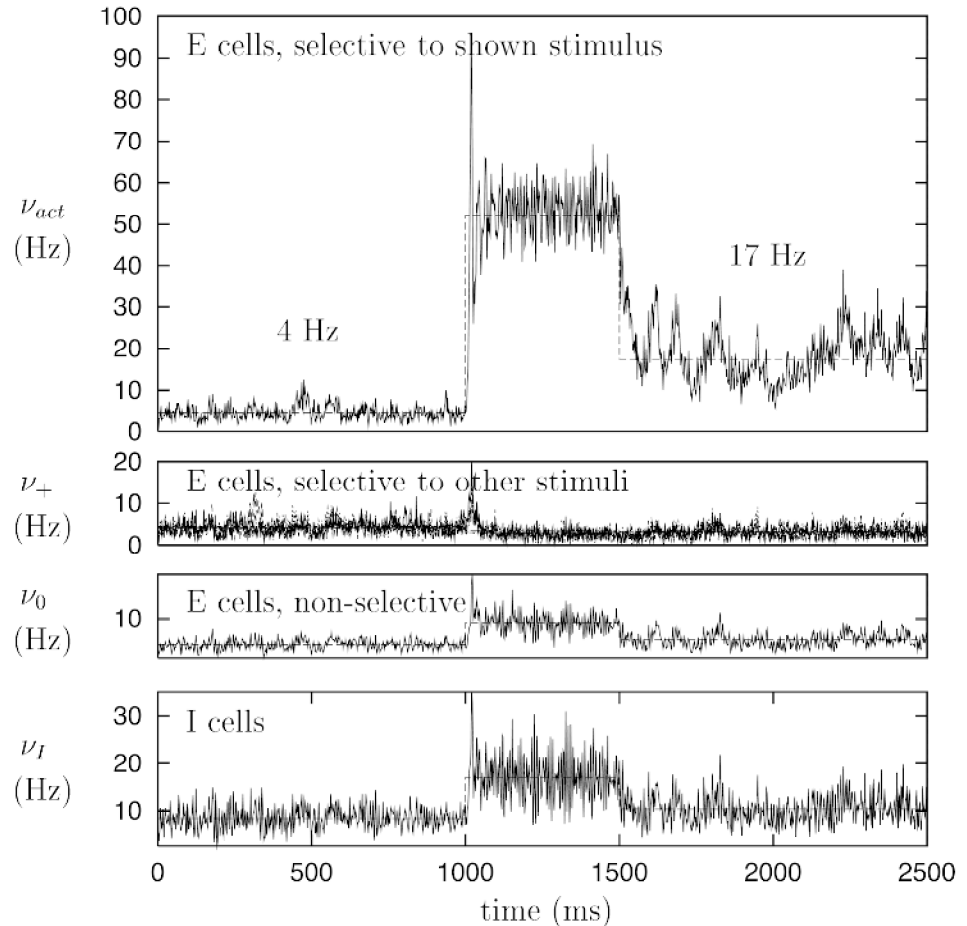


Figure 3. Bifurcation diagram: regions in potentiation-rate plane for stable stationary states of spontaneous activity and selective delay activity. Solid curves denote rates in stable states versus potentiation. Dashed curves indicate a stationary, unstable state, with separating basins of attraction of the two solutions (when both exist). For low potentiation (<2.05), there is only spontaneous activity. For intermediate potentiation (≥ 2.05 , ≤ 2.3), there are two stable states. With high potentiation (≥ 2.3), spontaneous activity is unstable (after Brunel, 2000).

Figure 4. Simulations: time evolution of the average emission rate in four neural populations. From the top: an excitatory subpopulation activated by the stimulus (*selective*); excitatory neurons selective to *other stimuli* (grouped together); excitatory neurons nonselective to any stimulus; and the inhibitory population (I). After 1,000 ms of spontaneous activity, a stimulus is presented for 500 ms. When it is removed, the selective subpopulation continues its delay activity at a high rate. Horizontal lines indicate the average rate as given by the dynamic mean-field theory (after Brunel, 2000).



first 1,000 ms. What is plotted is the instantaneous average rate in each population. For the first second, all neurons have low rates. Then a stimulus is presented for 500 ms, by raising the rate of the external afferents to the neurons in one of the populations of selective neurons. The rate in that population rises sharply. Following the removal of the stimulus, one observes 1 s of propagation of a DAD, in which the selective population maintains an elevated rate (lower than under stimulation). The horizontal lines give the mean rate in each population as predicted by mean-field theory.

Discussion

The extended mean-field approach described in this chapter has been applied in richer situations. In particular, it has provided an account of context correlations generated in learning the DMS paradigm with images organized in fixed sequences (Amit, Brunel, and Tsodyks, 1994). Attractor dynamics in localized circuits has also been applied to the gaze fixation problem (Seung, 1996), after having been extended to include line attractors. Recently, such models have been made much more biologically plausible with respect to robustness and realistic rates (Wang, 1999) by the introduction of receptor dynamics, leading to a successful description of working memory immune to distractors (Brunel and Wang, 2001), and to the modeling of spatial working memory (Compte et al., 2000).

Although fruitful, even extended mean-field theory has its limitations. It deals with uniform populations and stationary states, and hence it is insensitive to transmission delays and nonuniform

instabilities. The approach has seen significant extensions capable of satisfactorily handling oscillatory instabilities (Brunel and Hakim, 1999). With the extension of mean-field theory to situations of rich structural complexity, another important step will have been accomplished.

Simulations remain a central tool, functioning as a crucial check on theory. And because the simulation represents an underlying model of the cortical module, the validity of the model as a description of biological reality must often be confronted on levels that theory does not reach, such as simultaneous multicell data, spike emission statistics, and much more.

Road Map: Biological Networks

Background: Hebbian Synaptic Plasticity

Related Reading: Dynamic Remapping; Short-Term Memory; Statistical Mechanics of Neural Networks

References

- Amit, D. J., 1995, The Hebbian paradigm reintegrated: Local reverberations as internal representations, *Behav. Brain Sci.*, 18:617. ♦
- Amit, D. J., and Brunel, N., 1997a, Dynamics of a recurrent network of spiking neurons before and following learning, *Network*, 8:373. ♦
- Amit, D. J., and Brunel, N., 1997b, Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex, *Cereb. Cortex*, 7:237. ♦
- Amit, D. J., Brunel, N., and Tsodyks, M. V., 1994, Correlations of cortical Hebbian reverberations: Experiment versus theory, *J. Neurosci.*, 14:6445.

- Braitenberg, V., and Schüz, A., 1991, *Anatomy of the Cortex*, Berlin: Springer-Verlag.
- Brunel, N., 1996, Hebbian learning of context in recurrent neural networks, *Neural Computat.*, 8:1677.
- Brunel, N., 2000, Persistent activity and the single cell f-I curve in a cortical network model, *Network*, 11:261. ♦
- Brunel, N., and Hakim, V., 1999, Fast global oscillations in networks of integrate-and-fire neurons with low firing rates, *Neural Computat.*, 11:1621.
- Brunel, N., and Sergi, S., 1998, Firing frequency of leaky integrate-and-fire neurons with synaptic currents dynamics, *J. Theor. Biol.*, 195:87.
- Brunel, N., and Wang, X.-J., 2001, Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition, *J. Comput. Neurosci.*, 11:63. ♦
- Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J., 2000, Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model, *Cereb. Cortex*, 10:910.
- Fusi, S., Annunziato, M., Badoni, D., Salamon, A., and Amit, D. J., 2000, Spike-driven synaptic plasticity: Theory, simulation, VLSI implementation, *Neural Computat.*, 12:2227.
- Fusi, S., and Mattia, M., 1999, Collective behavior of networks with linear (VLSI) integrate and fire neurons, *Neural Computat.*, 11:633.
- Fuster, J., 1995, *Memory in the Cerebral Cortex*, Cambridge, MA: MIT Press.
- Goldman-Rakic, P., 1987, Circuitry of primate prefrontal cortex and regulation of behavior by representational memory, in *Handbook of Physiology*, vol. 5, *The Nervous System* (Editor), chap. 9, Bethesda, MD: American Physiological Society. ♦
- Seung, H. S., 1996, How the brain keeps the eye still, *Proc. Natl. Acad. Sci. USA*, 93:13339.
- Softky, W. R., and Koch, C., 1992, Cortical cells should spike regularly but do not, *Neural Computat.*, 4:643.
- Tuckwell, C. T., 1988, *Introduction to Theoretical Neurobiology*, vol. 2, Cambridge, Engl.: Cambridge University Press.
- Wang, X. J., 1999, Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory, *J. Neurosci.*, 19:9587. ♦

Cortical Memory

Joaquín M. Fuster

Introduction

The representation of cognitive functions in the cerebral cortex has been the subject of continuous debate since Broca identified a cortical area involved in spoken language. Essentially, the debate has been taking place between two major schools of thought. On one side are those who propose the parcellation of cortex into discrete regions or modules dedicated to special functions, such as language, memory, and perception, or to their specific contents. This is the “localizationist” point of view, which is essentially similar to that of phrenology but legitimized by the scientific method. On the other side of the debate is the “holistic” position, adopted by those who propose the distribution of cognitive functions in wide and overlapping expanses of cortex. As in any theoretical debate, we also find here the eclectics and compromisers, who espouse an intermediate position: some functions or contents are localized and others are distributed. It is increasingly recognized that memory is one such function, with some of its components localized in neuronal networks that are circumscribed to discrete domains of cortex and others widely distributed in networks that extend beyond the boundaries of cortical areas defined by cellular architecture. Thus, the aggregate of experience and knowledge about oneself and the surrounding world would be represented in cortical networks of widely varying size and distribution. This concept does not exclude extracortical structures from memory storage and function or the possibility that, after being acquired in the cortex, some memory is relegated to some of those structures, such as the basal ganglia. Historically, the concept of cortical network memory has two roots. The first is the indirect empirical evidence, gathered by physicians and experimentalists in the past two centuries, that discrete lesions of the cerebral cortex rarely result in deficits of memory or any of its behavioral manifestations, while commonly affecting sensory or motor functions. Karl Lashley (1950) was the first to obtain systematically that kind of indirect evidence by ablations of cortical areas in animals; from their results, he inferred that memory must be widely distributed in the cortex, and further, that dispersed neuronal assemblies could represent the same memories, or *engrams*. The second root of the concept is theoretical. Hayek (1952) was the first to formalize it by postulating large-scale cortical networks (he called them maps) that would represent all the experience ob-

tained through the senses. Subsequently, that concept gained further theoretical support from the fields of artificial intelligence and connectionism. The fundamental idea is that mnemonic information is stored in distributed, net-like patterns of cortical connectivity that are established by experience. In more recent times, neuroscientists have developed several theoretical variants of that idea by adding to it structural and functional constraints and by extending it to other cognitive functions, such as perception.

This article presents in broad outline a model of network memory in the neocortex that is supported by a large body of empirical evidence from neuropsychology, behavioral neurophysiology, and neuroimaging. Its essential features are (1) the acquisition of memory by the formation and expansion of networks of neocortical neurons through changes in synaptic transmission, and (2) the hierarchical organization of memory networks, with a hierarchy of networks in posterior cortex for perceptual memory and another in frontal cortex for executive memory.

Formation of Memory

Toward the end of the nineteenth century, Cajal proposed that memory is essentially formed and stored by changes in the connections between nerve cells. That notion was subsequently expressed by many others, and for many years it remained widely accepted but unproved. It was theoretically formulated in considerable detail by Hebb (1949), and more recently received substantial support by the discovery of two general categories of facts (see Kandel, 1991, for a review): (1) the electrical induction of persistent synaptic changes in cellular assemblies of the hippocampus—a phylogenetically ancient sector of cortex—and (2) the induction of similar changes in the neural circuits of invertebrate animals by behavioral conditioning. Less direct evidence indicates that synaptic plasticity is at the foundation of memory in the cerebral cortex (Fuster, 1999) and the cerebellum (Thompson, 1986). It is now widely accepted that the acquisition of memory consists essentially in the modulation of synaptic transmission, and also, to some extent, in the elimination of synapses.

Hebb (1949), referring to the cortex, postulated that “two cells or systems that are repeatedly active at the same time will tend to become associated, so that activity in one facilitates activity in the

other." Thus, temporally coincident inputs would tend to associate the neurons that receive them, by facilitating the synapses between them. This principle, which has been called synchronous convergence (Fuster, 1999), would lead to the formation of the hebbian "cell assembly," a basic neural net of cortical representation in sensory and parasensory cortex. Based on data bearing on the plasticity of responses of visual cortical cells to optic stimuli, Stent (1973) made a strong theoretical argument for the operation of that principle in the neocortex.

Whereas simple sensory memories may be formed and represented in cell assemblies, nets, or modules of sensory and parasensory cortex, the neuropsychological evidence from humans and animals indicates that the more complex memories of individual experience, as well as abstract knowledge, extend into areas of the cortex of association. Temporally coincident or near-coincident experiences of one or more sense modalities will modulate synaptic contacts between cells in those areas, thus leading to the formation of wider networks that represent assorted items of individual memory and, at higher cortical levels, of knowledge, which is the conceptual or semantic form of memory. The boundaries of those larger networks extend beyond those of any given cytoarchitectonic area, however defined.

The formation of the associative neuronal networks that support and contain memory follows gradients of corticocortical connectivity, which have been most thoroughly investigated in the non-human primate (Pandya and Yeterian, 1985). That connectivity departs from primary sensory and motor areas and flows into progressively higher areas of unimodal and multimodal association. The connectivity is reciprocal throughout, such that each connective step, from one area to the next, is reciprocated by fibers running in the opposite direction. Some connections converge and others diverge. Thus, at all levels and between levels, three basic structural features can be recognized in connective networks: convergence, divergence, and feedback or recurrence. In addition, cortical areas of one hemisphere are connected, again reciprocally, with homologous areas of the other hemisphere.

To some extent, the connectivity mediating memory formation follows also maturational gradients, in that it proceeds from area to area following the order in which cortical areas have myelinated in early ontogeny (Fuster, 1997). It also follows gradients of sensory and motor processing. As memories increase in complexity, in terms of the variety and complexity of associated experiences and the sensory inputs that convey them, their networks become progressively wider, and thus span progressively more associative cortical areas of polymodal convergence.

Whereas synaptic modulation and synchronous convergence seem essential features of the process of memory network formation, the process in more general terms is one of self-organization (Kohonen, 1984). The networks and their connective substrate are auto-constituted as the result of the interactions of the organism with its environment. Through these interactions, and mostly by synchronous convergence, new cortical nets are formed and old ones expanded in a dynamic process that persists throughout the life of the organism. In the formation of a memory network, synaptic facilitation is produced not only by the simultaneity of external inputs but also by the simultaneity of these inputs with "inputs" internally generated by the concomitant activation or retrieval of preexisting components of the network. Thus, new memory is formed from old memory.

In recent years, there has been increasing neuropsychological evidence that the hippocampus plays a critical role in the formation of neocortical memory networks (Squire, 1987). It has been established that the hippocampus is reciprocally connected with the cortical areas of association, though not with primary sensory or motor cortex (Amaral, 1987). That connectivity courses through and under the cortex of the peri- and entorhinal region, in the medial and

inferior aspects of the temporal lobe. In this manner, reciprocal connections link the hippocampus with the associative cortices of the posterior (postrolandic) regions of the cerebral hemispheres as well as with those of the frontal lobe (prerolandic), notably the prefrontal cortex. The mechanisms by which the hippocampus mediates the formation of memory networks in the neocortex are not known. Possibly, long-term potentiation (LTP) is one of those mechanisms. Further, there is reason to suspect that certain excitatory glutaminergic receptors, notably NMDA receptors, take part in the process. That process would result in protein changes in the membrane of cortical cells, which in turn would strengthen their synapses and thus imprint memories in cortical networks. Inputs from the amygdala, a limbic structure essential for the evaluation of the emotional and motivational significance of sensory information, may also intervene in the process.

Phyletic Memory

The primary sensory and motor cortices lie in the interface between the environment and the cortex of association. Primary sensory cortices provide the inlets of sensory information for the formation of perceptual memory (episodic, semantic, etc.) in posterior cortex of association; primary motor cortex, on the other hand, provides the outlet of frontal association cortex (executive memory) to lower motor structures (e.g., basal ganglia, cerebellum). Thus, primary cortices—ontogenetically the first to myelinate—constitute the cortical gate from sensation to memory, and from memory to action.

It is physiologically plausible to view primary sensory and motor cortices as the foundation of all individual memory, themselves constituting a form of universal memory. These cortices constitute a basic form of memory that I have termed *phyletic memory*. According to this view, phyletic memory is simply the structure of the primary sensory and motor cortex at birth, common to all animals of the same species. Thus, phyletic memory is the most basic of all memories, the memory that the organism has formed through millions of years and countless generations in its interactions with the surrounding world. Those genetically predetermined structures of cortex devoted to the analysis of elementary sensory features and to the integration of the elementary primitives of movement would form the basic template on which all individual memory would grow. In summary, primary cortices can legitimately be called memory, because they retain a form of basic and retrievable information: the memory of the species, the sensory and motor information that the organism has acquired and stored in the course of evolution for survival and procreation. Thus, phyletic memory, the primary sensory and motor cortices, contains already at birth much of the adaptive power of the species. After necessary "rehearsal" in the critical periods of early ontogeny, phyletic memory remains ready for "recall" through a lifetime—ready, that is, to recognize the essential features of the world, and to retrieve and organize the basic patterns of movement for adaptation to the environment. On that foundation of phyletic memory, all the memory of the individual will be developed and hierarchically organized by experience and according to the principles mentioned in the previous section: a hierarchy of perceptual memories will be formed in posterior cortex and a hierarchy of motor memories in frontal cortex.

Perceptual Memory

All memory acquired through the senses qualifies as perceptual memory. This includes a vast fund of individual experience, from the simplest forms of sensory memory to abstract knowledge—in other words, to all that we commonly understand as the memory or knowledge of events, objects, persons, animals, facts, names, and concepts. Perceptual memories, and their cortical substrate,

appear hierarchically organized, as the diagram of Figure 1 schematically represents. Such an organization can be inferred from the gradients of cortical connectivity and from neuropsychological evidence, namely, from the study of the psychological effects of cortical lesions. This evidence, however, is somewhat confounded by the variability between subjects and by the profuse relationships of association between the various categories of memory—in other words, by the heterarchy within the hierarchy.

The base of the perceptual hierarchy is sensory phyletic memory, the structure of primary sensory cortices. Immediately above it, in cortex of sensory association, are the memories of the sensory qualities of objects and experiences, unimodal and polymodal. Further up in the hierarchy, in higher associative cortex, memories become more personal and complex, with specific temporal and spatial tags. These memories fall under the category of what is commonly designated declarative memory, the memory of events and experiences. Finally, at the highest levels of the hierarchy, in areas of the temporal and parietal lobes called transmodal (Mesulam, 1998), resides the knowledge of facts, concepts, and names.

At every stage of the hierarchy of perceptual memory, memory networks are essentially made of connections between neuronal aggregates at that stage and by convergence of inputs from below,

ultimately from the senses. Thus, at the first associative sensory stage beyond phyletic memory, networks are formed by associations between sensory representations of the same modality to form assemblies or networks of unimodal sensory memory. Above, in polysensory association cortex, and with inputs from unimodal areas, more complex networks of polymodal memory are formed that associate sensory features of multiple origins. Those polymodal networks constitute the substrate for diverse forms of episodic and semantic memory, with wide distribution in higher association areas of posterior cortex. Networks of the transmodal cortex of areas 39 and 40, including cortex of the superior temporal gyrus (Wernicke's area), probably represent the highest forms of conceptual and semantic knowledge. Lesion of these areas induces certain forms of sensory aphasia and agnosia, including the loss of perceptual memory of speech and objects, or object categories.

In general, memory networks expand as they gain in hierarchical level. They fan up from their sensory base as they penetrate progressively higher layers of cortical organization with progressively broader and higher associations. Once established, memories and their networks are linked not only horizontally, within their hierarchical layer, but vertically, between layers (symbolized by vertical bidirectional arrows in Figure 1). A low-level sensory network, for

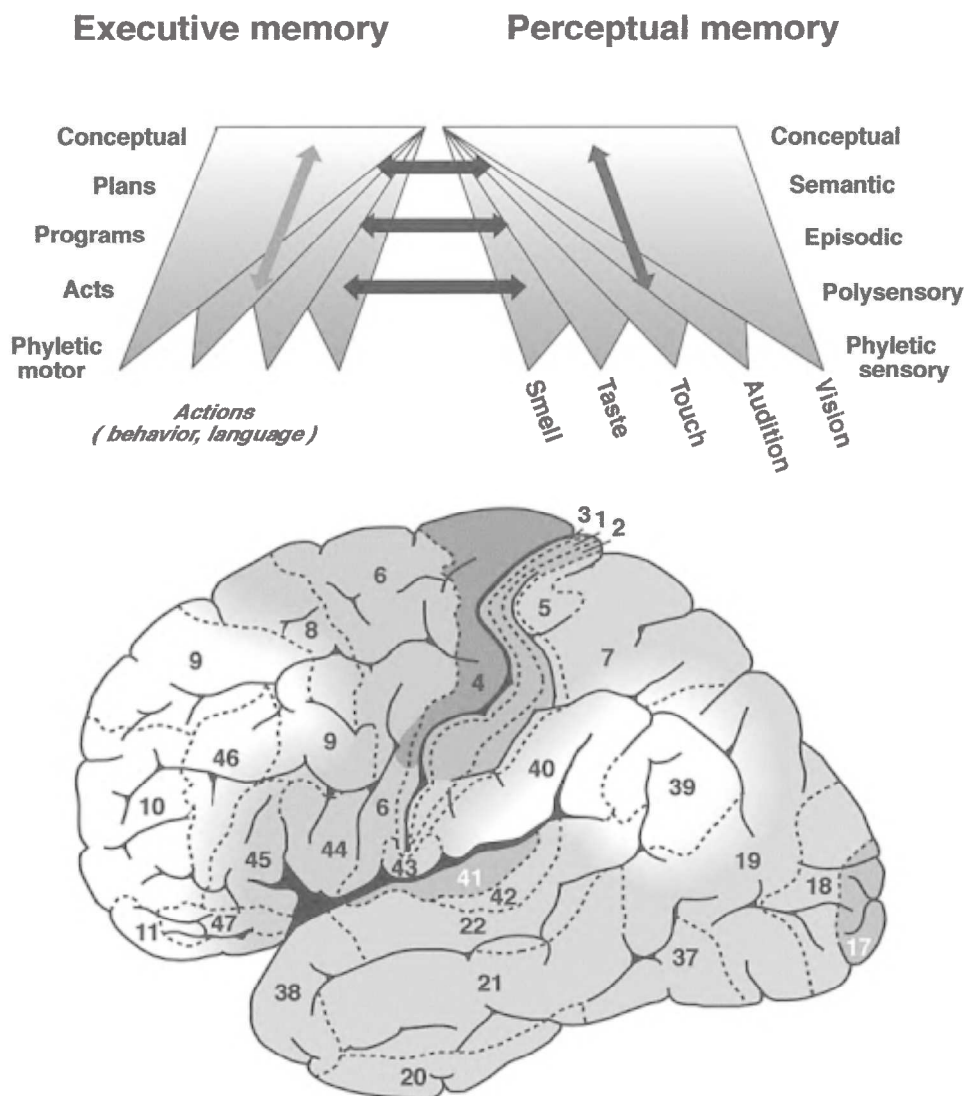


Figure 1. Schematic diagram of the hierarchical organization of memories in accord with the model presented in this article.

example, will establish associations with high-level semantic or conceptual networks. This will result from the co-occurrence of new sensory experiences with the activation of the high-level networks they elicit. In any event, the latter networks will be largely formed by ascending connections from lower levels. Assorted sensory experiences with common attributes will contribute to the formation of conceptual networks at the higher, semantic level. As a consequence of these dynamic interactions, networks of lower and concrete content will be nested within the higher-level conceptual and semantic networks that they have helped establish and that represent the higher and more abstract aspects of perceptual memory. As a consequence of these interactions within and between layers, memory networks profusely share cells and fibers. Conversely, any cortical cell and its connections may be part of many memory networks.

Further, as memory networks gain in hierarchical category, they become not only broader but more resilient, perhaps from increased number and redundancy of associations. Whereas sensory memories are especially vulnerable to discrete local damage (notably the phyletic structure of primary sensory cortices), high-level associative memories are less vulnerable. Since the latter are anchored in multiple sensory experiences, abstract knowledge is less liable to cortical damage than the concrete aspects of declarative memory, such as dates, names, faces, and places. The clinical evidence shows that these concrete elements of memory are especially liable to cortical injury. To some extent, which varies with the individual, these concrete elements of memory are subject to gradual loss as a function of normal aging. On the other hand, it takes a massive cortical lesion for the loss of conceptual perceptual memory to occur, that is, for the loss of what Kurt Goldstein called the "abstract attitude," meaning the utilization of conceptual knowledge in daily life.

Executive Memory

Motor memory, the representation of motor acts and behaviors, is widely distributed throughout the central nervous system. The spinal cord, the brainstem, and the cerebellum constitute the lowest levels of the hierarchy of brain structures harboring motor memories. These structures are the depositories of much of the motor phyletic memory of the organism at birth. This memory is largely innate, stereotypical, and dedicated to the fulfillment of essential drives. Some of it is conditionable and subject to cortical control and modulation.

The highest levels of motor memory, both phyletic and acquired, are supported by the cortex of the frontal lobe (Figure 1). The cortical networks of motor memory are formed essentially in accord with the same principles that guide the formation of perceptual memory, notably synchronous convergence. Here, however, the simultaneously converging signals are of both sensory and motor origin. This includes visual and auditory stimuli that coincide with motor action, or release it by one mechanism or another, and kinesthetic stimuli that accompany the action; in addition, some of the signals are "efferent copies" of the action as it is executed, copies that are provided by recurrent or collateral inputs from the motor system (corollary discharge). Consequently, many motor networks of the frontal cortex are extensions of perceptual networks of posterior cortex. Long corticocortical connections—in the uncinate fasciculus—would provide the functional substrate for those extensions (horizontal arrows in Figure 1).

At the base of the cortical executive or motor hierarchy is the primary motor cortex of the precentral gyrus, the substrate of phyletic motor memory in the neocortex. It supports the representation and execution of elementary motor acts that are defined by the contraction of particular muscles and muscle groups. Hierarchically above primary cortex is the premotor cortex (area 6), which several

lesion and unit studies implicate in the representation of motor acts and programs defined by goal and trajectory (Wiesendanger, 1981; Alexander and Crutcher, 1990). This cortex also participates in the elementary structuring of language. The more complex and novel programs of acquired behavior and language are represented in the prefrontal cortex, which is the highest level of the executive hierarchy. The prefrontal cortex is the cortex of association of the frontal lobe. Both phylogenetically and ontogenetically, it is one of the latest neocortical regions to develop (Fuster, 1997). In the human, it does not reach full maturation until the second decade of life. It receives abundant connections from posterior cortex, limbic formations, and the brainstem.

Neuropsychological studies implicate the prefrontal cortex in the representation of schemas of goal-directed action. Human subjects who have sustained lesions of lateral prefrontal cortex have difficulty remembering and formulating new plans of behavior and structures of language. Monkeys have difficulty learning and executing behavioral tasks that require the sequencing of motor acts, especially if the sequencing contains temporal gaps that have to be bridged by active memory ("working memory"). This kind of behavior is epitomized by the so-called delay tasks (e.g., delayed response, delayed matching). In both the human and nonhuman primate, lesions of the lateral prefrontal cortex impair performance on such tasks. There is some apparent specificity within the lateral prefrontal areas for the kinds of sensory information and motor activity that are processed in those tasks. Regional specificity, however, is secondary to temporal factors. Prefrontal lesions cause deficits in the formation and execution of temporal sequences of behavioral action ("temporal gestalts"), whatever the sensory and motor components of those sequences.

The practice and repeated execution of behavioral sequences seems to lead to the relocation of their representation from prefrontal cortex to lower stages of the motor hierarchy, especially the basal ganglia. Frontal lesions in human subjects induce deficits in the performance of complex voluntary movements without impairing automatic ones. This occurs even if these automatic movements are just as complex and require just as much effort as when they were originally learned. By neuroimaging, it is possible to follow to some degree the migration of executive memories from the prefrontal cortex to other structures of lower hierarchical rank. In the initial stages of the learning of a sequential motor task, certain areas of the dorsolateral prefrontal are activated (Jenkins et al., 1994). As the task becomes routine, parts of the cerebellum and the basal ganglia become more active, and the prefrontal cortex less. Still represented in prefrontal cortex, presumably, are those aspects of the task that are subject to uncertainty or ambiguity. Such is the case with delay tasks, where stimuli and responses contain both uncertainty and ambiguity. This is the reason why the correct performance of these tasks, even after learning, continues to depend on the functional integrity of the prefrontal cortex, and prefrontal cells continue to be involved in that performance.

Retrieval of Memory

Just as the formation of a memory is an associative phenomenon, so is the retrieval of that memory from permanent storage. Both associative phenomena are interdependent. New memory networks are formed by association between co-occurring sensory inputs and between these inputs and older networks that they activate by association. Therefore, new memory is in many respects the expansion of old memory. Whereas individual memories and their networks are expansions of phyletic memory, new individual networks are expansions of old ones. The essential point is that every act of memory formation is accompanied by retrieval of established memory. Retrieval is indispensable for memory formation. That may be

the reason why the hippocampus appears to play an important role in both the acquisition and the recall of memory in the neocortex.

Electrophysiological studies in the monkey and neuroimaging in the human indicate that the retrieval of a memory, whether spontaneous or evoked, essentially consists in the activation of the cortical network that represents it (reviewed in Fuster, 1999). The neuronal mechanisms of retrieval are not known but can be presumed to involve the correlated activation of all the neuronal elements of a perceptual network by the sensory or internal (mental) activation of one of its associated components. If the network has executive or motor components, then the activation will extend to an associated executive network of frontal cortex. Above a certain threshold, that activation of an executive network may lead to the execution of action. If the action is sequential and dependent on serial perceptual inputs, then cell prefrontal assemblies will interact with posterior cortical assemblies in the mechanisms of working memory and preparatory set at the foundation of the perception-action cycle (Fuster, 1997; see also PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION). Those mechanisms are still poorly understood. However, it seems increasingly plausible that they include the reverberation of activity in memory networks through recurrent circuits (Zipser et al., 1993).

Discussion

This article presents a general connectionist model of memory. It is proposed that memories consist in, and are represented by, widely distributed and profusely interactive networks of neocortical neurons. Memory networks are formed by modulation of synaptic contacts between concomitantly activated neurons representing discrete features of the external and internal environment. This takes place in the neocortex under the agency of limbic structures, notably the hippocampus. Memory formation is a highly dynamic process closely interdependent with retrieval, and memory networks remain in constant change throughout life, subject to expansion by new associations—and to age-related attrition. Memory networks are hierarchically organized in two broad sectors of neocortex: perceptual memory in posterior (postrolandic) cortex and executive memory in frontal (prerolandic) cortex. At the base of each cortical hierarchy there is a layer of phyletic memory (the structure of primary sensory or motor cortex). Above that layer, in cortex of association, lie the memories of the individual, from the lowest perceptual and motor representations to the highest conceptual representations of perception and action. The semantic and conceptual representations of perception lie in transmodal areas of posterior cortex, whereas the schematic and conceptual representations of action lie in areas of prefrontal cortex. Both are interconnected by long corticocortical fibers, thus forming high-level

sensorimotor networks. In the organization of complex behavior, the two sectors of neocortex, posterior and prefrontal, interact dynamically at the summit of the perception-action cycle.

Road Map: Cognitive Neuroscience

Related Reading: Hebbian Synaptic Plasticity; Sequence Learning; Short-Term Memory; Visual Scene Perception

References

- Alexander, G. E., and Crutcher, M. D., 1990, Functional architecture of basal ganglia circuits: Neural substrates of parallel processing, *Trends Neurosci.*, 13:266–271.
- Amaral, D. G., 1987, Memory: Anatomical organization of candidate brain regions, in *Handbook of Physiology: Nervous System*, vol. V: *Higher Functions of the Brain*, Part 1 (F. Plum Ed.), Bethesda, MD: American Physiological Society, pp. 211–294.
- Fuster, J. M., 1997, *The Prefrontal Cortex* (3rd ed.), Philadelphia: Lippincott-Raven.
- Fuster, J. M., 1999, *Memory in the Cerebral Cortex*, Cambridge, MA: MIT Press. ♦
- Hayek, F. A., 1952, *The Sensory Order*, Chicago: University of Chicago Press.
- Hebb, D. O., 1949, *The Organization of Behavior*, New York: Wiley. ♦
- Jenkins, I. H., Brooks, D. J., Nixon, P. D., Frackowiak, R. S. J., and Passingham, R. E., 1994, Motor sequence learning: A study with positron emission tomography, *J. Neurosci.*, 14:3775–3790.
- Kandel, E. R., 1991, Cellular mechanisms of learning and the biological basis of individuality, in *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds.), Norwalk, CT: Appleton and Lange, pp. 1009–1031.
- Kohonen, T., 1984, *Self-Organization and Associative Memory*, Berlin: Springer-Verlag. ♦
- Lashley, K. S., 1950, In search of the engram, *Symp. Soc. Exp. Biol.*, 4:454–482.
- Mesulam, M., 1998, From sensation to cognition, *Brain*, 121:1013–1052. ♦
- Pandya, D. N., and Yeterian, E. H., 1985, Architecture and connections of cortical association areas, in *Cerebral Cortex*, vol. 4 (A. Peters and E. G. Jones, Eds.), New York: Plenum Press, pp. 3–61. ♦
- Squire, L. R., 1987, *Memory and Brain*, New York: Oxford University Press.
- Stent, G. S., 1973, A physiological mechanism for Hebb's postulate of learning, *Proc. Natl. Acad. Sci. USA*, 70:997–1001.
- Thompson, R. F., 1986, The neurobiology of learning and memory, *Science*, 233:941–947.
- Wiesendanger, M., 1981, Organization of secondary motor areas of cerebral cortex, in *Handbook of Physiology* (S. R. Geiger, Ed.), Bethesda, MD: American Physiological Society, pp. 1121–1147.
- Zipser, D., Kehoe, B., Littlewort, G., and Fuster, J., 1993, A spiking network model of short term active memory, *J. Neurosci.*, 13:3406–3420. ♦

Cortical Population Dynamics and Psychophysics

Udo A. Ernst and Christian W. Eurich

Introduction

Visual cortex is one of the most extensively studied regions in the mammalian brain. Over the past decade, much anatomical, physiological, and psychophysical knowledge about its properties has been accumulated experimentally. Considerable effort has been expended on theoretical and computational work to reproduce basic phenomena and to explain their underlying mechanisms.

In this article, we discuss one specific computational approach that has been successfully applied to a variety of problems on different levels of cortical information processing. The approach describes the cortical population dynamics in the form of structurally simple differential equations for the neurons' firing activities. The model class was introduced by Wilson and Cowan (1972, 1973) and is still very popular for, in our opinion, two reasons: first, it is

powerful enough to reproduce a variety of cortical phenomena, and it captures the dynamics of neuronal populations seen in numerous experiments; and second, its degree of complexity is still low enough to allow analytical treatment that yields an understanding of the mechanisms leading to the observed behavior.

In the next section, we introduce the model class and discuss some of its basic properties. The following sections show how this model can be applied to explain dynamical properties of the primate visual system on different levels, from single neuron properties like selectivity for the orientation of a stimulus to higher cognitive functions related to the binding and processing of stimulus features in psychophysical discrimination experiments.

The goal of this article is to show that a model that abstracts from biophysical details is often sufficient to identify possible neuronal mechanisms of cortical information processing. The diversity of the examples we mention demonstrates that even a simplifying approach can place seemingly unrelated or even controversial findings in one coherent, unifying framework.

The Wilson-Cowan Model Class

A basic introduction to differential equations in the context of neural systems and the class of models described here can be found in Wilson (1999).

Single Units

The basic unit of the model is a neuronal population. The dynamics of an uncoupled population is described by an ordinary differential equation for its activity $A(t)$, which consists of a decay term and the synaptic input $I(t)$ (τ is a time constant),

$$\tau \frac{dA(t)}{dt} = -A(t) + h(I(t)) \quad (1)$$

The gain function, h , describes the activation of a population dependent on its synaptic input I . Wilson and Cowan (1972) derived this equation from a more general integro-differential equation by applying a temporal filter (*time coarse graining*). Therefore, the resulting Equation 1 is structurally simple, but not exact: one has to bear in mind that temporal variations on a small time scale have been averaged out. In the original publication of Wilson and Cowan, the activity A was identified by the *proportion* of active neurons in a population. For this reason, the gain function h was chosen to be a sigmoid function that saturates at 1 for high input levels (no more than all neurons can be active at a specific time). However, A can equally well be interpreted as the population's firing rate (this will be the case throughout this article). Then h is no longer bound to saturate at 1, and one of the simplest choices is a function that is 0 up to some threshold, and then increases linearly. This piecewise linear function may allow for an analytical treatment by considering appropriate case distinctions. Omitting the saturation regime at high I does not impose serious restrictions, because cortical neurons usually operate within the linear regime of their gain function. Furthermore, diverging network activity in the model will be an indication of an unphysiological parameter regime in which the overall network activity is not dynamically regulated, as it is likely to be in the cortex.

Columns

Cortical nervous tissue contains both excitatory and inhibitory neurons in a dense network. A general model of this network necessarily has to include both cell types, forming one excitatory population (e) and one inhibitory population (i). Each population typically represents some hundreds of single neurons. The populations are mutually connected with weights w_{ee} , w_{ie} , w_{ei} , and w_{ii} . Index pairs like ei are interpreted as a connection originating at the

excitatory population and targeting the inhibitory population. We will identify the resulting dynamical system (Wilson and Cowan, 1972) with the concept of a cortical column:

$$\tau_e \frac{dA_e(t)}{dt} = -A_e(t) + h_e(w_{ee}A_e(t) - w_{ie}A_i(t) + I_e(t)) \quad (2)$$

$$\tau_i \frac{dA_i(t)}{dt} = -A_i(t) + h_i(w_{ei}A_e(t) - w_{ii}A_i(t) + I_i(t)) \quad (3)$$

At this point, we would like to note that two very similar model classes exist in the literature. Their dynamics differ in the sense that in the first class, A describes the activation or the firing rate of a population (in this case the nonlinearity in h is applied to the total synaptic input, $\dot{A} = -A + h(wA + I)$), while in the second class, A denotes the membrane potential (in that case h is applied directly to A , $\dot{A} = -A + wh(A) + I$). A reader of the original publications should not be confused, because both variants lead to qualitatively similar results and are often equally well suited to tackle a specific modeling problem.

Layers

A neuronal layer may be described as a multitude of columns arranged in a topographically ordered space. This space may have a varying number of dimensions. For example, some authors have used a one-dimensional chain representing the orientation preference axis; others identify a two-dimensional layer with the surface of the cortical tissue. With $\vec{x}, \vec{x}' \in C$ denoting positions within such a layer, the columns are coupled by appropriately chosen functions $W_{\{ei, ie, ee, ii\}}(\vec{x}, \vec{x}')$ (so-called *lateral couplings*). Mathematically, the neuronal layer is described as a pair of coupled partial differential equations (Wilson and Cowan, 1973):

$$\begin{aligned} \tau_e \frac{\partial A_e(\vec{x}, t)}{\partial t} = & -A_e(\vec{x}, t) \\ & + h_e \left(w_{ee} \int_C A_e(\vec{x}', t) W_{ee}(\vec{x}, \vec{x}') d\vec{x}' \right. \\ & \left. - w_{ie} \int_C A_i(\vec{x}', t) W_{ie}(\vec{x}, \vec{x}') d\vec{x}' + I_e(\vec{x}, t) \right) \end{aligned} \quad (4)$$

$$\begin{aligned} \tau_i \frac{\partial A_i(\vec{x}, t)}{\partial t} = & -A_i(\vec{x}, t) \\ & + h_i \left(w_{ei} \int_C A_e(\vec{x}', t) W_{ei}(\vec{x}, \vec{x}') d\vec{x}' \right. \\ & \left. - w_{ii} \int_C A_i(\vec{x}', t) W_{ii}(\vec{x}, \vec{x}') d\vec{x}' + I_i(\vec{x}, t) \right) \end{aligned} \quad (5)$$

The inputs $I_{\{e,i\}}$ are typically calculated by integrating a stimulus $S(\vec{x}, t)$ over an *afferent* coupling function $V_{\{e,i\}}(\vec{x}, \vec{x}')$:

$$I_{\{e,i\}}(\vec{x}, t) = \int_R S(\vec{x}', t) V_{\{e,i\}}(\vec{x}, \vec{x}') d\vec{x}' \quad (6)$$

where \vec{x}, \vec{x}' are elements of an input space R . A convenient choice for the lateral as well as the afferent couplings are functions decaying with the distance between two populations, as has been shown in anatomical and physiological studies. Choosing W satisfying $W(\vec{x}, \vec{x}') = W(|\vec{x} - \vec{x}'|)$, the model becomes translationally and rotationally invariant. A commonly used prototype for these kernels is an n -dimensional Gaussian function defined as

$$W_{\{ee, ei, ie, ii\}}(|\vec{x} - \vec{x}'|) = \frac{1}{(\sqrt{2\pi}\sigma_{\{e,i\}})^n} \exp \left(-\frac{(\vec{x} - \vec{x}')^2}{2\sigma_{\{e,i\}}^2} \right) \quad (7)$$

The computational advantage of these kernels is that the integration reduces to a multiplication in Fourier space, which speeds up computation time considerably.

For most of this article, we will assume that connections originating from inhibitory populations will be longer than those originating from excitatory populations. Following this scheme, the effective coupling between two columns will have the shape of a Mexican hat (difference of Gaussians). This assumption, which is often made in modeling studies, is questionable insofar as long-ranging patchy excitatory connections exist, at least in the mature primary visual cortex. This may not be a problem, because the layout of primary visual cortex revealed by the structure of the orientation preference map suggests that inhibitory couplings dominate, at least over intermediate distances. Nevertheless, we will also discuss which different or additional phenomena are observed in the presence of long-range axons. For a more detailed introduction to neural layers, see **LAYERED COMPUTATION IN NEURAL NETWORKS**.

Dynamical Regimes and Orientation Preference

Quasi-linear and Marginally Stable Regimes

With different choices of the system parameters in Equations 4 and 5, almost all model variants exhibit one of two different dynamical behaviors: if the strength of the afferent input dominates over the lateral feedback, a homogeneous and constant input will lead to an activation pattern that is also spatially and temporally constant. This steady state is stable against noise. The parameter regime where this behavior occurs is called the *quasi-linear regime* (upper region in Figure 1). As soon as the inhibitory feedback gets weaker or the excitatory feedback gets stronger, the system enters a second regime, called the *marginally stable regime* (Ben-Yishai, Bar-Or, and Sompolinsky, 1995). Now the steady state is unstable, and even the smallest perturbation leads to the emergence of a pattern of activation clusters commonly called *blobs* (central portion of Figure 1). The mechanism for this type of pattern formation is easy to understand: if the input at one position is slightly increased, this perturbation of the steady state will be amplified by the dominating

excitatory feedback, while the longer-ranging inhibition will suppress the activity in the surround of the emerging blob. For related reading, see **WINNER-TAKE-ALL NETWORKS**.

Lateral Feedback and Orientation Selectivity

The existence of a marginally stable regime could have consequences for the emergence of orientation selectivity in primary visual cortical neurons. Ben-Yishai et al. (1995) observed that the shape of the blobs remains invariant against different input levels. In a one-dimensional model of a cortical hypercolumn, where $\bar{x} \in C$ is identified with the orientation preference Φ , they demonstrated that the response behavior of neurons to oriented gratings is accurately reproduced: the orientation tuning width remains largely invariant under changes of the stimulus amplitude (contrast). This finding indicates that cortical dynamics may be dominated by lateral feedback rather than by feedforward excitation. A weak afferent orientation bias, such as that emerging from a Hubel-Wiesel arrangement of LGN receptive fields, would then suffice to induce a sharply tuned orientation tuning curve. This idea, with its pros and cons and also the experimental evidence for the origin of orientation tuning, is discussed in detail in **ORIENTATION SELECTIVITY (q.v.)**.

Inhomogeneities and Cortical Maps

Localization of Activation Clusters

In the marginally stable regime, each perturbation lays the seed for the emergence of an activation cluster. This perturbation could be induced by the afferent input, but also by structural inhomogeneities in the model. For example, the lateral coupling function may not be perfectly translationally and rotationally invariant but could be subject to small random jitter. Then, even a homogeneous and constant input would lead to the emergence of activation clusters. Preferentially, these clusters will be located at positions where by

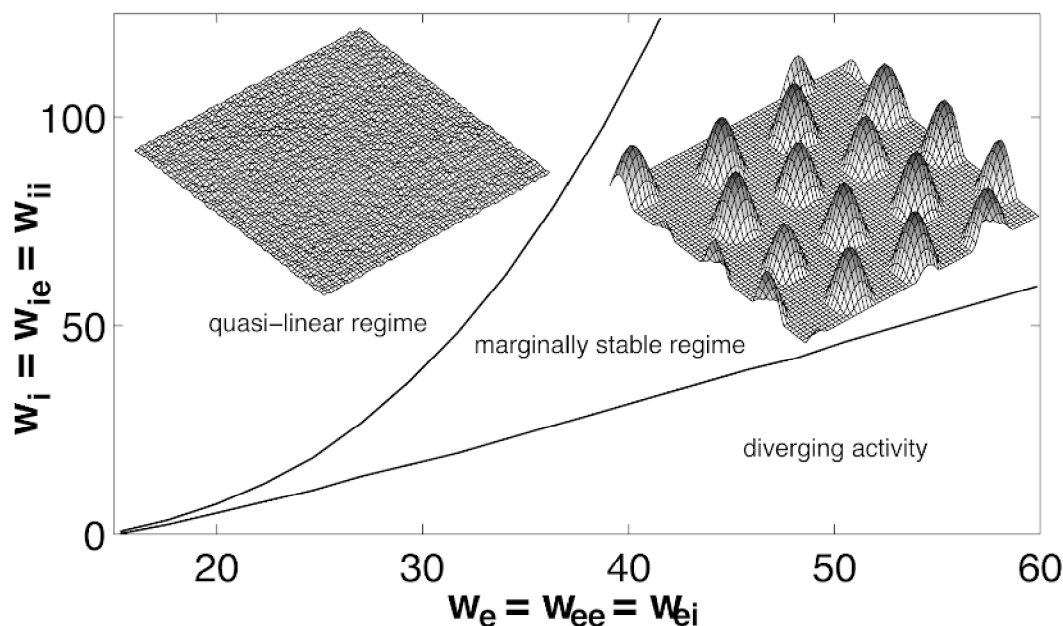


Figure 1. Phase diagram of the spatiotemporal Wilson-Cowan equations (Equations 4 and 5) for a two-dimensional sheet of neurons with threshold-linear h . Three dynamical regimes can be distinguished, depending on the weights w_e and w_i : the quasi-linear and the marginally stable regimes, as

described in the text, and a biologically implausible regime where neural firing rates diverge. Drawings in the quasi-linear and marginally stable regimes show a typical steady-state activity pattern of the excitatory layer for a constant input I plus a small amount of noise.

chance the lateral excitatory feedback is slightly stronger than at other positions nearby. If this jitter comes together with inhomogeneous afferent inputs, both effects will add up, and blobs will choose positions where the afferent input plus the lateral feedback will be largest. Note that in this model, the noise breaks the symmetry of the coupling kernels, and the model will no longer be rotationally and translationally invariant.

Instantaneous Emergence of Cortical Maps

Ernst et al. (2001) simulated a two-dimensional Wilson-Cowan model in which they put a small amount of static noise on the lateral coupling matrix, as can be expected in a biological system, with all its irregularities. They presented moving gratings or bars as stimuli, generating an inhomogeneous afferent input. By recording the model's response to differently oriented gratings (Figure 2A), they found that orientation and direction preference maps naturally emerged when the blobs localized at the spatial inhomogeneities in the model cortex (Figure 2B).

This model has several advantages over other approaches to map development because it reproduces seemingly controversial findings from experimental studies. First, the structure of the maps shows up within milliseconds and does not require any learning. Second, due to the intracortical origin of the map structure seeded by the random jitter of the lateral connections, the feature maps are identical for stimulation of either of the two eyes. Third, the gratings induce an oscillatory movement of the blobs around their preferred positions, which is different for opposite directions of movement. This suggests a new mechanism for directional selectivity of the neuronal response (for a detailed discussion, see DIRECTIONAL SELECTIVITY). And finally, the model reproduces the known relationships between different kinds of feature maps. Taken together, these properties qualify this approach as a model for the initial phase in cortical development where the coarse layout of the maps is determined, which then could get subsequently refined and rearranged by self-organizing mechanisms (see Swindale, 1996, for an extensive review and discussion).

Long-Range Connections and Contour Integration

Up to now, the couplings have been chosen as if there were no long-ranging excitatory connections in the brain. However, those

connections exist, and they preferentially link neurons having similar orientation preferences. What dynamical phenomena can one expect if these connections are included?

Long-Range Connections

Several authors employ a connection scheme that locally has the shape of a Mexican hat but extends over that region, sending out additional excitatory connections targeting inhibitory and excitatory populations with a similar orientation preference. The columns in these models have a position (x_1, x_2) within the nervous tissue and an orientation preference Φ ; thus, $C \ni \vec{x} = (x_1, x_2, \Phi)$. While the response of the *classical* model without long-range interactions would follow the dynamics described in the previous sections, the addition of long-range connections opens up the possibility that spatially extended stimuli modulate this response. The modulation will depend on the strength, or contrast, and the orientation of the stimuli presented. One important aspect of this excitatory modulation is that the net effect on the column's firing rate depends on activation of the target column, especially in cases where the populations have different thresholds or gains. The reason for this is that long-range input converges onto inhibitory and excitatory target populations; thus, the excitatory target population receives direct excitation and indirect inhibition. The balance between those two sources determines whether the total input inhibits or excites the target population (*differential interaction*).

Nonclassical Receptive Fields

Models with long-range connections have been examined to find an explanation for the so-called nonclassical receptive fields of neurons. Most visual cortical cells have shown dramatic changes in their response to a stimulus within their normal receptive field, when an additional stimulus has been presented outside that region (this additional stimulus alone would elicit no response). Typical phenomena include an increase in the response, if the two stimuli have orthogonal orientations, and a decrease if the stimuli are parallel in orientation (Sillito et al., 1995). The latter modulatory effect may change its sign when the stimuli are presented at a lower luminance level. These findings can largely be explained by population models with long-range interactions; in particular, it is easy to explain the sign change with dynamical properties relying on

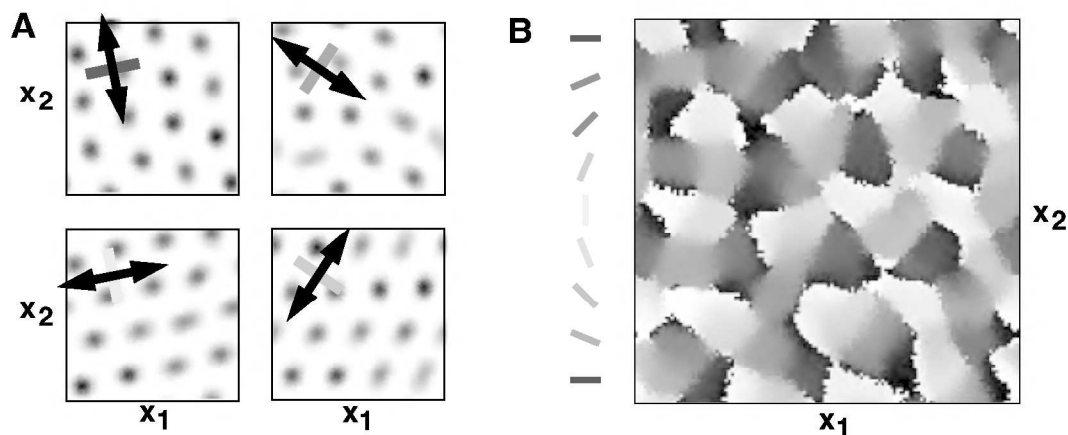


Figure 2. Orientation preference arises from the interaction of jitter in the neuronal connections with an oriented moving stimulus in a two-dimensional neural layer (i.e., $\vec{x} = (x_1, x_2)$). A, Blob pattern emerging on presentation of moving gratings, covering the whole input space R , and

having different orientations, as shown by the bars. B, Vectorial sum of the single-condition blob patterns in A for different orientations, coded in scales of gray. The picture strongly resembles orientation maps obtained experimentally with voltage-sensitive dyes. (Adapted from Ernst et al., 2001.)

the differential interaction scheme (see Stetter, Bartsch, and Obmayer, 2000, and references therein). More information on non-classical receptive fields can be found in *VISUAL CORTEX: ANATOMICAL STRUCTURE AND MODELS OF FUNCTION* (q.v.).

Association Fields

Another type of cortical coupling function is motivated by *association fields* measured in psychophysical experiments. Association fields quantify how the presentation of a bar at position (x_1, x_2) with orientation Φ will increase or decrease the threshold for detecting a bar at position (x'_1, x'_2) with orientation Φ' . The coupling matrix and model dimension are similar to the models employed in the last paragraph, with one important difference: the coupling function W is chosen according to the association field and therefore is not only orientation selective, but also directionally biased. In other words, two columns best responding to oriented bars being aligned in succession will be connected with a positive weight, while two columns best responding to oriented bars being aligned in parallel will be connected with a negative weight (Figure 3A) or will remain unconnected.

Contour Integration

An aspect of cortical information processing that can be examined and understood in this type of model is the dynamics of contour integration. Contours can be interpreted as a succession of aligned bars; thus, a coupling matrix based on the association field is especially suited to enhance the activity of columns stimulated by elements of the contour, whereas the activity of columns stimulated by distractors becomes suppressed. Li (1999, 2001) has accumulated evidence that contour integration may be explained by this kind of cortical model (Figures 3B and 3C). A close relation of

modeling work and psychophysical experiment shows that the structural simplicity of the Wilson-Cowan model class allows making specific predictions about certain experiments while opening the door to an understanding of the mechanisms at work behind the scenes.

Transient Dynamics and Feature Binding

The modeling approaches discussed so far have focused on the long-term behavior of solutions of Wilson-Cowan type of equations. In particular, steady states of the system and their stability have been associated with phenomena of cortical physiology and psychophysics. In this section, we study the transient dynamics of coupled neural populations and link it to perceptual phenomena in the context of feature binding.

Feature Inheritance and Shine-through

The spatiotemporal behavior of the visual system can be assessed psychophysically through experiments in which stimuli are presented successively for short time intervals. The visual system is thus forced to work at its spatial and temporal limit, resulting in illusions that elucidate cortical mechanisms of signal processing.

Two such illusions have recently been described (Herzog and Koch, 2001; Herzog, Fahle, and Koch, 2001). In the so-called *feature inheritance effect* (Figure 4A), a single vernier—two bars that are slightly displaced—is presented for a brief time (i.e., 10–30 ms, depending on the individual performance of the subject). The vernier is followed by a double grating of five nondisplaced bars that is presented for 300 ms. Psychophysically, subjects are not aware of the vernier but perceive a displaced grating. That is, the vernier is masked by the grating, which inherits the vernier's displacement. The inheritance effect has also been demonstrated for other fea-

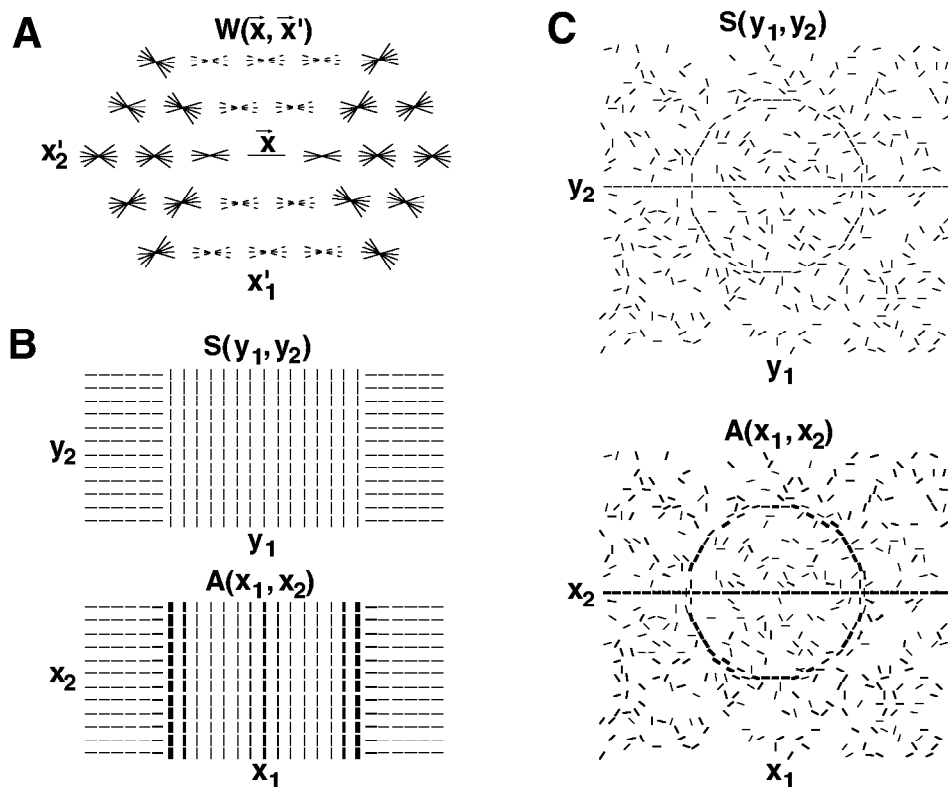


Figure 3. A, Coupling scheme W connecting the center column at \bar{x} with the surround columns at \bar{x}' ($\bar{x} = (x_1, x_2, \Phi)$). Excitatory connections are marked with thin bars, inhibitory connections with broken bars. The orientation of the bars denotes the difference in orientation preference of the connected columns. B and C, Sample stimuli $S(\bar{y})$, $\bar{y} \in R$, and the corresponding activation pattern $A(\bar{x})$, $\bar{x} \in C$. The activation level and the orientation preference of the corresponding column are coded by the thickness and the orientation of the bar, respectively. The connection scheme together with the model's dynamics lead to the enhancement of (orientation) discontinuities (*edge detection*, B) and to an increase in the activity of detectors stimulated by a closed contour (*contour integration*, C). (Adapted from Li, 1999.)

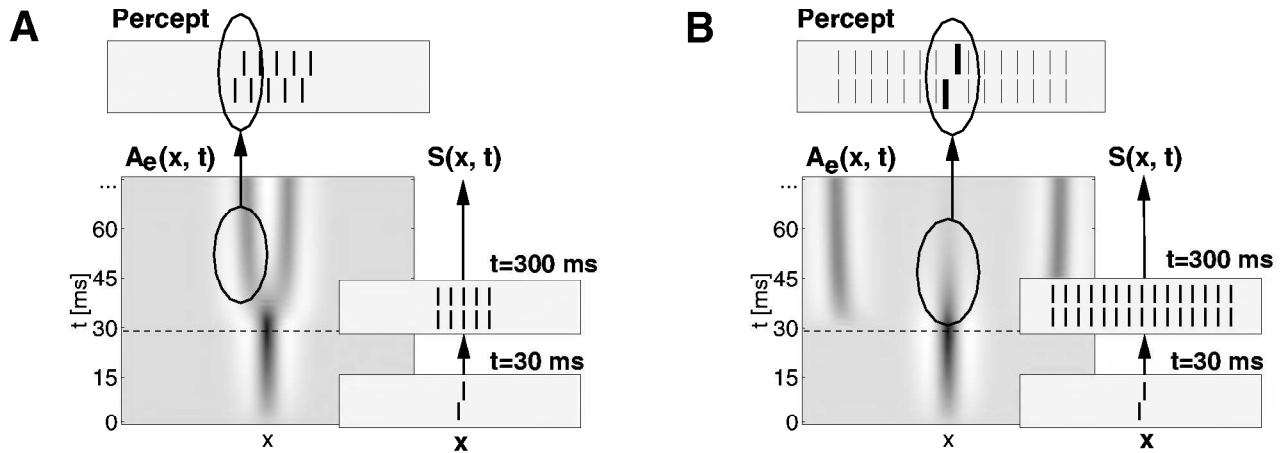


Figure 4. Stimuli, visual percept, and simulation result for the feature inheritance and shine-through effects. In the feature inheritance condition (A), a single vernier is followed by presentation of a grating of five bars (right panels). The percept is a displaced grating (top). The simulation shows the activity of the excitatory population in gray-scale coding. The central peak resulting from the vernier is rapidly suppressed by the edge activity of the

grating. In the shine-through condition (B), the vernier is followed by presentation of an elongated grating of 25 bars (right). Perceptually, the vernier looks superimposed on the grating (top). The simulation of the excitatory population reveals that in this case, the central vernier activity persists for a longer time, leading to conscious perception of the shine-through element.

tures, such as orientation and apparent motion (Herzog and Koch, 2001).

Changes in the geometrical arrangement of the grating can modify or even abolish the feature inheritance effect. An example is shown in Figure 4B, where an extended grating of 25 bars follows the presentation of the vernier. In this case, subjects are aware of the vernier, which appears superimposed on the grating (*shine-through effect*).

Vernier Visibility as a Transient Effect

The spatiotemporal version of the Wilson-Cowan equations (Equations 4 and 5) can be used to account for the vernier visibility in the different masking conditions. To elucidate the underlying neural mechanisms, a simple, one-dimensional version without the property of orientation tuning is employed. Consider an excitatory and an inhibitory population of cortical neurons arranged along a one-dimensional axis, $C = \mathbb{R}$. The input space is also taken to be one-dimensional, $R = \mathbb{R}$. For reasons of simplicity, we assume symmetry in the weights ($w_{ee} = w_{ei}$; $w_{ie} = w_{ii}$) and in the interaction kernels ($W_{ee} = W_{ei} = W_e$; $W_{ie} = W_{ii} = W_i$). The latter are modeled as Gaussians (cf. Equation 7). The external input current is identical for both populations, $I_e(x, t) = I_i(x, t) \equiv I(x, t)$, and is given by a convolution of the presented spatiotemporal stimulus intensity $S(x, t)$ with a Mexican hat type of filter $V(x - x')$ whose integral vanishes,

$$V(x - x') = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{(x - x')^2}{2\sigma_e^2}\right) - \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x - x')^2}{2\sigma_i^2}\right)$$

resembling on-off receptive field properties of LGN neurons. The stimulus $S(x, t)$ takes the value 1 if it is part of the vernier or a bar element, and 0 otherwise.

The system parameters—kernel widths, synaptic weights, population time constants, and gain functions—are adjusted considering symmetries and relations in cortical anatomy and physiology.

Numerical results for the feature inheritance and shine-through conditions as described above are given in Figure 4. The gray-

scale-coded activities of the excitatory populations show peaks at the position of the vernier and at the edges of the gratings, whereas almost no activity emerges in the bulk of the gratings. A comparison of the central peaks reveals that in the feature inheritance condition (Figure 4A), the vernier activity decays earlier than in the shine-through condition (Figure 4B). This is due to a strong inhibition by the active neurons representing the nearby edges of the grating. However, in the extended grating comprised of 25 bars, the edges are too far away to exert an influence on the center. The fast suppression of the vernier activity by the small grating shown in Figure 4A leads to a complete masking of the vernier element and a subsequent erroneous binding of its feature, the displacement, to the grating. On the other hand, conditions that allow a longer persistence of the vernier activity, like the one shown in Figure 4B, result in a conscious perception of the vernier and its displacement. Thus, the occurrence of feature inheritance or shine-through is explained by the transient dynamics of a Wilson-Cowan type of model.

The model can be applied to a number of further stimulus conditions and provides quantitative predictions for the visibility of the vernier element with a single set of model parameters. The model is robust with respect to parameter changes, and the overall results are the same no matter whether the dynamical equations are formulated for the population firing rates or the average membrane potentials. In fact, Li was also able to see the described transient behavior in her cortex model (Li, personal communication). The reduced one-dimensional model presented here also yields an analytical access and allows the identification of neural mechanisms responsible for the observed psychophysical effects (Herzog, Ernst, and Eurich; see <http://www-neuro.physik.uni-bremen.de/institute/research/vernier.html>)

Discussion

The Wilson-Cowan model class yields a description of the behavior of coupled neural populations on a coarse time scale. Versions of the model include purely temporal behavior and spatiotemporal behavior in one and two spatial dimensions, and may incorporate further stimulus features such as the orientation of edges as additional model dimensions. The relatively simple structure of the

equations allows an analytical and thorough numerical access to the system dynamics. In recent years the model class has been successfully employed to account for various physiological and psychophysical phenomena of the visual system such as orientation selectivity, cortical map formation, figure-ground segregation, feature binding, and masking effects.

Phenomena outside the visual system are beyond the scope of this article. The same holds for several dynamical aspects of the population equations that have not been addressed; among these are hysteresis phenomena and limit-cycle activity (Wilson and Cowan, 1972). For example, Tsodyks et al. (1997) have modeled oscillatory neural activity in rat hippocampus.

An important extension of the Wilson-Cowan model class is obtained if the simplification of time coarse graining is dropped. The search for appropriate equations describing the behavior of neural populations also on fast time scales and under the consideration of noise is a topic of much current interest. A suggestion that has been put forward in this context is the use of a Fokker-Planck equation; see INTEGRATE-AND-FIRE NEURONS AND NETWORKS and Knight (2000) and references therein for a framework of a variety of such approaches.

Road Map: Neural Coding

Related Reading: Direction Selectivity; Integrate-and-Fire Neurons and Networks; Layered Computation in Neural Networks; Orientation Selectivity; Pattern Formation, Neural; Visual Cortex: Anatomical Structure and Models of Function; Winner-Take-All Networks

References

- Ben-Yishai, R., Bar-Or, R., and Sompolinsky, H., 1995, Theory of orientation tuning in visual cortex, *Proc. Natl. Acad. Sci. USA*, 92:3844–3848.
- Ernst, U. A., Pawelzik, K. R., Sahar-Pikielny, C., and Tsodyks, M. V., 2001, Intracortical origin of visual maps, *Nature Neurosci.*, 4:431–436.
- Herzog, M. H., Fahle, M., and Koch, C., 2001, Spatial aspects of object formation revealed by a new illusion, shine-through, *Vision Res.*, 41:2325–2335.
- Herzog, M. H., and Koch, C., 2001, Seeing properties of an invisible object: Feature inheritance and shine-through, *Proc. Natl. Acad. Sci. USA*, 98:4271–4275.
- Knight, B. W., 2000, Dynamics of encoding in neural populations: Some general mathematical features, *Neural Computat.*, 12:473–518.
- Li, Z., 1999, Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex, *Netw. Comput. Neural Syst.*, 10:187–212.
- Li, Z., 2001, Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computat.*, 13:1749–1780.
- Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davis, J., 1995, Visual cortical mechanisms detecting focal orientation discontinuities, *Nature*, 378:492–496.
- Stetter, M., Bartsch, H., and Obermayer, K., 2000, A mean field model for orientation tuning, contrast saturation and contextual effects in area 17, *Biol. Cybern.*, 82:291–304.
- Swindale, N. V., 1996, The development of topography in the visual cortex: A review of models, *Netw. Comput. Neural Syst.*, 7:161–247. ♦
- Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., and McNaughton, B. L., 1997, Paradoxical effects of external modulation of inhibitory interneurons, *J. Neurosci.*, 17:4382–4388.
- Wilson, H. R., 1999, *Spikes, Decisions, and Actions*, Oxford: Oxford University Press. ♦
- Wilson, H. R., and Cowan, J. D., 1972, Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.*, 12:1–24.
- Wilson, H. R., and Cowan, J. D., 1973, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik*, 13:55–80.

Covariance Structural Equation Modeling for Neurocognitive Networks

Anthony Randal McIntosh

Introduction

Covariance structural equation modeling (CSEM) has proven to be an important analytic tool in functional neuroimaging research. Along with other network analytic methods that are focused on relating the interactions between brain areas with cognition, CSEM can provide important insights into the operation of large-scale neurocognitive systems. This review is divided into three sections. The first outlines the theoretical and technical basis for the examination of large-scale neural systems. The second provides a brief summary of some applications of CSEM to human neuroimaging data collected using either positron emission tomography (PET) or functional magnetic resonance imaging (fMRI). There are several applications of CSEM to studies of normal aging and patient populations that will not be reviewed here (Maguire, Vargha-Khadem, and Mishkin, 2001). The article concludes with some speculation on the utility of network analyses in developing brain theory and how it may be used in conjunction with synthetic neural modeling approaches. For related points on this topic, see SYNTHETIC FUNCTIONAL BRAIN MAPPING.

Theoretical Basis of Network Analysis and the Tools

The driving assumption behind the use of CSEM is that the covariances/correlations of activity are measures of neural interac-

tions. The activity measures may be electromagnetic (e.g., field potentials) or hemodynamic (e.g., cerebral blood flow, oxygen consumption). Neural interactions refer to influences that different elements in the nervous system have on each other via synaptic communication (the term “elements” refers to any constituent of the nervous system, either a single neuron or collections thereof). Activity changes in one neural element usually result from a change in the influence of other connected elements, so focusing only on activity in one area cannot identify the change in afferent influence. Furthermore, it is logically possible for the influences on an element to change without an appreciable change in measured activity. The simplest example would be one in which an afferent influence switches from one source to another without a change in the strength of the influence. Monitoring regional activity alone would miss this shift, but measures of the relation of activity between elements—e.g., the covariance—would not.

Structural Equation Modeling

The foundation for covariance analysis in neuroimaging was laid by Horwitz in a number of papers that looked at regional interrelations in a pairwise manner (Horwitz, Duara, and Rapoport, 1984). Covariance structural equation modeling (CSEM), or path analysis, is a logical extension from this; it uses the anatomical connections

between areas in an attempt to characterize the effects between areas, called path coefficients, within a larger functional network (for a complete review, see Bollen, 1989; McIntosh and Gonzalez-Lima, 1994). The covariances among the variables, computed either across subjects within a specific cognitive task or across tasks within a subject, are used to provide weights for the anatomical links in a manner similar to a multiple linear regression.

Covariances used in multivariate analyses can identify the dominant functional and/or effective connections during the performance of a cognitive or behavioral operation. In the context of neuroimaging, *functional connectivity* is a statement that two regions show some non-zero correlation of activity but does not specify how this correlation comes about, while *effective connectivity* is a statement about the direct effect one region has on another, accounting for mutual or intervening influences (Friston, 1994). For the present review, measures of covariances are estimations of functional connectivity while the path coefficients derived from CSEM are estimates of effective connectivity.

In terms of basic equations, consider a four-node network that can be expressed by a series of structural equations (regression equations) as follows:

$$\begin{aligned} A &= \psi_A \\ B &= wA + \psi_B \\ C &= vA + yB + \psi_C \\ D &= xA + zB + \psi_D \end{aligned}$$

where w , x , y , and z are parameters to be estimated (the path coefficients) and ψ denotes a residual influence unique to that region. The residual term is best conceived of unique effects on the region that are not part of the network under consideration.

The network can also be expressed as a series of matrices:

$$\begin{pmatrix} \eta \\ A \\ B \\ C \\ D \end{pmatrix} = \begin{pmatrix} \beta \\ 0 & 0 & 0 & 0 \\ w & 0 & 0 & 0 \\ v & y & 0 & 0 \\ x & z & 0 & 0 \end{pmatrix} * \begin{pmatrix} \eta \\ A \\ B \\ C \\ D \end{pmatrix} + \begin{pmatrix} \psi \\ \psi_A & 0 & 0 & 0 \\ 0 & \psi_B & 0 & 0 \\ 0 & 0 & \psi_C & 0 \\ 0 & 0 & 0 & \psi_D \end{pmatrix}$$

where η contains the variances of the regions, β contains the path coefficients, and ψ contains the residuals.

Estimates of the path coefficients and residuals are derived from the original interregional covariances. Let the covariance matrix among regions be σ , and let the covariances implied by the estimates for β and ψ be Σ , which is computed by

$$\Sigma = \text{inv}(I - \beta)^T * (\psi^T * \psi) * \text{inv}(I - \beta)$$

where inv denotes matrix inversion, I is an identity matrix, and T denotes a matrix transpose. Initial estimates for β and ψ are usually done by using a variation of least squares, and then the estimates are improved through iterative fitting to minimize the difference between σ and Σ . There are a number of different iterative search algorithms and fitting functions that have been used in CSEM, which are reviewed by Bollen (1989).

The path coefficients can also be compared statistically between different tasks or groups. This has been the primary application of CSEM to neuroimaging, where the goal is to determine whether the effective connections within the same anatomical network vary depending on cognitive challenge. Some recent extensions have used CSEM to test hypothesized models of effective connections for specific cognitive operations (Bullmore et al., 2000).

Empirical Examples

Dorsal and Ventral Stream Processing in Perceptual Matching

One well-established functional distinction in the brain is between object and spatial visual pathways. The foundation for this dual organization can be traced at least as far back as the 1930s, and one of its strongest expressions to date is in the dorsal and ventral cortical processing streams, described by Ungerleider and Mishkin (1982), which corresponds to object and spatial processing pathways, respectively. A similar duality was identified in humans with the aid of PET (Haxby et al., 1991). In an attempt to characterize the interactions within these cortical pathways, McIntosh et al. (1994) used CSEM to explore cortical interactions that were specific to object and spatial processing.

A match-to-sample task for faces was used to explore object vision. For spatial vision, a match-to-sample task for the location of a dot within a square was used. The results from the right hemisphere analysis are presented in Figure 1 (left hemisphere interactions did not differ between tasks). Path coefficients along the ventral pathway from cortical area 19v extending into the frontal lobe were stronger in the face-matching model, while interactions along the dorsal pathway from area 19d to the frontal lobe were relatively stronger in the location-matching model. Among posterior areas, the differences in path coefficients were mainly in magnitude. Occipitotemporal interactions between area 19v and area 37 were stronger in the face-matching model, while the impact of area 17/18 to 19d and the occipitoparietal influences from area 19d to area 7 were stronger in the location matching model.

The anatomical connections allowed for interactions between the dorsal and ventral pathways with connections from area 37 to area 7 and from area 7 to area 21. The interactions among these areas showed task-dependent differences in magnitude and sign. The functional networks show that while the strongest positive interactions in each model may have preferentially been in one pathway, the parallel pathways were not functioning independently. Strong interactions between parallel pathways have been a consistent finding in all CSEM applications to imaging data. Therefore, while a certain pathway or area may be critical for a particular function, operations in the intact brain involve interactions among many regions.

Spatial Attention

Network interactions that underlie cognitive operations are expressed as a change in the effective connections between parts of the network. As is illustrated above, if visual attention is directed to the features of an object, effective connections among ventral posterior cortical areas tend to be stronger, whereas visual attention directed to the spatial location leads to stronger interactions among dorsal posterior areas. Another way that cognitive operations may express is through the modulation of effective connections between regions. In this case, an area may provide an enabling condition to foster communications between regions. Such an effect was the focus of a model of visuospatial attention by Buchel and Friston (1997).

In this fMRI study, subjects alternated between periods of overt attention for a change to a moving visual dot pattern and periods when they were not attending to the display. Two models were evaluated. The first was a feedforward network from primary visual (V1) to dorsal occipital cortex (V5) to posterior parietal (PP) cortices. In this model, the "attend" condition showed stronger path coefficients than when there was no direct attention to the display. The second model assessed whether prefrontal

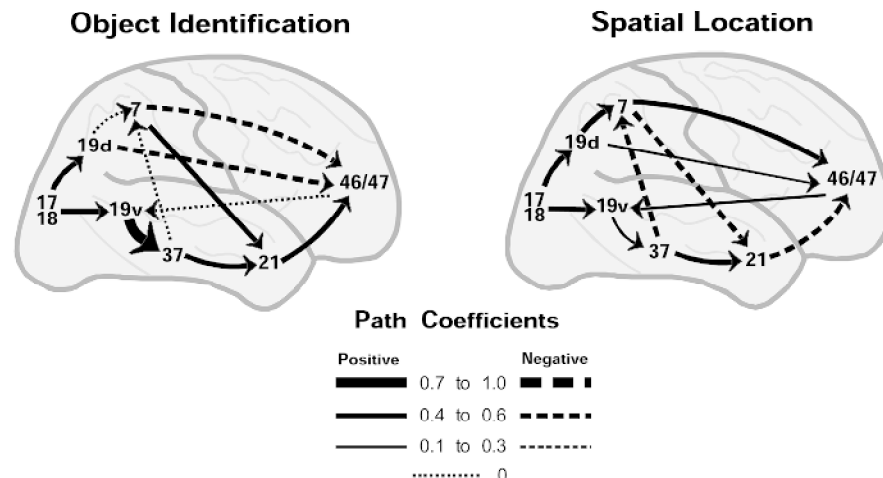


Figure 1. Functional interactions between cortical areas in the right hemisphere for object and spatial vision operations. The numbers on the cortical surface refer to Brodmann areas (d = dorsal, v = ventral). The arrows represent the anatomical connections between areas, and the magnitude of the direct effect from one area to another is proportional to the arrow width for each path.

cortex (PFC) had a modulatory effect in the effective connections between V5 and PP. This model had a direct influence from V5 to PP as well as the interaction term where the effect from V5 to PP depended on PFC (PFCmod). The expectation was that if the attentional effects manifest only on the direct connection from V5 to PP, then the estimate for PFCmod would be zero. The authors demonstrated that the modulatory effect was significant for all subjects. This modulatory effect seemed to vary in an activity-dependent manner in which the relation between V5 and PP was stronger during periods when PFC activity was highest.

This model demonstrated two important points. From the methodological point, it shows how more complex neurobiological effects can be modeled in CSEM. Buchel and Friston modeled a simple interaction term, but that same approach can be used to model true nonlinear effects (Bollen, 1989). From the neurobiological perspective, it demonstrates that cognition can be expressed in the brain either as changes in the direct effect between regions, as a modulatory effect, or in some cases both. What CSEM provides is a quantitative method to examine these possibilities.

Associative Learning

One example of the use of network analysis to test specific hypotheses about regional interactions comes from study of sensory learning (McIntosh, Cabeza, and Lobaugh, 1998). The task had subjects learn an association between a tone and a visual stimulus. Using PET, brain activity was measured in response to the tone across acquisition of the association. The expectation was that as the tone acquired behavioral significance, presentation of the tone would elicit activity in visual areas. Because the activation of visual areas would occur without overt visual stimulation, the second hypothesis was that this activation would be mediated through effects from higher-order cortical areas, likely posterior association or prefrontal cortices. The effective connections between visual cortex and anterior regions were quantified by using CSEM to determine which of these candidates exerted the strongest influence on these visual areas.

Activation of occipital cortex in area 18 to the tone was observed as training proceeded, confirming the first expectation. CSEM of influences on area 18 suggested that two areas in particular seem to exert the dominant influence as the association was learned. Superior temporal cortex (auditory association, area 41/42) and prefrontal cortex near area 10 both changed their ef-

fect on area 18 from suppressive to facilitory as the association was learned.

This study demonstrated learning-related changes in effective connectivity. Although the areas selected for CSEM were related to behavior, it is not clear whether the changes in effective connectivity impacts directly on learning-related performances changes. Buchel, Coull, and Friston (1999) provided a convincing demonstration that changes in effective connectivity related directly to learning. Subjects learned to associate visually presented objects with their location in space. To confirm that the behavioral changes across the experiment related directly to network dynamics, the changes in effective connections between dorsal and ventral processing stream were correlated with subject's learning rate. A robust correlation was found between the learning rate and the change in the influence of posterior parietal cortex (dorsal stream area) on inferotemporal cortex (ventral stream area). This is consistent with the task demands and demonstrates that network dynamics can be directly related to overt changes in behavior.

Awareness and Prefrontal Cortex Interactions

As a follow-up to the simple associative learning study described above, McIntosh and colleagues further investigated the neural interactions subserving cross-modal learning using differential conditioning (McIntosh, Rajah, and Lobaugh, 1999). Two tones were used that had differential relations to the visual stimuli. One tone was a strong predictor of the presentation of a visual stimulus (Tone+), and the other tone was a weak predictor (Tone-). The sample of subjects was divided perfectly in half into those who were aware of the stimulus associations and those who were not. Furthermore, only *aware* subjects learned the differentiation between the tones, while *unaware* subjects showed no behavioral evidence of learning.

The strongest group difference in brain activity elicited by the tones was in left prefrontal cortex (LPFC) near area 9. In aware subjects, LPFC activity showed progressively greater activity to Tone- than Tone+. In unaware subjects, no consistent changes were seen in LPFC or in any of the other regions. At first, these results seem to confirm the prominent role of PFC in monitoring functions and especially its putative role in awareness. However, PFC activation has also been found in tasks in which there was no overt awareness (Berns, Cohen, and Mintun, 1997). It was thus possible that interactions of PFC with other brain regions,

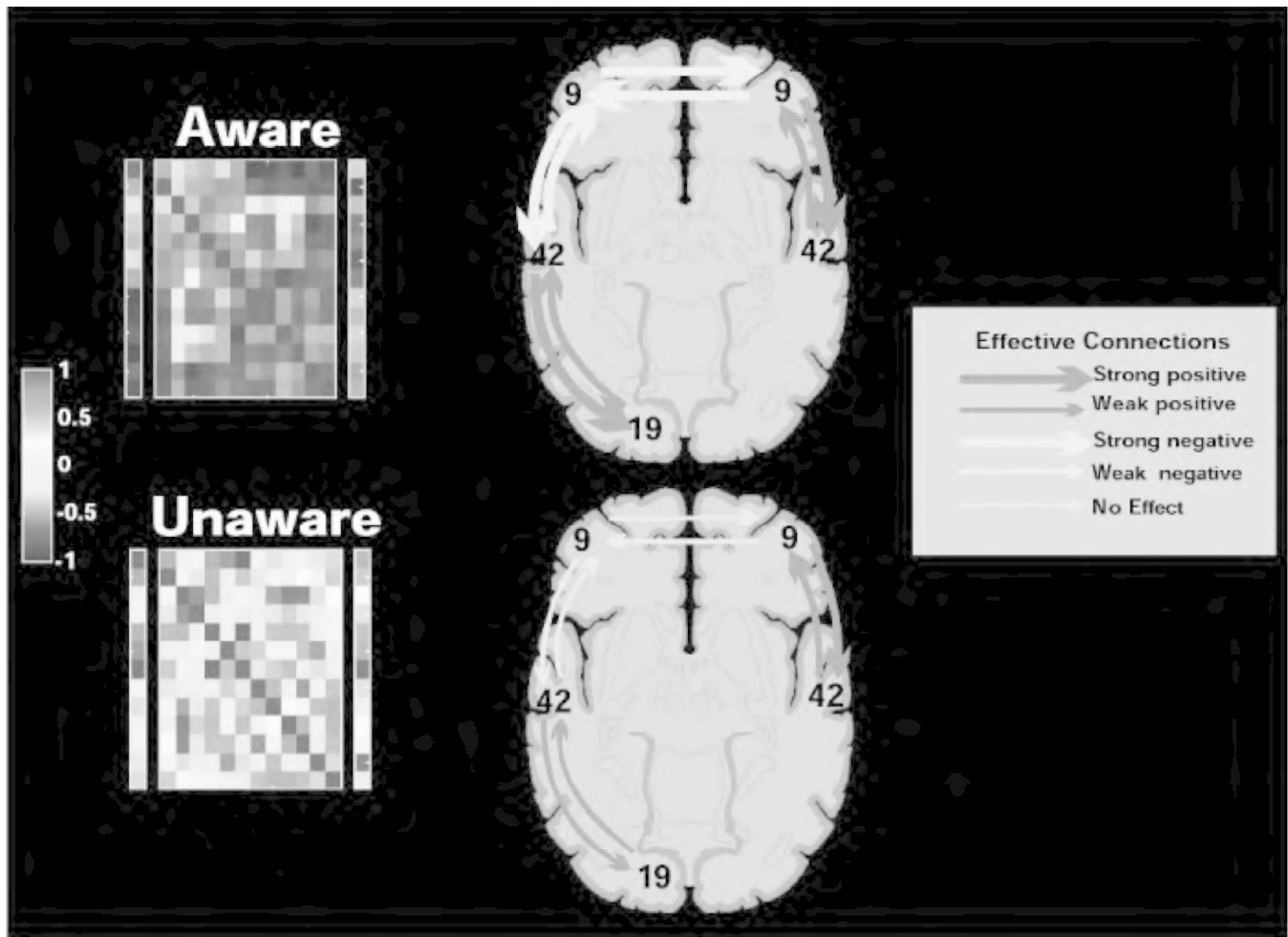


Figure 2. Correlation plots and functional networks from late phases of training in a differential sensory conditioning task. The plots on the left of the figure are pseudo-colored matrices of the correlations of left prefrontal cortex area 9 (first column) with sensory and association cortices, and the correlations of all regions with behavior are shown in the last column of the matrix. The correlation values are represented as color gradations, ac-

cording to the scale on the far left of the figure. The matrices are symmetric about the major diagonals, which are all ones. The right of the figure shows the functional networks for the two groups. Aware subjects showed strong effective connections involving left prefrontal area 9 and other regions. Conversely, the network for the unaware subjects showed no strong left prefrontal involvement.

present in aware but not in unaware subjects, would better describe the neural system underlying awareness in this task.

When the interactions of LPFC was assessed between the two groups, large differences were found in the strength and pattern of functional connections among several brain areas, including right PFC, bilateral superior temporal cortices (auditory association), occipital cortex, and medial cerebellum. These areas were much more strongly correlated in aware subjects than unaware subjects (Figure 2). To explore some of the network interactions within an anatomical reference, CSEM was applied to a subset of these regions.

As may be expected from the correlation matrices, there were significant differences in the effective connections for aware subjects, including robust interactions involving LPFC. The functional network for unaware subjects differed from that for aware subjects, but there were no significant changes in effective connections across the experiment for the unaware group, and the involvement of LPFC was weak. This confirms that LPFC was not interacting systematically across subjects in the unaware group.

Discussion: Implications for Computational Modeling and Theoretical Efforts

Object and Spatial Vision

Aside from the confirmation of the dorsal/ventral distinction, it was noted that these functionally independent parallel systems could interact during different operations. While there were dominant interactions along dorsal and ventral streams in spatial and object tasks, respectively, there was cross-talk between the two streams, which implies that they need not be functionally independent. The study of Buchel et al. (1999) demonstrates that the interactions between pathways can depend on experience, and the study of Buchel and Friston (1997) suggests mechanisms whereby other cortical areas may modulate the interactions between posterior sensory regions.

Neural Context and Network Operations

A consistent observation in CSEM applications to neuroimaging is that similar areas are engaged in many cognitive functions, and

what discriminates these cognitive operations are the pattern of interactions within large-scale networks rather than the involvement of a particular area. This has led to the concept of a *neural context*, in which the functional relevance of a particular region is determined by its interactions with other areas (McIntosh, 1999). Systems-level neuroanatomy shows that brain regions receive projections from many areas and send projections to many others. Neural context emphasizes that the precise pattern of these functionally engaged structural connections defines the translation of brain operations into cognitive operations. Consequently, the same region may show identical levels of activity across many different tasks. Because the pattern of interactions with other connected areas differs, the resulting cognitive operations vary, yet the same region is involved in each task. Stated differently, the neural context within which an area is active embodies the cognitive operation. The implications for computational modeling would be to develop models that can carry out more than one cognitive function through a change in the pattern of interactions among nodes. Thus, the focus would go from trying to model the computations in a single task to deriving a single model configuration that can do multiple tasks.

The relationship between neural modeling and experimental neuroscience continues to mature. The primary utility of modeling comes when they are derived to support specific hypotheses, and serve as frameworks for understanding complex data. Modeling efforts, whether empirically based like CSEM, theoretical, or computational, are all approximations of reality and as such are ultimately false. However, it is not necessarily the model that furthers our knowledge base. Rather, when a model is falsified and subsequently reformulated, scientific progress is made.

Road Map: Cognitive Neuroscience

Related Reading: Imaging the Visual Brain; Object Recognition, Neurophysiology; Schema Theory; Synthetic Functional Brain Mapping

References

- Berns, G. S., Cohen, J. D., and Mintun, M. A., 1997, Brain regions responsive to novelty in the absence of awareness, *Science*, 276:1272–1275.
- Bollen, K. A., 1989, *Structural Equations with Latent Variables*, New York: Wiley. ♦
- Buchel, C., and Friston K., 1997, Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modeling and fMRI, *Cereb. Cortex*, 7:768–778.
- Buchel, C., Coull, J. T., and Friston, K. J., 1999, The predictive value of changes in effective connectivity for human learning, *Science*, 283:1538–1541.
- Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S., and Sharma, T., 2000, How good is good enough in path analysis of fMRI data? *Neuroimage*, 11:289–301.
- Friston, K. J., 1994, Functional and effective connectivity: A synthesis, *Hum. Brain Mapp.*, 2:56–78.
- Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., and Rapoport, S. I., 1991, Dissociation of object and spatial visual processing pathways in human extrastriate cortex, *Proc. Natl. Acad. Sci. USA*, 88:1621–1625.
- Horwitz, B., Duara, R., and Rapoport, S. I., 1984, Intercorrelations of glucose metabolic rates between brain regions: Application to healthy males in a state of reduced sensory input, *J. Cereb. Blood Flow Metabol.*, 4:484–499.
- Maguire, E. A., Vargha-Khadem, F., and Mishkin, M., 2001, The effects of bilateral hippocampal damage on fMRI regional activations and interactions during memory retrieval, *Brain*, 124:1156–1170.
- McIntosh, A. R., 1999, Mapping cognition to the brain through neural interactions, *Memory*, 7:523–548.
- McIntosh, A. R., and Gonzalez-Lima, F., 1994, Structural equation modeling and its application to network analysis in functional brain imaging, *Hum. Brain Mapp.*, 2:2–22. ♦
- McIntosh, A. R., Cabeza, R. E., and Lobaugh, N. J., 1998, Analysis of neural interactions explains the activation of occipital cortex by an auditory stimulus, *J. Neurophysiol.*, 80:2790–2796.
- McIntosh, A. R., Rajah, M. N., and Lobaugh, N. J., 1999, Interactions of prefrontal cortex related to awareness in sensory learning, *Science*, 284:1531–1533.
- McIntosh, A. R., Grady, C. L., Ungerleider, L. G., Haxby, J. V., Rapoport, S. I., and Horwitz, B., 1994, Network analysis of cortical visual pathways mapped with PET, *J. Neurosci.*, 14:655–666.
- Ungerleider, L. G., and Mishkin, M., 1982, Two cortical visual systems, in *Analysis of Visual Behavior* (D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds.), Cambridge, MA: MIT Press, pp. 549–586.

Crustacean Stomatogastric System

Scott L. Hooper

Introduction

The stomatogastric nervous system (STNS) of decapod crustacea generates the rhythmic motor patterns of the four areas of the crustacean stomach—the esophagus, cardiac sac, gastric mill, and pylorus—and contains some of the best-understood central pattern-generating networks in neurobiology. This work has identified four widely applicable properties of STNS networks. First, rhythmicity in these highly distributed networks depends on both network synaptic connectivity and slow (tens to hundreds of milliseconds), active neuronal membrane properties. Second, modulatory influences can induce individual STNS networks to produce multiple outputs, “switch” neurons between networks, or fuse individual networks into single larger networks. Third, modulatory neuron terminals receive network synaptic input. Modulatory inputs can be sculpted by network feedback, and become integral parts of the networks they modulate. Fourth, network synaptic strengths can vary as a

function of pattern cycle period and duty cycle. Similar complex properties are present in many biological neural networks (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION; HALF-CENTER OSCILLATORS UNDERLYING RHYTHMIC MOVEMENTS; OSCILLATORY AND BURSTING PROPERTIES OF NEURONS). Introducing such neurons into artificial neural networks may afford significant functional advantages.

Background

Morphology

All known STNS motor neurons and interneurons are monopolar and have inexcitable somata; synaptic contacts and the voltage-dependent channels underlying slow-wave activity are located in neuropil processes. Spike initiation zones are located near the axon

and are electrically distant from the cell body (Selverston et al., 1976).

Synapses

Input and output synapses are located side by side on neuropil processes. Neuronal output thus results from local integration in the neuropil instead of the whole-cell integration present in neurons with distinct pre- and postsynaptic regions (Selverston et al., 1976). In the two best-studied STNS networks, the pyloric and gastric networks, synaptic release is a graded (analogue) function of membrane potential (Graubard, Raper, and Hartline, 1983). Multiple transmitters and receptors, each with their own time course, are present in STNS networks. In particular, many STNS synapses induce slow responses with characteristic times as long as 100 ms (Miller, 1987).

Cellular Properties

STNS neurons display a variety of active, long-duration cellular properties. Many show postinhibitory rebound (PIR); i.e., inhibition induces a subsequent, postsynaptic membrane-generated, active depolarization above rest (Selverston et al., 1976). Some are endogenous oscillators—neurons that spontaneously depolarize and hyperpolarize in a rhythmic fashion. Others have “plateau properties.” A plateau neuron has, in addition to a stable rest potential, a quasi-stable depolarized plateau membrane potential. Plateau transitions are regenerative; depolarization from rest above a threshold voltage activates voltage-dependent conductances that drive the neuron to the plateau, and small hyperpolarizations from the plateau induce active repolarization to rest (Russell and Hartline, 1978). Plateau properties nonlinearly transform inputs in amplitude and time in that small-amplitude inputs can induce a full transition, and brief inputs (tens of microseconds) induce long-lasting responses (uninhibited plateaus can last for seconds). STNS models suggest that oscillation, plateaus, and PIR are essential for reproducing the dynamic properties of STNS networks (Golowasch et al., 1992; Marder and Selverston, 1992; Nadim et al., 1998).

Neuromodulation

Synaptic strength and cellular properties in STNS networks are altered by modulatory influences. These influences can induce individual neurons to express plateaus or to become endogenous oscillators (Harris-Warrick, Nagy, and Nusbaum, 1992), and dramatically increase the functional repertoire of STNS neurons and networks.

Mechanisms of Central Pattern Generation

The pyloric network is the best-understood STNS network and will be used to illustrate several general principles that underlie rhythmic pattern generation in this and other systems. The pyloric network is a central pattern generator, i.e., it endogenously generates rhythmic patterns without timing cues from the rest of the nervous system. The right portion of Figure 1 shows simultaneous recordings from all six pyloric neuron types. Each cycle consists of a sequence of bursts of action potentials from the pyloric neurons (Figure 1 shows two cycles). Pyloric dilator (PD) motor neuron activity is classically used to define the beginning of the sequence; they and the anterior burster (AB) interneurons fire together. The lateral pyloric (LP) and inferior cardiac (IC) motor neurons fire next, and finally the pyloric (PY) and ventricular dilator (VD) motor neurons fire. The network's synaptic connectivity (Miller, 1987) is shown in Figure 1, left. Small circles represent inhibitory chemical synapses, resistors represent bidirectional electrical coupling, and the diode represents rectifying electrical coupling.

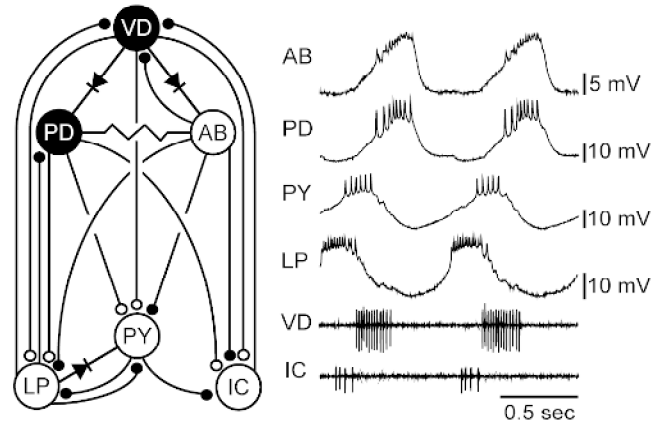


Figure 1. Pyloric network synaptic connectivity (left) and output (right). Small circles represent chemical synapses, resistors represent bidirectional electrical coupling, and the diode represents rectifying electrical coupling. Calibration bar: 20 mV for PD and LP neurons, 10 mV for other neurons.

and the diode represents rectifying electrical coupling (current flows only in the direction of the arrow). The white neurons are glutamatergic and induce rapid, short-lasting inhibition in their followers; the black neurons are cholinergic and induce slow, long-lasting inhibition.

It might be difficult initially to understand how a network dominated by inhibitory synapses is rhythmically active. One solution would be for all network neurons to be endogenous oscillators, and under some circumstances all pyloric neurons can be (Bal, Nagy, and Moulins, 1988). However, the network also produces a slow, but correctly ordered, rhythmic output even in cases in which none of its neurons are endogenous oscillators (Miller, 1987). This rhythmicity arises because the network has multiple locations in which two neurons inhibit each other. Mutual inhibition can lead, in neurons possessing PIR and plateau potentials, to a synaptically based rhythmicity called half-center oscillation. Rhythmicity arises in this synaptic arrangement because the inhibition caused by one neuron's firing induces a delayed PIR, plateau, and firing in the second. The inhibition of the first neuron caused by the second neuron's firing induces in turn a delayed PIR, plateau, and firing in the first neuron, after which the process repeats. This multiplicity of mechanisms makes it difficult to ascribe specific aspects of network function to specific network neurons. Network rhythmicity and pattern instead arise emergently from a combination of neuronal cellular properties and the network's distributed synaptic connectivity.

Nonetheless, a qualitative understanding of pyloric activity can be achieved by considering the cellular properties of the network's neurons and the network's synaptic interconnectivity. The AB neuron is generally the fastest oscillator and drives the PD neurons through their electrical coupling. These neurons inhibit all other pyloric neurons. The LP and IC neurons rebound most quickly after this inhibition and fire next. These neurons inhibit the VD and PY neurons, but eventually the PY neurons rebound and fire. PY neuron firing inhibits the IC and LP neurons and releases the VD neuron from inhibition. The PY and VD neurons are turned off by the next AB/PD neuron burst, and the cycle repeats.

Multifunctional Networks

The pyloric network produces different outputs in response to modulatory inputs (Harris-Warrick et al., 1992). Pyloric network complexity may thus exist to allow construction of many functional

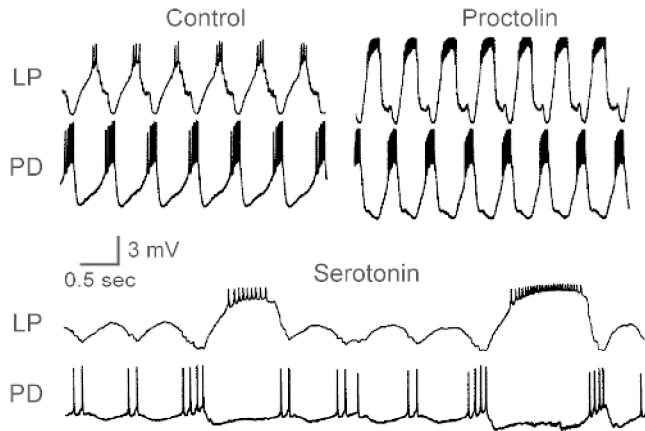


Figure 2. Proctolin (top right panel, 10^{-6}M) and serotonin (bottom panel, 10^{-4}M) induce the pyloric network to produce different motor patterns. (Modified from Marder, E., Hooper, S. L., and Eisen, J. S., 1987, Multiple neurotransmitters provide a mechanism for the production of multiple outputs from a single neuronal circuit, in *Synaptic Function* (G. M. Edelman, W. E. Gall, and M. W. Cowan, Eds.), New York: Wiley, pp. 305–327.)

configurations from a single anatomical network. Figure 2 shows the pyloric outputs induced by the neuromodulators proctolin and serotonin. Proctolin increases LP neuron activity (top trace). Serotonin induces a pattern in which, instead of the LP and PD neurons bursting in alternation, the LP neuron bursts once every two to three PD neuron bursts. Network multifunctionality is not limited to the pyloric network; the gastric mill network also produces multiple outputs. Work in *in vivo* and semiintact preparations shows that these networks produce multiple outputs in the animal as well, and thus the multiple outputs observed *in vitro* (and the network switches and fusions noted below) may be behaviorally relevant. Direct evidence for this contention exists in two cases. The first is modulation of the gastric mill rhythm by a cholecystokinin-like hormone whose hemolymph concentration rises after feeding. The second is feedback from stomach proprioceptive neurons that contain acetylcholine and serotonin and that have short-acting, classical (cholinergic) and long-lasting modulatory (serotonergic) effects that would increase pyloric and gastric mill activity (Turrigiano and Heinzel, 1992).

The cellular targets of several modulators are known (Harris-Warrick et al., 1992). This work has two generally relevant results. First, many modulators alter the inherent cellular properties of their target neurons. Second, because directly affected neurons alter the activity of other network neurons, and because these changes modify the responses of directly affected neurons, network activity changes cannot be explained solely by considering the directly affected neurons. Network rearrangements are instead global responses of the entire network, and so not only network rhythmicity but also network modulatory responses are *distributed* functions in these networks.

Neuron Switching and Network Fusion

Until now the four STNS networks have been treated as separate entities. However, the stomach is anatomically compact, adjoining stomach regions are mechanically coupled, and coordination of the motor patterns these networks produce is likely critical for behavior. Considerable internetwork connectivity exists, and network boundaries are variable. These interactions range from simple coordination between networks to cases in which neurons are switched between networks to cases in which networks fuse to form

new networks. Figure 3 shows an example of fusion (Meyrand, Simmers, and Moulins, 1991). The upper panel shows simultaneous recordings, obtained under control conditions, of the VD and PD neurons (pyloric network) and the gastric mill (GM) and lateral posterior gastric (LPG) neurons (gastric mill network). The bottom panel shows the activity of these neurons after discharge of an identified modulatory neuron; the two networks now produce a single conjoint rhythm different from either the pyloric or the gastric rhythm.

The mechanisms underlying network switching and fusion are known in two cases. In one (a switch), the change was due to plateau suppression in the switching neuron (Hooper and Moulins, 1989); in the other (a fusion), it was due to increased strength of an internetwork synaptic connection (Dickinson, Meccas, and Marder, 1990). Nevertheless, the effects on total network activity again could be understood only by considering the entire network. Long-lasting cellular properties and distributed network architecture thus underlie not only STNS central pattern generation and multifunctionality, but also many aspects of internetwork interaction and restructuring.

Sculpting of Modulatory Input by Network Local Feedback

Many STNS modulatory fibers arise from physically distant cell bodies. The terminals of these fibers receive presynaptic input from, and can serve as integral members of, the networks they modulate. Figure 4 shows one such input, modulatory commissural neuron 1 (MCN1) (Coleman, Meyrand, and Nusbaum, 1995). MCN1 chemically excites the lateral gastric (LG) neuron and interneuron 1 (Int1) of the gastric mill network, and is electrically coupled to the LG neuron (Figure 4A). When MCN1 is active, Int1 fires first, owing to MCN1's fast chemical synapse onto it. Int1 firing inhibits the LG neuron, and this reduces the effectiveness of MCN1's electrical input to the LG neuron (left, Figure 4B). MCN1's slow chemical excitation eventually brings the LG neuron to threshold. LG neuron firing inhibits Int1, presynaptically inhibits

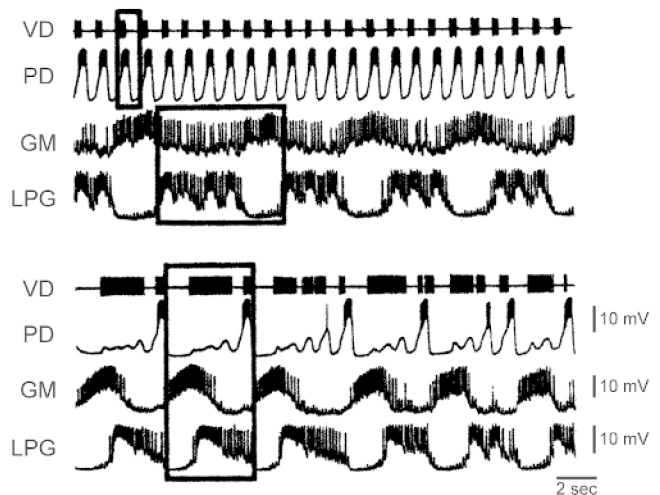


Figure 3. A defined modulatory input fuses networks that are generally distinct (top panel; pyloric, VD, PD neuron traces; gastric mill, GM, LPG neuron traces) into a single network that produces a novel output (bottom panel). (From Meyrand, P., Simmers, J., and Moulins, P., 1991, Construction of a pattern-generating circuit with neurons of different networks, *Nature*, 351:60–63, Figure 3. © 1991, Macmillan Magazines Ltd. Reprinted with permission.)

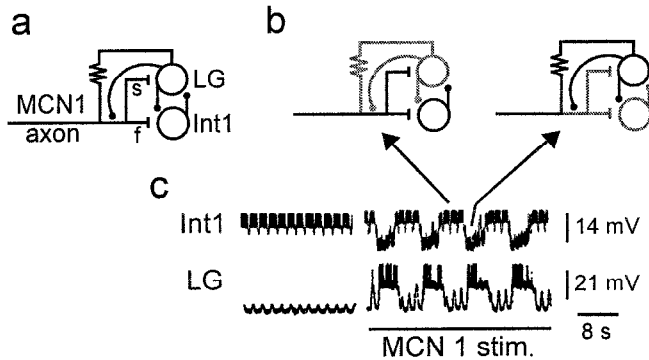


Figure 4. Presynaptic inhibition can sculpt modulator input, and make input terminals integral parts of the networks they modulate. *A*, Synaptic connectivity pattern. MCN1 (a descending modulatory input) makes a fast excitatory synapse onto Int1 and a slow excitatory synapse onto the lateral gastric (LG) neuron. The LG neuron is electrically coupled to, and presynaptically inhibits, the MCN1 axon. The LG neuron and Int1 inhibit each other. *B*, Local feedback rhythmically switches the active mode of MCN1 input. *C*, Tonic MCN1 input (MCN1 active in right panel) activates the gastric mill network. (From Coleman, M. J., Meyrand, P., and Nusbaum, M. P., 1995, A switch between two modes of synaptic transmission mediated by presynaptic inhibition, *Nature*, 378:502–505. © 1995, Macmillan Magazines Ltd. Reprinted with permission.)

MCN1's chemical input to the LG neuron and Int1, and increases the effectiveness of MCN1's electrical input to the LG neuron (right, Figure 4*B*). This electrical input maintains LG neuron activity for a period, but eventually, owing to the lack of MCN1 excitatory chemical input, the LG neuron ceases firing. This removes the LG neuron's presynaptic inhibition of MCN1 input to Int1, Int1 is driven to fire, and the cycle repeats. This switching between different modes of MCN1 input to the gastric mill network, itself driven by gastric activity, continues throughout gastric cycling (Figure 4*C*, right, arrows; the left part shows gastric mill activity before MCN1 activity).

Dependence of Synaptic Strength on Pattern Activity

As noted earlier, modulatory effects cannot be fully understood without considering the entire network. A striking example is the observation that synaptic strength depends on pattern cycle period (Figure 5; Manor et al., 1997). The panels show PD neuron response to rhythmic, square-wave LP neuron depolarizations (recall that pyloric neurons release transmitter as a graded function of membrane potential) at 2.5-s (left) and 0.75-s (right) periods. The PD neuron response is much less at the faster period: Inputs that

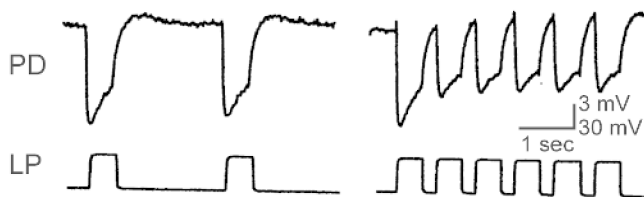


Figure 5. LP to PD neuron synaptic strength varies as a function of LP neuron cycle period. PD neuron postsynaptic response is much less at a 0.75-s LP neuron cycle period (right) than at a 2.5-s period (left). (Modified from Manor, Y., Farzan, N., Abbott, L. F., and Marder, E., 1997, Temporal dynamics of graded synaptic transmission in the lobster stomatogastric ganglion, *J. Neurosci.*, 17:5610–5621.)

alter pattern period can thus alter synaptic strength without directly affecting synaptic transmission. These activity-dependent synaptic strength changes clearly would be fundamentally important in shaping network response to modulation, and provide additional evidence of the interlinking of direct and indirect responses that allow these networks to respond to modulatory input in a global fashion.

Discussion

Work in a well-defined invertebrate nervous system has revealed several cellular and synaptic properties essential for biological function that often are not included in artificial neural networks. Chief among these are the following:

1. Synapses with different characteristic time courses present in the same network.
2. Neurons with long-lasting (tens to hundreds of milliseconds) voltage-dependent conductances that give rise to complex, long-duration cellular characteristics (oscillation, plateaus, PIR).
3. Neuromodulatory inputs that alter synaptic strength and inherent cellular properties.
4. Local feedback that alters neuromodulatory input activity and makes modulatory inputs an integral part of the network they modulate.
5. Network synaptic strengths that depend on network activity (cycle period, duty cycle).

Modeling work shows that many of the functional characteristics (rhythmic pattern production, multiple pattern production, neuron switching, network fusion) of STNS networks require model neurons and networks that incorporate these properties (Golowasch et al., 1992; Marder and Selverston, 1992; Nadim et al., 1998). Including model neurons with similar properties in neural network simulations may enhance artificial neural network capability and deepen understanding of both artificial and biological neural network function.

Road Maps: Motor Pattern Generators; Neuroethology and Evolution

Related Reading: Half-Center Oscillators Underlying Rhythmic Movements; Ion Channels: Keys to Neuronal Specialization; Neuromodulation in Invertebrate Nervous Systems; Oscillatory and Bursting Properties of Neurons; Respiratory Rhythm Generation

References

- Bal, T., Nagy, F., and Moulins, M., 1988, The pyloric central pattern generator in crustacea: A set of conditional neuronal oscillators, *J. Comp. Physiol.*, 163:715–727.
- Coleman, M. J., Meyrand, P., and Nusbaum, M. P., 1995, A switch between two modes of synaptic transmission mediated by presynaptic inhibition, *Nature*, 378:502–505.
- Dickinson, P. S., Mecsas, C., and Marder, E., 1990, Neuropeptide fusion of two motor pattern generator circuits, *Nature*, 344:155–158.
- Golowasch, J., Buchholtz, F., Epstein, I. R., and Marder, E., 1992, Contribution of individual ionic currents to activity of a model stomatogastric ganglion neuron, *J. Neurophysiol.*, 67:341–349.
- Graubard, K., Raper, J. A., and Hartline, D. K., 1983, Graded synaptic transmission between identified spiking neurons, *J. Neurophysiol.*, 50:508–521.
- Harris-Warrick, R. M., Nagy, F., and Nusbaum, M. P., 1992, Neuromodulation of stomatogastric networks by identified neurons and transmitters, in *Dynamic Biological Networks: The Stomatogastric Nervous System* (R. M. Harris-Warrick, E. Marder, A. I. Selverston, and M. Moulins, Eds.), Cambridge, MA: MIT Press, pp. 87–138. ♦
- Hooper, S. L., and Moulins, M., 1989, A neuron switches from one network to another by sensory induced changes in its membrane properties, *Science*, 244:1587–1589.

- Manor, Y., Farzan, N., Abbott, L. F., and Marder, E., 1997, Temporal dynamics of graded synaptic transmission in the lobster stomatogastric ganglion, *J. Neurosci.*, 17:5610–5621.
- Marder, E., and Selverston, A. I., 1992, Modeling the stomatogastric nervous system, in *Dynamic Biological Networks: The Stomatogastric Nervous System* (R. M. Harris-Warrick, E. Marder, A. I. Selverston, and M. Moulins, Eds.), Cambridge, MA: MIT Press, pp. 161–196.
- Meyrand, P., Simmers, J., and Moulins, M., 1991, Construction of a pattern-generating circuit with neurons of different networks, *Nature*, 351:60–63.
- Miller, J. P., 1987, Pyloric mechanisms, in *The Crustacean Stomatogastric System* (A. I. Selverston and M. Moulins, Eds.), Berlin: Springer-Verlag, pp. 109–136. ♦
- Nadim, F., Manor, Y., Nusbaum, M. P., and Marder, E., 1998, Frequency regulation of a slow rhythm by a fast periodic input, *J. Neurosci.*, 18:5053–5067.
- Russell, D. F., and Hartline, D. K., 1978, Bursting neural networks: A reexamination, *Science*, 200:453–456.
- Selverston, A. I., Russell, D. F., Miller, J. P., and King, D. G., 1976, The stomatogastric nervous system: Structure and function of a small neural network, *Prog. Neurobiol.*, 7:215–290. ♦
- Turrigiano, G. G., and Heinzel, H.-G., 1992, Behavioral correlates of stomatogastric network function, in *Dynamic Biological Networks: The Stomatogastric Nervous System* (R. M. Harris-Warrick, E. Marder, A. I. Selverston, and M. Moulins, Eds.), Cambridge, MA: MIT Press, pp. 197–220. ♦

Data Clustering and Learning

Joachim M. Buhmann

Introduction

Intelligent data analysis extracts symbolic information and relations between objects from quantitative or qualitative data. A prominent class of methods are clustering or grouping principles which are designed to discover and extract structures hidden in data sets (Jain and Dubes, 1988). The parameters that represent the clusters are either estimated on the basis of quality criteria or cost functions or, alternatively, they are derived by local search algorithms that are not necessarily following the gradient of a global quality criterion. This approach to inference of structure in data is known as unsupervised learning in the neural computation literature. Clustering as a fundamental pattern recognition problem can be characterized by the following design steps:

1. *Data representation*: What data types represent the objects in the best way to stress relations between the objects, e.g., similarity?
2. *Modeling*: How can we formally characterize interesting and relevant cluster structures in data sets?
3. *Optimization*: How can we efficiently search for cluster structures?
4. *Validation*: How can we validate selected or learned structures?

It is important to note that the data representation issue predetermines what kind of cluster structures can be discovered in the data. Vectorial data, proximity or similarity data, and histogram data are three examples of a wide variety of data types that are analyzed in the clustering literature. On the basis of the data representation, the modeling of clusters defines the notion of groups of clusters in the data and separates desired group structures from unfavorable ones. We consider it mandatory that the modeling step yields a quality measure that is either optimized or approximated during the search for hidden structure in data sets. Formulating the search for clusters as an optimization problem allows us to validate clustering results by large deviation estimates; i.e., robust cluster structures in data should vary little from one data set to a second data set generated by the same data source.

The reader should note that the clustering literature in pattern recognition and applied statistics as well as in communications and information theory pursues two apparently different goals, namely, either (1) *density estimation* or (2) *data compression*. Both goals, however, are tightly related by the fact that the correct identification of the probability model of the source yields the best code for data compression. Mathematically, data compression aims at optimal

partitionings of the data, and stochastic sampling of partitions yields density estimates that are optimized in the maximum entropy sense. This issue is addressed in more detail later in this article.

Data Representations for Clustering

Various data types have been introduced in the pattern recognition literature. Mathematically, we distinguish between the object or design space \mathcal{O} that contains different object configurations $\mathbf{o} \in \mathcal{O}$ and the measurement space \mathcal{F} with measurements $\mathbf{x} \in \mathcal{F}$. Objects might be faces of people, and the corresponding measurements might be intensity or range images. A datum is defined as a relation between a design space \mathcal{O} and a measurement space \mathcal{F} . This relation $(\mathbf{o}, \mathbf{x}) \in \mathcal{O} \times \mathcal{F}$ can represent a functional dependence $\mathbf{x} : \mathbf{o} \mapsto \mathbf{x}(\mathbf{o})$ between objects and measurements or a stochastic dependence $\mathbf{P}\{\mathbf{x}|\mathbf{o}\}$. The following categories of data types are most commonly used in data analysis problems:

- *Vectorial data* characterize an object \mathbf{o} by a number of attributes which are combined to a d -dimensional feature vector $\mathbf{x}(\mathbf{o}) \in \mathbb{R}^d$.
- *Distributional data* of an object \mathbf{o} are described by an empirical probability distribution or histogram $\mathbf{P}\{\mathbf{x}|\mathbf{o}\}$ over features $\mathbf{x} \in \mathcal{F}$.
- *Proximity data* are characterized by pairwise comparisons between objects according to a proximity measure, e.g., $\mathbf{x}(\mathbf{o}_i) := \{D(\mathbf{o}_i, \mathbf{o}_j) \in \mathbb{R}; 1 \leq j \leq n\}$. Dissimilarity measures $D(\cdot, \cdot)$ often fulfill additional properties, e.g., nonnegativity, vanishing self-dissimilarity, symmetry, and the triangular inequality.

Various polyadic data types such as co-occurrence data (e.g., word bigrams in linguistics, consumer behavior data in economics) or even more complex data types (trigrams) are occasionally considered in the empirical sciences but are not further discussed here.

Modeling Cluster Structure

The goal of clustering is to assign objects with similar properties to the same clusters and dissimilar objects to different clusters. Mathematically, assignments of objects to clusters is represented by an assignment function $m : \mathcal{O} \rightarrow \{1, 2, \dots, k\}$, $\mathbf{o} \mapsto m(\mathbf{o})$. Data clustering pursues the goal to determine a partitioning of object space into subsets $\mathcal{G}_\alpha \equiv \{\mathbf{o} \in \mathcal{O} : m(\mathbf{o}) = \alpha\}$, $1 \leq \alpha \leq k$. The space of all clustering solutions is the set of all assignment functions $\mathcal{M} = \{m : \mathcal{O} \rightarrow \{1, 2, \dots, k\}\}$. The quality of these partitions

is evaluated by an appropriate homogeneity measure for the respective data type. Depending on the clustering goal, either the quality is optimized or the optimum is approximated that might yield a unique solution or a set of approximating solutions. The most commonly used clustering costs are invariant under permutations of the cluster indices. Hierarchical and topological clustering methods impose additional structure on the partitions. These principles are briefly sketched for all three data types.

Central Clustering or Vector Quantization

Clustering n objects that are represented as vectorial data $\mathcal{X} := \{\mathbf{x}_i \in \mathcal{F} : 1 \leq i \leq n\}$ induces a partitioning of the feature space $\mathcal{F} \subset \mathbb{R}^d$. The set of objects is partitioned into clusters in such a way that the average distance from data points to their cluster centers $\mathcal{Y} = \{\mathbf{y}_v \in \mathcal{F} : 1 \leq v \leq k\}$ is minimized. The representation of data \mathbf{x}_i by the centroid $\mathbf{y}_{m(i)}$ induces distortion/quantization costs $D_{i,m(i)}$ due to information loss. The functional form of $D_{i,m(i)}$ depends on the weighting of data distortions, in which quadratic costs $D_{i,m(i)} = \|\mathbf{x}_i - \mathbf{y}_{m(i)}\|^2$ and k -means $\mathbf{y}_\alpha = \sum_{\mathbf{o}_i \in \mathcal{G}_\alpha} \mathbf{x}_i / |\mathcal{G}_\alpha|$ is the most common choice. More general distortion measures like l_p -norms are occasionally considered. The cost function for k -means clustering is defined as

$$H^{cc}(m; \mathcal{X}) = \sum_{i \in n} \|\mathbf{x}_i - \mathbf{y}_{m(i)}\|^2 \quad (1)$$

In neural networks, the centroids $\mathbf{y}_{m(i)}$ can be implemented by neural feature detectors that are equipped with radially symmetric receptive fields and a global activity normalization. The size k of the cluster set, i.e., the complexity of the clustering solution, has to be determined a priori or by a problem-dependent complexity measure (Buhmann and Kühnel, 1993). A minimum of the cost function formulated in Equation 1 can be found by varying the assignments $m(i)$, which effectively is a search in a discrete space with exponentially many states. The optimization procedure implicitly yields the cluster means $\{\mathbf{y}_v\}$ by estimating optimized assignments $\{m(i)\}$. A supervised version of central clustering is discussed as LEARNING VECTOR QUANTIZATION (q.v.) in the literature.

Distributional Clustering

Distributional data represent the co-occurrence of objects and features by histograms (Pereira, Tishby, and Lee, 1993). Denote by $\mathcal{D} = \mathcal{O} \times \mathcal{F}$ the data space, i.e., the product space of objects $\mathbf{o} \in \mathcal{O}$ and features $\mathbf{x} \in \mathcal{F}$. In information retrieval, objects might be documents and features might be keywords. The $\mathbf{o} \in \mathcal{O}$ is characterized by an empirical conditional distribution $\hat{\mathbf{P}}\{\mathbf{x}|\mathbf{o}\}$ (histograms); i.e., we count the occurrence of a feature value \mathbf{x} given the object \mathbf{o} .

In distribution clustering, objects are grouped according to the similarity of their histograms $\hat{\mathbf{P}}\{\mathbf{x}|\mathbf{o}\}$, with a cluster-specific prototypical distribution of features $\mathbf{P}\{\mathbf{x}|\theta_{m(\mathbf{o})}\}$ that is parametrized by θ_α . The natural distortion measure between two histograms is defined by the Kullback-Leibler divergence (see LEARNING AND STATISTICAL INFERENCE), i.e., $D_{\mathbf{o},m(\mathbf{o})} = \hat{\mathbf{P}}\{\mathbf{o}\} D^{\text{KL}}(\hat{\mathbf{P}}\{\mathbf{x}|\mathbf{o}\} \|\mathbf{P}\{\mathbf{x}|\theta_{m(\mathbf{o})}\})$. This idea behind distributional clustering closely resembles k -means clustering when the Euclidian distance is replaced by the KL-divergence. The costs of distributional clustering sum up all distortions of objects, i.e.,

$$H^{\text{hc}}(m, \theta; \mathcal{X}) = |\mathcal{O}| \sum_{\mathbf{o} \in \mathcal{O}} \hat{\mathbf{P}}\{\mathbf{o}\} D^{\text{KL}}(\hat{\mathbf{P}}\{\mathbf{x}|\mathbf{o}\} \|\mathbf{P}\{\mathbf{x}|\theta_{m(\mathbf{o})}\}) \quad (2)$$

The costs formulated in Equation 2 define the log-likelihood of a statistical model that explains the data \mathcal{X} by a mixture of data sources: (1) select an object $\mathbf{o} \in \mathcal{O}$ with probability $\mathbf{P}\{\mathbf{o}\}$; (2) choose the cluster α according to the cluster membership of $\alpha =$

$m(\mathbf{o})$; (3) select $\mathbf{x} \in \mathcal{F}$ according to the class-conditional distribution $\mathbf{P}\{\mathbf{x}|\theta_{m(\mathbf{o})}\}$. A very insightful relation to rate distortion theory with side information (Cover and Thomas, 1991) has been described in Tishby, Pereira, and Bialek (1999) and it is called the information bottleneck method.

Pairwise Clustering

Clustering nonmetric data that are characterized by proximity information and not by explicit Euclidean coordinates can be formulated as a graph optimization problem. Given is a graph (v, ε) with weights $\mathcal{D} := \{D_{ij}\}$ on the edges (i, j) . The vertices denote the objects to be grouped and the edge weights encode dissimilarity information. Compact clusters are represented by a partition of the vertex set with small dissimilarities between all objects that belong to the same cluster. To simplify the notation, the subset of edges with one vertex in cluster α and one vertex in cluster β is denoted by $\varepsilon_{\alpha\beta} = \{(i, j) \in \varepsilon : \mathbf{o}_i \in \mathcal{G}_\alpha \wedge \mathbf{o}_j \in \mathcal{G}_\beta\}$. A meaningful cost function for pairwise clustering that primarily avoids grouping dissimilar objects into one cluster is defined by

$$H^{\text{pc}}(m; \mathcal{D}) = \sum_{v \leq k} \left(|\mathcal{G}_v| \sum_{(i,j) \in \varepsilon_{vv}} \frac{D_{ij}}{|\mathcal{G}_{vv}|} \right) \quad (3)$$

Preferred clusters are subsets of objects with minimal average intracluster dissimilarities, weighted by the cluster size $|\mathcal{G}_v|$. This cost function has the remarkable and, for applications, extremely valuable invariance that the assignments do not change if all dissimilarities are changed by the same off-diagonal offset D_0 ; i.e., $\arg \min_m H^{\text{pc}}(m; D) = \arg \min_m H^{\text{pc}}(m; \tilde{D})$, with $\tilde{D} = D_{ij} + D_0(1 - \delta_{ij})$.

An alternative to the pairwise clustering cost function in Equation 3 has been proposed by Shi and Malik (2000) in the context of image segmentation. The clustering criterion

$$H^{\text{nc}}(m; \mathcal{D}) = \sum_{v \leq k} \left(\frac{\sum_{(i,j) \in \varepsilon_{vv}} D_{ij}}{\sum_{(i,j) \in \varepsilon: i \in \mathcal{G}_v, j \in \mathcal{G}_v} D_{ij}} \right) \quad (4)$$

weighs the intracluster compactness with the integrated dissimilarities between objects $i \in \mathcal{G}_v \vee j \in \mathcal{G}_v$ in cluster v and all other objects. Good approximate solutions to Equation 4 can be found by spectral graph theory. Originally, this method was developed for similarity data $S_{ij} = \exp(-D_{ij}/\Delta)$ and it was formulated as a minimization of the normalized cut of similarities between two clusters.

A pairwise clustering principle based on local object similarity rather than global cluster compactness has been suggested in Blatt, Wiseman, and Domany (1997) exploiting an analogy to granular magnets. Locality is achieved by the exponential transformation $S_{ij} = \exp(-D_{ij}/\Delta)$, which effectively decouples objects with $D_{ij} \gg \Delta$. The granular magnet concept has been abstracted to the method of typical (multiway) cuts (Gdalyahu, Weinshall, and Werman, 2001), which produces randomized approximations to clustering solutions with small intercluster similarity. The costs for this percolation-type model is given by

$$H^{\text{mc}}(m; S) = \sum_{(i,j) \in \varepsilon} S_{ij} - \sum_{v \leq k} \sum_{(i,j) \in \varepsilon_{vv}} S_{ij} = \sum_{v \leq k} \sum_{\mu \neq v} \sum_{(i,j) \in \varepsilon_{v\mu}} S_{ij} \quad (5)$$

The identity in Equation 5 relates similarity to neighbors of the same cluster to the combinatorial optimization problem to find a multicut in the graph (v, ε) . $H^{\text{mc}}(m; S)$ stresses local consistency in clustering, whereas $H^{\text{pc}}(m; D)$ emphasizes global compactness of clusters. The granular magnet clustering model generalizes the nearest neighbor linkage method, a popular graph theoretic clustering technique (Duda, Hart, and Stork, 2001).

Topological and Hierarchical Clustering

The four proposed clustering criteria for vectorial, histogram, and proximity data evaluate cluster configurations in a permutation-invariant way. In neurobiology, a topological organization of neurons is often imposed to ensure that nearby neurons process related information in feature space. Respective network structures are known as SELF-ORGANIZING FEATURE MAPS (q.v.). All the introduced clustering principles (Equations 1–5) can be generalized to topology-preserving clustering methods by replacing the unique assignments $\mathbf{o} \mapsto m(\mathbf{o})$ with probabilistic assignments $\mathbf{o} \mapsto \alpha$ with probability $\mathbf{T}_{m(\mathbf{o}),\alpha}$. The cost function for clustering vectorial data translates to

$$H^{\text{som}}(m; \mathcal{X}) = \sum_{i \leq n} \sum_{v \leq k} \mathbf{T}_{m(i),v} \|\mathbf{x}_i - \mathbf{y}_v\|^2 \quad (6)$$

In case that \mathbf{T} defines a \tilde{d} -dimensional neighborhood system, the costs in Equation 6 prefer arrangements of centroids on a \tilde{d} -dimensional smooth manifold in the d -dimensional feature space. The additional quantization errors inflicted by the probabilistic assignments are also known in the information theory literature (Cover and Thomas, 1991) as index confusion $v \rightarrow \alpha$ due to channel noise. The same principle can be formulated for histogram clustering and for pairwise clustering.

In many application areas a hierarchical partitioning of the object space is favored over unconstrained or topological partitions. The reasons for this preference either are computational, since tree-like structures support rapid data access and efficient algorithmics, or else the data source is hypothesized to be of a tree-like nature. The assignments of objects to clusters have to observe an inclusion principle (Jain and Dubes, 1988) that ensures that subpartitions at a fine level observe the partitions higher up in the tree. It is important to include information of the tree topology in the clustering criterion. Cluster trees resemble decision trees in classification, and they can be naturally implemented in biological brains by layered neural networks.

Optimization

The clustering cost functions can be minimized by various deterministic or stochastic methods from combinatorial and continuous optimization. A widely used class of methods estimates the clustering parameters in an iterative way by first keeping the continuous cluster parameters (centroids, prototype histograms) fixed and optimizing the assignments. In a second step, the continuous parameters are calculated from the new assignments.

The k-Means Algorithm

A well-known algorithm for on-line estimation of prototypes in central clustering is the k -means algorithm (MacQueen, 1967). The k means $\{\mathbf{y}_v : 1 \leq v \leq k\}$ are initialized by the first k data points $\{\mathbf{x}_i : 1 \leq i \leq k\}$. A new data vector \mathbf{x}_{t+1} , $t \geq k$ is assigned to the closest mean \mathbf{y}_α according to the nearest neighbor rule

$$m(t+1) = \arg \min_v \|\mathbf{x}_{t+1} - \mathbf{y}_v\| \quad (7)$$

with ties being handled appropriately. The new mean vector is adjusted in response to data vector \mathbf{x}_{t+1} according to the learning rule

$$\mathbf{y}_{m(t+1)}^{(t+1)} = \mathbf{y}_{m(t+1)}^{(t)} + \frac{1}{|\mathcal{G}_{m(t+1)}|} (\mathbf{x}_{t+1} - \mathbf{y}_{m(t+1)}^{(t)}) \quad (8)$$

All other means \mathbf{y}_v , $v \neq m(t+1)$ remain unchanged. The learning rule in Equation 8 adjusts the closest mean \mathbf{y}_α proportional to the deviation $(\mathbf{x}_{t+1} - \mathbf{y}_\alpha^{(t)})$ normalized by the number of data vectors that have already been assigned to this cluster center.

Probabilistic Partitioning Algorithms

The class of stochastic optimization algorithms with SIMULATED ANNEALING AND BOLTZMANN MACHINES (q.v.) as the most prominent techniques plays an eminent role in pattern recognition. The variables of the optimization problem, e.g., the assignments in clustering, are treated as random variables of a stochastic (Markovian) process. The Markov chain Monte Carlo algorithm samples from a set of data partitionings, all of which are considered to be compatible with the data. The size of this set has to be controlled by a cluster validation scheme. Robust clustering methods are derived from the maximum entropy principle, which states that assignments are distributed according to the Gibbs distribution

$$\mathbf{P}(m; \mathcal{X}) = \exp(-(H^{\text{cc}}(m; \mathcal{X}) - F)/T), \quad (9)$$

$$F = -T \log \sum_{m \in \mathcal{M}} \exp(-H^{\text{cc}}(m; \mathcal{X})/T) \quad (10)$$

The “computational temperature” T serves as a Lagrange parameter for the expected costs. The free energy F in Equation 10 can be interpreted as a smoothed version of the original cost function H^{cc} (Rose, Gurewitz, and Fox 1990). The cost function H^{cc} , which is linear in the assignments, yields a factorized Gibbs distribution

$$\mathbf{P}(m; \mathcal{X}) = \prod_{i \leq n} \frac{\exp(-D_{i,m(i)}/T)}{\sum_{\mu \leq k} \exp(-D_{i,\mu}/T)} = \prod_{i \leq n} \mathbf{P}_{i,m(i)} \quad (11)$$

$$\mathbf{P}_{i,v} := \frac{\exp(-D_{i,v}/T)}{\sum_{\mu \leq k} \exp(-D_{i,\mu}/T)}, \quad \forall v \in \{1, \dots, k\} \quad (12)$$

$\mathbf{P}_{i,v}$ denote expectation values of assignments. The centroids have to maximize the entropy of the Gibbs distribution, which yields the centroid constraint

$$0 = \sum_{i \leq n} \mathbf{P}_{i,v} \frac{\partial}{\partial \mathbf{y}_v} D_{i,v}, \quad \forall v \in \{1, \dots, k\} \quad (13)$$

The Gibbs distribution in Equation 11 can also be interpreted as the complete data likelihood for mixture models with parameters \mathcal{Y} . Basically, the Gibbs distribution of the k clusters describes a mixture model with equal priors for each component and equal, isotropic covariances. The assignments $m(i)$ and their expectations $\mathbf{P}_{i,v}$ correspond to the unobservable variables in mixture models and the component densities, respectively. Algorithmically, the centroids and the expected assignments are estimated in an iterative fashion by solving the centroid Equation 13 for fixed expected assignments and, subsequently, inserting the centroids in Equation 12 (Rose et al., 1990; Buhmann and Kühnel, 1993).

The temperature parameter T controls the uncertainty of the clustering problem; i.e., in the limit $T \rightarrow 0$, the solution of Equation 12 corresponds to hard clustering with Boolean assignments $\mathbf{P}_{i,v} \in \{0, 1\}$ of a data vector \mathbf{x}_i to the closest cluster center \mathbf{y}_v . Large temperature represents the uncertain limit with partial assignments of data vectors to several clusters ($0 \leq \mathbf{P}_{i,v} \leq 1$). The reader should note that the iterative search for solutions of Equations 12 and 13 is guaranteed to yield a local minimum of the costs that, however, could be far from the global minimum.

Gaussian Mixture Models

Natural clusters in data sets are modeled by a mixture of stochastic data sources (McLachlan and Basford, 1988). Each component of this mixture, a data cluster, is described by a univariate probability density that is the stochastic model for an individual cluster. The sum of all component densities forms the probability density of the mixture model

$$P(\mathbf{x}; \Theta) = \sum_{v \leq k} \pi_v p(\mathbf{x}; \theta_v) \quad (14)$$

Let us assume that the functional form of the probability density $P(\mathbf{x}; \Theta)$ is completely known up to a finite and presumably small number of parameters $\Theta = (\theta_1, \dots, \theta_k)$. For the most common case of Gaussian mixtures, the parameters $\theta_v = (\mathbf{y}_v, \Sigma_v)$ are the coordinates of the mean vector and the covariance matrix. The a priori probability π_v of the component v , which is assumed to be known, is called the mixing parameter.

Adopting this framework of parametric statistics, the detection of data clusters reduces mathematically to the problem of how to estimate the parameters Θ of the probability density for a given mixture model. A powerful statistical technique for finding mixture parameters is the maximum likelihood method (Duda et al., 2001); i.e., one maximizes the probability of the independently, identically distributed data set $\{\mathbf{x}_i : 1 \leq i \leq n\}$ given a particular mixture model. For analytical purposes it is more convenient to maximize the log-likelihood

$$L(\Theta) = \sum_{i \leq n} \log P(\mathbf{x}_i; \Theta) = \sum_{i \leq n} \log \left(\sum_{v \leq k} p(\mathbf{x}_i; \theta_v) \pi_v \right) \quad (15)$$

which yields the same maximum likelihood parameters Θ . The straightforward maximization of Equation 15 results in a system of transcendental equations with multiple roots. The ambiguity in the solutions originates from the lack of knowledge of which mixture component v has generated a specific data vector \mathbf{x}_i , and therefore which parameter θ_v should be influenced by \mathbf{x}_i in the estimation procedure.

An efficient solution to overcome the computational problem of how to estimate parameters of mixture models with the maximum likelihood method is provided by the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977). The EM algorithm estimates the unobservable assignments in a first step. The estimates are denoted by $\mathbf{P}_{i,v}$. On the basis of these maximum likelihood estimates $\{\mathbf{P}_{i,v}\}$, the parameters Θ are calculated in a second step. An iteration of these two steps renders the following algorithm for Gaussian mixtures:

- *E-step*: The expectation value of the complete data log-likelihood is calculated conditioned on the observed data $\{\mathbf{x}_i\}$ and the parameter estimates $\hat{\Theta}$. This yields the expected assignments of data to mixture components, i.e.,

$$\begin{aligned} \mathbf{P}_{i,\alpha}^{(t+1)} &= \frac{p(\mathbf{x}_i; \hat{\mathbf{y}}_\alpha^{(t)}, \hat{\Sigma}_\alpha^{(t)}) \pi_\alpha}{\sum_{v \leq k} p(\mathbf{x}_i; \hat{\mathbf{y}}_v^{(t)}, \hat{\Sigma}_v^{(t)}) \pi_v} \\ &= \frac{\pi_\alpha |\hat{\Sigma}_\alpha^{(t)}|^{-1/2} \exp(-1/2 (\mathbf{x}_i - \hat{\mathbf{y}}_\alpha^{(t)})^T (\hat{\Sigma}_\alpha^{(t)})^{-1} (\mathbf{x}_i - \hat{\mathbf{y}}_\alpha^{(t)}))}{\sum_{v \leq k} \pi_v |\hat{\Sigma}_v^{(t)}|^{-1/2} \exp(-1/2 (\mathbf{x}_i - \hat{\mathbf{y}}_v^{(t)})^T (\hat{\Sigma}_v^{(t)})^{-1} (\mathbf{x}_i - \hat{\mathbf{y}}_v^{(t)}))} \end{aligned} \quad (16)$$

- *M-step*: The likelihood maximization step estimates the mixture parameters, e.g., centers and variances of the Gaussians (Duda et al., 2001, p. 89):

$$\hat{\mathbf{y}}_\alpha^{(t+1)} = \frac{\sum_{i \leq n} \mathbf{P}_{i,\alpha}^{(t+1)} \mathbf{x}_i}{\sum_{i \leq n} \mathbf{P}_{i,\alpha}^{(t+1)}}, \quad (17)$$

$$\hat{\Sigma}_\alpha^{(t+1)} = \frac{1}{\sum_{i \leq n} \mathbf{P}_{i,\alpha}^{(t+1)}} \sum_{i \leq n} \mathbf{P}_{i,\alpha}^{(t+1)} (\mathbf{x}_i - \hat{\mathbf{y}}_\alpha^{(t+1)}) (\mathbf{x}_i - \hat{\mathbf{y}}_\alpha^{(t+1)})^T \quad (18)$$

Note that Equations 17 and 18 have a unique solution after the expected assignments $\{\mathbf{P}_{i,\alpha}^{(t+1)}\}$ have been estimated. The monotonic increase in the likelihood up to a local maximum guarantees the convergence of the EM algorithm.

Mean Fields for Pairwise Clustering

Minimization of the quadratic cost function formulated in Equation 3 turns out to be algorithmically complicated because of pairwise, potentially conflicting correlations between assignments. The deterministic annealing technique, which produces robust reestimation equations for central clustering in the maximum entropy framework, is not directly applicable to pairwise clustering since there is no analytical technique known to capture correlations between assignments $m(i)$ and $m(j)$ in an exact way. Mean field annealing, however, approximates the intractable Gibbs distribution by the best factorial distribution. The influence of the random variables $m(j)$, $j \neq i$, on $m(i)$ is treated by a mean field that measures the average feedback on $m(i)$. Mathematically, the approximation problem to calculate the Gibbs distribution is replaced by a minimization of the Kullback-Leibler divergence between the approximating factorial distribution and the Gibbs distribution (Hofmann and Buhmann, 1997). A maximum entropy estimate of the mean fields $h_{i,v}$ yields the transcendental equations

$$\mathbf{P}_{i,v} = \frac{\exp(-h_{i,v}/T)}{\sum_{\mu \leq k} \exp(-h_{i,\mu}/T)} \quad (19)$$

$$h_{i,v} = \frac{1}{n\pi_v} \sum_{(i,j) \in \varepsilon} \mathbf{P}_{j,v} \left(D_{ij} - \frac{1}{2n\pi_v} \sum_{(j,r) \in \varepsilon} \mathbf{P}_{r,v} D_{jr} \right) \quad (20)$$

The variables $h_{i,v}$ depend on the given distance matrix D_{ik} , the averaged assignment variables $\{\mathbf{P}_{i,v}\}$, and the cluster weights $\pi_v := \sum_{i \leq n} \mathbf{P}_{i,v}$. Equation 20 suggests an algorithm for learning the optimized cluster assignments that resembles the EM algorithm. In the E-step, the assignments $\{\mathbf{P}_{i,v}\}$ are estimates for given $\{h_{i,v}\}$. In the M-step, the $\{h_{i,v}\}$ are reestimated on the basis of new assignment estimates. This iterative algorithm converges to a consistent solution of assignments for the pairwise data clustering problem that locally maximizes the entropy.

Cluster Validation

One of the most important problems in data analysis, besides proper modeling, is to test solutions of pattern recognition problems for robustness. The influence of noise in the data on the estimates of cluster centers in central clustering should be minimal in the large data limit. The data analyst, therefore, requests that the clustering solution should be similar if the clustering algorithm partitions a second data set from the same data source. This robustness requirement limits the number of clusters that can be reliably inferred from the data. If too many clusters are supposed to be estimated, then the noise will strongly influence the values of the clustering parameters.

The field of statistical learning theory addresses robustness questions and model complexity issues in the context of supervised learning, in particular for classification and regression. The same trade-off between the complexity of the hypothesis class and the amount of data limits the inference precision in unsupervised learning tasks as data clustering. Theoretical results have to estimate the probability of large deviations between solutions found on two different sample sets. Assume that the cluster solution is quantified by the costs $H(m; \mathcal{X})$ and that we have two data sets, $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$. The costs $H(m; \mathcal{X})$ might have the combinatorial forms of Equations 1–5, or they could be the log-likelihood (Equation 15) for density estimation. The optimal cluster assignments with re-

spect to the two data sets are $m^{(1)}$ and $m^{(2)}$. A robustness criterion based on large deviation theory should bound from above the probability

$$P\{H(m^{(1)}; \mathcal{X}^{(2)}) - H(m^{(2)}; \mathcal{X}^{(2)}) > \varepsilon\} \leq \delta \quad (21)$$

i.e., we require that the optimal solution on the first data set also perform well on a second data set $\mathcal{X}^{(2)}$ with high probability $1 - \delta$. Statistical learning theory relates this deviation to the complexity of the solution space \mathcal{M} . Algorithmically, such a complexity control can be implemented by cross-validation, where the algorithm stops to further improve the solution in the optimization process when the costs on the second sample set increase again. This procedure tries to avoid overfitting of model parameters to the data.

Classical strategies to validate clustering solutions are based on Bayesian model selection. The number of modes in mixture model inference can be limited by regularization, e.g., by the minimum description length principle (Duda et al., 2001), which implements the Occam's razor principle for inference (see MINIMUM DESCRIPTION LENGTH ANALYSIS). Alternatives to this criterion are the Bayesian information criterion (BIC), Akaike's information criterion (AIC), or the network information criterion (NIC) (Ripley, 1996), all of which penalize the complexity of the model, e.g., the number of clusters and their cluster parameters. These criteria are asymptotic ($n \rightarrow \infty$) in nature.

Discussion

Data clustering as one of the most fundamental information processing procedure to extract symbolic information from subsymbolic data follows the four design steps of pattern recognition: (1) data representation, (2) structure definition, (3) structure optimization, and (4) structure validation. It is still very speculative how these mathematical structures are implemented in biological brains. Neurons with localized receptive fields might provide the neural correlate of centroids in Euclidian feature spaces. Relational data might be represented by neurons that function as comparators. The stochastic dynamics of neurons naturally introduces randomness in the search process for compact cluster structures, and thereby introduces robustness against fluctuations in the data analysis process.

Databases for Neuroscience

Jeffrey S. Grethe

Introduction

The advancement of science depends heavily on researchers' ability to share, exchange, and organize large quantities of heterogeneous data in an efficient manner. This is especially true for researchers who are involved in the construction of simulations of the nervous system. Data spanning a wide range of disciplines (e.g., anatomy, physiology, behavior) are needed to conceptualize, design, and validate such simulations. The need for informatics research related to databases is highlighted by the increasingly important role it has played in various scientific communities to help achieve their goals. This is most evident in the molecular biology community (Persidis, 1999; Roos, 2001), where since the beginning of the 1980s, a large collection of specialized data repositories has been constructed. As part of these developments, the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>)

Road Map: Learning in Artificial Networks

Background: Learning Vector Quantization

Related Reading: Principal Component Analysis; Stochastic Approximation and Efficient Learning

References

- Blatt, M., Wiseman, S., and Domany, E., 1997, Data clustering using a model granular magnet, *Neural Comput.*, 9:1805–1842.
- Buhmann, J., and Kühnel, H., 1993, Vector quantization with complexity costs. *IEEE Trans. Inform. Theory*, 39:1133–1145.
- Cover, T. M., and Thomas, J. A., 1991, *Elements of Information Theory*, New York: Wiley. ♦
- Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. Ser. B*, 39:1–38.
- Duda, R. O., Hart, P. E., and Stork, D. G., 2001, *Pattern Classification*, New York: Wiley. ♦
- Gdalyahu, Y., Weinshall, D., and Werman, M., 2001, Self-organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization, *IEEE Trans. Pattern Anal. Machine Intell.*, 23:1053–1074.
- Hofmann, T., and Buhmann, J. M., 1997, Pairwise data clustering by deterministic annealing, *IEEE Trans. Pattern Anal. Machine Intell.*, 19:1–14.
- Jain, A. K., and Dubes, R. C., 1988, *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall. ♦
- MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations, in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.
- McLachlan, G. J., and Basford, K. E., 1988, *Mixture Models*, New York: Marcel Dekker.
- Pereira, F., Tishby, N., and Lee, L., 1993, Distributional clustering of English words, in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, pp. 183–190.
- Ripley, B. D., 1996, *Pattern Recognition and Neural Networks*, Cambridge, Engl.: Cambridge University Press.
- Rose, K., Gurewitz, E., and Fox, G., 1990, A deterministic annealing approach to clustering, *Pattern Recognit. Lett.*, 11:589–594.
- Shi, J., and Malik, J., 2000, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Machine Intell.*, 22:888–905.
- Tishby, N., Pereira, F., and Bialek, W., 1999, The information bottleneck method, in *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*, Champaign: University of Illinois Press, pp. 368–377.

was established to develop and maintain public databases and tools for searching and analyzing genomic information. These tools have resulted in incredible scientific advancement within the genomic community, most notably the dissemination of information concerning the human genome (<http://www.ncbi.nlm.nih.gov/genome/guide/human/>). The success of the Human Genome Project has influenced the application of informatics research to other disciplines, namely, neuroscience.

The amount of neuroscience information being collected is increasing dramatically. This information explosion makes it difficult to search and organize basic research data (Huerta, Koslow, and Leshner, 1993). To that end, the Human Brain Project (<http://neuroinformatics.nih.gov>) was initiated in September 1993 to provide informatics tools to help manage this data explosion. A main component of the initiative is to provide neuroscientists access to information at all levels of integration through a collection of net-

work-accessible databases. Just as bioinformatics proved to be indispensable for the growth of the molecular biology community, neuroinformatics will become a key component in future neuroscience research. However, the neuroscience community faces some unique challenges.

Neuroscience data are extremely varied and complex. A myriad of structured, textual, graphical, and other information captures experimental results, hypotheses, scientific assumptions, conclusions, formal presentations of theories, and observations. This poses a particularly interesting and important problem for informatics research: to build on state-of-the-art techniques and mechanisms to accommodate the characteristically distinct needs of neuroscience, develop solutions to those problems, and transfer these results to the neuroscience community. Over the last decade, neuroinformatics research has begun to provide mechanisms for the storage of data from varied neuroscientific disciplines (see Table 1 for an overview of database projects). Owing to the great diversity of the databases being developed, we cannot discuss each of these database projects in detail here. Therefore, this article presents and discusses several issues regarding the development of databases for the neuroscience community (for other reviews, see Chicurel, 2000; Kötter, 2001).

Issues in the Development of Neuroscience Databases

Diversity of Neuroscience Data

The interplay of anatomical structure, biochemical processes, and electrical signals gives rise to natural and disease processes. To investigate the underlying mechanisms responsible for these processes, information from varied sources must be integrated. For example, the search for clinical treatments for Parkinson's disease requires information from a wide variety of sources. Information concerning neurotransmitters and receptors, anatomical pathways, neuronal properties, and the information processing that occurs across areas of the basal ganglia must be synthesized and related to clinical, pharmacological, and genetic information. As this example illustrates, neuroscience is an extremely diverse field involving many disciplines. In addition to its diversity, the information of interest to neuroscience researchers is archived throughout the world and stored in a myriad of formats that are accessible through a variety of interfaces and retrieval languages. The data sources can include conventional databases that are accessed through a database query language to web archives accessed by a web browser. Compounding the problem for the neurosciences is that unlike the genomic community, in which data tend to be stored in a few large databases, the neurosciences community will have to be able to access a large, heterogeneous collection of databases.

To be able to integrate such diverse sources, the various communities within the neurosciences must begin to develop standards for their community's data. Currently, in these communities, there exist many different and incompatible data formats that do not allow for the free exchange of data. In addition to needing standards for the raw data themselves, these communities must also develop standards for the description of the actual data (i.e., a formalized description of the meta-data). A possible solution to this problem can be found in the adoption of extensible markup language (XML) technologies. XML (Bosak and Bray, 1999) is a markup language that allows one to describe semistructured representations of information. More specifically, it allows one to annotate data with semantic tags that describe the structure and content of the data. For example, in the modeling community, the Neural Open Markup Language (NeuroML) (Goddard et al., 2001) is being developed to allow researchers to describe models at varying levels of complexity, from cell membranes to large-scale neural networks. In addition

to the ability to exchange data within a community, integration of data from diverse biological disciplines will be needed. In contrast to integration problems found in typical database federations (i.e., different representations of the same information), the neurosciences must be able to integrate data from sources with heterogeneous data and representations. One possible solution for this type of integration has been to extend the conventional wrapper-mediator architecture with domain-specific knowledge (Gupta, Lüdäscher, and Martone, 2000).

Data Representation

Data representation is a critical issue for many of the database projects that are currently under way in the neurosciences. One of the first and most central issues in representing data from the brain is the manner in which one describes the fundamental objects of neuroscience. As an example, one need only to look at the classification of brain areas. Currently, there are a number of classification schemes that are not completely compatible. The problem becomes even worse when one considers the difficulty in classifying analogous brain regions across species. How, then, is one supposed to represent location of brain data within a database? Just specifying the textual location might not be sufficient. One partial solution to this problem would be to reference all locations as three-dimensional coordinates within a reference coordinate system (Mazziotta et al., 1995).

The complexity of the experimental paradigms is another critical issue for the representation of neuroscience data. The field of neuroscience is evolving rapidly, and the diversity of these descriptors does not allow one to know a priori what information will need to be stored for future experiments. To address this problem, many database projects stress the importance of an extensible foundation (Nadkarni et al., 1999; Arbib and Grethe, 2001; Gardner et al., 2001) that allows for varying information to be archived in a structured fashion.

User Access

For researchers to be able to navigate and integrate all the information to be provided, user interfaces need to be developed that will give the researcher a seamlessly integrated view of the information concerning a particular topic. These interfaces for neuroscience information would provide a number of novel opportunities that are not available when access is provided only to individual pieces of information. However, the vast amount of information available to researchers over the Internet poses several issues when one attempts to integrate this information. These issues include knowing where the relevant information is located, being able to access that information, integrating and transforming the data into a unified framework, and methods for visualizing the data in an appropriate way.

It is important to realize that the multidisciplinary nature of neuroscience dictates an important additional requirement for these interfaces. Researchers who wish to examine data related to a specific topic will not be experts in every aspect of that topic. For example, a neurochemist studying the kinetics and binding properties of various neurotransmitter receptors might not be aware of the latest clinical information regarding the neurotransmitter in question. Using structured review and summary information as a framework for the presentation of data from varied sources will be imperative when data are to be provided to such a diverse user community.

A Case Study

With support from the National Science Foundation, the National Institutes of Health, and the Keck Foundation, the fMRI Data Cen-

Table 1. Databases Being Developed for the Neurosciences. This table provides a list of databases covering a wide range of subdisciplines within the neurosciences. All databases listed include a brief description, a URL, and a notation as to what forms of data the database is primarily concerned with.

Database	Description	Data
3D Neuron Centered Database http://www.ncmir.ucsd.edu/NCDB/	System integrating three-dimensional cellular microscopic data characterizing the realistic locations, surface morphology, and cellular constituents	Cellular, morphology
Brain Image Database (BRAID) http://braid.rad.jhu.edu/	Archive of normalized spatial and functional neuroimaging data with an analytical query mechanism	Neuroimaging
BrainInfo http://braininfo.rprc.washington.edu/	Helps to identify structures in the brain and provides additional information about these structures	Anatomy
BrainMap http://ric.uthscsa.edu/projects/brainmap.html	Database for meta-analysis of author-supplied activations from select human functional brain-mapping literature	Neuroimaging
Brain Models on the Web http://www-hbp.usc.edu/Projects/bmw.htm	Database of neural models at the network level, synaptic and kinetic levels, and models that integrate across the levels	Simulation
BrainWeb Simulated Brain Database http://www.bic.mni.mcgill.ca/brainweb/	Database of realistic MRI data produced by a simulator to evaluate the performance of various image analysis methods	Neuroimaging
Cell Signaling Networks Database http://geo.nihs.go.jp/csndb/	A knowledge base for signaling pathways of human cells	Cellular
CoCoMac http://www.cocomac.org/	Database of anatomical connectivity in the macaque	Anatomy
Cortical Neuron Net Database http://cortex.med.cornell.edu/	Database of electrophysiological and other information describing cortical neurons	Physiology
Digital Anatomist Project http://www9.biostr.washington.edu/da.html	Anatomy information system that is available over the web	Anatomy
European Computerised Human Brain Database http://fornix.neuro.ki.se/ECHBD/Database/index.html	Database for relating function to microstructure of the cerebral cortex of humans	Neuroimaging
fMRI Data Center http://www.fmridc.org	Publicly accessible repository of peer-reviewed fMRI studies and their underlying data	Neuroimaging
GENESIS Modeler's Workspace http://www.genesis-sim.org/hbp/	System to help users create and organize models and interact with other databases and simulation software	Simulation
ICBM http://www.loni.ucla.edu/ICBM/	Developing a probabilistic reference system for the human brain	Neuroimaging
Ligand Gated Ion Channel Database http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html	Sequence database for neurotransmitter receptors	Sequence
NeuArt http://www-hbp.usc.edu/Projects/neuart.htm	Database to retrieve graphical data from a neuroanatomical database	Anatomy
NeuroCore http://www-hbp.usc.edu/Projects/neurocore.htm	Database framework for the storage of data from neuroscientific experiments that can be extended to meet a specific lab's requirements	Physiology, chemistry
NeuroGenerator http://www.neurogenerator.org/	Database generator for anatomical and functional images of the human brain	Neuroimaging
NeuroScholar http://chasseur.usc.edu/ns/	Knowledge management system for literature	Literature
NeuroSys http://cns.montana.edu/research/neurosys/	Lightweight peer-to-peer data-sharing system with a database front end	Cellular, physiology
NTSA Workbench http://soma.npa.uiuc.edu/isnpa/isnpa.html	Tools for the storage and retrieval of large neuronal time series data sets	Physiology
SenseLab http://ycmi-hbp.med.yale.edu/senselab/	Collection of six related databases that focus on information processing in nerve cells of the olfactory system	Anatomy, physiology, sequence, simulation
Surface Management System (SuMS) Database http://stp.wustl.edu/sums/	Surface-based database of neuroimaging data intended to aid in cortical surface reconstruction, visualization, and analysis	Neuroimaging
Virtual Neuromorphology Electronic Database http://www.krasnow.gmu.edu/L-Neuron/database/index.html	A database for three-dimensional neuronal structures	Morphology, cellular
XANAT http://stp.wustl.edu/resources/xanat.html	A graphical anatomical database of neural connectivity	Anatomy

ter (Van Horn et al., 2001) was established in the autumn of 1999 with the objective of creating a mechanism by which members of the neuroscientific community may more easily share functional neuroimaging data. By building a publicly accessible repository of raw data from peer-reviewed studies, the Data Center aims to create a successful environment for the sharing of neuroimaging data in the neurosciences. By increasing the number of scientists who can examine, consider, analyze, and assess the neuroimaging data that have already been collected and published, the center hopes to speed the progress of understanding cognitive processes, the neural substrates that underlie them, and the diseases that affect them. The development of such an archive for neuroimaging data posed certain challenges.

The most critical issue that needed to be addressed in the building of the Data Center was the representation of the data to be archived. In and of itself, raw functional imaging data may not be particularly useful. Proper analysis requires at least knowledge of how the data were acquired. Therefore, in addition to the data sets themselves, the Data Center must store all technical descriptions of the data necessary for someone other than the study's authors to accurately reproduce and interpret the original results. This includes both the information regarding the scanning parameters and the experimental protocols (information and timing of the stimuli) presented to the subject during the collection of the data. The Data Center cannot know a priori all the descriptors required for studies being submitted and therefore relies on an extensible framework for the storage of the data.

The descriptors that are used in describing a neuroimaging experiment constitute one type of meta-data (i.e., data that contain information about the data). These meta-data will allow the data to be organized in a suitable fashion for keyword-based queries. However, it is also important to be able to offer the researcher the ability to query the content of the imaging data directly. Querying the raw functional data directly is not feasible, owing to the amount of imaging data present in the archive (e.g., studies archived at the Data Center can be on the order of hundreds of megabytes to tens of gigabytes). To be able to perform content-based queries on the image data, concise meta-data need to be generated to describe the individual images and time series of images. The meta-data can then be used as a means to quickly find studies of interest based on features derived from the data themselves. Researchers may then further examine these data using more rigorous analysis methods.

Another critical challenge for the Data Center was the protection of human subjects data. This problem is not unique to the Data Center; any database archiving data from human subjects must comply with U.S. government regulations on the protection of such data. To that end, the Data Center makes every reasonable effort to ensure that all data included within the Data Center's archive are anonymized so that data being used by researchers cannot be linked back to the individual subjects who provided it. Neuroimaging data are also unique in that the data themselves can be used as an identifier. It is possible to reconstruct recognizable images of a subject's face from high-resolution structural magnetic resonance data. To that end, all structural data must be stripped (i.e., removal of facial features) before being included in the data archive.

Currently, the Data Center hosts a web site where authors from around the world have been submitting entire, raw data sets from peer-reviewed journals. Visitors to the Data Center's web site may search its holdings via a MEDLINE-inspired query interface. This allows researchers to easily access and search the Data Center's resources without the need to learn a new interface and query language. A researcher who finds data sets of interest can request that they be shipped to him or her on either CD (at no charge) or digital tape. As of July 2002, the Data Center had shipped over 400 data requests to researchers in over 35 countries.

By providing raw neuroimaging data, tools for screening the details of the experimental and scanner protocols, as well as techniques for data discovery and mining, the Data Center will give researchers access to a much larger pool of data than can be found in any single fMRI study. Thus, dynamic analyses across data sets and experiments may be performed, new statistical techniques can be developed, improved techniques for image processing can be investigated, and novel neuroscientific questions may be explored.

Conclusion

Providing scientists access to an ever increasing collection of information in a coherent framework is vital to the future advancement of science. Already in 1985, a National Academy of Sciences report (Morowitz, 1985) suggested that biological research had reached a point at which "new generalizations and higher order biological laws are being approached but may be obscured by the simple mass of data." It is hoped that the continued progress of neuroinformatics in the development of databases and tools for neuroscience will help researchers navigate the masses of data and information that are being generated. In addition to the development of tools for the community, the success of neuroinformatics will depend on how these technologies are accepted by the community at large. The future success of neuroscience databases and neuroinformatics in general will require the neuroscience community to make a paradigm shift similar to those that have already taken place in the molecular biology community.

Road Map: Implementation and Analysis

Related Reading: Neuroinformatics; Neurosimulation: Tools and Resources

References

- Arbib, M. A., and Grethe, J. S., 2001, *Computing the Brain: A Guide to Neuroinformatics*, San Diego: Academic Press. ♦
- Bosak, J., and Bray, T., 1999, XML and the second generation web, *Scientific American*, 280:89–93. ♦
- Chicurel, M., 2000, Databasing the brain, *Nature*, 406:822–825.
- Gardner, D., Knuth, K. H., Abato, M., Erde, S. M., White, T., DeBellis, R., and Gardner, E. P., 2001, Common data model for neuroscience data and data model interchange, *J. Am. Med. Inform. Assoc.*, 8:17.
- Goddard, N. H., Hucka, M., Howell, F., Cornelis, H., Shankar, K., and Beeman, D., 2001, Towards NeuroML: Model description methods for collaborative modeling in neuroscience, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 356:1209–1228.
- Gupta, A., Ludäscher, B., and Martone, M. E., 2000, Knowledge-based integration of neuroscience data sources, *12th International Conference on Scientific and Statistical Database Management (SSDBM)*, Berlin, Germany, IEEE Computer Society, July 2000.
- Huerta, M. F., Koslow, S. H., and Leshner, A. I., 1993, The human brain project: An international resource, *Trends Neurosci.*, 16:436–438.
- Kötter, R., 2001, Neuroscience databases: Tools for exploring brain structure-function relationships, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 356:1111–1120. ♦
- Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P. T., and Lancaster, J., 1995, A probabilistic atlas of the human brain: Theory and rationale for its development, *NeuroImage* 2:89–101.
- Morowitz, H. J., 1985, *Models for Biomedical Research, A New Perspective*, Washington, DC: National Academy Press.
- Nadkarni P. M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., and Miller, P., 1999, Organization of heterogeneous scientific data using the EAV/CR representation, *J. Am. Med. Inform. Assoc.*, 6:478–493.
- Persidis, A., 1999, Bioinformatics, *Nat. Biotechnol.*, 17:828–830.
- Roos, D. S., 2001, Bioinformatics: Trying to swim in a sea of data, *Science*, 291:1260–1261. ♦
- Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., and Gazzaniga, M. S., 2001, The fMRIDC: The challenges and rewards of large scale databasing of neuroimaging studies, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 356:1323–1339.

Decision Support Systems and Expert Systems

Nikola Kasabov

Introduction

The complexity and the dynamics of many real-world problems, such as decision making and the prediction of economic and financial indexes, decision making in medicine, knowledge discovery in bioinformatics, adaptive intelligent control of industrial processes, on-line decision making based on a large amount of continuous and dynamically changing information on the World Wide Web, and many other problems, impose certain requirements on the information systems, methods, and tools used for this purpose. The information systems and tools must:

- Deal with different types of data and knowledge.
- Adjust incrementally to dynamic changes in the operating environment, accommodating new data and introducing new variables and features as needed without the requirement of redesigning or retraining the whole system. Such adaptation may have to occur in an on-line, real-time mode.
- Update the system's knowledge in a dynamic way.
- Explain in an appropriate way what knowledge is contained in the system, or what knowledge has been learned during the system's operation.

The area of information science and artificial intelligence that is concerned with these issues is called *decision support systems*. A decision support system is an information system that helps humans make a decision about a given problem, under given circumstances and constraints. A decision-making system is a system that makes final decisions and takes actions. Some examples are automated trading systems on the Internet, or systems that grant loans through electronic submissions.

Expert systems belong to the subject area, as they are information systems that contain expert knowledge about a particular problem and perform inference when new data are entered that may be partial or inexact. They provide a solution that is expected to be similar to the solution provided by experts in the field. An expert system usually consists of several parts: (1) a knowledge base, where the expert knowledge resides; (2) a database, where historical and new data are stored; (3) an inference machine, which provides different types of inferences; (4) an efficient interface to users; (5) an explanation module, which provides an explanation of *how* and *why* a certain decision was recommended by the system; and (6) a module that learns and accumulates new knowledge, based on the system's operation and on new incoming data (see Duda and Shortliffe, 1983; Kasabov, 1996).

In the rest of this article we will use the collective term *decision system* to refer to either a decision support system, or a decision-making system, or the most sophisticated among them, the expert system.

Because decision systems deal with different types of data and problem knowledge in different modes (e.g., static versus dynamic data, fixed versus adaptive knowledge, off-line versus on-line mode, and so on), different methods and approaches have been used to build them. These methods include statistical methods (e.g., clustering, principal components analysis, hidden Markov models), mathematical modeling, finite automata, methods used in artificial intelligence (e.g., logic systems, rule-based systems, case-based reasoning, different types of neural networks), and hybrid methods that combine all of the foregoing (Duda and Shortliffe, 1983; Kasabov, 1996).

Decision systems, as presented here, are both human-oriented and human-like systems. The human brain is the ultimate decision

system. This article discusses how neural networks, which are vague analogues of the human brain, can be employed in a decision system. We will be concerned with the human-like implementation of decision systems, which is one of their aspects. But the most important aspect of decision systems is their functionality. Because decision systems help humans in the decision process, they should be comprehensible by humans; they should incorporate elements of human-like thinking and human-like information and knowledge processing; and they must be human oriented.

Neural networks and connectionist-based systems have been applied in decision systems as either learning machines or knowledge representation machines, or both. The following listing enumerates some of these applications.

1. Neural networks are used as low-level data-processing and pattern-matching modules, while rule-based modules are used for higher-level decision making (Fu, 1989; Kasabov, 1990). Neural networks are incorporated as part of a production system or of a first-order logic system for decision support.
2. A base consisting of a fixed set of flat rules is incorporated into a connectionist structure with a predefined built-in inference mechanism (Gallant, 1993). No learning is applied.
3. All elements of a production system (e.g., rules, facts, inference machine) are represented in different connectionist modules (Touretzky and Hinton, 1988). No learning is applied on the structure.
4. Trainable knowledge-based neural networks that have fixed structures in terms of number of neurons and connections are used to accommodate both knowledge (rules) and data (Fu, 1989; Towell and Shavlik, 1993; Cloete and Zurada, 2000). Such knowledge-based neural networks are the fuzzy neural networks (Kasabov, 1996).
5. Knowledge-based neural networks that develop (evolve) their structure, their functionality, and their knowledge in time from incoming data, starting from an initial set of knowledge (if such is available), are used for building adaptive, incremental, on-line learning decision systems (Kasabov, 2002).

We will call a decision system that has neural network modules in its structure a *connectionist-based decision system* (CBDS). The following discussion presents different architectures and applications of such systems, starting with the general framework of a CBDS.

General Framework of a Connectionist-Based Decision System

The framework of a CBDS is shown in Figure 1. It consists of the following parts:

- Preprocessing part (e.g., to perform data filtering and feature extraction).
- Neural network part, consisting of neural network modules that are trained on data and incorporate knowledge.
- Higher-level knowledge-based part (e.g., rules for producing final decisions).
- Adaptation part. This part evaluates the system's performance and makes changes to the system's structure and functionality. For example, output error can be used to adjust relevant neural network modules.

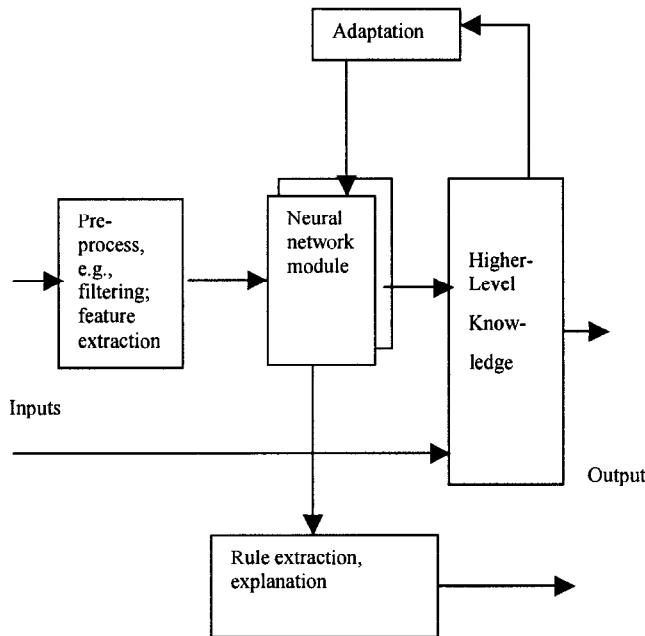


Figure 1. A framework of a connectionist-based decision system (CBDS).

- Rule extraction and explanation part. This part uses both extracted rules from the neural network modules and rules from the decision part to explain *what* the system “knows” about the problem it is designed to solve, and *why* a particular decision for a concrete input vector has been made.

Different types of neural networks (e.g., multilayer perceptrons, self-organizing maps, radial basis functions, fuzzy neural networks) can be used as part of a CBDS. The most commonly used networks are the knowledge-based neural network.

Knowledge-Based Neural Networks

Knowledge-based neural networks are prestructured neural networks that allow for data and knowledge manipulation, including learning from data, rule insertion, rule extraction, adaptation, and reasoning (Towell and Shavlik, 1993; Cloete and Zurada, 2000; Mitra and Hayashi, 2000). Knowledge-based neural networks have been developed either as a combination of symbolic AI systems and neural networks, or as a combination of fuzzy logic systems and neural networks, or as other hybrid systems.

The knowledge represented in knowledge-based neural networks is mainly in the form of different types of IF-THEN rules. Some of them are listed below:

1. Simple propositional rules (e.g., IF x_1 is A AND/OR x_2 is B, THEN y is C, where A, B, and C are constants, variables, or symbols of true/false type) (Gallant, 1993).
2. Propositional rules with certainty factors (e.g., IF x_1 is A (CF_1) AND x_2 is B (CF_2), THEN y is C (CF_c)) (Fu, 1989; Touretzky and Hinton, 1988).
3. Zadeh-Mamdani fuzzy rules (e.g., IF x_1 is A AND x_2 is B, THEN y is C, where A, B, and C are fuzzy values represented by their membership functions).
4. Takagi-Sugeno fuzzy rules (e.g., IF x_1 is A AND x_2 is B, THEN y is $a.x_1 + b.x_2 + c$, where A, B, and C are fuzzy values and a , b , and c are constants)

5. Rules that have associated degrees of importance and certainty degrees (e.g., IF x_1 is A (DI_1) AND x_2 is B (DI_2), THEN y is C (CF_c), where DI_1 and DI_2 represent the importance of each of the condition elements for the rule output, and the CF_c represents the strength of this rule) (Kasabov, 1996).
6. Rules that represent associations of clusters of data from the problem space (e.g., Rule j : IF [an input vector x is in the input cluster defined by its center (x_1 is A_j , to a membership degree of MD_{1j} , AND x_2 is B_j , to a membership degree of MD_{2j}) and by its radius R_{j-in}], THEN [y is in the output cluster defined by its center (y is C, to a membership degree of MD_c) and by its radius R_{j-out} , with $Nex(j)$ examples represented by this rule] (Kasabov, 2002).
7. Temporal rules (e.g., IF x_1 is present at a time moment t_1 (with a certainty degree and/or importance factor of DI_1) AND x_2 is present at a time moment t_2 (with a certainty degree/importance factor DI_2), THEN y is C (CF_c)).
8. Temporal, recurrent rules (e.g., IF x_1 is A(DI_1) AND x_2 is B(DI_2) AND y at the time moment ($t - k$) is C, THEN y at a time moment ($t + n$) is D(CF_c)).

There are several methods for rule extraction from a knowledge-based neural network. Three of them are explained below:

1. Activating a trained knowledge-based neural network on input data and observing the patterns of activation.
2. Rule extraction through analysis of the connections in a trained knowledge-based neural network.
3. Methods that combine 1 and 2.

In terms of applying knowledge-based neural networks to make new inferences, three types of methods can be used:

1. Rules extracted from a knowledge-based neural network are interpreted in another inference machine (e.g., fuzzy inference, production system).
2. The rule-based learning and reasoning modules constitute an integrated connectionist structure, so that reasoning is performed in the connectionist structure.
3. The two options above are combined in one CBDS.

In terms of learning and rule extraction in a knowledge-based neural network, we can differentiate the following cases: (1) off-line learning and static rule set extraction: first, learning is performed, and then rules are extracted, which is an one-off process; (2) on-line learning: rules can be extracted at any time during a continuous on-line learning process.

In terms of learning and optimization in a knowledge-based neural network, there are three cases: (1) globally optimized networks: for every new example all the connections change during learning; (2) locally optimized networks: for a new example only few connections change. Local optimization in a knowledge-based neural network would allow for adjusting the network to accommodate new data through tuning a small number of elements, and also for extracting locally meaningful rules. The rules can be tuned as the system works.

(3) Fuzzy neural networks are neural networks that can be interpreted in terms of fuzzy rules; neuro-fuzzy inference systems are fuzzy systems that can be structurally represented in a connectionist way (Kasabov, 1996; Cloete and Zurada, 2000). A review of neuro-fuzzy systems for rule generation has been published by Mitra and Hayashi (2000).

One example of a fuzzy neural network is shown in Figure 2 (Kasabov, 1996, 2002). The architecture consists of five layers of neurons and four layers of connections. The first layer of neurons receives the input information. The second layer calculates the

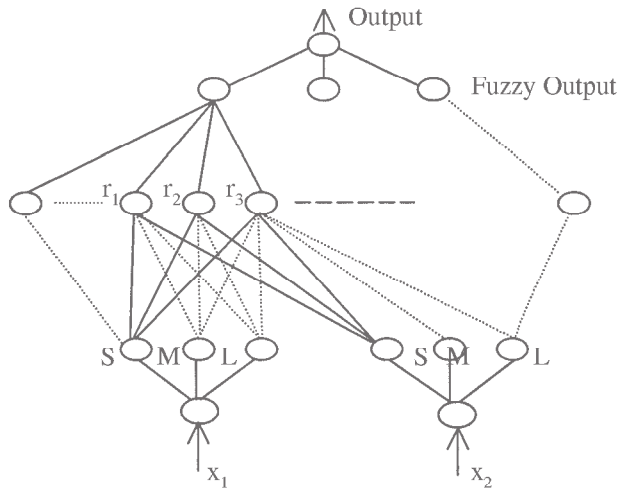


Figure 2. A simple neuro-fuzzy inference system with two inputs, three membership functions, and one output. Rule node r_1 can be represented as a rule: IF x_1 is S and x_2 is S, THEN output is S (certain statistical and linguistic parameters are attached).

fuzzy membership degrees to which the input values belong to predefined fuzzy membership functions, e.g., small, medium, or large. The membership functions can be kept fixed or can change during training. The third layer of neurons represents associations between input and output variables, or fuzzy rules. The fourth layer calculates the degrees to which output membership functions are matched by the input data, and the fifth layer does defuzzification and calculates values for the output variables.

Static Versus Dynamic, Evolving CBDS

A static CBDS is trained on a data set, but it is not continuously updated and adjusted on any new data during its operation. It may be trained regularly on new data plus the previously used data so that the system learns the new data and maintains the previous information as well (the plasticity/stability dilemma). An example is given below.

Example: A CBDS for mortgage approval. A neural network is trained on data for decision making on mortgage approval (the data are also available from the web site <http://divcom.otago.ac.nz/infosci/KEL/data>). The following attributes are used: Input1: character (0—doubtful; 1—good); Input2: total asset; Input3: equity; Input4: mortgage loan; Input5: budget surplus; Input6: gross income; Input7: debt-servicing ratio; Input8: term of loan; Output: decision (disapprove; approve). In a particular experimental CBDS, neural networks are trained on ten different sets of data, each of them containing both positive (applications are approved) and negative (applications are rejected) examples. Each trained neural net-

work is tested on another data set of positive and negative examples. The generalization error of the neural network is evaluated. Rules are extracted from the trained neural network that explains the decision process. The trained neural network is not incrementally trained on new data and it is used as part of a CBDS.

Dynamic, evolving CBDS systems evolve their structure, their functionality, and their knowledge from incoming data rather than having a predefined structure and a fixed set of rules. They learn and improve continuously over time, as is shown in Figure 3. An example of a set of methods that can be used to build an evolving CBDS is the evolving connectionist systems (ECOS) paradigm (Kasabov, 2002).

Evolving CBDS systems adapt to a changing environment, possibly in real time. They can learn in a “lifelong” learning mode, and they are able to explain what they have learned in terms of rules and other types of knowledge. An evolving CBDS would have a modular, open structure evolving over time. Initially, the neural network can be a mesh of nodes (neurons) with very few connections between them, predefined through prior knowledge or “genetic” information. An initial set of rules can be inserted in this structure. Gradually, through self-organization, the system becomes more and more “wired.” The network learns different patterns (exemplars, prototypes) from the training examples. For example, in an ECOS architecture, a node is created and designated to represent an individual example if this example is significantly different from the previous ones (with the level of differentiation established through dynamically changing parameters). In addition to a growing procedure, a pruning procedure is defined. It allows for removing neurons and their corresponding connections that are not actively involved in the functioning of the ECOS, thus making space for new input patterns. Different modes of learning in ECOS are possible, among them are an active learning mode (learning is performed when a stimulus—input pattern—is presented and kept active) and a passive (e.g., sleep) learning mode (learning is performed when there is no input pattern presented at the input of the ECOS).

Adaptive CBDS usually operate as on-line decision systems that make decisions and adjust their knowledge through an incremental, continuous learning from incoming data. Such systems, for example, learn the dynamic changes in a stock index, or adjust their knowledge to include new gene discoveries in Bioinformatics.

Example: On-line financial decision making based on on-line prediction of the MIB30 stock index. Figure 4 shows the use of a neural network for on-line learning and prediction 5 days ahead of the moving average values of the MIB30 stock index (the Milan stock index). Here, an evolving fuzzy neural network (Kasabov, 2002) is used as the neural network. The neural network is trained on-line on consecutive data vectors. Inputs to the neural network are 5-day moving averages of the following variables: $DJ(t)$, $DJ(t-1)$, $MIB30(t)$, $MIB30(t-1)$, euro/US\$(t), euro/US\$($t-1$).

At any time of the functioning of the evolving fuzzy neural network, rules can be extracted. The rules represent the current asso-

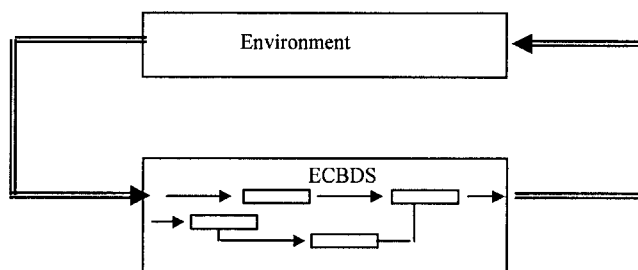


Figure 3. Adaptive, evolving CBDS learn their structure and functionality through interaction with the environment

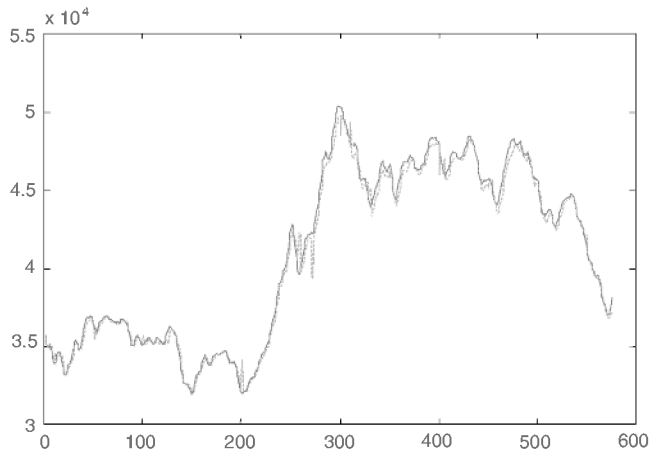


Figure 4. Adaptive CBDS in financial decision making, showing the process of on-line learning and prediction 5 days ahead of the MIB30 index. Chart shows the desired versus the predicted output value for 576 consecutive days.

ciation between the input variables and the output variable, as is illustrated in the rule below:

Rule: IF [DJ(t) is (Medium 0.828) and DJ($t - 1$) is (Medium 0.840) and MIB30(t) is (Low 0.885) and MIB30($t - 1$) is (Low 0.887)] (receptive field = 0.808), THEN MIB30($t + 5$) will be (Low 0.852) (accommodated training examples 182 out of 576).

Connectionist-Based Decision Systems in Economics and Finance

Beltraffi, Margarita, and Terna (1996) describe a variety of CBDS systems in the area of finance and economics. These include CBDS for solving problems such as stock trading, portfolio decision, exchange rate prediction, fraud detection, credit scoring, and many more. Some CBDS systems contain expert rules and make decisions similar to the decisions made by experts, as illustrated in the next example.

Example: A CBDS for simulation and prediction of decisions made by the European Central Bank (ECB) on interest rate intervention (Rizzi et al., 2002). The ECB meets regularly to make a decision on the interest rate for the European Union countries. Modeling this decision-making process is extremely difficult, as it requires both the comprehensive knowledge the ECB members have and fast adaptation to new situations. The system developed by Rizzi et al. (2002) consists of two parts. The first part has several neural networks that use six groups of economic indicators (total of 17 variables) and produces intermediate outputs that include the predicted values for some time series. These values, as well as other economic variables, are fed into a rule-based system that constitutes the second part of the system. The system finally suggests what ECB decisions on the interest rate might be expected at the next three consecutive meetings. Figure 5 shows some test results. The system learns and improves over time after each new datum is entered.

Connectionist-Based Decision Systems in Bioinformatics

Processing a large amount of data, such as in medicine, biochemistry, or molecular biology, and making decisions based on this

information is a task where CBDS systems have been successfully applied. Baldi and Brunak (2001) have demonstrated the application of machine learning techniques, including neural networks, to solving difficult problems, such as DNA promoter recognition, RNA splice junction identification, secondary structure protein prediction, and so on.

Example: CBDS for RNA splice junction identification. Figure 6 shows a simple three-layer perceptron neural network for identifying a biological feature, such as a splice junction from an input RNA sequence. The neural network has 60 inputs, which is the length of the RNA window sequence, five intermediate neurons (nodes), and one output neuron to encode for the feature (1—present, 0—not present). (The data set used for training is available at <http://www.ics.uci.edu/~mllearn/MLRepository.html>.) The splice junction predicted by the trained neural network from new input data should be further analyzed before a final decision is made.

An on-line training and prediction system for the same problem that is based on an evolving fuzzy neural network is available at <http://divcom.otago.ac.nz/infosci/kel/CBIIS/GenIn/>. The system can also extract rules, such as the rule shown below:

Rule: IF -----C--C-C-TCC-G--CTC-GT-C--GGTGAGTG--GGC---C---G-GG-C--CC- THEN [Junction Exon-Intron] Receptive field = 0.216. Examples covered by the rule are 26 out of 1,000.

Future Directions for Connectionist-Based Decision Systems

CBDS systems are a fast-growing area. New applications expected to be developed include new techniques for learning, data mining, and knowledge discovery, and practical applications in finance and economics, bioinformatics and life sciences, process control and manufacturing, medicine, and the social sciences.

Road Map: Artificial Intelligence; Applications

Related Reading: Forecasting; Hybrid Connectionist/Symbolic Systems

ECB Intervention	Forecasts of the Expert System		
Date (t)	Date ($t + 1$)	Date ($t + 2$)	Date ($t + 3$)
3 Feb. 2000 0.25	16 Mar. 2000 0.25 (0.25)	27 Apr. 2000 0 (0.25)	8 Jun. 2000 0 (0.50)
16 Mar. 2000 0.25	27 Apr. 2000 0.25 (0.25)	8 Jun. 2000 0.25 (0.50)	31 Aug. 2000 0.25 (0.25)
27 Apr. 2000 0.25	8 Jun. 2000 0.25 (0.5)	31 Aug. 2000 0.25 (0.25)	19 Oct. 2000 0.25 (0.25)
8 Jun. 2000 0.5	31 Aug. 2000 0.25 (0.25)	19 Oct. 2000 0.25 (0.25)	30 Nov. 2000 0.25
31 Aug. 2000 0.25	19 Oct. 2000 0.25 (0.25)	30 Nov. 2000 0.25	18 Jan. 2001 0.25

Figure 5. CBDS in economics: the interest rate decision intervention by the European Central Bank (ECB) at five ECB meetings, versus the calculated ahead intervention by an expert system. After each intervention index decided by the ECB is made available, the expert system adjusts to this decision and then suggests what the ECB decision at three consecutive meetings ahead might be. The real intervention indexes are the numbers in parentheses.

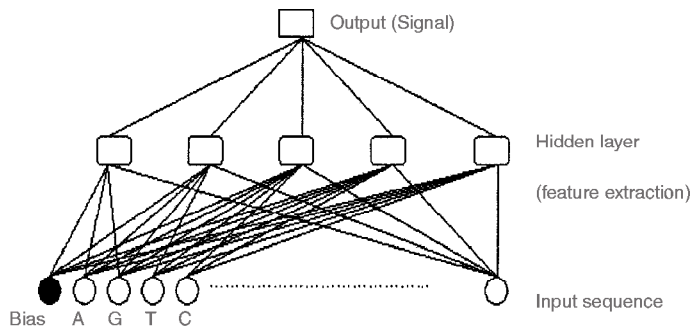


Figure 6. CBDS in bioinformatics. A neural network that takes an input vector of 60 nucleotides from an RNA data and evaluates the probability for having exon-intron, or intron-exon splice junction in the middle of the sequence, or no junction at all.

References

- Baldi, P., and Brunak, S., 2001, *Bioinformatics: A Machine Learning Approach*, Cambridge, MA: MIT Press. ♦
- Beltraffi, A., Margarita, S., and Terna, P., 1996, *Neural Networks for Economics and Financial Modelling*, New York: International Thomson Computer Press.
- Cloete, I., and Zurada, J., Eds., 2000, *Knowledge-Based Neurocomputing*, Cambridge, MA: MIT Press.
- Duda, R. O., and Shortliffe, E. H., 1983, Expert systems research, *Science*, 220:261–268.
- Fu, L. M., 1989, Integration of neural heuristics into knowledge-based inference, *Connect. Sci.*, 1:325–340. ♦
- Gallant, S. I., 1993, *Neural Network Learning and Expert Systems*, Cambridge, MA: MIT Press. ♦
- Kasabov, N. K., 1990, Hybrid connectionist rule-based systems, in *Artificial Intelligence: Methodology, Systems, Applications* (P. Jorrand and V. Sgurev, Eds.), Amsterdam, North-Holland: Elsevier, pp. 227–235.
- Kasabov, N., 1996, *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering*, Cambridge, MA: MIT Press. ♦
- Kasabov, N., 2002, *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines*, London: Springer-Verlag.
- Mitra, S., and Hayshi, Y., 2000, Neuro-fuzzy rule generation: Survey in soft computing framework, *IEEE Trans. Neural Net.*, 11 No. 3
- Rizzi, R., Bazzana, F., Kasabov, N., Fedrizzi, M., and Erzegovesi, L., 2002, A connectionist-based decision support system for modelling the interest rate intervention made by the European Central Bank, *Eur. J. Operat. Res.*, Special Issue on Decision Support Systems.
- Sima, J., and Cervenka, J., 2000, Neural knowledge processing in expert systems, in *Knowledge-Based Neurocomputing* (I. Cloete and J. Zurada, Eds.), Cambridge, MA: MIT Press, pp. 419–466.
- Towell, G. G., and Shavlik, J. W., 1993, Extracting refined rules from knowledge-based neural networks, *Machine Learn.*, 13:71–101.
- Touretzky, D., and Hinton, G., 1988, A distributed connectionist production system, *Cognit. Sci.*, 12:1423–1466.

Dendritic Learning

Bartlett W. Mel

Introduction

In most neurons of the vertebrate central nervous system (CNS), dendritic trees are the primary input surfaces of neurons, receiving thousands to hundreds of thousands of synaptic contacts from other neurons. During development, a period during which the brain is first “wired up,” the physical interface between axons and dendrites is extremely dynamic (Cline, 1999), involving large-scale growth, retraction, and remodeling of axonal and dendritic arbors on time scales of minutes to hours. In the adult nervous system, contrary to the connectionist view that neural plasticity is limited to the strengthening and weakening of existing synaptic connections, evidence suggests that physical remodeling of axons and dendrites, including the formation of new synaptic contacts and elimination of existing ones, continues to some degree throughout life (Klintsova and Greenough, 1999). Given that dendrites are highly complex structures both anatomically and physiologically and are the principal substrates for information processing within the neuron itself (Stuart, Spruston, and Häusser, 1999), the question arises as to the consequences of axodendritic structural plasticity for learning and memory. In other words, does experience-dependent remodeling of the physical interface between axons and dendrites have any special role in the long-term storage of information in the brain, beyond that associated with the establishment of the brain’s basic neuron-to-neuron wiring diagram?

A critical assumption of most models of learning and development holds that the *neuron* is the appropriate level of granularity for analysis, where the outcome of learning is expressed in terms of changes in the strengths of connections w_{AB} between point neurons *A* and *B* (see HEBBIAN SYNAPTIC PLASTICITY). The enormous physical complexity of individual neurons compels us to consider other types of plasticity, however. For example, activity-dependent rules appear to modulate the density and spatial distribution of the ion channels governing a cell’s basic electrical behavior (see ACTIVITY-DEPENDENT REGULATION OF NEURONAL CONDUCTANCES), and can modulate neurite outgrowth and branching (van Ooyen et al., 2002). Both examples involve changes in the neural substrate that are not naturally described in terms of changes in neuron-to-neuron connection strengths.

In this article, we explore the possibility that long-term learning in the mature brain may depend on structural remodeling at the interface between axons and dendrites, a process that continues throughout life. In particular, we cite evidence for the view that long-term storage may involve the correlation-based sorting of synaptic contacts onto the *many separate dendrites* of a target neuron, just as conventional models of neural development typically involve the sorting of synaptic contacts onto the *many separate neurons* of a target population (see DEVELOPMENT OF RETINOTECTAL MAPS; OCULAR DOMINANCE AND ORIENTATION COLUMNS; SELF-ORGANIZING FEATURE MAPS). This shift in granularity is justified

by the assumption that individual dendrites, or parts of dendritic trees, act as separately thresholded neuron-like subunits, functionally analogous to the point neurons that populate coarser-grained models of learning and development. We focus on the main projection neurons of cortical tissue, pyramidal cells, although our discussion likely applies to other types of cells as well.

The Neuron as Two-Layer Neural Network

The dendrites of pyramidal cells contain a large number and variety of voltage-dependent channels that profoundly influence their integrative behavior (Häusser, Spruston, and Stuart, 2000). A variety of evidence from intracellular recordings and imaging studies suggests that active spike-like responses can be localized within individual thin branches (Schiller et al., 2000), supporting the idea that individual dendritic branches can act as surrogate “neurons” capable of separately thresholding their synaptic inputs.

Anatomical hints are similarly supportive of such a possibility. If advantages accrue to a cell that maintains multiple integrative subregions within its dendritic tree, one might expect pyramidal cell morphologies to maximize the number of subunits available for independent synaptic processing, subject to practical constraints such as that the cell remain of manageable size, that nonlinear synaptic interactions remain confined to individual subunits, and so on. Pace, Tieman, and Tieman (2000) found that among layer 4 stellate cells of the cat striate cortex, the number of long, thin, terminal sections is nearly constant (around 40) and is independent of the number of primary dendritic branches emanating from the cell body (Figure 1). Given that most of the synapses onto basal dendrites lie on the long, thin, unbranched terminal sections (Elston and Rosa, 1998), the data of Pace et al. suggest a developmental program that tightly regulates the production of mutually isolated dendritic subunits, and then arranges for synapses to be formed primarily there.

What are the implications of this type of morphology for synaptic integration? Cable theory (Koch, 1999) suggests that a dendritic arbor consisting of many thin-branched subtrees radiating outward from a much larger-diameter cell body or dendritic trunk is ideally suited to isolate voltage responses within individual thin branches. Indeed, the possibility that dendritic trees could support complex multisite nonlinear operations, including logic-like operations, has been explored in a number of modeling studies (see Mel, 1999; Segev and London, 2000).

One recent study utilized a simplified model pyramidal cell whose dendrites contained AMPA/NMDA synapses and low concentrations of voltage-dependent Na^+/K^+ channels capable of generating dendritic spikes (Archie and Mel, 2000). When total synaptic drive to the cell was held constant but was distributed in varying spatial patterns to two dendritic branches, the average firing rate of the cell was approximated by a sum-of-squares model. The finding is intriguing in that it suggests a possible connection between the computation carried out within the dendrites and the quadratic “energy” models used to describe several types of visual receptive field properties (Mel, 1999). This study provided the first direct test of the “sum of subunits” model for synaptic integration, in which (1) the thin dendritic branches act like separately thresholded nonlinear subunits and (2) the outputs of the thin branch subunits are summed linearly via the main trunks and cell body prior to global thresholding:

$$y(\mathbf{x}) = g\left(\sum_{i=1}^m \alpha_i b\left(\sum_{j=1}^k w_{ij} x_{ij}\right)\right) \quad (1)$$

where m is the number of subunits, k is the number of synapses per subunit, w_{ij} is the weight, and x_{ij} is the activity of the j th input to the i th subunit, b is the subunit nonlinearity, α_i is the coupling of the i th subunit to the cell body, and g is a global output nonlinearity. Of interest, Equation 1 also describes the input-output relation of a conventional two-layer neural network.

Structural Plasticity at the Axodendritic Interface

The possibility that neurons contain multiple, separately thresholded dendritic subunits has profound implications for the mechanisms governing the formation and remodeling of the axodendritic interface, and for the physical substrate for long-term information storage in neural tissue. The point may be illustrated from the perspective of axon i in the process of “choosing” which subunit $s \in \{1 \dots k\}$ to enervate on postsynaptic neuron j during learning or development. The subunit thresholding function b , which generates nonlinear interactions among the set of inputs to each subunit, ensures that i ’s effectiveness in driving cell j depends not just on its own activity x_i and associated weights w_{ijs} , but also on the activity and weights of the other axons providing input to the same subunits. Thus, given compartmentalized neurons, the “receptive field” of the neuron changes, in general, when any single axon withdraws

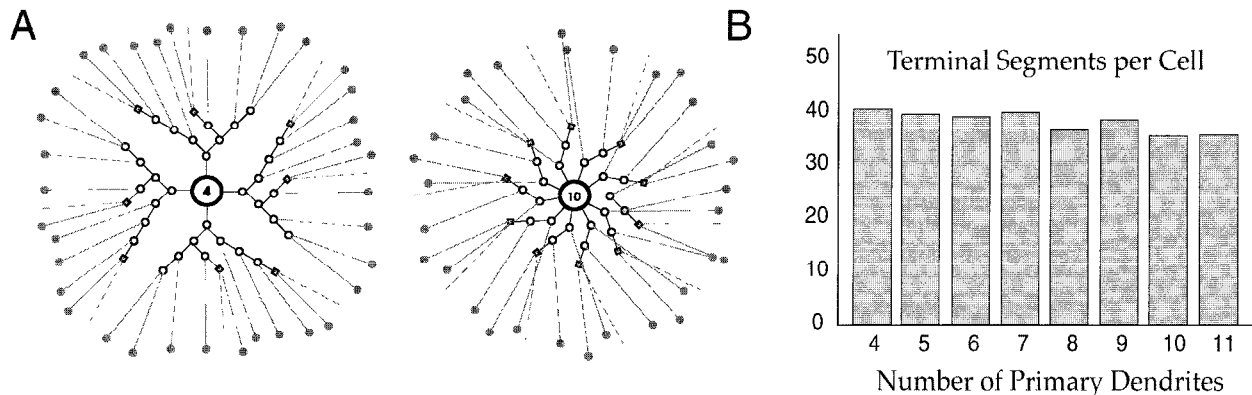


Figure 1. Evidence for regulation of number of functional dendritic subunits. *A*, Schematics of spiny-stellate cell morphology in cat visual cortex derived from 3D reconstructions, shown for cells with four or ten primary dendrites. *B*, Number of unbranched terminal segments per cell is nearly

constant (around 40) for cells with widely varying numbers of primary dendrites. (Adapted with permission from Pace, Tieman, and Tieman, 2000.)

a synaptic contact from one subunit and forms a new contact on another, even when the “change of address” involves two branches of the same postsynaptic cell. By contrast, learning models operating at neuron-level granularity, which encode only the overall connection strength between pairs of neurons, lack the parameters needed to represent such changes.

Partnership Combinatorics

The possibility that learning-related mechanisms could orchestrate the correlation-based sorting of synaptic contacts, not just onto whole neurons but a level down, on to specific dendritic subunits, raises the question of whether the physical interface between axons and dendrites in cortical tissue is amenable to this type of fine-scale structural plasticity. The question is important in that, when a neuron contains subunits, its capacity to absorb learned information is closely tied to the *addressing flexibility* of the tissue, that is, the flexibility to establish arbitrary partnerships between presynaptic axons and postsynaptic dendritic subunits (Poirazi and Mel, 2001).

It is now well established that axons, dendrites, and spines are highly dynamic structures, both during development and in the adult brain (Cline, 1999; Klintsova and Greenough, 1999). One model of neural development holds that (1) synapses are initially formed between axons and dendrites in a random, activity-independent fashion, and (2) synapses that are frequently co-activated with their neighbors within the same postsynaptic compartment are structurally stabilized and retained, while those that are poorly correlated with their neighbors are eliminated. If, however, the relevant postsynaptic unit is the individual thin dendrite rather than the whole neuron, then these same neurobiological mechanisms could also drive the separate mapping of like-activated synaptic cohorts onto distinct dendritic subregions.

Although the idea is intriguing, a serious practical difficulty arising from the need to establish on-demand partnerships between arbitrary pairs of axons and dendrites is that of physical proximity, or the lack thereof: it is unreasonable to expect that during the

course of learning, particularly in the densely packed neuropil of the mature brain, axons or dendrites should be regularly required to advance and retract over long distances in search of appropriate partnerships.

What physical properties of axons and dendrites might enhance their partnership flexibility, minimizing the need for long-distance travel to form arbitrary pairings between presynaptic axons and dendritic subunits? Axonal and dendritic arborizations are heavily interdigitated within the three-dimensional (3D) volume of the cortical neuropil, an arrangement that maximizes the probability of a close approach between any given axon and any given compartment of a postsynaptic cell. This qualitative observation is illustrated by the montage of an axon and a dendrite shown in Figure 2.

Implications of Dendritic Subunits and Structural Plasticity for Long-Term Memory Storage

We recently set out to quantify the excess trainable storage capacity contained in the selective mapping of synaptic contacts onto dendritic subunits, and to characterize how this excess capacity depends on dendritic geometry (Poirazi and Mel, 2001). We compared the capacity of a subunit containing neuron with m branches (subunits) and k synapses per branch (Equation 1) to that of a point neuron with the same number of synaptic sites but a linear summation rule, that is, with $b(x) = x$ in Equation 1. The two neuron models were thus identical except for the presence or absence of a fixed subunit nonlinearity.

Assuming synaptic contacts of unit weight—although any of the d input lines could form multiple connections to the same or different branches—we derived upper bounds on the capacity of a linear (B_L) versus nonlinear (B_N) cell:

$$B_L = 2 \log_2 \binom{s+d-1}{s}$$

$$B_N = 2 \log_2 \left(\binom{k+d-1}{k} + m - 1 \right) \quad (2)$$

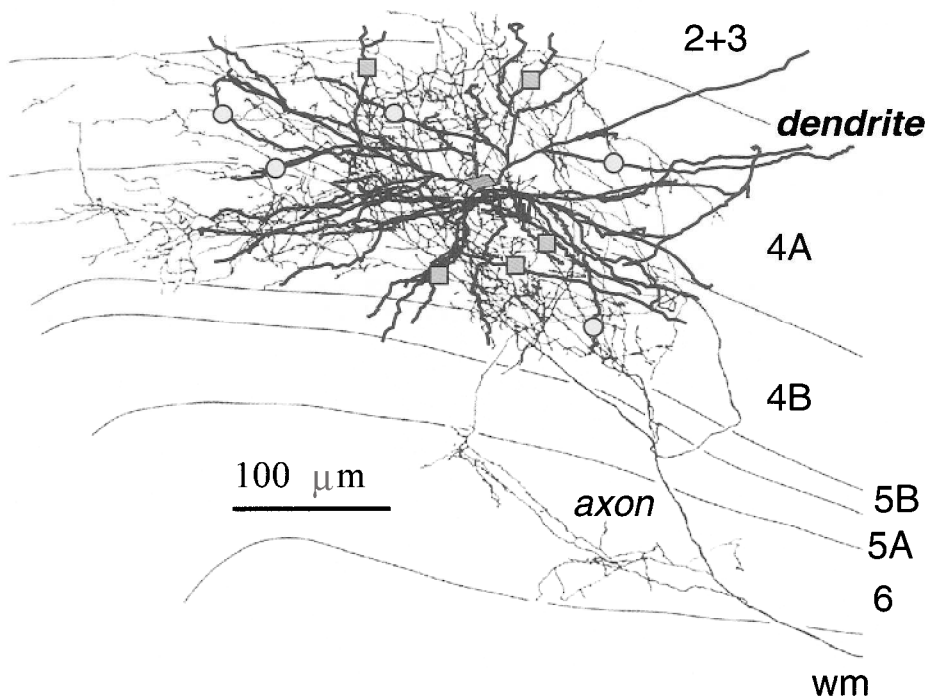


Figure 2. Interdigitated axonal and dendritic trees provide an ideal interface for flexible partnering between a presynaptic axon and the several dozen dendritic subunits of a single postsynaptic neuron. The picture was created by superimposing a dendritic arbor from a cat layer 4 spiny stellate cell (courtesy of Judith Hirsch) on a thalamocortical afferent taken from Freund et al. (1985, *J. Comp. Neurol.*, 242:263). Geometric symbols illustrate locations where minor extensions or retractions could lead to the establishment of new contacts (circles) and the elimination of old contacts (squares) between the axon and dendrite during the course of learning.

The expressions in each case estimate the number of distinct input-output functions that can be expressed by the respective model when assigning $s = m \cdot k$ synaptic contacts with replacement from d distinct input lines. The combinatorial terms take into account the redundancies associated with the two models, that is, the changes in synaptic connectivity that have no effect on the cell's "receptive field"; the logarithm converts the raw function counts into *bits*. B_L and B_N are plotted in Figure 3A for a cell with 10,000 synaptic contacts and three values of d . Capacity is shown on the y-axis for a range of cell geometries represented along the x-axis. The values of B_L are shown on the left and right edges of the plot, since the capacity of a point neuron is equivalent to a subunitized neuron in a degenerate state with either a single branch containing 10,000 synapses or with 10,000 branches containing one synapse each. The peak capacity occurs for cells containing approximately 1,000 subunits of size 10, where the optimal geometry depends little on d over the order-of-magnitude range tested.

The optimal subunit size predicted by Equation 2 ($k = 10$) is considerably smaller than would be expected for a pyramidal cell, whose individual thin terminal branches typically contain hundreds of synapses. This discrepancy could be explained in part by our assumption here that synaptic weights are binary valued; in pilot runs with multivalued weights, the optimal subunit size was pushed to substantially larger values ($k = 25$ for 4-level weights). Furthermore, the assumption that all input axons have ready access to all dendrites is unlikely to hold in neural tissue. Wherever this assumption is violated, pressure would exist to grow longer dendritic subunits, thereby permitting each subunit to gain access to a larger fraction of the "input vector."

Empirical Testing of Memory Capacity

To validate the analytical model, we trained both linear and nonlinear cells on random old/new classification problems using a stochastic gradient descent learning rule (Poirazi and Mel, 2001). Memory capacity was measured for cells with different dendritic geometries by determining the size of the training set that could be internalized with a recognition error rate of 2%. A comparison of analytical versus empirical capacities for both linear and nonlinear cells is shown in Figure 3B. The curves are remarkably similar in form, with peak capacity occurring for cells of similar shape. The optimal nonlinear cell with 10,000 synapses outperformed its size-matched linear counterpart by a factor of 46, learning 27,400 versus 600 patterns at the 2% error criterion. In further experiments with a population of independently trained cells, we found the excess storage capacity available to a structural learning rule could easily approach two orders of magnitude (Poirazi and Mel, 2001).

Discussion

Several types of evidence were cited that call into question the classical "point neuron" as a model for a pyramidal cell or other large dendritic neuron of the CNS. We presented an alternative model supported by physiological, anatomical, and modeling studies, in which the output of the cell represents the sum of a moderately large set of separately thresholded dendritic subunits—a formulation that at a mathematical level is identical to that of a

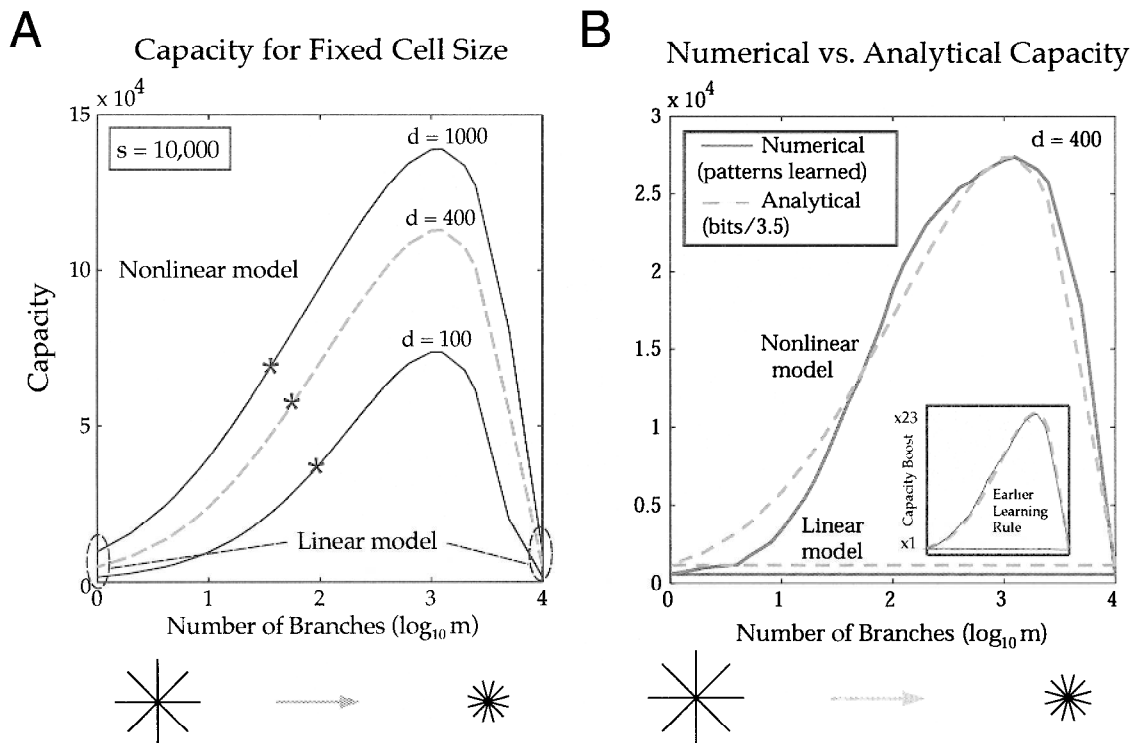


Figure 3. Linear versus nonlinear cell capacity as a function of branch geometry. *A*, Capacity of a nonlinear cell with 10,000 sites for three values of d . Cells at both ends of the x-axis have a capacity equivalent to that of the linear model. Asterisks indicate the half-maximum capacity. *B*, Comparison of memory capacity predicted by analysis with that found empirically in random memorization problems. Dashed lines show analytical

curves for linear and nonlinear cells (nonlinear capacity curve corresponds to dashed curve in *A*). Solid curves show capacity measured empirically at 2% error criterion. Analytical curves were scaled down together by a factor of 3.5, to align peak analytical and empirical capacity values for nonlinear model. (Adapted from Poirazi and Mel, 2001.)

conventional artificial neural network built from two layers of point neurons. Multisubunit neurons, if and where they are used in the brain, could help to minimize hardware-associated costs, including brain size, processing delays, error-signal management overhead, and so on.

Although the validity of the subunitized-neuron hypothesis remains to be proved empirically, we have nonetheless explored some of its major consequences for learning and memory. The two main consequences are as follows. First, in a functionally compartmentalized neuron, the formation of new synapses and the elimination of old ones during learning can no longer be viewed simply as a means to increase or decrease the overall connection strength between two neurons, a common interpretation of new synapse or spine formation. Indeed, the concept of “overall connection strength between two neurons” is no longer well-defined, in the sense that the interaction between two neurons can no longer be captured by a single positive or negative coefficient. The granularity has changed: we must now worry about the role of learning-related mechanisms in tuning the connection strengths between a many-fingered presynaptic fiber and the multiple dendritic subunits of a given postsynaptic cell.

Second, from the perspective of learning theory, this change in granularity brings with it a large increase in the number of modifiable parameters available to the neural tissue. And this is not a purely theoretical construct: we have found in simulation studies that these extra parameters translate directly into additional long-term storage capacity, hidden in the fine structure of the axodendritic interface. Interestingly, other sources of “hidden” capacity have been proposed to exist, based on still other kinds of modifiable parameters, including a cell’s capacity to discriminate among time-varying signals by varying passive time constant and spike threshold (Zador and Pearlmuter, 1996), or to maximize mutual information between input and output firing rates by varying properties of the resident voltage-dependent ion channels (Stemmler and Koch, 1999). Future experiments will ultimately determine the extent to which these reservoirs of structure-, channel-, or time-based capacity are in fact exploited within the living brain.

Dendritic Processing

Idan Segev and Michael London

Introduction

We are fortunate to be in the midst of the “dendritic revolution” that emerged when the first edition of the *Handbook* appeared in 1995. For the first time ever, systematic and intimate electrical and optical visits became possible to the dendrites (Figure 1)—the largest component of the mammalian brain in both surface area and volume. These ongoing visits have yielded a much more fascinating picture of the electrical behavior and chemical properties of dendrites than one could have imagined only a few years ago. It is now clear that the dendritic membrane hosts a variety of nonlinear voltage-gated ion channels that endow dendrites with potentially powerful computing capabilities. Our century-old perception of dendrites as electrical devices that carry information unidirectionally, from the many dendritic (input) synapses to the soma and (output) axon, has undergone a dramatic revision. The surprising finding is that dendrites of many central neurons also carry information “backward,” via active propagation of action potentials (APs) from the axon to the dendrites. These “backpropagating APs”

Road Map: Neural Plasticity

Related Reading: Dendritic Processing; Development of Retinotectal Maps; Ocular Dominance and Orientation Columns

References

- Archie, K. A., and Mel, B. W., 2000, An intradendritic model for computation of binocular disparity, *Nature Neurosci.*, 3:54–63.
- Cline, H. T., 1999, Development of dendrites, in *Dendrites* (G. Stuart, N. Spruston, and M. Häusser, Eds.), Oxford, Engl.: Oxford University Press, pp. 35–67. ♦
- Elston, G. N., and Rosa, M. G., 1998, Morphological variation of layer III pyramidal neurones in the occipitotemporal pathway of the macaque monkey visual cortex, *Cereb. Cortex*, 8:278–294.
- Häusser, M., Spruston, N., and Stuart, G. J., 2000, Diversity and dynamics of dendritic signaling, *Science*, 290:739–744. ♦
- Klintsova, A. Y., and Greenough, W. T., 1999, Synaptic plasticity in cortical systems, *Curr. Opin. Neurobiol.*, 9:203–208. ♦
- Koch, C., 1999, *Biophysics of Computation*, Oxford, Engl.: Oxford University Press.
- Mel, B. W., 1999, Why have dendrites? A computational perspective, in *Dendrites* (G. Stuart, N. Spruston, and M. Häusser, Eds.), Oxford, Engl.: Oxford University Press, pp. 271–289. ♦
- Pace, C. J., Tieman, D. G., and Tieman, S. B., 2000, Neuronal form: Patterns of dendritic branching in layer 4 stellate cells, *Soc. Neurosci. Abstr.*, 2(794.2):489.
- Poirazi, Y., and Mel, B. W., 2001, Impact of active dendrites and structural plasticity on the memory capacity of neural tissue, *Neuron*, 29:779–796.
- Schiller, J., Major, G., Koester, H. J., and Schiller, Y., 2000, NMDA spikes in basal dendrites of cortical pyramidal neurons, *Nature*, 404:285–289.
- Segev, I., and London, M., 2000, Untangling dendrites with quantitative models, *Science*, 290:744–750. ♦
- Stemmler, M., and Koch, C., 1999, How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate, *Nature Neurosci.*, 2:521–527.
- Stuart, G., Spruston, N., and Häusser, M., Eds., 1999, *Dendrites*, Oxford, Engl.: Oxford University Press.
- van Ooyen, A., Corner, M., Kater, S., and van Pelt, J., 2002, Activity-dependent neurite outgrowth and network development, in *Modeling Neural Development* (A. van Ooyen, Ed.), Cambridge, MA: MIT Press.
- Zador, A., and Pearlmuter, B. A., 1996, VC dimension of an integrate-and-fire model neuron, *Neural Computat.*, 8:611–624.

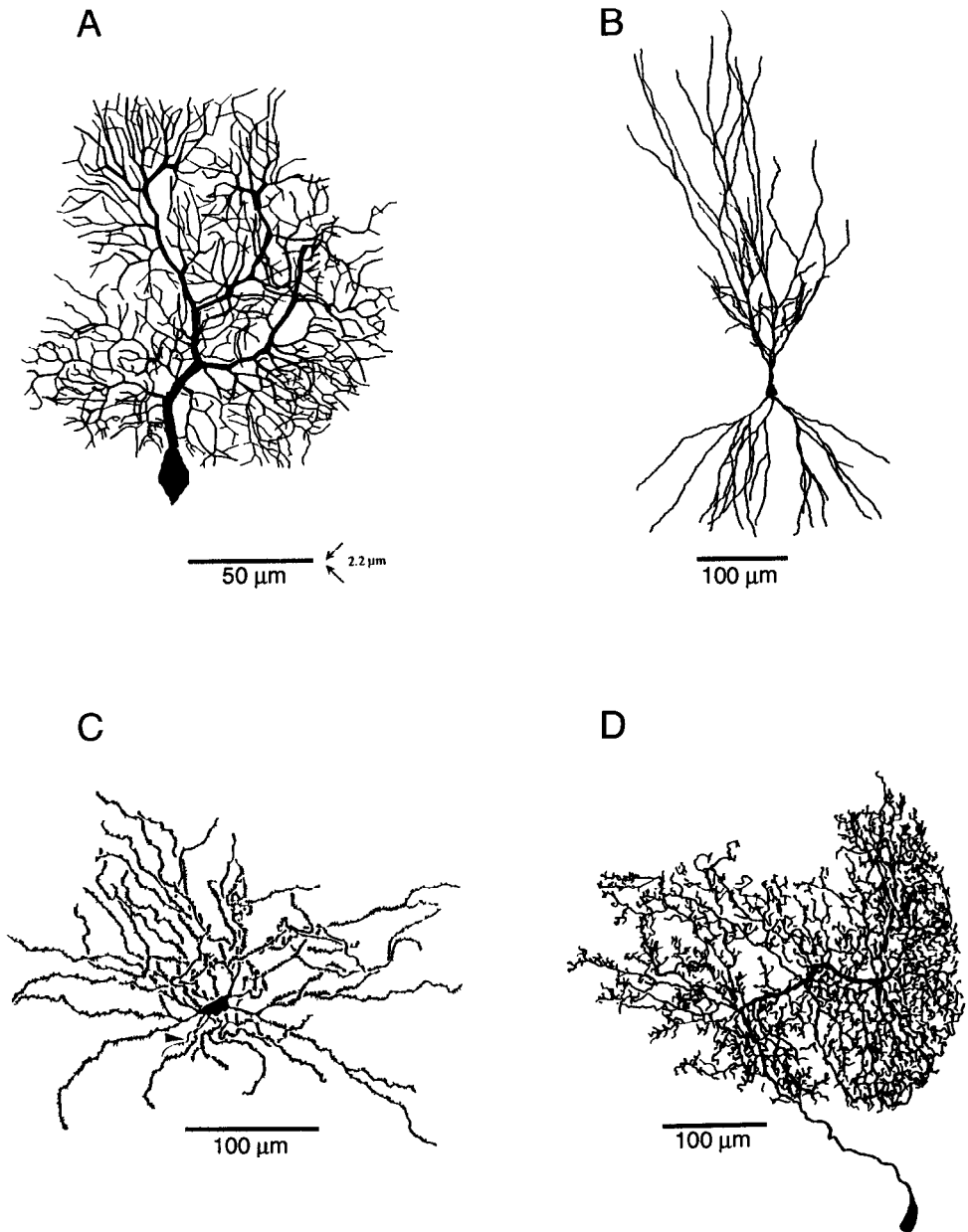
trigger plastic changes in the dendritic input synapses, and these changes are believed to be the fundamental processes that underlie learning and memory in the brain. Our view of dendrites as static elements has also changed: we now know that the fine morphology as well as the electrical properties of dendrites change dynamically, in an activity-dependent manner. Indeed, one of the major challenges in the years to come is to understand how these changes are correlated with the behavior and learning of the animal.

This article affords a brief introduction to the newly discovered charms of dendrites. For the dendritic-minded reader, several recent reviews are recommended, including the book *Dendrites*, edited by Stuart, Spruston, and Häusser (1999), the quartet of reviews of work on dendrites in *Science* (October 27, 2000), and the review by Euler and Denk (2001).

Dendrites—What Made the Revolution Possible

A combination of several new technologies has made it possible to view dendrites with unprecedented acuity, notably differential-

Figure 1. Dendrites have unique shapes, a feature that is used to characterize neurons into types. In many neuron types, synaptic inputs from a given source are preferentially mapped into a particular region of the dendritic tree. *A*, Cerebellar Purkinje cell of the guinea pig (reconstructed by M. Rapp). *B*, CA1 pyramidal neuron from the rat (reconstructed by B. Claiborne). *C*, Neostriatal spiny neuron from the rat. (Courtesy of C. Wilson.) *D*, Axonless interneurons of the locust (reconstructed by G. Laurent). Typically, synaptic inputs are distributed nonrandomly over the dendritic surface. (*C'* from Wilson, in McKenna, T., Davis, J., and Zornetzer, S. F., Eds., 1992, *Single Neuron Computation*, Boston Academic Press. Reproduced with permission.)



interference contrast (DIC) video microscopy and two-photon microscopy (for a review, see Euler and Denk, 2001). These methods allow researchers to systematically record electrically (using microelectrodes) and optically (using ion-dependent dyes) from identified dendritic locations. Consequently, the effect of a single synaptic input can be monitored in both its dendritic origin and simultaneously in the soma/axon region. Direct measurements of the electrical properties of the dendritic ion channels are now feasible (Johnston et al., 1996; Reyes, 2001); and, with molecular methods (Crick, 1999), specific membrane proteins in dendrites can be marked and manipulated to identify the type and distribution of the various receptors and ion channels in the dendritic membrane. When combined with sophisticated analytic and numerical models, the beginnings of a functional picture emerge from these diverse experimental data. Some of the key experimental and theoretical findings are described below.

Fundamental Facts About Dendrites: Morphological Face, Electrical Character, and Synaptic Inputs

Table 1 summarizes the functionally important facts about dendrites. Because dendrites come in many shapes and sizes, such a summary inevitably presents only a rough range of values. Nonetheless, several important functional conclusions can be drawn from this table.

1. *Morphologically*, dendrites tend to ramify, creating large and complicated trees. Dendrites are thin processes, starting with a diameter of a few micrometers near the soma; the branch diameter typically falls below $1\ \mu\text{m}$ with successive branching, often reaching a distance of 1 mm from the soma. Many (but not all) types of dendrites are studded with abundant tiny branches, or appendages, called dendritic spines. When present,

Table 1. Range of Values for Dendritic Machinery

Morphology	Physiology	Synaptology
Diameter near soma: 1–6 μm	<i>Passive properties of dendrites^a</i>	No. of synapses/neuron: 500–200,000
Diameter at distal tips: 0.3–1 μm	Membrane resistivity (R_m): 1–100K Ωcm^2	Type I (excitatory): 60%–90% ^c ; distributed, majority on spines
Average path length: 0.15–1.5 mm	Axial resistivity (R_i): 70–300 Ωcm	Type II (inhibitory): 10%–40% ^c ; near soma, some on spines
Total dendritic length: 1–10 mm	Membrane capacitance (C_m): 1–2 $\mu\text{F}/\text{cm}^2$	
Dendrite area: 2,000–750,000 μm^2	Membrane time constant (τ_m): 1–100 ms	
Dendritic trees/neuron: 1–16	Dendrite space constant (λ): 0.2–1 mm	<i>Excitatory synaptic input^d</i>
Dendritic tips/neuron: 10–400	Electrotonic length ($L = x/\lambda$): 0.2–2	AMPA: g_{peak} : 0.1–0.3 ns; t_{peak} : 0.3–1 ms (may increase with distance from soma)
Dend. spines/neuron: 300–200,000	Soma input resistance (R_{in}): $1\text{--}10^3$ M Ω	NMDA: g_{peak} : 0.05–0.5 ns; t_{peak} : 5–50 ms
Spine density/1 μm dendrite: 0.5–14	Input resistance at tips (R_T): $10^2\text{--}10^3$ M Ω	
Spine length: 0.1–2 μm	Steady-state attenuation factor:	<i>Inhibitory synaptic input^e</i>
Spine neck diameter: 0.04–0.5 μm	Soma \rightarrow tip: 1.1–2	GABA _A : g_{peak} : 0.4–1 ns; t_{peak} : 0.2–1.2 ms
Spine head diameter: 0.3–1 μm	Tip \rightarrow soma: 2–15	GABA _B : g_{peak} : 0.1–0.3 ns; t_{peak} : 40–150 ms
Spine volume: 0.005–0.3 μm^3		
	<i>Excitable properties of dendrites^b</i>	
	Ca ²⁺ channels (L, N, P type)—local dendritic Ca ²⁺ AP:	
	Ca ²⁺ concentration in spines	
	Na ⁺ channels: Fast activating/inactivating—supports soma \rightarrow dendritic backpropagating AP	
	K ⁺ channels, I _A , and mixed current, I _h —Increased density with distance from soma—“shock absorbers,” linearization, and temporal normalization.	

^aThe passive properties of dendrites can be strongly modulated by synaptic activity as well as by voltage-gated channels.

^bCharacterization of the voltage-gated ion channels in dendrites (i.e., type, distribution, density, and kinetics) is only beginning to emerge through the use of molecular probes and patch clamp techniques. For reviews, see Koch and Segev (1998, chap. 5) and Reyes (2001).

^cBased on data from cortical and hippocampal pyramidal neurons.

^dSee, e.g., Stern, P., Edwards, F. A., Sakmann, B., 1992, Fast and slow components of unitary EPSCs on stellate cells elicited by focal stimulation in slices of rat visual-cortex, *J. Physiol. (Lond.)*, 449:247–278.

^eSee, e.g., Otis, T. S., De Koninck, Y., and Mody, I., 1993, Characterization of synaptically elicited GABA_B responses using patch-clamp recordings in rat hippocampal slices, *J. Physiol. (Lond.)*, 463:391–407.

DENDRITIC SPINES (q.v.) are the major postsynaptic targets for *excitatory* synaptic inputs.

2. *Electrically*, dendrites can be characterized by their *passive* properties (the passive “skeleton”), to which the (nonlinear) synaptic- and voltage-dependent ion channels are added. The passive (near resting potential) skeleton can be characterized by the specific membrane resistivity, R_m , of dendrites, which is relatively high (R_m is on the order of 1,000–100,000 Ωcm^2), implying that the dendritic membrane is a good electrical insulator. With a specific capacitance, C_m , of approximately 1 $\mu\text{F}/\text{cm}^2$, the dendritic membrane time constant, τ_m (which sets the range for the time window for the integration of synaptic inputs), is on the order of $\tau_m = R_m C_m = 10\text{--}100$ ms. The axial (longitudinal) resistivity of the dendritic cytoplasm, R_i , ranges between 70 and 300 Ωcm , and this, together with the small dimensions of distal arbors, implies a large input resistance (impedance) in the dendrites. The increase in dendritic diameter as one approaches the soma implies a large attenuation factor (on the order of 100) of the peak synaptic potential as it spreads from its origin at the distal dendritic site to the soma (Figure 3A, B).

The *active* properties of dendrites are the outcome of the excitable channels embedded in their membrane. It has been known for years that some types of dendrites, most notably the cerebellar Purkinje cell, can even generate local dendritic AP. More recently, data have indicated that many central neurons can generate dendritic spikes under favorable conditions, and indeed, most of the different types of excitable ion channels in neurons are hosted by the dendrites. Dendrites of different neuron types (e.g., hippocampal CA1, layer V pyramids) bear different combinations of ion channel types, which to a large extent dictate the input-output style of these neurons. A large amount of new information about these ion channels has accumulated in the past few years (see, e.g., Koch and Segev, 1998; Reyes,

2001). These nonlinear ion channels have significant consequences for the computational capabilities of dendrites.

3. *Synaptically*, dendrites of central neurons are covered with synaptically activated (transmitter-gated) receptors. The AMPA and NMDA receptors for glutamate are typically associated with excitatory inputs, whereas GABA_A and GABA_B receptors are typically associated with inhibition. These receptors are not randomly distributed over the dendritic surface (Pettit and Augustine, 2000), and their density can be dynamically modified in an activity-dependent manner. In many, but not all, types of central neurons, the inhibitory-mediated receptors are more proximal than the excitatory synapses (see Shepherd, 1998).

Both the excitatory and the inhibitory synaptic inputs operate in most cases by locally increasing the conductance of the post-synaptic membrane (opening specific ion channels). Therefore, the synaptic input itself perturbs the electrical properties of dendrites, and thus the synaptic input is inherently nonlinear. The time course of the synaptic conductance change may vary by one or two orders of magnitude. The fast excitatory (AMPA) and inhibitory (GABA_A) inputs operate on a time scale of a few milliseconds and have a peak conductance on the order of 1 ns; this peak conductance is approximately ten times larger than the slow excitatory (NMDA) and inhibitory (GABA_B) inputs, which both act on a time scale of 10–100 ms.

Dendritic Modeling

In two groundbreaking studies, Rall (1959, 1964) established the theoretical foundation that allowed the morphological, electrical, and synaptic properties of dendrites to be linked together in a functionally meaningful framework (see Segev, Rinzel, and Shepherd, 1995). Rall’s passive cable theory for dendrites, complemented by his compartmental modeling approach, laid the groundwork for a

quantitative exploration of the integrative (input-output) function of dendrites.

Passive Cable Theory for Dendrites

Rall described current flow (and the spread of the resultant voltage) in morphologically and physiologically complicated passive dendritic trees using the one-dimensional cable equation

$$\frac{\partial^2 V}{\partial X^2} = \frac{\partial V}{\partial T} + V(X, T) \quad (1)$$

where V is the voltage across the membrane (interior minus exterior, relative to the resting potential); $X = x/\lambda$, where x is the distance along the core conductor (cm) and the *space constant*, λ , is defined as $\sqrt{r_m/r_i}$; and $T = t/\tau_m$, where the *time constant*, τ_m , is $r_m c_m$. Further, r_m is the membrane resistance for unit length (in Ωcm), c_m is the membrane capacitance per unit length (in F/cm), and r_i is the cytoplasm resistance per unit length (in Ω/cm). A complete derivation of the cable equation can be found in the chapter by Rall (in Koch and Segev, 1998, chap. 2; see also Jack, Noble, and Tsien, 1983, and PERSPECTIVE ON NEURON MODEL COMPLEXITY).

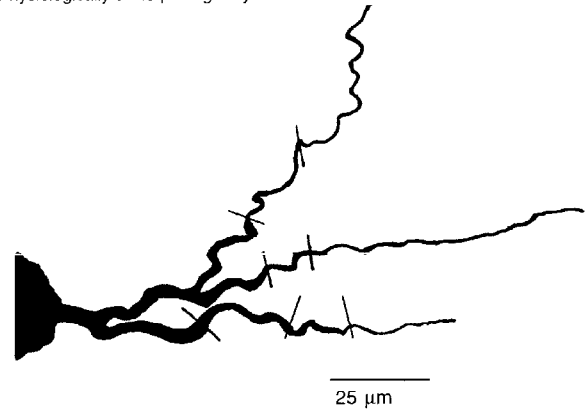
The solution to Equation 1 depends on, in addition to the electrical properties of the membrane and cytoplasm and dimensions of the dendritic tree, the boundary condition at the ends of the segment toward which the current flows. Rall showed that Equation 1 could be solved analytically for passive dendritic trees with arbitrary branching. He modeled the dendritic tree as a collection of short cylindrical segments (Figure 2B), where the tree attached to the end of each segment acts as a sink for the longitudinal current (i.e., a “leaky end”).

An example of such a solution for the steady-state case is depicted in Figure 3A and for the transient case in Figure 3B. Several important implications of Figure 3 are discussed later in this article and in Segev and London (2000). Recent extensions of the passive cable theory for dendrites have used the different moments of the transient synaptic potential to explicitly define the notion of input synchrony and propagation delay in dendrites (reviewed in Koch and Segev, 1998, chap. 2). The correspondence between cable theory and the Schrödinger eigenvalue problem for the one-dimensional bounded motion of a particle in quantum mechanics was also used to study the effect of spatially nonuniform membrane conductance on signal transfer in dendrites (see Stuart et al., 1999, chap. 9).

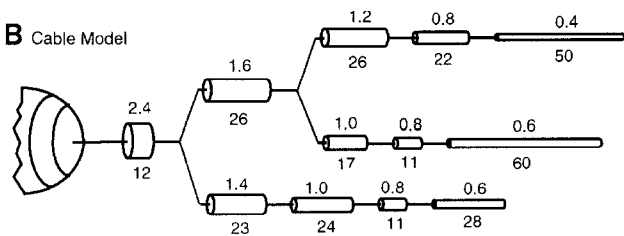
Compartmental Modeling Approach

The compartmental modeling approach complements cable theory by overcoming the assumption that the membrane is passive and that the input is a current source. Mathematically, the compartmental approach is a finite-difference (discrete) approximation to the (nonlinear) cable equation. It replaces the continuous cable equation (Equation 1) with a set, or a matrix, of ordinary differential equations; typically, numerical methods are employed to solve this system (which can include thousands of compartments and thus thousands of equations) for each time step. In the compartmental model, dendritic segments that are electrically short are assumed to be isopotential and are lumped into a single RC (circuit of resistors and capacitors) (either passive or active) membrane compartment (Figure 2C). Compartments are connected to each other via a longitudinal resistivity according to the topology of the tree. Hence, differences in physical properties (e.g., diameter, membrane properties) and differences in potential occur between compartments rather than within them. When the dendritic tree is

A Physiologically & Morphologically Characterized Neuron



B Cable Model



C Compartmental Model

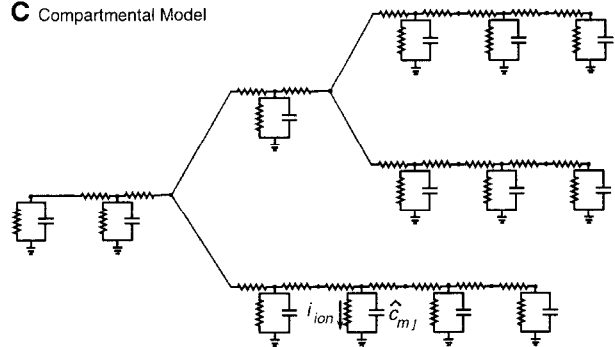


Figure 2. Dendrites (A) are modeled either as a set of cylindrical membrane cables (B) or as a set of discrete, isopotential RC compartments (C). In the cable representation (B), the voltage at any point in the tree is computed from Equation 1 and the appropriate boundary conditions are imposed by the tree. In the compartmental representation the tree is discretized into a set of interconnected RC compartments; each is a lumped representation of a sufficiently small dendritic segment. Membrane compartments are connected via axial, cytoplasmic resistances. Here the voltage can be computed at each compartment for any nonlinear input and voltage- and time-dependent membrane properties.

divided into sufficiently small segments (compartments), the solution to the compartmental model converges with that of the continuous cable model. Compartments can represent a patch of membrane with a variety of voltage-gated (excitable) and synaptic (time-varying) channels. Popular public domain computer programs have been developed to simulate compartmental modeling; the most notable are NEURON (NEURON SIMULATION ENVIRONMENT) and GENESIS (GENESIS SIMULATION SYSTEM). A review of the compartmental modeling approach can be found in Koch and Segev (1998, chap. 3).

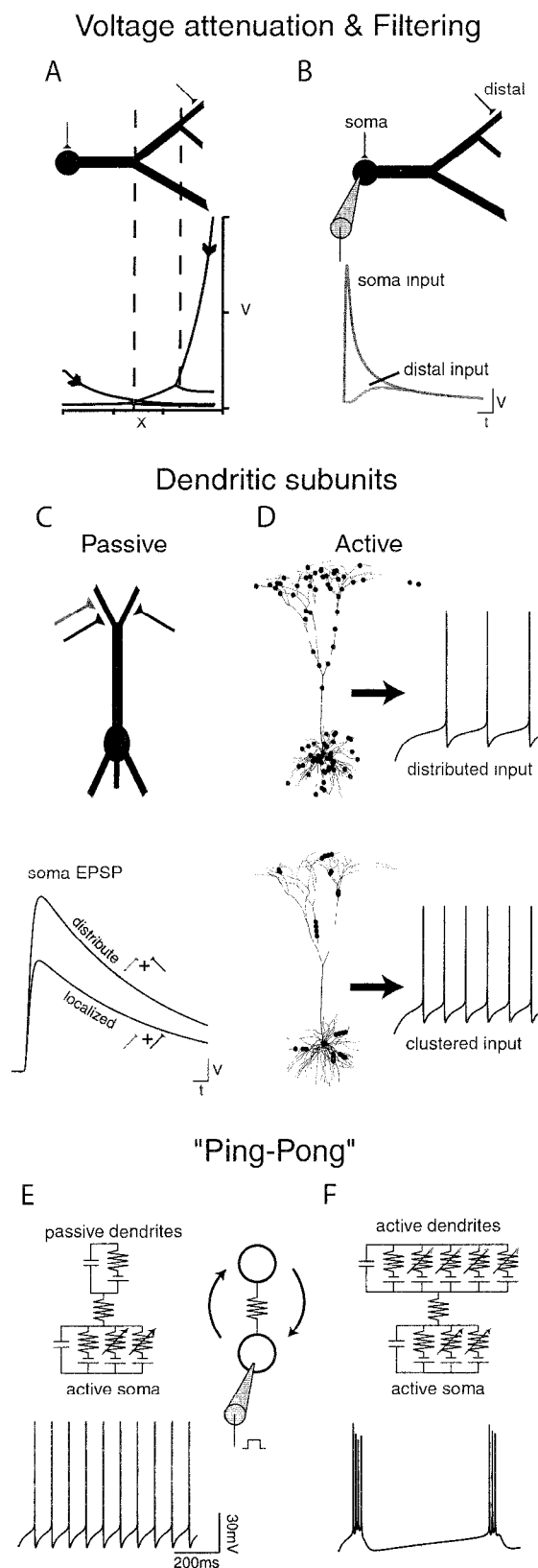


Figure 3. Effects of dendrites on synaptic inputs processing. *Top panel,* Passive dendrites impose voltage attenuation and low-pass filtering on their synaptic inputs. (A) Voltage response to a brief current pulse in a simple branched dendritic model (top). Attenuation of voltage peak is plotted for two cases, somatic input and dendritic input. The attenuation is asymmetric and is much steeper in the dendrite-to-soma direction. The voltage response at the input site is much larger for the distal dendritic input (large input impedance). (B) Voltage is transient at the soma for somatic input and dendritic input. The filtering effect of the dendrite gives rise to a temporal delay and to an increase in half-width of the distal dendritic input. *Middle panel,* The soma output depends on the degree of synaptic clustering on dendritic subunits. (C) In passive dendrites, sublinear summation of synaptic inputs is less pronounced (saturation is reduced) when the inputs are distributed in different dendritic arbors. (D) In excitable dendrites, a certain degree of spatial clustering of excitatory synapses (bottom) may result in a significant boosting of the synaptic charge that reaches the soma, because it produces larger local dendritic depolarization that may activate the local excitable channels. As a consequence, the axon fires more vigorously (right). One hundred excitatory synapses were used in both cases: *top*, ten clusters of ten synapses each; *bottom*, 100 clusters of one synapse each. Each synapse was activated 40 times per second for 1 s. *Bottom panel,* Backward-forward “Ping-Pong” interaction between the ion channels in the soma/axon region and the excitable channels in dendrites shapes the output pattern of spikes firing in the axon. Two models of a cortical pyramidal neuron were used, one with passive dendrites (E) and the other with excitable dendrites (F). For passive dendrites, the axon fires regularly in response to steady soma depolarization, whereas in the model with excitable dendrites, it fires repeated spike bursts. The geometry of the dendritic tree plays a crucial role in this “Ping-Pong” interaction.

Interaction Between Dendrites and Synapses: Main Insights

The theoretical background just outlined and the many results from modeling and experimental studies on dendrites over the last 40 years provide important insights into the input-output properties of dendrites. These properties can be summarized as follows:

Attenuation. Dendritic trees are electrically distributed (rather than isopotential) elements. Consequently, voltage gradients exist over the tree when synaptic inputs are activated locally. In passive dendrites, the voltage tends to attenuate much more severely in the dendrite-to-soma direction (up to a few hundredfold for brief synaptic inputs at distal sites) than in the opposite direction (Figure 3A). This attenuation implies that many (several tens) of excitatory inputs should be activated within the integration time window of about $1 \tau_m$, in order to build up enough depolarization to reach the threshold for spike firing at the soma and axon. Several experimental as well as theoretical studies (e.g., Reyes, 2001) have proposed that voltage-dependent currents might amplify or “boost” distal synaptic inputs while propagating toward the soma, but it is still an open question to what extent these mechanisms operate in the *in vivo* situation.

In contrast to the severe attenuation of the voltage *peak amplitude* in passive dendrites, the attenuation of the synaptic charge is relatively small. In many situations the charge is the relevant parameter for determining the firing rate at the axon, and thus, in these cases, the “cost” of placing the synapse at the dendrites rather than at the soma is quite small. Active dendritic currents can boost the synaptic charge, and this boosting is expected to be larger at distal dendritic sites, where the local input resistance (and local depolarization) is large. For example, synapses that are located on excitable dendritic spines (see DENDRITIC SPINES) can trigger regenerative activity that may spread and indirectly activate nearby dendritic regions (a “chain reaction” between active dendritic spines). Synaptic inputs that are mediated by NMDA receptors may implement a similar mechanism for amplification, in which a certain degree of spatial clustering of excitatory inputs is more likely to activate these NMDA receptors and enhance charge transfer to the soma. In this case, the output at the axon depends sensitively on the size and site of the “cluster” (Figure 3D).

Filtering. With passive (RC) properties, dendrites behave like low-pass filters for their synaptic inputs. As a result, synaptic potentials are delayed and become significantly broader as they spread away from the dendritic input site (Figure 3B). The large sink imposed by the tree at distal arbors enhances the decay of local synaptic potentials; smaller enhancement (and thus broader local potentials) is expected at more proximal input sites. The difference in the width of the synaptic potential at different parts of the dendritic tree implies multiple time windows for synaptic integration in the tree. At the soma, the time window for synaptic integration is primarily governed by τ_m , whereas at distal dendritic arbors it may be as short as $0.1 \tau_m$ or less (Koch and Segev, 1998, chap. 2, p. 65). The massive synaptic activity present *in vivo* greatly affects the apparent membrane conductance and effectively changes the properties of the dendritic filter (Koch and Segev, 1998, chap. 3, p. 125). Extensive background activity can reduce the effective membrane time constant by tenfold. The presence of voltage-activated currents in the dendrites also affects dendritic filtering. Slow currents might turn the dendrites into bandpass filters by suppressing very slow frequencies (Hutcheon and Yarom, 2000). These bandpass filters can be sharpened by the presence of amplifying currents, such as persistent sodium. Other currents might change the apparent capacitance of the filter (Reyes, 2001).

Dendritic inhibition—local veto operation. Inhibitory synapses (which are associated with a conductance change in series with a battery whose value is near the resting potential) are more effective when located on the path between the excitatory input and the “target” region (soma) than when placed distal to the excitatory input. Thus, when strategically placed, inhibitory inputs can specifically veto parts of the dendritic tree and not others (see Rall, 1964; Koch, Poggio, and Torre, 1982; Jack et al., 1983; Segev et al., 1995).

Spatiotemporal integration. Because of dendritic delay, the somatic depolarization that results from activation of excitatory inputs in the dendrites is very sensitive to the temporal sequence of the synaptic activation. It is largest when the synaptic activation starts at distal dendritic sites and progresses proximally. Activation of the same synapses in the reverse order in time will produce smaller somatic depolarization. Thus, the output of neurons with dendrites is inherently *directionally selective* (see PERSPECTIVE ON NEURON MODEL COMPLEXITY).

Gain control. Active dendritic currents (both inward and outward) may serve as a mechanism for synaptic gain control. The membrane potential dynamically controls active conductances in the neuron (e.g., its input resistance, electrotonic length); hence the neuron output depends on its state (its membrane potential). Active currents (e.g., outward K^+ current) can act to counterbalance excitatory synaptic inputs (negative feedback) and thus stabilize the input-output characteristics of the neuron. At other voltage regimes, active currents may effectively increase the input resistance and reduce the electrotonic distance between synapses (positive feedback), with the consequence of nonlinearly boosting a specific group of coactive excitatory synapses (Stuart et al., 1999, and references in Segev and London, 2000).

Ping-Pong interactions between synaptic inputs and dendritic spikes. The presence of dendritic spikes (both backpropagating Na^+ spikes as well as local Ca^{2+} spikes) in some neuron types may lead to interesting and nontrivial interactions in the tree. For example, a precise temporal coincidence between the backpropagating Na^+ spike and a distal synaptic input in the apical tree of cortical pyramids may trigger a broad dendritic Ca^{2+} spike that, in turn, leads to a burst of Na^+ spikes at the axon (Reyes, 2001). Thus, a burst of Na^+ spikes in the axon reflects the co-occurrence of input to a basal tree (which generates a backpropagating spike) and an excitatory input to the distal apical tree. Theoretical studies show that slow ion currents in the dendritic tree may interact with the fast spike-generating mechanism at the soma/axon hillock, and this “Ping-Pong” interaction can give rise to many complex spiking patterns in the axon, ranging from regular, high-frequency spiking to spike bursting (Figure 3E; for details, see Koch and Segev, 1998, chap. 5, p. 206).

Computational Function of Dendrites

Different brain areas are specialized in computing specific functions, and each of these areas consists of different types of dendritic structures. Do the unique morphology and electrical properties of the dendrites in a given brain area play a key role in implementing the computational task of this particular piece of brain? Unfortunately, in most cases it is hard to define the specific computational function the neuron executes. Furthermore, we rarely know the nature of the synaptic output that a particular neuron receives while the system computes. Therefore, current theoretical efforts are focusing on exploring the kinds of computations that neurons could potentially implement with their dendrites and synaptic inputs. Sev-

eral such computations were mentioned earlier; here the major ones are discussed more fully.

Selectivity for Direction of Motion

Because the depolarization in the axon is sensitive to the spatio-temporal order of synaptic activation over the dendritic surface, neurons with dendrites can compute the direction of motion (see Figure 4 and PERSPECTIVE ON NEURON MODEL COMPLEXITY; for details, see Segev et al., 1995; Koch and Segev, 1998). Already at the vertebrate retina some ganglion cells show directional selectivity such that their firing rate is significantly higher when a visual stimulus moves in one direction (preferred direction) rather than the opposite direction (null direction; Figure 4). Whether this computation is implemented postsynaptically by the dendrites of the ganglion cells or is already computed in neurons that are presynaptic to the ganglion cells is still under debate (see Taylor et al., 1999, and Borg-Graham, 2000—references in Euler and Denk, 2001). This debate highlights the need to record from dendrites during the processing of sensory input; recent developments in optical recording from dendrites are likely to provide a direct means of exploring the computational role of dendrites (Single and Borst, 1998; Euler and Denk, 2001).

Indeed, the most direct evidence for dendritic computation comes from work by Single and Borst (1998; see also reviews in Segev and London, 2000, and VISUAL COURSE CONTROL IN FLIES). This study explored the processing of visual information in the fly, in which a population of large interneurons, the tangential cells (TCs), spatially integrates the output signals of many thousands of columnar neurons, each of which is sensitive to a very small part of the visual scene. These TCs are all motion-sensitive: they are excited by motion in one direction and are inhibited by motion in the opposite direction. By combining intracellular recordings and calcium imaging from dendrites *in vivo*, two major processing steps that are implemented by the TC dendrites were identified. Through the processing of opponent input elements having opposite preferred directions, the directional selectivity of presynaptic neurons is significantly enhanced in the TCs. It was also shown that dendritic filtering helps distinguish a change in contrast due to stimulus motion from changes due to purely local patterns of the stimulus. The result of this integration is a graded depolarization in the axon of the tangential cells; this depolarization represents information about image velocity with high fidelity (Single and Borst, 1998).

Coincidence Detection

Neurons with dendrites can function simultaneously in multiple time windows. Distal arbors act more like coincidence detectors, whereas the soma acts more like an integrator when brief synaptic inputs (i.e., AMPA and GABA_A) are involved (Koch and Segev, 1998, chap. 2). Active dendrites provide an additional mechanism for coincidence detection based on local dendritic Na⁺ spikes (see SINGLE-CELL MODELS) or on the temporal coincidence between backpropagating Na⁺ spikes and local excitatory synaptic input, which gives rise to a broad local dendritic Ca²⁺ spike and a corresponding burst of Na⁺ spikes in the axon (see Reyes, 2001, and Figure 4). Ca²⁺ accumulation in dendritic spines can also serve as a detector for coincidence between back propagating Na⁺ spikes and excitatory input to the spine (see DENDRITIC SPINES).

The coincidence detector (CD) neurons in the auditory brainstem constitute a special case (Agmon-Snir, Carr, and Rinzel, 1998). These neurons possess bipolar dendrites, each of which receives excitatory synaptic inputs from only one ear. The neuron fires only if the inputs arriving from both ears occur within a very narrow time window (tens of microseconds); this is used by these cells to

detect interaural time differences, and therefore the location of the sound source. The problem is how to distinguish a strong input arriving from one ear and an input that arrives simultaneously from both ears. This task cannot be achieved by a neuron that sums its inputs linearly. Agmon-Snir et al. (1998) have shown, using a simplified model of CD neurons, that a strong input arriving at one dendrite (from one ear) cannot drive the cell to fire, because of synaptic saturation, whereas inputs that reach both dendrites summate more linearly with each other at the soma and can trigger an output spike. Thus, segregation of the inputs from each ear to different dendrites, together with the inherent nonlinearity of synaptic summation, improves sound localization.

Feature Extraction, Input Classification, and Logical Operations

Neurons with dendrites can implement a multidimensional classification task where the neurons' output is sensitive to a specific combination of active synapses over the dendritic tree (Stuart et al., 1999, chap. 11). For example, biophysical modeling of cortical pyramidal neurons shows that, with appropriate mapping of synaptic inputs from the lateral geniculate body (LGN) to the dendritic tree, and using local dendritic nonlinearities (mediated via NMDA receptors or excitable channels), these model neurons exhibit several nonlinear features of visual neurons, including phase invariance orientation tuning (a critical feature of complex cells), binocular disparity, and nonlinear boosting of tuning curves (Figure 4).

AND-NOT type of logical operations could be implemented by strategically placed excitatory and inhibitory synaptic inputs over the dendritic tree (Koch et al., 1982, and Figure 4). Because neurons with dendrites can function as many quasi-independent functional subunits, the dendritic tree as a whole can implement a rich repertoire of logical operations. Also, within each of these functional subunits, very localized plastic processes can take place (e.g., on a single dendritic spine; see DENDRITIC SPINES).

Dendritic Learning

Over the past few years, it has become evident that dendrites are not static elements. Rather, there is a continuous change in the morphology of dendritic spines (Segev and London, 2000) and a dynamic change in the distribution of ion channels and synapses on the dendritic surface. Two recent computational studies suggest that these plastic properties of dendrites might play a role in learning. Stemmler and Koch (1999) demonstrated that by changing the distribution of membrane ion channels in a dendritic compartment via a learning rule, the neurons' output can optimize the representation of sensory information. The segregation of ion channels between the dendritic compartment and the soma axon compartment is critical for this mechanism to work. Poirazi and Mel (see DENDRITIC LEARNING) have shown that by spatially reallocating (reshuffling) synapses that are temporally correlated so that they will contact nearby dendritic locations, active dendritic subunits are created whereby the coactivation of synapses within a subunit results in significant synaptic boosting by local nonlinear dendritic mechanisms. This spatial clustering of temporally correlated inputs over the dendritic tree greatly increases the input classification capacity of the neuron.

Discussion

Nature apparently uses dendrites as the basic building blocks for a wide range of nervous systems, from the simplest organism to the most sophisticated one (us?). Dendrites, therefore, need to provide

Computation	Implementation (Biophysical mechanism)	Example
Direction selectivity	<ol style="list-style-type: none"> 1. Dendritic delay + temporal sequence of synaptic activation 2. Asymmetric mapping of inhibition and excitation 3. Integration of many local direction selective detectors 	
Coincidence detection	<ol style="list-style-type: none"> 1. Backpropagating Na spike + Dendritic Ca spike 2. Input segregation on dendrites + synaptic saturation 3. Backpropagating Na spike + NMDA (Ca in spines) 4. Local dendritic Na spike generated by precisely timed excitation 5. Short local delays in thin dendrites 	
Logical operation	<ol style="list-style-type: none"> 1. "On path" dendritic inhibition + distal excitation (AND-NOT) 2. Local spike in dendritic spines (AND, OR) 	
Feature extraction	<ol style="list-style-type: none"> 1. Mapping synaptic inputs to distinct nonlinear dendritic subunits 	

Figure 4. Summary of dendritic computation.

a good solution for the real-life problems that behaving animals must deal with, and the challenge is to explore how the intricacies of dendrites provide such credible solutions. In principle, dendrites could optimize several constraints simultaneously, from the wiring requirements of neural networks to chemical and electrical constraints imposed by the nervous system.

In terms of brain connectivity, dendrites enable the spatial segregation of different types of inputs at different dendritic regions, thus providing a natural means of preserving strong local interactions among specific types of input sources and, at the same time, integrating different input sources onto the same postsynaptic neuron.

The electrical distributed structure of dendrites can give rise to functional subtrees that are partially electrically decoupled from each other. In each of these subtrees, local operations could take place, including local synaptic boosting, local dendritic spiking, local dampening of the backpropagating spike, and so forth. The spatiotemporal (backward-forward) interaction between the different dendritic subtrees could result in a wide range of firing repertoires that would be hard to replicate with spherical (isopotential) neurons. The dendritic tree also behaves like a sink for the excitable channels in the axon; thus it tends to stabilize network activity.

From a chemical perspective, dendrites with their fine diameters and their specialization, such as dendritic spines, allow rapid and very specific activity-dependent local changes to take place in intracellular ion concentrations (e.g., of Ca^{2+} ions). This chemical compartmentalization provides the means for triggering local plastic processes (local “synaptic learning,” local morphological changes in dendritic spines). Moreover, the maintenance of ion gradients between the two sides of the dendritic membrane consumes energy. In this context, dendrites may optimize the ratio between membrane area (required for connecting a large number of synapses onto a single neuron) and cell volume, thus reducing energy consumption.

Whether or not dendrites contribute to network computation should be assessed within a wider framework, namely, in terms of the effect of the complexity of the computation executed by a single neuron on the network function. Artificial neural networks typically assume that neurons are simple computational units and that the computation that the nervous system performs emerges as a collective phenomenon. In this framework, the computation at the single neuron level (e.g., orientation selectivity) merely reflects network dynamics. However, as was shown in this article, real neurons can potentially do much more than integrate their input and compare it to a fixed threshold. Moreover, in some cases (mainly in invertebrates), it is clear that single identified neurons actually perform many of the computations required for accomplishing a specific behavioral task.

Thus, the challenge to neural modelers is to better understand whether, and how, the complexity of the single neuron significantly

enhances the computational capacity of the network. Once theoretical studies are able to proceed hand in hand with direct measurements obtained from dendrites during actual behavior (and while their nervous system is computing), we will gain a better understanding of the computational role of dendrites. At this point the “dendritic revolution” will come to a close.

Road Map: Biological Neurons and Synapses; Grounding Models of Networks

Background: Single-Cell Models

Related Reading: Dendritic Learning; Dendritic Spines; Perspective on Neuron Model Complexity

References

- Agmon-Snir, H., Carr, C. E., and Rinzel, J., 1998, The role of dendrites in auditory coincidence detection, *Nature*, 393:268–272.
- Crick, F., 1999, The impact of molecular biology on neuroscience, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 354:2021–2025.
- Euler, T., and Denk, W., 2001, Dendritic processing, *Curr. Opin. Neurobiol.*, 11:415–422. ♦
- Hutcheon, B., and Yarom, Y., 2000, Resonance, oscillation and the intrinsic frequency preferences of neurons, *Trends Neurosci.*, 23:216–222.
- Jack, J. J. B., Noble, D., and Tsien, R. W., 1983, *Electrical Current Flow in Excitable Cells*, Oxford, Engl.: Oxford University Press.
- Johnston, D., Magee, J. C., Colbert, C. M., and Cristie, B. R., 1996, Active properties of neuronal dendrites, *Annu. Rev. Neurosci.*, 19:165–186.
- Koch, C., Poggio, T., and Torre, V., 1982, Retinal ganglion cells: A functional interpretation of dendritic morphology, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 298:227–263.
- Koch, C., and Segev, I., Eds., 1998, *Methods in Neuronal Modeling: From Ions to Networks*, Cambridge, MA: MIT Press. ♦
- Pettit, D. L., and Augustine, G. J., 2000, Distribution of functional glutamate and GABA receptors on hippocampal pyramidal cells and interneurons, *J. Neurophysiol.*, 84:28–38.
- Rall, W., 1959, Branching dendritic trees and motoneuron membrane resistivity, *Exp. Neurol.*, 2:503–532.
- Rall, W., 1964, Theoretical significance of dendritic trees for neuronal input-output relations, in *Neural Theory and Modeling* (R. F. Reiss, Ed.), Stanford, CA: Stanford University Press.
- Reyes, A., 2001, Influence of dendritic conductances on the input-output properties of neurons, *Annu. Rev. Neurosci.*, 24:653–675. ♦
- Segev, I., and London, M., 2000, Untangling dendrites with quantitative models, *Science*, 290:744–750. ♦
- Segev, I., Rinzel, J., and Shepherd, G., Eds., 1995, *The Theoretical Foundation of Dendritic Function*, Cambridge, MA: MIT Press.
- Shepherd, G. M., Ed., 1998, *The Synaptic Organization of the Brain*, New York: Oxford University Press. ♦
- Single, S., and Borst, A., 1998, Dendritic integration and its role in computing image velocity, *Science*, 281:1848–1850.
- Stemmler, M., and Koch, C., 1999, How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate, *Nat. Neurosci.*, 2:521–527.
- Stuart, G., Spruston, N., and Häusser, M., Eds., 1999, *Dendrites*, New York: Oxford University Press. ♦

Dendritic Spines

William R. Holmes and Wilfrid Rall

Introduction

The function of dendritic spines has been debated ever since they were discovered by Ramón y Cajal in 1891. Although it was widely believed that spines were important for intercellular communication, this was not demonstrated until 1959 when Gray, using the

electron microscope, showed that synapses are present on spines. Why should synapses exist on spines, rather than on dendrites? What role does spine morphology play in their function? Early theoretical studies of these questions focused on the electrical resistance provided by the thin spine stem and suggested that changes in stem diameter might be important for synaptic plasticity. Later

investigations found that if voltage-dependent conductances were present on spines, then spines could increase the computational possibilities of a cell. Recent studies suggest that spines are isolated compartments in which highly localized calcium signals can occur. The amplitude and time course of these calcium signals can initiate localized cascades of biochemical reactions important for synaptic plasticity. (For reviews, see Harris, 1999; Nimchinsky, Sabatini, and Svoboda, 2002; and Sabatini, Maravall, and Svoboda, 2001.)

Spine Morphology

Dendritic spines are short, appendage-like structures found on many different cell types. Spines are composed of a bulbous "head" connected to the dendrite by a thin "neck" or "stem." An excitatory synapse is usually found on the spine head, and some spines also have a second, usually inhibitory, synapse located near or on the spine neck. Spines typically are small in size, but because they occur in densities of 1–2 spines/ μm or more, spine membrane area can comprise 40–60% of the total neuron membrane area.

Attempts have been made to classify spines based on their size, shape, and dendritic location. Jones and Powell (1969) categorized spines as sessile (stemless) or pedunculated (having a peduncle, or stem), with sessile spines more common in proximal regions and pedunculated spines in distal regions. Peters and Kaiserman-Abramof (1970) classified spines as (1) stubby, (2) mushroom shaped, or (3) thin or long-thin (Figure 1). Stubby spines were most numerous in proximal regions, long-thin spines dominated distal regions, and mushroom-shaped spines were distributed almost uniformly. These categories are arbitrary since spine shape varies con-

tinuously and all types of spines are found in all areas. In some brain regions, it has not been possible to categorize spines in any systematic manner, but categories and a range of dimensions are useful for models.

Passive Models of Spine Function

Models in which the spine is represented as a passive electric circuit show that the resistance of a thin spine stem can attenuate a synaptic input delivered to the spine head. This can be seen by considering the circuit pictured in Figure 2.

Assuming a constant synaptic conductance, currents flowing in the spine head can be described by Kirchhoff's law as:

$$V_{SH}/R_{SH} + g_{syn}(V_{SH} - V_{EQ}) + V_{SH}/(R_{SS} + R_{BI}) = 0 \quad (1)$$

where the first term is leakage current across the spine head membrane, the second is the synaptic current, and the third is the flow of current through the spine stem to ground. Because R_{SH} is large, the first term is negligible compared to the other two terms and can be ignored. With this simplification, the steady-state spine head potential is

$$V_{SH} = V_{EQ}/[1 + 1/(g_{syn}(R_{SS} + R_{BI}))] \quad (2)$$

Similarly, the currents in the dendrite are described as

$$V_{BI}/R_{BI} + (V_{BI} - V_{SH})/R_{SS} = 0 \quad (3)$$

which can be rearranged as

$$V_{BI} = V_{SH}R_{BI}/(R_{BI} + R_{SS}) \quad (4)$$

Combining Equations 2 and 4, the voltage at the dendrite is

$$V_{BI} = V_{EQ}/[1 + 1/(g_{syn}R_{BI}) + R_{SS}/R_{BI}] \quad (5)$$

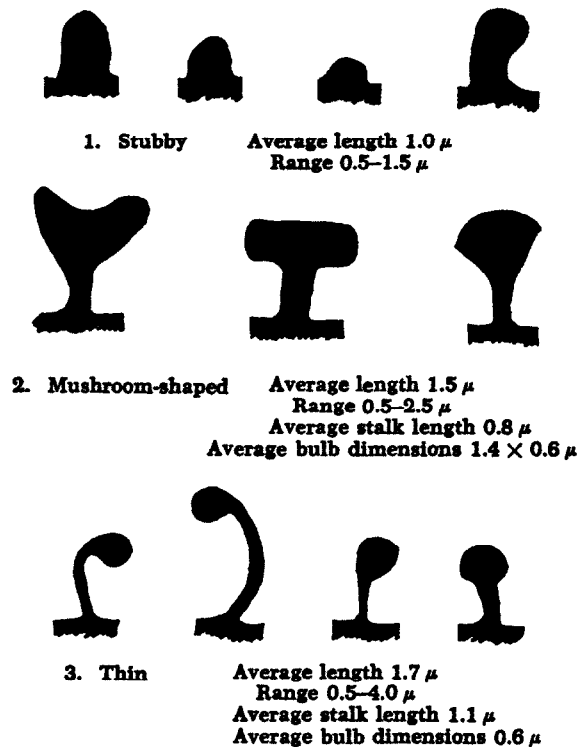


Figure 1. Variety of spine shapes in parietal cortex. (Adapted from Peters and Kaiserman-Abramof, 1970, Table 1. Copyright © 1970 by the Wistar Institute of Anatomy and Biology. Used by permission of John Wiley & Sons, Inc.)

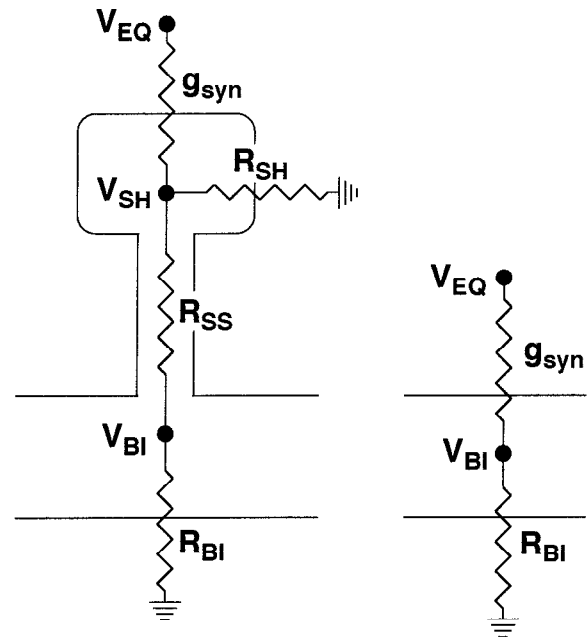


Figure 2. Electrical circuit of a dendritic spine. V_{EQ} is the synaptic reversal potential, V_{SH} is the voltage in the spine head, and V_{BI} is the voltage in the dendrite at the base of the spine. R_{SH} is the spine head resistance, R_{SS} is the spine stem resistance, and R_{BI} is the branch input resistance at the base of the spine. g_{syn} is the synaptic conductance. The corresponding circuit for input on a dendrite is shown on the right.

If, however, the synapse were on the dendrite, Kirchhoff's current law says that

$$g_{\text{syn}}(V_{\text{BI}} - V_{\text{EQ}}) + V_{\text{BI}}/R_{\text{BI}} = 0 \quad (6)$$

Rearranging, we have

$$V_{\text{BI}} = V_{\text{EQ}}/[1 + 1/(g_{\text{syn}}R_{\text{BI}})] \quad (7)$$

The only difference between Equations 5 and 7 is the presence of the ratio $R_{\text{SS}}/R_{\text{BI}}$ in the denominator of Equation 5, and this term accounts for voltage attenuation when the synapse is on the spine instead of the dendrite.

Spines also attenuate synaptic current. The synaptic currents for spine and dendritic inputs can be computed by substituting expressions for V_{SH} and V_{BI} in Equations 2 and 7 in place of V in $g_{\text{syn}}(V - V_{\text{EQ}})$. The resulting currents are given by the right side of Equations 5 and 7 divided by R_{BI} . The size of the synaptic current entering the spine is attenuated because, for identical g_{syn} , the voltage in the spine head owing to input there is closer to V_{EQ} than the voltage change with a dendritic input.

The $R_{\text{SS}}/R_{\text{BI}}$ ratio in these equations suggests another possible function for dendritic spines. If a neuron can adjust spine stem morphology (and hence R_{SS}), then spines provide a mechanism to allow synaptic weights to be modified (Rall, 1978). For this mechanism to be important, the value of $R_{\text{SS}}/R_{\text{BI}}$ should lie between 0.1 and 10 times $1 + 1/(g_{\text{syn}}R_{\text{BI}})$, because within this range a small change in $R_{\text{SS}}/R_{\text{BI}}$ can have a significant effect on V_{BI} (see Equation 5). Early experimental estimates of $R_{\text{SS}}/R_{\text{BI}}$ suggested that it might lie in this effective operating range. This encouraged investigators to look for spine dimension changes in various experimental situations. However, recent estimates of g_{syn} , R_{SS} , and R_{BI} suggest that $R_{\text{SS}}/R_{\text{BI}}$ may not fall in this range. For example, if $g_{\text{syn}} = 0.5$ nS and $R_{\text{BI}} = 200$ M Ω , then $1/(g_{\text{syn}}R_{\text{BI}}) = 10$. This means that R_{SS} should be 220–2,200 M Ω ; morphological measurements and imaging experiments indicate, however, that R_{SS} is 5–150 M Ω in many neuron types (Svoboda, Tank, and Denk, 1996). This would seem to rule out the possibility that the function of spines can be explained by changes in passive electrical properties caused by changes in spine dimensions.

Models of Excitable Spines

If voltage-dependent conductances exist on spines, then spines might exist to amplify synaptic inputs. Early theoretical studies showed that postsynaptic potential and charge transfer were five- to ten-fold larger for synaptic input on spines with voltage-dependent conductances than for input to passive spines. Subsequent studies found that interactions among excitable spines could create a number of interesting computational possibilities for information transfer (reviewed in Segev and Rall, 1998). Chain reactions of spine head action potentials, spreading a certain distance proximally and distally, were theoretically possible. Considerable amplification of the initial input could occur even if only a small percentage of spines possessed voltage-dependent channels. Alternatively, subthreshold depolarization could inactivate channels and prevent amplification of later inputs.

Although the prediction that spines exist to amplify synaptic input has received some limited experimental support, the level of amplification is likely to be much less than predicted by the models for two reasons. First, the models assume large densities of voltage-dependent ionic channels in spines (~500–1000 channels per spine) while recent estimates place the number of calcium channels in spines at 1–20 (Sabatini and Svoboda, 2000). Second, the models require a large R_{SS} to get the interesting interaction effects. The latest models reduce the required R_{SS} value to 95 M Ω by assuming kinetics for a low threshold calcium action potential in the spine

head, but this value is still at the high end of the 5–150 M Ω range quoted earlier.

Models of Calcium Diffusion in Spines

The theoretical studies described above searched for a function for spines that depended on the electrical resistance of the spine neck. Besides providing an electrical resistance to current flow, the thin spine neck provides a *diffusional* resistance to the flow of ions and molecules. The spine neck, by restricting the flow of materials out of the spine head, might effectively isolate the spine head and provide a localized environment where reactions specific to a particular synapse can occur.

Calcium is a prime candidate for a substance that might be selectively concentrated in the spine head. Calcium is important for a large number of metabolic processes and has been shown to be necessary for the induction of long-term potentiation (LTP), but high concentrations of calcium can lead to cell death. Spines might provide isolated locations where high concentrations of calcium can be attained safely without disrupting other aspects of cell function.

Compartmental models of dendritic spines that included calcium diffusion, buffering, and extrusion tested this idea. The models predicted that calcium influx at the spine head could produce transient spine head calcium concentrations greater than 10 μ M, while an equivalent influx at the dendrite would produce only a small concentration change. Spine head calcium concentration changes greater than 10 μ M were subsequently observed experimentally.

Large increases in calcium concentration occur because of the small volume of the spine head and because incoming calcium cannot be buffered or pumped out instantaneously. These large transient increases are restricted to the spine head because of the diffusional resistance of the spine neck. The thin spine neck acts as a constriction that slows calcium diffusion. Any calcium that does enter the spine neck is further hindered from diffusing to the dendrite by the presence of calcium buffer and pumps in the spine neck. Calcium that diffuses through the spine neck to the dendrite has little effect on dendritic calcium concentration because of the large dendritic volume.

Spines Allow Coincidence Detection

Amplification of spine head calcium concentration can occur with backpropagating action potentials, allowing spines to detect temporal coincidence of pre- and postsynaptic activity. Although calcium can enter spines via voltage-gated calcium channels or be released from internal stores, in most cells, most spine head calcium enters through NMDA receptor channels. These ligand-gated channels are subject to voltage-dependent magnesium block. Theoretical studies predict a sharp increase in spine head calcium influx when synaptic input occurs slightly before action potential invasion of spines. This occurs because the voltage boost provided by the action potential relieves magnesium block of the NMDA receptor channels. Increases in spine head calcium concentration with coincident pre- and postsynaptic activity has been observed experimentally in a number of labs (e.g., Yuste et al., 1999).

Modeling Calcium-Induced Plasticity in Spines

Spine head calcium concentration changes are important because synaptic plasticity involves cascades of biochemical reactions that are activated or deactivated to different degrees depending on temporal characteristics of the calcium signal. For example, large increases in spine head calcium concentration allow calcium to bind to calmodulin and the calcium-calmodulin complex then activates

a number of kinases including calcium-calmodulin-dependent protein kinase II (CaMKII). CaMKII constitutes 20% of the total protein in the postsynaptic density and is thought to play a key role in LTP induction. Recent spine models include predictions of levels of CaMKII activation in the spine head for various input conditions (Holmes, 2000).

The CaMKII pathway is just one of many pathways affected by calcium. Future spine models will study how dozens of different molecules and proteins interact, form signaling complexes, and produce LTP. This is a daunting task, but progress has been made defining network pathways and compiling the necessary rate constants.

Factors Affecting Spine Calcium Dynamics

A number of factors affect the dynamics of the spine head calcium transient. The most important of these are buffer concentration, the magnitude and duration of the calcium current, and spine shape. Buffer concentration and calcium currents in spine heads have been difficult to quantify experimentally, but the range of spine shapes is known for several neuron types. Simulations suggest that spines with fat spine necks are not able to concentrate calcium as quickly or to the same high levels as spines with long-thin necks. Thus, spine shape changes that occur with LTP (Yuste and Bonhoeffer, 2001) are likely to affect the rates and types of calcium-dependent reactions that can occur in a particular spine. During development spines are particularly motile and experimental advances now allow these spine morphology changes to be observed in real time (Matus, 2000). These motility changes can continuously alter calcium decay kinetics in spines.

Discussion

Early theoretical work with passive spine models showed the importance of the electrical spine stem resistance for synaptic transmission and demonstrated how synaptic weights might be modified by changes in this resistance. Later modeling showed that input could be amplified or transformed in a variety of interesting ways if spines have excitable membrane. However, experimental measurements suggest that spine stem *electrical* resistance is too small to play a significant role in electrical signaling in all but a small percentage of spines.

The current hypothesis is that spines, by restricting diffusion of substances away from the synapse, provide a safe, local, and isolated environment in which specific biochemical reactions can occur. In particular, the spine stem provides a *diffusional* resistance that allows calcium to become concentrated in the spine head and

calcium-dependent reactions to be localized to the synapse. This could be very important for plasticity changes, such as those that occur with long-term potentiation. Spine morphology may determine the magnitude of the diffusional resistance and play a role in determining, or restricting, the types of biochemical reactions that can take place at a synapse.

Road Map: Biological Neurons and Synapses

Related Reading: Dendritic Processing; Hebbian Synaptic Plasticity; Ion Channels: Keys to Neuronal Specialization; Neocortex: Chemical and Electrical Synapses

References

- Gray, E. G., 1959, Axo-somatic and axo-dendritic synapses of the cerebral cortex: An electron-microscopic study, *J. Anat.*, 93:420–433.
- Harris, K. M., 1999, Structure, development, and plasticity of dendritic spines, *Curr. Opin. Neurobiol.*, 9:343–348. ♦
- Holmes, W. R., 2000, Models of calmodulin trapping and CaM kinase II activation in a dendritic spine, *J. Computat. Neurosci.*, 8:65–86.
- Jones, E. G., and Powell, T. P. S., 1969, Morphological variations in the dendritic spines of the neocortex, *J. Cell. Sci.*, 5:509–529.
- Matus, A., 2000, Actin-based plasticity in dendritic spines, *Science*, 290:754–758.
- Nimchinsky, E. A., Sabatini, B. L., and Svoboda, K., 2002, Structure and function of dendritic spines, *Annu. Rev. Physiol.*, 64:313–353. ♦
- Peters, A., and Kaiserman-Abramof, I. R., 1970, The small pyramidal neuron of the rat cerebral cortex: The perikaryon, dendrites and spines, *Am. J. Anat.*, 127:321–356.
- Rall, W., 1978, Dendritic spines and synaptic potency, in *Studies in Neurophysiology* (R. Porter, Ed.), New York: Cambridge University Press, pp. 203–209.
- Ramón y Cajal, S., 1891, Sur la structure de l'écorce cerebrale de quelques mammifères, *Cellule*, 7:124–176.
- Sabatini, B. L., Maravall, M., and Svoboda, K., 2001, Ca^{2+} signaling in dendritic spines, *Curr. Opin. Neurobiol.*, 11:349–356. ♦
- Sabatini, B. L., and Svoboda, K., 2000, The number and properties of calcium channels in single dendritic spines determined by optical fluctuation analysis, *Nature*, 408:589–593.
- Segev, I., and Rall, W., 1998, Excitable dendrites and spines: Earlier theoretical insights elucidate recent direct observations, *Trends Neurosci.*, 21:453–460. ♦
- Svoboda, K., Tank, D. W., and Denk, W., 1996, Direct measurement of coupling between dendritic spines and shafts, *Science*, 272:716–719.
- Yuste, R., and Bonhoeffer, T., 2001, Morphological changes in dendritic spines associated with long-term synaptic plasticity, *Annu. Rev. Neurosci.*, 24:1071–1089. ♦
- Yuste, R., Majewska, A., Cash, S. S., and Denk, W., 1999, Mechanisms of calcium influx into hippocampal spines: Heterogeneity among spines, coincidence detection by NMDA receptors, and optical quantal analysis, *J. Neurosci.*, 19:1976–1987.

Development of Retinotectal Maps

Geoffrey J. Goodhill

Introduction

A common feature of many axonal projections between different regions of the nervous system is their organization into topographic maps, whereby nearby cells in the input structure project to nearby cells in the output structure. How do such maps form? The best studied example of this phenomenon is the formation of maps from the retina to more central targets. In frogs, fishes, and chicks, the main visual center is the optic tectum. During development, fibers

grow from each retina to the opposite tectum, crossing at the optic chiasm, to form a “retinotopic” map. This map is oriented such that the nasal-temporal and dorsal-ventral axes of each retina map to the caudal-rostral and lateral-medial axes of each tectum, respectively (Figure 1A). In mice, there are topographic projections from the retina to both the superior colliculus (SC) and the lateral geniculate nucleus (LGN). The SC receives connections only from the contralateral eye, as in the retinotectal projection in frogs, fishes, and chicks, whereas the LGN receives in addition a small

	Perturbation	Outcome
a	Normal	
b	Tectal rotation	
c		
d	Retinal rotation	
e	Expansion	
f	Compression	
g	Mismatch	
h	Compound eye	
i	Tectal transplant	
j		

Figure 1. Summary of the outcomes of surgical manipulation experiments in the retinotectal system. The circle represents the retina and the ellipse represents the tectum. N, nasal; T, temporal; R, rostral; C, caudal. See text for more details.

ipsilateral projection. For many purposes, this system can be considered as a sheet of retinal ganglion cells connected to a sheet of tectal/SC/LGN cells by a bundle of ganglion cell fibers. Although there is wide variation among species in the degree of order existing in the optic nerve, it is almost always the case that the final map in the tectum is ordered to a greater extent than is the optic nerve (reviewed in Udin and Fawcett, 1988). The development of such maps appears to proceed in two stages: the first involves axon guidance independent of neural activity, while the second involves the refinement of initially crude patterns of connections by processes dependent on neural activity. This article focuses on the former events (see also AXONAL PATH FINDING); for a discussion of the latter see SELF-ORGANIZING FEATURE MAPS and OCULAR DOMINANCE AND ORIENTATION COLUMNS.

Experimental Data

The Chemospecificity Hypothesis

The main hypothesis driving experimental work in this area has been the idea of “chemospecificity,” proposed by Sperry in 1963:

The establishment and maintenance of synaptic associations [is] conceived to be regulated by highly specific cytochemical affinities that arise systematically among the different types of neurons involved via self-differentiation, induction through terminal contacts, and embryonic gradient effects. . . . [I propose] an orderly cytochemical mapping in terms of two or more gradients of embryonic differentiation that spread across and through each other with their axes roughly perpendicular. These separate gradients successively superimposed on the retinal and tectal fields and surroundings would stamp each cell with its appropriate latitude and longitude expressed in a kind of chemical code with matching values between the retinal and tectal maps (p. 707).

The most obvious test of this hypothesis is to uncover the molecular identity of these gradients and show that the pattern of these molecules is crucial for proper map formation. However, such a molecular approach has borne fruit only very recently (see below). Instead, much research immediately following Sperry until the 1980s adopted the approach, more rooted in classical embryology, of disrupting the normal development (or, more typically, regeneration) of this system and comparing the outcome with the predictions of Sperry’s hypothesis. These experiments are summarized in the next section and in Figure 1. For more detailed discussions, including original references, see Udin and Fawcett (1988), Holt and Harris (1993), and Goodhill and Richards (1999).

Surgical Manipulations

Shifting connections. In fishes and frogs, the retina expands radially during development, while the tectum expands mostly along one dimension. The retinotectal map remains ordered throughout this time, indicating that the retinotectal projection is continually shifting.

Ectopic targeting. Retinal axons entering the tectum via abnormal trajectories can still find their appropriate termination sites.

Rotation. If in *Xenopus* a presumptive tectum is rotated early enough during development, a map is formed that is normal relative to the whole animal (Figure 1B), whereas later rotations lead to a rotated map (Figure 1C). Initially it was thought that eye rotation could also lead to both a normal outcome if performed early enough and a rotated outcome if performed later. However, more recent experiments have always found rotated maps (Figure 1D).

Retinal ablation (“expansion”). The map formed after removal of half of the retina initially covers half the tectum, but then gradually expands to fill the whole tectum (Figure 1E). Axon terminal density remains the same. If the optic nerve is then made to regenerate again, an expanded map is immediately formed.

Tectal ablation (“compression”). If half of the tectum is ablated, the regenerated map is compressed into the remaining tectal space (Figure 1F). If “mismatched” halves of the retina and tectum are ablated, a topographic map still forms (Figure 1G).

Compound eye experiments. When a whole eye is created by fusing together two half-eye rudiments before connections are made, with the two halves being from opposite eyes but of the same type

(i.e., nasal, ventral, or temporal), they each map across the whole tectum in the mirror image of each other (Figure 1H). When fragments smaller than half a retina are substituted early in development ("pie-slice" eyes), the retinal fragments map appropriately for their original position, although they can also show some degree of reprogramming.

Translocation. If two parts of the tectum are reciprocally translocated, regenerating retinal axons innervate their normal piece of tectum, and also appropriately reverse their order if the tectal fragment is rotated (Figure 1I). However, in some cases a map can be formed that ignores the translocation; i.e., the fibers tend to align with fibers in the surrounding tectum, regardless of the orientation of the transplant (Figure 1J).

Branching. In frogs and fishes, retinal axons grow to their final termination zones directly. In chicks and rodents, however, retinal axons grow past their final termination zone initially, form axon collaterals along the whole axon shaft (but preferentially in their topographically correct region), and then select their specific topographic termination zone by stabilizing an axon collateral while the distal part of the axon is pruned back.

Direct Evidence for Molecular Gradients

A major breakthrough in the understanding of the molecular basis of retinotectal map formation came in the mid-1990s, with discoveries centering on the erythropoietin-producing hepatocellular (Eph) family of receptor tyrosine kinases and their associated ligands, the ephrins (reviewed in Flanagan and Vanderhaeghen, 1998, and O'Leary, Yates, and McLaughlin, 1999). The Eph/ephrins come in two families, A and B, with promiscuous binding within a family but little affinity between families. It now appears that the A family is important for mapping along the rostral-caudal axis of the tectum, while the B family may be important for mapping along the dorsal-ventral axis.

In chick retina (and on the axons of retinal ganglion cells), Eph A3 is expressed in an increasing nasal to temporal gradient, while Eph A4 and A5 are uniformly expressed (Figure 2). In chick tectum, levels of both ephrin-A2 and ephrin-A5 rise from rostral to caudal, with the latter being restricted to more caudal locations and rising more steeply. In mouse retina, Eph A4 and Eph A5 are expressed, with Eph A4 being expressed uniformly and Eph A5 expressed in an increasing nasal to temporal gradient. In mouse SC, levels of ephrin-A5 rise from rostral to caudal, while levels of ephrin-A2 drop off at both ends of the SC.

A GPI-linked molecule, possibly ephrin-A5, in rostral tectum can preferentially induce branch formation of temporal retinal axons in chicks, and ephrin-A5 has been shown to act as a promotor of axonal branching in the formation of layer-specific circuits in the cortex. Ephrin-A2 and ephrin-A5 are also expressed in the retina in an increasing temporal to nasal gradient, though the functional implications of this phenomenon are still unclear. Several lines of evidence suggest that these receptor-ligand interactions are *repulsive*.

Theoretical Models

The development of retinotectal maps has inspired a large number of theoretical models. Some models have attempted to closely match the large array of data specific to the system in particular species; others have postulated more abstract and generic mechanisms for topographic map formation that can be applied across many systems. This section reviews in historical order some of the models designed to account for the initial activity-independent

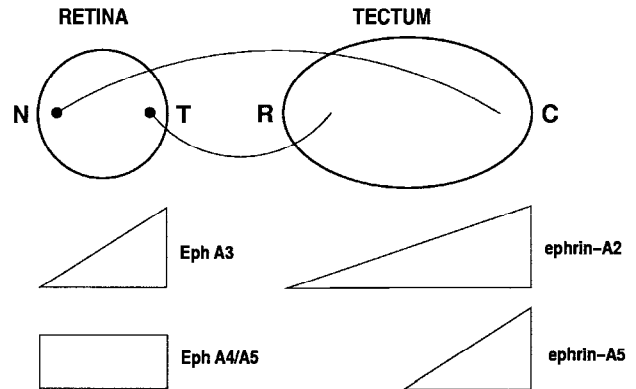


Figure 2. The normal retinotectal mapping and the distribution of Eph receptors and ephrin ligands in the chick. Note that the gradients run in opposite directions. Gradient shapes are not precisely linear. N, nasal; T, temporal; R, rostral; C, caudal.

stage of map formation and specific results, such as expansion, compression, translocation, and rotation of maps.

Prestige and Willshaw's Model

Prestige and Willshaw (1975) were the first to formalize notions of chemospecific matching suggested by Sperry's hypothesis. They distinguished between two forms of chemical matching, which they termed group I and group II (or type I and type II). In group I matching, each presynaptic cell j has an affinity for just a small neighborhood of postsynaptic cells, with peak affinity for the cell topographically matching to j in the postsynaptic sheet. Although such a rigid scheme can form a map under normal conditions, it cannot account for certain experimental data without invoking the unsatisfactory notion of respecification. In group II matching,

all axons have maximum affinity for making and retaining contacts at one end of the postsynaptic sheet of cells, and progressively less for the cells at greater distances from that end. Similarly, all postsynaptic cells have maximum affinity for axons from one end of the postsynaptic set, and axons remote from this end have correspondingly less likelihood of retaining any contact. We may thus talk of graded affinity within both pre- and postsynaptic sets . . . (pp. 82–83).

Simulations showed that this mechanism is capable of forming an ordered map if competition is introduced, in the form of normalization. Prestige and Willshaw specified that each presynaptic cell could make only a fixed number of contacts N_{pre} among the postsynaptic cells, and similarly each postsynaptic cell could only support a fixed number of contacts N_{post} from presynaptic cells. This ensures an even spread of connections: without competition for postsynaptic sites, every presynaptic cell would connect only to the highest-affinity end of the postsynaptic set, while without competition for presynaptic sites, every postsynaptic cell would receive connections only from cells at the highest-affinity end of the presynaptic set. However, to explain compression and expansion of retinotectal maps, Prestige and Willshaw found that it was necessary to make the additional assumption ("regulation") that N_{pre} and N_{post} were also altered. For instance, if half the postsynaptic set is removed, then, unless N_{post} is increased, connections will only be made from the highest-affinity end of the presynaptic set.

The Arrow Model

In the “arrow” model of Hope, Hammond, and Gaze (1976), it is assumed that (1) retinal fibers that terminate next to each other in the tectum are able to compare their relative positions of origin, and (2) retinal fibers can identify both rostral-caudal and medial-lateral axes of the tectum (this could be implemented in terms of a polarity gradient of marker along each of these axes). The procedure is then basically the bubblesort algorithm, although a biological implementation for this is not described. Starting from initially random connections in which each retinal fiber contacts one tectal cell, two retinal fibers that terminate next to each other in the tectum are chosen at random. Their retinal positions are compared, and if they are appropriate, the sites of termination of the fibers are exchanged. This simple algorithm is capable of forming topographic maps under normal conditions, and also rotated maps when a piece of the tectal array is rotated. However, it fails to account for the translocated map experiments.

The Tea Trade Model

Willshaw and von der Malsburg (1979) proposed a model based on Sperry’s idea that map formation is somehow dependent on induction of molecules from the retina into the tectum. The model is expressed in terms of molecular markers, but markers intrinsic only to the presynaptic sheet. There are no preexisting markers in the postsynaptic sheet: it is assumed that presynaptic markers are transported to the postsynaptic sheet via induction along retinal axons. Several markers exist in the presynaptic sheet, the sources of which are spaced out in the presynaptic sheet at fairly regular intervals. An analogy presented to explain the working of the model was the import of tea from plantations in India to British towns; hence the name “tea trade model.”

Initially, markers diffuse in the presynaptic sheet until a stable distribution is set up. It is assumed that each presynaptic axon then induces the vector of markers existing at that point in the presynaptic sheet into the postsynaptic sheet, where the markers diffuse according to the same rule as in the presynaptic sheet. The rate of induction at each synapse is proportional to the strength of that synapse. Synaptic strengths are updated periodically according to the degree of similarity between the vector of markers each fiber carries and the vector of markers already existing at those points it contacts in the postsynaptic sheet. Synaptic updating occurs with a molecular analogue of Hebb’s rule: the strength of connection between presynaptic cell i and postsynaptic cell j is increased in proportion to the similarity of their vectors of molecular markers. A slight overall bias in the connection strengths is specified initially in order to provide a global orientation for the map. Synaptic strengths are normalized by division so that each presynaptic cell can only support a certain total strength of connections. This ensures that every presynaptic axon makes contacts in the postsynaptic sheet. Willshaw and von der Malsburg (1979) showed that such a process reaches a stable state and can form appropriate maps under normal conditions, and also correctly predict the outcome of a range of results under abnormal conditions. In particular, some regeneration experiments can be accounted for under this scheme by assuming that when regeneration occurs, the postsynaptic sheet holds a “memory” of the previous pattern of innervation in terms of the previous stable distribution of markers. An algorithm similar to the tea trade model was analyzed mathematically for the one-dimensional case by Häussler and von der Malsburg (1983).

Gierer’s Model

Gierer (1983) proposed a model based on the matching of preexisting gradients in retina and tectum as in the chemoaffinity hy-

pothesis. He imagined that mapping is controlled by the concentration of an inhibitory substance $p(x, u)$, where x indicates tectal position and u indicates retinal position. p is produced by a reaction of the graded distribution of a retinal marker present on the tips of retinal axons with a graded distribution of a tectal marker. Axons then grow down the gradient of p to a minimum, stopping when $\partial p/\partial x = 0$.

For this to form a map, it is obviously required that the position of this minimum vary smoothly as a function of retinal origin u ; in the simplest case when $x = u$. There is an infinite number of combinations of gradient shapes and reaction rules that accomplish this; Gierer specifically considered exponential gradients. He also suggested a possible mechanism for gradient change in response to surgical manipulations such as retinal or tectal ablation. Reminiscent of the tea trade model, the idea is that

retinal fibre terminals induce, in the tectum, a slow increase in source (e.g. an enzyme) producing an additional contribution to p , and that the rate of increase is proportional to the local density of retinal fibre terminals. If the sources thus produced persist on the tectum while fibre terminals continuously move to respecified positions of minimal p , this process will eventually smooth out differences in the density of fibre terminals, giving rise to compression or expansion in the dimensions of ablation (p. 84).

The specific trajectories that axons might take to their targets under such a gradient matching scheme were also simulated.

Honda (1998) presented a simpler model in which axons grow to the tectal position where the retinal receptor concentration $R[u]$ times the tectal ligand concentration $L[x]$ equals a constant S , or $R[u]L[x] = S$, where S is the same for all axons. A brief analytical investigation of this type of model can be found in Goodhill and Richards (1999).

Fraser’s Model

In a more abstract vein, Fraser (e.g., Fraser and Perkel, 1990) introduced the notion that the state of the system could be described by an “adhesive free energy” G , which depends on how successfully a number of constraints are satisfied. In his model the form of the final mapping is found from minimizing G by simulated annealing. The constraints employed (in order of decreasing weighting in G) are (1) a position-independent adhesion between retinal and tectal cells, (2) a general competition among retinal axons for tectal space, (3) a tendency for neighboring axon terminations in the tectum to stabilize if those axons come from neighboring positions in the retina, (4) a dorsoventral gradient of adhesive specificity in retina and tectum, and (5) an anteroposterior gradient in retina and tectum. Although this model accounts (at least qualitatively) for a large proportion of the experimental literature, a developmental mechanism that could perform such a minimization was not provided.

Cowan’s Model

Whitelaw and Cowan (1981) attempted to integrate both marker- and activity-based mechanisms in map formation by combining a gradient of adhesive specificity with activity-dependent synaptic updating. Changes in synaptic strengths are multiplied by the degree of “adhesion” between the corresponding pre- and postsynaptic cells, and both pre- and postsynaptic normalization are employed. The model predicts a range of the experimental literature (including expansion and compression, mismatch, rotation, and compound eye experiments), and also draws attention to experimental evidence contradictory to the induction hypothesis of the tea trade model. More recent additions to this model, for instance

a tendency for fibers to stick to their retinal neighbors, have increased the range of data it can account for (see, e.g., Weber et al., 1997).

Discussion

Consideration of the wide variety of data described in the preceding sections—Eph/ephrin gradients, disturbances of normal development and regeneration, and branching—suggests two major limitations of previous models. First, few of them take into account recent discoveries regarding molecular gradients and their role in retinotectal mapping. Second, most of the previous models take synaptic strengths as their primary variable between arrays of retinal and tectal locations. They generally assume each retinal location is initially connected to all tectal locations, or at least to a topographically specific subset. Synaptic strengths are then updated according to rules that depend in various ways on correlated activity, competition for tectal space, molecular gradients, and fiber-fiber interactions. However, these last three effects enter only as terms modulating the development of synaptic strengths. With rare exceptions, actual movement or branching of axons to find their correct targets is not considered. The key experimental results that future computational models of retinotectal map formation should attempt to account for, besides normal map development, include the results of single and multiple knockout experiments of the relevant Eph receptors and ephrin ligands, results of experiments in which Eph/ephrins are misexpressed in retina or tectum, the correct guidance of retinal axons that enter the tectum by ectopic routes, and the results of retinal and tectal ablation and transplantation experiments.

Road Maps: Neural Plasticity; Vision

Related Reading: Axonal Path Finding; Collicular Visuomotor Transformations for Gaze Control; Ocular Dominance and Orientation Columns; Pattern Formation, Neural

References

- Flanagan, J. G., and Vanderhaeghen, P., 1998, The ephrins and Eph receptors in neural development, *Annu. Rev. Neurosci.*, 21:309–345. ♦
- Fraser, S. E., and Perkel, D. H., 1990, Competitive and positional cues in the patterning of nerve connections, *J. Neurobiol.*, 21:51–72.
- Gierer, A., 1983, Model for the retinotectal projection, *Proc. Roy. Soc. Lond. B*, 218:71–93.
- Goodhill, G. J., and Richards, L. J., 1999, Retinotectal maps: Molecules, models, and misplaced data, *Trends Neurosci.*, 22:529–534. ♦
- Häussler, A. F., and von der Malsburg, C., 1983, Development of retinotopic projections: An analytical treatment, *J. Theor. Neurobiol.*, 2:47–73.
- Holt, C. E., and Harris, W. A., 1993, Position, guidance, and mapping in the developing visual system, *J. Neurobiol.*, 24:1400–1422. ♦
- Honda, H., 1998, Topographic mapping in the retinotectal projection by means of complementary ligand and receptor gradients: A computer simulation study, *J. Theoret. Biol.*, 192:235–246.
- Hope, R. A., Hammond, B. J., and Gaze, R. M., 1976, The arrow model: Retinotectal specificity and map formation in the goldfish visual system, *Proc. R. Soc. Lond. B*, 194:447–466.
- O'Leary, D. D. M., Yates, P. A., and McLaughlin, T., 1999, Molecular development of sensory maps: Representing sights and smells in the brain cell, *Cell*, 96:255–269.
- Prestige, M. C., and Willshaw, D. J., 1975, On a role for competition in the formation of patterned neural connexions, *Proc. R. Soc. Lond. B*, 190:77–98.
- Sperry, R. W., 1963, Chemoaffinity in the orderly growth of nerve fiber patterns and connections, *Proc. Natl. Acad. Sci. USA*, 50:703–710.
- Udin, S. B., and Fawcett, J. W., 1988, Formation of topographic maps, *Annu. Rev. Neurosci.*, 11:289–327. ♦
- Weber, C., Ritter, H., Cowan, J., and Obermayer, K., 1997, Development and regeneration of the retinotectal map in goldfish: A computational study, *Philos. Trans. R. Soc. Lond. B*, 352:1603–1623.
- Whitelaw, V. A., and Cowan, J. D., 1981, Specificity and plasticity of retinotectal connections: A computational model, *J. Neurosci.*, 1:1369–1387.
- Willshaw, D. J., and von der Malsburg, C., 1979, A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem, *Philos. Trans. R. Soc. B*, 287:203–243.

Developmental Disorders

Annette Karmiloff-Smith and Michael S. C. Thomas

Introduction

Connectionist models have recently provided a concrete computational platform from which to explore how different initial constraints in the cognitive system can interact with an environment to generate the behaviors we find in normal development (Elman et al., 1996; Thomas and Karmiloff-Smith, 2002a). In this sense, networks embody several principles inherent to Piagetian theory, the major developmental theory of the twentieth century. By extension, these models provide the opportunity to explore how shifts in these initial constraints (or boundary conditions) can result in the emergence of the abnormal behaviors we find in atypical development. Although this field is very new, connectionist models have already been put forward to explain disordered language development in Specific Language Impairment (Hoeffner and McClelland, 1993), Williams syndrome (Thomas and Karmiloff-Smith, 2002b), and developmental dyslexia (Seidenberg and colleagues; see, e.g., Harm and Seidenberg, 1999) and to explain unusual characteristics of perceptual discrimination in autism (Gustafsson, 1997; Cohen, 1998). In this article, we examine the types of initial computational constraints that connectionist modelers typ-

ically build in to their models and how variations in these constraints have been proposed as possible accounts of the causes of particular developmental disorders. In particular, we examine the claim that these constraints are candidates for what will constitute innate knowledge. First, however, we need to consider a current debate concerning whether developmental disorders are a useful tool to explore the (possibly innate) structure of the normal cognitive system. We will find that connectionist approaches are much more consistent with one side of this debate than the other.

Developmental Disorders and Modularity

Cognitive neuropsychology assumes that the adult cognitive system has a modular structure, whereby the system can be decomposed into specialized functional components (although whether these components correspond to localized areas of the brain or can be captured by brain-imaging techniques remains an open question). In addition, cognitive neuropsychology assumes that that selective behavioral deficits in adults with brain damage can reveal this modular structure. Developmental disorders can also produce

apparently specific deficits in the end state of development. For example, Williams syndrome (WS), a developmental disorder caused by a microdeletion of contiguous genes on one of the alleles of chromosome 7, is characterized by a behavioral profile of relative proficiency in language, face processing, and theory-of-mind (attributing mental states to others) but severe deficits in other skills such as visuospatial processing, number, and problem solving (Karmiloff-Smith, 1998). In hydrocephalus with associated myelomeningocele (a protrusion of the membranes of the brain or spinal cord through a defect in the skull or spinal column), language can be the only area of proficiency. Individuals suffering from Specific Language Impairment (SLI) show the opposite pattern, often apparently performing within the normal range in all domains except language. In autism, even individuals with normal IQs are selectively impaired in tasks that require judging another's mental states (Baron-Cohen, Tager-Flusberg, and Cohen, 1993).

Evidence in many of these disorders of specific high-level deficits at the end of development has encouraged some researchers to view developmental disorders as offering the same theoretical insights into the static structure of the cognitive system that cases of adult brain damage provide. Where such disorders are of a genetic origin, selective developmental deficits are interpreted as revealing innate underpinnings for such structure. For example, Baron-Cohen et al. (1993) have argued that in individuals with autism, an apparent deficit in reasoning about mental states can be explained by the impairment of an innate, dedicated module for such reasoning (the theory-of-mind module). Van der Lely (1997) maintains that selective behavioral deficits in the language performance of children with grammatical SLI can be explained by damage to an underlying, innate module representing syntactic (rule-based) information. Clahsen and Almazan (1998) have proposed that a behavioral deficit in WS language supports the view that while their syntactic skills are "intact," individuals have a deficit to a specific aspect of their (modular) language knowledge: that of accessing information about words that are exceptions to syntactic rules.

However, there are a number of problems with the adult brain damage approach to developmental disorders (Karmiloff-Smith, 1998; Thomas and Karmiloff-Smith, in press). These boil down to the suspicion that such an approach massively underestimates the complexity of the path from gene to behavior. The prevalence of many-to-many, very indirect mappings in the relationship of genes to cognition undermines the claim that direct specific mappings will exist between particular genes and individual high-level cognitive abilities. To the extent that genes are involved in the causal chain of several cognitive domains, it will be less likely that they code anything specific to a single domain. Indeed, direct mappings are unlikely, given that spatial distributions of gene expression in the brain are rarely narrowly confined to subsequent areas of functional specialization in the adult brain and therefore seem unable to code directly for domain-specific developmental outcomes. Even if they could, the idea that behavioral deficits that are identified in the end state of a developmental disorder could reflect the impairment of a single functional module is predicated on the dubious assumption that the rest of the cognitive system could nevertheless develop normally. For this to be true requires either that functional modules develop independently of overall brain growth or that the content of modules is fixed in advance (i.e., the content is innately specified). But neither of these assumptions is likely to be true. With regard to the first, Bishop (1997) has argued persuasively that interactivity rather than independence is the hallmark of early development. With regard to the second, it seems likely that modular structure in the cognitive system and in the brain is an outcome of development rather than a precursor to it and that the neonate brain does not support innate representations with specific content (Elman et al., 1996; Johnson, 1997). A growing number of studies show how both neural localization and neural specialization for

biologically important functions, such as species recognition (Johnson, 1997) and language (Neville, 1991), take place gradually across development.

An alternative to the use of the adult brain damage model is the neuroconstructivist approach (Elman et al., 1996; Karmiloff-Smith, 1998). This approach views developmental disorders in terms of different developmental trajectories, caused by initial differences at a neurocomputational level. Thus, there might be differences in the microcircuitry of the brain or the firing properties of neurons, as opposed to discrete lesions to particular large-scale brain structures or pathways. In this view, development is an interactive process in which the cognitive system self-organizes in response to interactions with a structured environment. Interestingly, this approach suggests that people with developmental disorders may exhibit strengths as well as weaknesses. This prediction is consistent with the superior face recognition skills found in WS and the superior perceptual discrimination abilities found in autism. Neuroconstructivism further suggests that equivalent behavior across normal and abnormal phenotypes may mask different underlying cognitive processes. The notion that an ability is "intact" or "spared" because there is no apparent deficit at the behavioral level employs terminology from the adult brain damage model that may be misleading. To take an example, people with Williams syndrome can display scores on some language and face-processing tasks that are in the normal range. Nevertheless, closer examination suggests that different *cognitive* processes underlie the equivalent *behavioral* scores (Karmiloff-Smith, 1998).

Initial Constraints in Connectionist Models

Current connectionist models of developmental disorders employ alterations to the initial model constraints that, after training, lead to an end state exhibiting behavioral deficits. Connection weights are usually randomized so that the normal network has no specific knowledge prior to training. Thus, it follows that the atypical network has no specific deficit in knowledge prior to training. The behavioral deficits that emerge when these atypical networks are trained are often quite different from the effects of damaging a normal network model after training has been completed. This holds even when the network manipulations are the same in each case (Thomas and Karmiloff-Smith, in press). Thus, current connectionist models are more consistent with the neuroconstructivist approach to developmental disorders than with the adult brain damage approach.

What, then, are the initial constraints that connectionist modelers build into their models of cognitive development? And how do changes to these constraints alter the trajectory of development? In fact, the constraints that connectionist models build in are quite strong ones, and this may come as a surprise to some readers. Connectionist models are often mischaracterized as being unitary, homogeneous, seamless, or undifferentiated domain-general learning devices, whereby the environment is all-powerful in shaping the behavior of the final system. In fact, current connectionist models have a great deal of pre-existing structure built into them prior to any exposure to their training environment. What is general about connectionism are the principles of computation (Seidenberg, 1993). When the general principles of computation are combined with the boundary conditions for a specific domain, the result is a domain-specific model. It is the generality of these principles that gives the connectionist approach its explanatory power. That is, connectionism seeks not just to formulate descriptive generalizations about empirical phenomena, but also to show how they derive from underlying and independently motivated principles (Seidenberg, 1993). However, connectionist models are just as reliant on the constraints of a given domain as they are on the computational principles. Without justified limitations in the design of network

models, they become overpowered data-fitting devices that can at best provide descriptively adequate accounts of cognitive abilities. In short, connectionist models of development do include initial structure but not initial representational content (Elman et al., 1996). The point is that in interacting with a training environment, networks create representational content and become increasingly structured. This additional structure reflects the nature of the training environment.

The structure or boundary conditions that these models build in prior to training typically involve the following:

1. *The initial state of the network, in terms of the number of units, layers, connections, and the pattern of connectivity, collectively known as the network architecture.* The architecture determines the computational power of the network and the type of computations to which the network will be suited. For example, recurrent networks are suited to processing sequential information, whereas associative networks are suited to pattern recognition. The a priori choice of the architecture will have a central role in determining the adequacy of the network in modeling a given domain of cognitive development. A reduction in the number of processing units, or the elimination of internal (hidden) processing units, will restrict the computational power of the network and, depending on the nature of the domain, cause some or all parts of the problem to be learned inadequately. Addition of layers of internal units beyond a single layer tends to delay learning, without a marked increase in effective computational power. Increase of hidden units within a single layer tends to improve performance on the training set but may impair generalization beyond the training set.
2. *The way a particular cognitive problem is presented to the network, in terms of the input and output representations.* For a given domain, the representations determine the nature of the computational problem that the network will face. When the network has to extract a function from the training set (such as a general rule or regularity), the representational scheme will be crucial in determining how transparent or opaque this function is to the network. For instance, if a network is given a training set in a form that masks the relevant similarity between those items in the problem domain that obey a rule, the network will have difficulty in extracting this regularity.
3. *The learning algorithm that the network will use to change its connection weights (and, potentially, its architecture).* Most networks are trained by changing weights to minimize some cost function, such as the difference between the actual output and a target output. The rate at which weights are changed can have an impact on the success of a network in learning a problem. In particular, in complex domains, if weights are changed too quickly, the network may commit too early to a partial solution to the problem and be resistant to change with subsequent training. The learning algorithm is key in determining the plasticity of the network to further learning. Some algorithms allow on-line changes to network architecture depending on how well the network is learning a problem. The way in which the network's computational power is altered on-line will again have a considerable influence on the success of the network in capturing a cognitive ability (see 1 above).
4. *The regime of training that the network will undergo.* After modelers determine the network and the learning rule, they then expose the network to a training set. Often, the network is exposed to the entire training corpus from the start of training. However, in some cases, the network might be trained on an initially limited training set, perhaps based on assumptions about the nature of a child's early learning environment. This initial restriction will affect later training. It may aid learning if the smaller set is representative of the larger set or if it allows

the construction of internal representations that will be useful in learning the larger set. On the other hand, it may impair learning if the initial training set contains detail that is irrelevant to the full domain. Alterations in network parameters early in training may have the same effect as restricting the initial training set, as Elman's work on learning syntax with recurrent networks has demonstrated (Elman et al., 1996).

In the connectionist framework, these constraints represent some of the candidates for innateness (although, equally, in principle, any of the constraints could also be experience dependent). Alterations in one or more of these constraints may then lead to the emergence of disordered representations and impaired behavior in a model of atypical development.

We have noted that these models do not support innate representational content in that their weights are initially randomized. However, it is an open question as to whether initial computational constraints (along with sensory input determined by the individual's interaction with the environment) are sufficient to drive development. One possible addition is the postulation of innate attentional predispositions (Elman et al., 1996). In this theory, innate knowledge is built into the subcortical part of brain in the form of a low acuity predisposition to attend certain stimuli. This predisposition then ensures the representation of input that will subsequently drive learning in the more powerful cortical system. For example, such an innate predisposition in face recognition might encourage the newborn infant to attend preferentially to visual stimuli containing a single blob positioned centrally below two blobs (see Johnson, 1997). Innate predispositions provide another candidate factor that might be altered in the start state of the atypical system.

In general, current connectionist models of normal development do not restrict themselves to computational constraints and innate attentional predispositions in their start states, since these models incorporate high-level, domain-specific representations. These models must therefore be seen as a halfway house. In the future they need to be extended to show how their domain-specific representations emerge from some prior process operating over lower-level information (and with its own computational constraints).

We now turn to consider recent examples of connectionist models of developmental disorders.

Recent Models

Autism

Autism is a developmental disorder characterized primarily by a central triad of deficits in social interaction, communication, and imagination but also by a range of secondary deficits. These include a restricted repertoire of interests, an obsessive desire for sameness, excellent rote memory, improved perceptual discrimination, and an impaired ability to form abstractions or generalize knowledge to new situations. Evidence from neuropathological investigations of the brains of affected individuals is suggestive of abnormal wiring patterns in various brain regions, perhaps caused by deficits in neuronal migration during fetal development, curtailment of normal neuronal growth, and/or aberrant development. Cohen (1998) has argued that these structural deficits are consistent with too few neurons in some brain areas, such as the cerebellum, and too many neurons in other areas, such as the amygdala and hippocampus. Cohen has proposed that simple connectionist models trained on categorization tasks can link such differences in neurocomputational constraints to some of the secondary deficits found in autism. In some cases, children with autism have trouble acquiring simple discriminations and attend to a restricted range of stimuli, while in others, children with autism have good discrimination and indeed

very good memory but seem to rely on representing too many unique details of stimuli. Cohen showed that simple backpropagation networks with too few hidden units showed a failure to learn classifications tasks, while those with a surfeit of hidden units showed very fast learning, but subsequently generalization became poor, and the network increasingly responded according to particular details of the training set.

In a related proposal, Gustafsson (1997) speculated that the combination of a failure to generalize and heightened perceptual discrimination might be traced to the atypical development of cortical feature maps. In particular, he suggested that higher-than-normal levels of within-layer inhibition in the initial cortical sheet would lead to overly fine-tuned perceptual features, permitting good discrimination but preventing good generalization.

Although Gustafsson did not support this idea with implementations, related work by Oliver et al. (2000) offers an insight into how feature map formation can be developmentally disrupted by changes in the initial properties of a self-organizing connectionist network. Oliver et al. employed a neurobiologically constrained network in which a two-dimensional output layer received information from a single input retina. The network was shown a set of stimuli in the form of bars lying across the input retina. Oliver et al. showed that, using their initial parameter set, the output layer formed a topographic map of the possible inputs: Certain areas of the output layer specialized in responding to each input, and areas representing similar inputs were adjacent to each other in the output layer. Oliver et al. then re-ran the model, disrupting the network in different ways before exposing it to the training stimuli. They varied the threshold of the output units, disrupted the connectivity between the input and the output layers, disrupted the connectivity responsible for lateral inhibition in output layer, and changed the similarity of the input stimuli to each other. These manipulations demonstrated that tiny differences in the initial constraints under which the model developed could have a very significant impact on the outcome of development. The resulting topographic map suffered a range of disruptions, including output units failing to specialize at all or simply turning off, specialization emerging but not in organized areas, and organized areas emerging but without adjacent areas representing similar-looking bars.

Much work remains to be done to develop these proposals exploring the neurocomputational underpinnings of developmental disorders such as autism and, in particular, to relate disordered feature maps or hidden unit numbers to higher-level cognitive deficits such as those in social interaction, communication, and imagination that characterize autism. Nevertheless, such work importantly illustrates a new conception of such disorders in terms of development operating under atypical constraints, rather than in terms of deficits to high-level modules in a static model of the normal adult cognitive system.

Developmental Dyslexia

This disorder has been the focus of much connectionist research, given the success of models in capturing the normal processes of reading. A number of models by Seidenberg and colleagues have sought to change initial constraints in models of reading to simulate either surface dyslexia (in which the subject has difficulty reading words that are exceptions to normal rules of pronunciation), phonological dyslexia (in which the reading of novel words is impaired), or a combination of both. Typically, these models learn mappings between codes representing orthography, phonology, and semantics. Surface dyslexia has been simulated by employing “too few” hidden units in the model or by reducing the learning rate. Phonological dyslexia has been simulated by degrading the phonological representations in some way, for instance, in the type of coding scheme used. For example, Harm and Seidenberg (1999)

pretrained one part of their model to develop appropriate phonological representations prior to learning the reading task. When this “phonological” part of the model was impaired, either by reducing its initial computational power or by limiting the size of the connection weights it could develop, the result was a network exhibiting phonological dyslexia at the end of training on the reading task.

Specific Language Impairment

Hoeffner and McClelland (1993) sought to capture deficits found in the morphosyntax of subjects with SLI, specifically their difficulty in the learning of rule-based inflectional morphology in verbs. Hoeffner and McClelland employed an attractor network mapping between semantic codes and phonological codes. They simulated SLI by changing the initial phonological representations, in line with a hypothesis that SLI may be caused by early perceptual impairments. Specifically, they impaired the network’s ability to represent word-final stops and fricatives (including /t/, /d/, and /s/). Although the model they used didn’t show an ideal fit to the normal data when unimpaired, it nevertheless captured a number of the key deficits of SLI when trained with impaired representations, particularly a selective difficulty with the formation of regular (+ed) past tenses. In this case, the initial phonological deficit obscured precisely the information that the network required to be able to learn the relevant generalizations about regular past tense formation. However, it should be noted that the perceptual deficit account of SLI remains controversial, and this disorder may well turn out to be heterogeneous, with several different causes.

Williams Syndrome

Recent work in our laboratory has examined underlying deficits in the language of individuals with Williams syndrome (WS). Their language was initially thought to be “spared,” but closer examination revealed a number of subtle deficits. It had been reported that individuals with WS show difficulties in forming the past tense of irregular verbs while showing good performance on the regular, rule-based past tense formations (Clahsen and Almazan, 1998). Our recent empirical work suggests that much of this apparently selective deficit is due to an overall delay in language development (young children also find irregular verbs difficult). However, individuals with WS do appear to show a deficit in generalizing knowledge of inflectional patterns to novel forms. These two patterns (selective difficulty with irregular inflections versus reduced levels of generalization) continue to be argued for in the literature on WS language.

Thomas and Karmiloff-Smith (2002b) set out to explore whether alterations to the initial computational constraints of a connectionist model of past tense development could account for either of these patterns of data. The past tense network mapped from verb stem to past tense form in the presence of lexical-semantic information. Various theoretical claims have been made that the WS language system develops under different constraints. These include the proposals that their phonological representations may be atypical and perhaps rely on sensitive auditory processing, that their lexical-semantic representations may show anomalous organization, or that lexical-semantic information about words may be poorly integrated with phonology. To explore the viability of these different accounts to explain the pattern of performance in the past tense task, Thomas and Karmiloff-Smith altered the initial constraints of the network model to implement each type of lower-level deficit.

The results revealed that reduced generalization was consistent with atypical phonological representations (specifically with reduced similarity and redundancy) or interference in integrating

lexical-semantic and phonological knowledge. A range of computational constraints caused poorer acquisition of irregular verbs; these verbs are in the minority, and their acquisition can be disrupted under nonoptimal learning conditions. However, attenuated activation from lexical-semantics was able to selectively delay irregular acquisition, offering a link to one of the proposed deficits in WS. Finally, the model demonstrated for the first time precisely how different computational constraints interact in a system in the process of development: The atypical trajectory that is found in WS may arise from the combination of more than one altered constraint.

Further empirical work remains to be carried out to clarify which of the WS patterns is the correct one. However, this modeling work has determined which of the competing accounts are viable within an existing developmental framework and therefore provided a focus for future investigations. Once again, this contrasts with previous theoretical work that construed the WS language system in terms of selective deficits to a static model of the normal adult language system (e.g., Clahsen and Almazan, 1998).

Discussion

Developmental disorders can inform the study of normal development because they provide a broader view of the parameter space within which development takes place. The empiricist argues that the environment specifies a capacity so strongly that systems with a wide variety of initial structures must come to reflect it. The nativist argues that the environment specifies a capacity so poorly that the system must come equipped with pre-existing structure if it is to find the correct solution given the environmental input. The neuroconstructivist argues that the robustness of the cognitive system to changes in its initial setup (as long as we can come to understand precisely what these changes are) will tell us how evolution has placed its bets concerning the capacities that can be trusted to emerge through experience and the capacities that must be given a firmer guiding hand through development. Connectionist models provide a powerful tool with which to investigate the role of initial computational constraints in determining the trajectory of both typical and atypical development. For developmental disorders in which selective deficits in high-level behaviors are reported, the connectionist framework ensures that these deficits are properly seen in terms of the outcome of the developmental process itself.

Road Map: Psychology

Related Reading: Cognitive Development; Language Acquisition; Neurological and Psychiatric Disorders

References

- Baron-Cohen, S., Tager-Flusberg, H., and Cohen, D. J., 1993, *Understanding Other Minds: Perspectives from Autism*, Oxford, Engl.: Oxford University Press.
- Bishop, D. V. M., 1997, Cognitive neuropsychology and developmental disorders: Uncomfortable bedfellows, *Quarterly Journal of Experimental Psychology*, 50A:899–923.
- Clahsen, H., and Almazan, M., 1998, Syntax and morphology in Williams syndrome, *Cognition*, 68:167–198.
- Cohen, I. L., 1998, Neural network analysis of learning in autism, in *Neural Networks and Psychopathology* (D. Stein and J. Ludick, Eds.), Cambridge, Engl.: Cambridge University Press, pp. 274–315.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., 1996, *Rethinking Innateness: A Connectionist Perspective on Development*, Cambridge, MA: MIT Press. ♦
- Gustafsson, L., 1997, Inadequate cortical feature maps: A neural circuit theory of autism, *Biol. Psychiatry*, 42:1138–1147.
- Harm, M., and Seidenberg, M. S., 1999, Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychol. Rev.*, 106:491–528.
- Hoeffner, J. H., and McClelland, J. L., 1993, Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation, in *Proceedings of the 25th Child Language Research Forum* (E. V. Clark, Ed.), Stanford University, CA: Center for the Study of Language and Information, pp. 38–49.
- Johnson, M. H., 1997, *Developmental Cognitive Neuroscience*, Oxford, Engl.: Blackwell. ♦
- Karmiloff-Smith, A., 1998, Development itself is the key to understanding developmental disorders, *Trends Cogn. Sci.*, 2:389–398. ♦
- Neville, H. J., 1991, Neurobiology of cognitive and language processing: Effects of early experience, in *Brain Maturation and Cognitive Development: Comparative and Cross-Cultural Perspectives. Foundation of Human Behavior* (K. R. Gibson and A. C. Peterson, Eds.), New York: Aldine de Gruyter, pp. 355–380.
- Oliver, A., Johnson, M. H., Karmiloff-Smith, A., and Pennington, B., 2000, Deviations in the emergence of representations: A neuroconstructivist framework for analysing developmental disorders, *Dev. Sci.*, 3:1–23.
- Seidenberg, M., 1993, Connectionist models and cognitive theory, *Psychol. Sci.*, 4:228–235.
- Thomas, M. S. C., and Karmiloff-Smith, A., 2002a, Modeling typical and atypical cognitive development, in *Handbook of Childhood Development* (U. Goswami, Ed.), Oxford, Engl.: Blackwell, pp. 575–599. ♦
- Thomas, M. S. C., and Karmiloff-Smith, A., 2002b, Modeling language acquisition in atypical phenotypes. Manuscript submitted for publication.
- Thomas, M. S. C., and Karmiloff-Smith, A., in press, Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behav. Brain Sci.*
- Van der Lely, H. K. J., 1997, Language and cognitive development in a grammatical SLI boy: Modularity and innateness, *J. Neurolinguistics*, 10:75–107.

Diffusion Models of Neuron Activity

Luigi M. Ricciardi and Petr Lánský

Introduction

We offer a survey of one-dimensional stochastic diffusion models for the membrane depolarization of a single neuron, with emphasis on the related first-passage-time (FPT) problems, namely, on the determination of the neuronal output. In complex deterministic neuronal models, such as the Hodgkin-Huxley or the Fitzhugh-Nagumo models, the generation of action potentials is automatically included, being itself an essential component of the equations

defining and characterizing these models. By contrast, for simple models (usually stochastic, such as those reviewed here), a firing threshold has to be introduced; the generation times of the action potentials are then identified with the instants when the membrane depolarization, which is modeled on the analogy of a particle randomly diffusing in a liquid, equals the firing threshold. Although these times are unique for deterministic models, in the case of stochastic models, and apparently also for many types of real neurons, the time interval that elapses between an action potential and the

next one is at least partly random. In the sequel, such a time interval will be viewed as a random variable. The properties of this random variable are mathematically investigated by studying the related FPT problem.

Hereafter, we shall make use of the formalism of stochastic differential equations (SDEs). These equations can be heuristically described as ordinary differential equations in which a rapidly and irregularly fluctuating term (usually involving white noise, in additive or multiplicative forms) is also present; they are also referred to as Langevin equations, after the renowned physicist Paul Langevin. To elucidate the nature and the assumptions underlying diffusion models, in a couple of paradigmatic instances we shall sketch how neuronal stochastic diffusion models can be constructed from first principles, that is, without resorting to SDEs. We shall thus indicate how the much-celebrated Wiener and Ornstein-Uhlenbeck (OU) neuronal models stem from simple Markov processes with discrete state spaces after a suitable limiting procedure.

The broad applicability and intuitive appeal of the OU model follow from the circumstance that, if formulated by means of an SDE, up to the stochastic term it coincides with the most common phenomenological deterministic neuronal model, the so-called “leaky integrate-and-fire model.” Hence, the OU model represents a natural bridge between deterministic and stochastic modeling of neurons’ activity. The success of the Wiener model, which is a special case of the OU model, is instead based on the fact that it is largely solvable in closed form. For this reason, it may serve as a prototype for possible comparisons.

Besides the OU and Wiener models, which are characterized by additive noise terms appearing in the corresponding SDEs, diffusion models with multiplicative noise will also be reviewed. Although neuronal models of the diffusion type are primarily used for the description of steady-state firing under a constant stimulation, some results that are obtained in the case of periodic stimulation will also be mentioned. Irrespective of the type of applied noise, the approach employed here is based on frequency coding presumption.

Stochastic diffusion models are not primarily aimed at direct comparisons with experimental data, being mathematical abstractions of less tractable but more realistic or transparent models. They mainly serve to study the properties of other models by mathematical methods and to produce qualitative predictions, as in the extensive investigations performed on the possibility of stochastic resonance in neurons and neuronal models. For this reason, direct analyses of experimental data by diffusion models are relatively rare. For some references on this matter, see Inoue, Sato, and Ricciardi (1995). More references and expanded commentary can be found in a recent review article (Lánský and Sato, 1999). Additional references and detailed treatment of some models are in Ricciardi (1977) and Tuckwell (1988). For a review on the FPT problem for diffusion processes, see Ricciardi and Sato (1989).

Diffusion Processes and Neuronal Modeling

From a biophysical point of view, neuronal models of single cells reflect the electrical properties of the membrane via electric circuit models that contain energy storage elements. Such circuit models can be described by means of differential equations for the membrane potential. A consistent body of data, recorded from a large variety of different neuronal structures and under different experimental conditions, suggests that the presence of stochastic variables should in general be included in the mathematical models of the neuron’s input-output activity, even though the role and influence of these variables are expected to depend on the nature of the specific questions one wishes to answer. Hereafter, we assume that there is a random component, generally denoted as *noise*, embedded in the neuron’s input. A second source of noise can be attrib-

uted to the neuron itself, where a random component is added to the incoming input signal. Taking this fact into account, the differential equation describing the neuronal membrane potential then includes a noisy term, and hence becomes itself an SDE whose solution can sometimes be approximated by a diffusion process. In other cases—for instance, when the neuron has very few synaptic inputs near the trigger zone—a Poisson-driven differential equation may be a biologically more appropriate model.

Here we shall assume that the membrane potential between two consecutive neuronal firings (spikes) is represented by a scalar diffusion process $X(t)$. Such a process can be described by the SDE

$$\begin{aligned} dX(t) &= \mu[X(t), t]dt + \sigma[X(t), t]dW(t), \\ P\{X(t_0) = x_0\} &= 1 \end{aligned} \quad (1)$$

where μ and σ are real-valued functions of their arguments satisfying certain regularity conditions, and $W(t)$ is the standard Wiener process. (See Ricciardi, 1977, for an expository discussion of SDEs.) The Wiener process has historically been exploited as the first mathematical model of Brownian motion, namely, the highly irregular and ceaseless motion characterizing all particles immersed in a fluid, irrespective of their nature, that was discovered by Robert Brown as early as 1827. Only in the year 1905 was a mathematical explanation of this mysterious phenomenon provided by Albert Einstein. In the mid-1930s Einstein’s theory was refined, completed, and made mathematically more rigorous and general by various physicists and mathematicians, including Norbert Wiener, after whom a fundamental stochastic diffusion process was named. In short, the foundations of a new branch of the theory of stochastic processes were laid, based on a formalism involving equations similar to those describing the time change of the temperature in each point of a metal rod initially heated at one point, or the change of density in each point of a liquid after dropping in it a specified amount of salt. Thus were the “diffusion equations” born.

A stochastic process $W(t)$ is said to be a Wiener process if

- (a) $W(0) = 0$.
- (b) $W(t)$ has stationary independent increments.
- (c) For every $t > 0$, $W(t)$ is normally distributed.
- (d) For all $t > 0$, $E[W(t)] = 0$.

These axioms imply $\text{Var}[W(t)] = a^2t$, where a^2 is a positive parameter, usually representing an empirical characteristic of the process, to be determined by observations. In this paper, we refer to the so-called *standard Wiener process* for which $a = 1$. The reference level for $X(t)$ in Equation 1 is usually taken to be equal to the resting potential. The initial voltage—namely, the reset value instantly attained following a spike—is often assumed to be equal to the resting potential: $X(t_0) = x_0$, where $t_0 \in \mathbb{R}$ denotes the initial time after spike generation.

Because of the simplicity of Equation 1, the process of the action potential generation is not an inherent part of the model, as it is in more complex models, and so the existence of a firing threshold potential $S(t)$ —customarily assumed to be a deterministic function of time such that $S(t_0) > x_0$ —must be imposed here. The model neuron fires whenever such a threshold potential is reached; then $X(t)$ is instantly reset to its initial value x_0 . This situation corresponds to an FPT problem for the associated diffusion process. The interspike intervals are identified with the FPT of $X(t)$ across $S(t)$, namely, with the random variable

$$T = \inf_{t \geq t_0} \{t : X(t) > S(t)\}, \quad X(t_0) = x_0 \quad (2)$$

The importance of interspike intervals is due to the generally accepted hypothesis that the information transferred within the nervous system is usually encoded by the timing of neuronal spikes.

In addition, the reciprocal relationship holding between the firing frequency and the interspike interval naturally leads to the problem of determining the probability density function (pdf) of T , namely, the function $g[S(t), t|x_0, t_0] = \partial_t P\{T \leq t\}$. Particularly when this density cannot be obtained analytically (which is practically the rule) or when it is too difficult to give sufficiently precise estimations of it, the analysis is focused on its statistical moments, and primarily on the mean $E(T)$ and variance $D^2(T)$. The coefficient of variation, $D(T)/E(T)$, is also used, as it is a measure of the relative spread of the distribution and of its deviation from exponentiality. The reciprocal relationship between the firing frequency, on the one hand, and the interspike interval on the other hand suggests plotting the inverse of the value of $E(T)$ versus the intensity of stimulation (reflected by μ) as a stochastic counterpart of the input-output frequency curve.

In the following, we shall restrict our considerations to constant thresholds, and hence we shall set throughout $S(t) = S$, with $S > x_0$. The more realistic models based on time-dependent thresholds, aimed to simulate various aspects of the time varying behavior of the neuron, are mainly used to mimic the relative refractory period, namely, the time change of sensitivity of the neuron to incoming stimuli after a spike has been released. Hence, time-dependent thresholds should be characterized by a very large initial value (possibly infinity), followed by a decay to the constant value S . A realistic example of such a threshold is provided by the function $S(t) = S + S_1 \exp(-t/\gamma)$, where S_1 denotes a very large positive constant and $\gamma > 0$ determines how fast the asymptotic value S is approached.

An alternative description of the process $X(t)$ is obtained via the so-called *diffusion equations approach*. First, one defines the transition pdf of $X(t)$ conditional on $X(t_0) = x_0$:

$$f(x, t|x_0, t_0) = \frac{\partial}{\partial x} P\{X(t) \leq x | X(t_0) = x_0\} \quad (3)$$

It can then be seen that in most cases, f satisfies the Fokker-Planck (FP) equation

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x} [A_1(x, t)f] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [A_2(x, t)f] \quad (4)$$

with initial condition $\lim_{t \rightarrow t_0} f(x, t|x_0, t_0) = \delta(x - x_0)$, and the Kolmogorov (K) equation

$$\frac{\partial f}{\partial t_0} + A_1(x_0, t_0) \frac{\partial f}{\partial x_0} + \frac{1}{2} A_2(x_0, t_0) \frac{\partial^2 f}{\partial x_0^2} = 0 \quad (5)$$

with initial condition $\lim_{t_0 \rightarrow t} f(x, t|x_0, t_0) = \delta(x_0 - x)$, where δ denotes the Dirac delta function. Here the coefficients $A_1(x, t)$ and $A_2(x, t)$ are the *infinitesimal moments*, or the *drift*, and *infinitesimal variance*, defined as

$$\begin{aligned} A_i(x, t) &:= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} E\{[X(t + \Delta t) - X(t)]^i | X(t) = x\} \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int (y - x)^i f(y, t + \Delta t | x, t) dy \quad (6) \\ &\quad (i = 1, 2) \end{aligned}$$

It is essential to mention that the quantities $A_1(x, t)$ and $A_2(x, t)$ are related in the following way (Ricciardi, 1977) to the functions μ and σ considered in Equation 1:

$$A_1(x, t) = \mu(x, t) \quad (7)$$

$$A_2(x, t) = \sigma^2(x, t) \quad (8)$$

Equations 1, 4, and 5 characterize the class of diffusion processes. While the description by Equation 1 is more intuitive and suitable for computer simulations, Equations 4 and 5 are suitable

for analytical and numerical treatments of the model. The parameter $A_1(x, t)$ (Equation 7) reflects the strength of the signal impinging on the neuron, while the parameter $A_2(x, t)$ (Equation 8) characterizes the level of noise associated with the signal. If we are interested in modeling the *spontaneous* or *resting* activity of a neuron or its *steady-state* response to a constant stimulus—which is usually required before attempting to model the neuron's response to time-varying stimuli—independence of time can be assumed in Equations 7 and 8. Then, the neuronal output modeled by FPT is a renewal process (intervals between firings are independent and identically distributed random variables), so that without any loss of generality, we can set $t_0 = 0$.

Neuronal Models with Additive Noise

Wiener Process

The year 1964 marks the beginning of the history of neuronal models based on diffusion processes. Indeed, in a much-celebrated article, Gerstein and Mandelbrot (1964) postulated that for a number of experimentally monitored neurons subject to spontaneous activity, the firing pdf could be modeled by the FPT pdf of a Wiener process. Actually, these authors were able to show that, by suitably choosing the parameters of the model, the experimentally recorded interspike interval histograms of many units could be fitted to an excellent degree of approximation by means of the FPT pdf of a diffusion process, characterized by the constant infinitesimal moments

$$A_1(x) = \mu, \quad \mu \in \mathbb{R} \quad (9)$$

$$A_2(x) = \sigma^2, \quad \sigma \in \mathbb{R}^+ \quad (10)$$

As is well known, in this case the transition pdf (Equation 3) is normal, with mean $\mu t + x_0$ and variance $\sigma^2 t$. By means of the methods outlined, for instance, in Ricciardi (1977), one can prove that the FPT pdf of such a process is given by

$$g(S, t|x_0) = \frac{S - x_0}{\sigma \sqrt{2\pi t^3}} \exp \left[-\frac{(S - x_0 - \mu t)^2}{2\sigma^2 t} \right] \quad x_0 < S \quad (11)$$

which in statistical literature is known as the Inverse Gaussian distribution. For $\mu \geq 0$, neuronal firing is a sure event, since the function in Equation 11 is normalized to unity. If one takes $\mu < 0$, the above FPT pdf can be interpreted as the firing pdf conditional on the event “firing occurs.” The case $\mu = 0$ is also of interest, since Equation 11 expresses a so-called “stable law.” This case provides an interpretation of numerous experimental results indicating that the shapes of histograms are sometimes preserved when the adjacent interspike intervals are summed. Making use of Equation 11, the moments of the density $g(S, t|x_0)$, i.e., the moments of the interspike intervals (that exist for all $\mu > 0$), can be calculated (Ricciardi et al., 1999).

To provide an interpretation of the model based on assumptions formulated in Equations 9 and 10, let us imagine that the neuron's membrane potential undergoes a simple random walk under the effect of excitatory and inhibitory synaptic actions. For simplicity, let us assume that the neuronal dynamics develops on a discrete time scale $0, \tau, 2\tau, \dots$, with $\tau > 0$ as an arbitrary time unit. Passing to the limit as $\tau \rightarrow 0$, it can be shown (Ricciardi, 1977) that the random walk can be made to converge to the diffusion process having drift μ and infinitesimal variance σ^2 , in other words, to a Wiener process. This procedure provides the simplest example in which the Wiener model for neuronal activity can be constructed.

The assumptions underlying this model are undoubtedly oversimplified, as some well-known electrophysiological properties of neuronal membrane are not taken into account. Therefore, the Wiener process should be viewed as a statistical descriptor of data

rather than as a biologically sound model. However, it must be stressed that the fittings of some experimental data by the FPT pdf (Equation 11) is truly remarkable (see Gerstein and Mandelbrot, 1964).

We should finally point out that if a time-varying threshold is introduced to account for relative refractoriness—which is mainly relevant in high firing rate conditions—the determination of the firing pdf for the Wiener neuronal model cannot be generally accomplished analytically. Hence, ad hoc numerical methods had to be devised.

Ornstein-Uhlenbeck Process

Despite the excellent fitting of some data, the neuronal model based on the Wiener process has been the object of various criticisms, chiefly because it does not include the well-known spontaneous decay of the membrane depolarization. A diffusion model that includes this specific feature is the Ornstein-Uhlenbeck (OU) neuronal model. This is defined as the diffusion process characterized by the following drift

$$A_1(x) = -\frac{x}{\vartheta} + \mu \quad (12)$$

and by the constant infinitesimal variance given in Equation 10. The new parameter $\vartheta > 0$ in Equation 12 reflects the required decay (the membrane time constant). This model is referred to as the “leaky integrator stochastic neuronal model.” Comparing Equations 12 and 9, we see that now the drift is state dependent. However, in the limit as $\vartheta \rightarrow \infty$, the moments of Equation 12 are identical to those in Equation 9, meaning that the OU model yields the Wiener (or perfect integrator) model in the limit of an infinitely large time constant. Recalling Equations 1, 7, and 8, we can interpret the OU model as generated by the following SDE:

$$dX(t) = \left[-\frac{X(t)}{\vartheta} + \mu \right] dt + \sigma dW(t), \quad P\{X(t_0) = x_0\} = 1 \quad (13)$$

where $W(t)$ is the standard Wiener process.

Equation 13 can be taken as defining the OU model. However, such a model can also be obtained from first principles by using the formalism of diffusion equations. To this end, let us initially assume that the neuron is subject to a sequence of excitatory and inhibitory postsynaptic potentials of constant magnitudes $e > 0$ and $i < 0$ occurring in time in accordance with two independent Poisson processes of parameters α_e and α_i , respectively. The membrane potential is thus viewed as a stochastic process $X(t)$ in continuous time with a discrete space consisting of the lattice $x_0 + ki + he$ ($h, k = 0, \pm 1, \dots$) with the points of discontinuity randomly occurring in time. Again, let x_0 ($x_0 < S$) denote the fixed initial depolarization at which the sample paths start at the fixed initial time $t_0 = 0$. The firing pdf is thus modeled as the pdf of the instants when the sample paths for the first time reach, or cross in the upward direction, the threshold S . In Ricciardi et al. (1999) it is shown that if the input rates are made larger and larger and the postsynaptic potentials are simultaneously made smaller and smaller (with a suitable constraint), the membrane potential “converges” to the diffusion process having infinitesimal moments given by Equations 10 and 12 with $\mu = 0$. The case $\mu \neq 0$ can be obtained by a slightly more complicated model in which multiple Poisson-distributed excitatory and inhibitory inputs impinge on the neuronal membrane. Equation 13, in turn, can be obtained by a limit procedure starting from the differential equation expressing the spontaneous exponential decay of the membrane depolarization after including in it excitatory and inhibitory Poisson-distributed independent input processes.

Solving either the FP equation (Equation 4) or the K equation (Equation 5) with parameters of Equations 10 and 12 under the

appropriate delta-initial conditions, and then setting $t_0 = 0$ (since the process is temporally homogeneous), one obtains (cf. Ricciardi and Sacerdote, 1979):

$$f(x, t|x_0) = [2\pi V(t)]^{-1/2} \exp \left\{ -\frac{[x - M(t|x_0)]^2}{V(t)} \right\} \quad (14)$$

Hence, at each time t the transition pdf (Equation 14) is normal with mean $M(t|x_0) = \mu \vartheta [1 - \exp(-t/\vartheta)] + x_0 \exp(-t/\vartheta)$ and variance $V(t) = \sigma^2 \vartheta [1 - \exp(-2t/\vartheta)]/2$.

It must be pointed out that the OU model differs from the Wiener model in several respects. First, an equilibrium regime exists, since in the limit as $t \rightarrow \infty$, the pdf (Equation 14) becomes normal with mean $\mu \vartheta$ and variance $\sigma^2 \vartheta/2$. Furthermore, attainment of the firing threshold is a sure event, irrespective of the value of μ . However, in contrast to the Wiener model, the FPT problem is in general very complicated, even in the case of constant thresholds.

Two distinct firing regimes can be established for the OU model. In the first one, the asymptotic mean depolarization $\lim_{t \rightarrow \infty} M(t|x_0) = \mu \vartheta$ is far above the firing threshold S , and the interspike intervals are relatively regular (deterministic firing). In the second one, it is $\mu \vartheta \ll S$, and firing is caused only by random fluctuations of the depolarization (stochastic or Poissonian firing). The term Poissonian firing is motivated by the circumstance that as the thresholds move farther and farther above the steady-state depolarization $\mu \vartheta$, the firing patterns achieve the characteristics of a Poisson point process.

Equally important to the construction of a model is its verification. Although the Wiener model has often been tested by with experimental data, only relatively recently has the OU model been systematically fitted to interspike histograms (Inoue et al., 1995), owing to the very complicated and cumbersome procedure required for estimating the parameters of the model from interspike intervals data. The available—and as yet unexploited—methods for estimating parameters are based on intracellular recording of the sample paths of $X(t)$ (Lánský, 1983).

The models encountered in the application of diffusion processes to theoretical neuroscience have been predominantly time homogeneous, as reflected in the fact that the functions μ and σ in Equation 1 do not depend on t . Recently, however, an interest in stochastic resonance (a cooperative effect that may arise out of the coupling between deterministic and random dynamics in a nonlinear system) has motivated studies on diffusion neuronal models with time-dependent parameters. As already mentioned, in the OU model two distinct regimes can be identified, deterministic and Poissonian firing. In the first regime, the signal (μ term) is large enough that firing events occur even in the absence of noise. The noise-activated regime corresponds to the situation in which the drift term alone is not sufficient to cause firings, which are instead induced by the noise. The methods of stochastic resonance extend this view mainly to subthreshold periodic signals. Two sources of periodicity can be expected in the signal: either an *endogenous periodicity* or a periodicity of input intensities, the *exogenous periodicity* (Lánský, 1997). Both these instances are included in the following model of the membrane depolarization:

$$dX(t) = \left[-\frac{X(t)}{\vartheta} + \mu + A \cos \omega t \right] dt + \sigma dW, \quad P\{X(t_i) = x_0\} = 1, \quad t \geq t_i \quad (15)$$

where the notation is the same as in Equation 13, $A > 0$ is a constant characterizing the amplitude of the input signal, t_i is the time of the last release of an action potential, and ω is the angular frequency of the driving force modulation ($2\pi/\omega$ is the modulation period). After each firing the membrane potential is instantly reset to its initial value x_0 . For exogenous periodicity, the phase of the signal

continues after a spike, while in the endogenous case it is always reset, which simplifies the calculations. In the case of endogenous periodicity, the intervals between firing form a renewal process, which quite naturally leads us to consider an FPT problem for the OU process (Equation 13) and a periodic boundary. It is intuitive that by a suitable transformation, an FPT problem for the process modeled by Equation 15 through a constant threshold S can be changed into an FPT problem for a time-independent OU model through a periodic boundary. For exogenous periodicity a different method has to be invoked. In both cases there exists an optimum level σ of noise for which the input frequency ω is best reflected in the output signal (Bulsara et al., 1996; Shimokawa, Pakdaman, and Sato, 1999; Plesser and Geisel, 2001).

It should not pass unnoticed that the deterministic versions of Equations 13 and 15, obtained in the limit $\sigma \rightarrow 0$, have long been known in the literature as the Lapicque model, or *leaky integrator* neuronal model (Tuckwell, 1988), presently revived within the context of artificial neural networks.

We conclude this section by pointing out that, especially for experimental purposes, it is sufficient to obtain information only on shape and location of the firing pdf, without finer details. This can be achieved via some of the moments of the firing time. The knowledge of the moments can also, at times, provide some extra valuable information. For example, after a systematic computations of mean and variance of the firing time and of the skewness (a measure of the deviation from symmetry of the firing pdf expressed as the ratio of the third-order central moment to the cube of the standard deviation), for a variety of thresholds and initial states a striking and unsuspected feature of the OU model has emerged: for boundaries of the order of a couple of units or more above the steady-state depolarization $\mu\vartheta$, the variance of the firing time equals the square of its mean value, to an excellent degree of approximation. Moreover, the skewness is equal to 2 to a very high precision. Finally, the goodness of these approximations increases with increasing values of the threshold. When all this information was put together, the conjecture emerged that the firing pdf is susceptible to an excellent exponential approximation for a wide range of thresholds. In addition, these “experimental observations” have led to quantitative results concerning the asymptotic exponential behavior of the FPT pdf not only for the OU process, but also for a wider class of diffusion processes, both for constant and for time-varying boundaries (see Ricciardi et al., 1999, and references therein).

As expected, the convergence to the exponential distribution for increasing thresholds is accompanied by a large increase in the mean firing time. This is in agreement with the finding that for some neurons, the histograms of the recorded interspike intervals are increasingly better fitted by exponential functions as the firing rates decrease.

Neuronal Models with Multiplicative Noise

In order to embody additional physiological features of real neurons, several alternative models have been proposed. It is well known (and as reflected in the Hodgkin-Huxley model), the change in membrane depolarization caused by a synaptic input depends on its current value. Basically, the depolarization of the membrane caused by an excitatory postsynaptic potential decreases linearly with decreasing distance of the membrane potential from the excitatory reversal potential, V_E . In the same manner, the hyperpolarization caused by an inhibitory postsynaptic potential is smaller if the membrane potential is closer to the inhibitory reversal potential, V_I . In this way the depolarization $X(t)$ is confined to the interval (V_I, V_E) , whereas in the models presented in the preceding section it was considered unrestricted. Inequalities $V_I < x_0 < S <$

V_E express obvious conditions relating reversal potentials, initial depolarization, and firing threshold. Although the only diffusion model of linear summation (with exponential decay) of the input signal is the OU process (Equation 13), there is a whole class of diffusion processes that are appropriate if reversal potentials are taken into account. In all these models the drift takes the form

$$A_1(x) = -\frac{x}{\vartheta} + \mu_1(V_E - x) + \mu_2(x - V_I) \quad (16)$$

where the parameter $\vartheta > 0$ is the membrane time constant, as in Equation 12, and $\mu_1 > 0$, $\mu_2 < 0$ are new parameters reflecting excitation and inhibition, now separately specified, in contrast to the case of Equation 12. Although both moments given by Equations 12 and 16 are linear, the interpretations underlying the two drifts are significantly different. Indeed, in Equation 12 there is a constant “leakage term” ϑ^{-1} , while for the models with reversal potentials the leakage ($\vartheta^{-1} + \mu_1 - \mu_2$) is explicitly input dependent.

The diffusion models that take into account the existence of the reversal potentials always lead to a multiplicative noise effect. This is in agreement with the general notion that an additive noise is generated by events independent of the transmitted message, whereas the multiplicative noise arises inside the processing unit, viz., inside the system. Common forms of infinitesimal variance in models with reversal potentials are the following:

$$A_2(x) = \sigma^2(x - V_I) \quad (17)$$

$$A_2(x) = \sigma^2(x - V_I)(V_E - x) \quad (18)$$

$$A_2(x) = \sigma_1^2(x - V_E)^2 + \sigma_2^2(x - V_I)^2 \quad (19)$$

The first one stresses the importance of the inhibitory reversal potential, which restricts the state space of $X(t)$ from below; the second and third form of the infinitesimal variance attribute equal relevance to both reversal potentials. The main difference between infinitesimal variances in Equations 17 and 18, on the one hand, and Equation 19 on the other lies in the fact that for model 19, the reversal potentials lose their natural role of boundaries of the depolarization (Hanson and Tuckwell, 1983). This fact strongly handicaps model 19. In models characterized by Equations 17 and 18, the behavior at the boundaries V_I and V_E is mathematically rather subtle.

Similarly to the OU model, the statistical moments of T can be calculated for the models defined by Equations 17 and 18. As for the OU process, here also the asymptotic exponentiality holds for low excitation levels. Thus, the only (though very substantial) effect stemming from the models we have considered with reversal potentials consists of the parameter interpretation. From the modeling point of view, the variety of forms identified for the infinitesimal variance and the linear form of the drift are not unexpected. Indeed, these models are meant to reflect, through an “equivalent” noisy ordinary differential equation, the properties, at a single location, of a spatially distributed neuron with noisy inputs, thus corresponding to a stochastic partial differential equation. The linear drift describes the passive electrical circuit properties of the membrane at the trigger zone and the mean effect of the noisy input. The infinitesimal variance, on the other hand, must take into account not only the diversity of spatial configurations for different neurons, but the location and type of synaptic input on that neuron as well. Hence, a variety of forms for this term in the diffusion equation are conceivably appropriate.

It must be emphasized that for neuronal diffusion models originating from SDEs that include either additive or multiplicative noise terms, the FPT problem is in general intractable with analytical tools. However, some efficient procedures have been devised to obtain accurate numerical evaluations of g for the general case

of time-varying firing thresholds and for arbitrary diffusion models (not necessarily of the OU type). This is an important target, because to calculate the unknown FPT pdf, one would have to solve Equations 4 or 5 under the appropriate initial delta conditions, and usually, as well, in the presence of complicated boundary conditions when the resulting equations are singular—a very difficult task that rarely leads to analytical solutions. Efficient numerical algorithms are thus especially desirable if one has to deal with time-varying neuronal thresholds. The standard procedure is based on the remark that the FPT pdf $g[S(t), t|x_0, t_0]$ can be proved to be a solution of the following integral equation:

$$f[S(t), t|x_0, t_0] = \int_0^t f[S(t), tS(\tau), \tau]g[S(\tau), \tau|x_0, t_0]d\tau \quad (20)$$

with $x_0 < S(t_0)$. This is a first-kind Volterra integral equation whose solution is made complicated by the circumstance that the kernel $f[S(t), tS(\tau), \tau]$ exhibits a singularity of the type $1/\sqrt{t - \tau}$ as $\tau \uparrow t$. Hence, the problem of determining $g[S(t), t|x_0, t_0]$ from Equation 20 via numerical methods is by no means trivial. Furthermore, all classic available algorithms necessitate the use of large computation facilities and sophisticated library programs. As a consequence, they are expensive to run and not suitable for suggesting to the modeler in real time how to identify the various parameters to fit the recorded data.

An entirely different approach that is quite general, although specifically useful for handling neuronal firing problems, was therefore developed. The guiding idea is to prove that the singular Equation 20 can be replaced by a nonsingular second-kind Volterra integral equation for g that possesses an extra degree of freedom, which can be used to remove the singularity of the kernel or, in some instances, to directly obtain a closed-form expression for $g[S(t), t|x_0, t_0]$. (See Ricciardi et al., 1999, and references therein for a description of the method and of related computational algorithms.)

Discussion

In this article we have outlined a few stochastic models for single neurons' activity based on the theory of diffusion processes. As we have sketched out, prediction of the firing pdf is a difficult problem whose solution can usually be approached only by numerical or simulation procedures. It must be stressed that the neuronal behavior described by diffusion models ultimately implies that for time-constant input, the neuron's output is a renewal process. However, one can conceive models aimed, for instance, at simulating the clustering effect in spike generation. Serial dependence among interspike intervals can be modeled in various ways, for instance by adjusting the reset value after each spike in the OU process. Another possibility consists of the inclusion of some kind of feedback in the model, usually the often experimentally observed *self-inhibition*. A further generalization is achieved by taking into account the spatial properties of a neuron. In the simplest way, this can be done by assuming that the model neuron consists of two compartments: (1) the dendritic compartment, described by a standard diffusion model, and (2) the trigger-zone compartment, including the spontaneous decay of depolarization and the firing mechanism (Lánský and Rodríguez, 1999).

To conclude this bird's-eye view of the topic, we would like to mention an alternative approach to the construction of diffusion models that are able to fit experimental data: reversing the problem. That is, instead of formulating a neuronal model $X(t)$ based on some reasonable assumptions, and then trying to compute the firing pdf as an FPT pdf through a preassigned threshold, one can proceed as follows. First, construct the histogram of the experimentally recorded spike train and try to fit it by a function $g(S, t|x_0)$, with S

and x_0 parameters to be determined by the standard methods. Once this task has been accomplished, ask the following questions: Can $g(S, t|x_0)$ be viewed as the FPT pdf, through the threshold S , of a diffusion process conditioned on $X(0) = x_0$? If the answer is yes, can such a process be uniquely determined? This is, so to speak, the inverse of the FPT problem. Quite surprisingly, a precise answer to this question can be provided, at least in principle (Capocelli and Ricciardi, 1972).

In summary, for modeling purposes we have essentially looked at a neuron as a black box, characterized by an input and an output, for which two distinct problems can be posed: (1) to determine the output for a given input, which has led us to an FPT problem, and (2) to guess the input by analyzing the output, which could be viewed as the *inverse* of an FPT problem. In both cases the class of input functions had to be specified beforehand in order to make these problems mathematically sound. The need for such a specification, in conjunction with a large body of well-known experimental data, has led us to assume that input stimulations and random components are such that the neuron's membrane potential is ultimately modeled by one-dimensional diffusion processes. A sketch of some of the ensuing diffusion models for neurons' activity, and their place in the biological literature, was the object of this article.

Road Map: Biological Neurons and Synapses

Background: Single-Cell Models

Related Reading: Rate Coding and Signal Processing

References

- Bulsara, A. R., Elston, T. C., Doering, C. R., Lowen, S. B., and Lindberg, K., 1996, Cooperative behavior in periodically driven noisy integrate-and-fire models of neuronal dynamics, *Phys. Rev. E*, 53:3958–3969.
- Capocelli, R. M., and Ricciardi, L. M., 1972, On the inverse of the first passage time probability problem, *J. Appl. Prob.*, 9:270–287.
- Gerstein, G. L., and Mandelbrot, B., 1964, Random walk models for the spike activity of a single neuron, *Biophys. J.*, 4:41–68.
- Hanson, F. B., and Tuckwell, H. C., 1983, Diffusion approximations for neuronal activity including synaptic reversal potentials, *J. Theor. Neurobiol.*, 2:127–153.
- Inoue, J., Sato, S., and Ricciardi, L. M., 1995, On the parameter estimation for diffusion models of single neurons' activities, *Biol. Cybern.* 73:209–221.
- Lánský, P., 1983, Inference for diffusion models of neuronal activity, *Math. Biosci.*, 67:247–260. ♦
- Lánský, P., 1997, Sources of periodical force in noisy integrate-and-fire models of neuronal dynamics, *Phys. Rev. E*, 55:2040–2044.
- Lánský, P., and Rodríguez, R., 1999, Two-compartment stochastic model of a neuron, *Physica D: Nonlinear Phenomena*, 132:267–286.
- Lánský, P., and Sato, S., 1999, The stochastic diffusion models of nerve membrane depolarization and interspike interval generation, *J. Periph. Nerv. Syst.*, 4:27–42. ♦
- Plesser, H. E., and Geisel, T., 2001, Stochastic resonance in neuron model: Endogenous stimulation revisited, *Phys. Rev. E*, 63:Article 031916 (6 pages).
- Ricciardi, L. M., 1977, *Diffusion Processes and Related Topics in Biology, Lecture Notes in Biomathematics*, Berlin: Springer-Verlag. ♦
- Ricciardi, L. M., Di Crescenzo, A., Giorno, V., and Nobile, A. G., 1999, Theoretical and algorithmic approaches to first passage time problems, *Math. Japonica*, 50:247–322. ♦
- Ricciardi, L. M., and Sacerdote, L., 1979, The Ornstein-Uhlenbeck process as a model for neuronal activity, *Biol. Cybern.*, 35:1–9.
- Ricciardi, L. M., and Sato, S., 1989, Diffusion processes and first-passage-time problems, in *Lectures in Applied Mathematics and Informatics* (L. M. Ricciardi, Ed.), Manchester, Engl.: Manchester University Press, pp. 206–285.
- Shimokawa, T., Pakdaman, K., and Sato S., 1999, Time-scale matching in the response of a leaky integrate-and-fire neuron model to periodic stimulus with additive noise, *Phys. Rev. E*, 59:3427–3443.
- Tuckwell, H. C., 1988, *Introduction to Theoretical Neurobiology*, Cambridge: Engl.: Cambridge University Press. ♦

Digital VLSI for Neural Networks

Dan Hammerstrom

Introduction

The computational overhead required to simulate artificial neural network (ANN) models, whether simplistic or realistically complex, is a key problem in the field because of the computational complexity of these models. Network simulations are required both for research and for commercial products. Most researchers currently perform these simulations on standard computer technology, such as high-end workstations or personal computers. However, as the field progresses, researchers are moving to larger and ever more complex models that challenge even the fastest computers.

A reasonably realistic neural model could approach one million neurons and tens of billions of connections, where a “connection” is a data transfer path between two neurons. In addition to size, the models themselves are becoming more complex as we move from simple inner products to spiking neurons with temporal time course that require a convolution to be performed at each synapse.

For these reasons, there has been much interest in developing custom hardware for ANNs. The inherent parallelism in ANN and connectionist models suggests an opportunity to speed up the simulations. Their simple, low-precision computations also suggest an opportunity to employ simpler and cheaper, low-precision digital hardware implemented by full-custom silicon or by field-programmable gate arrays (FPGAs).

This chapter discusses digital electronic implementations of ANNs. First, we look at the differences between digital and analog design techniques with a focus on performance/cost trade-offs. Second, we consider the use of traditional processors in parallel configurations for ANN emulation. Third, to convey a sense of some of the issues involved in designing digital structures for ANN emulation, a custom digital ANN processor is discussed: the Adaptive Solutions CNAPS. Although this chip is no longer produced, it is still being used. Its simple architecture makes it a good vehicle to understand the trade-offs inherent in emulating neural structures digitally. Fourth, we look briefly at FPGA technology as a promising alternative for digital implementation of ANNs.

Why Digital?

Performance/Cost

One commonly held belief in the ANN research community is that analog computation, in which signals are transmitted and manipulated as strengths, generally voltage or current, is inherently superior to digital computation, in which signals are transmitted and manipulated as serial or parallel streams of 1s and 0s. But in fact, both technologies have advantages and disadvantages. The best choice depends on a variety of factors.

Why is analog appealing? An important reason is that it provides 10–100 times greater computational density than digital computation. *Computational density*—the amount of computation per unit area of silicon—is important because the cost of a chip is generally proportional to its total area. In analog circuitry, complex, nonlinear operations such as multiply, divide, and hyperbolic tangent can be performed by a handful of transistors. Digital circuitry requires hundreds or even thousands of wires and transistors to perform the same operations. Analog computation also performs these operations using far less power per computation than digital computation.

If analog is so good, why are people still building digital chips, and why are most commercial products digital? One important reason is familiarity. People know how to build digital circuits, and

they can do it reliably, no matter the size and complexity of the system. This is partly the legacy of having thousands of digital designers all over the world constantly tweaking and improving design techniques and software. It is also easier to create a digital version of a computation, in which a computer program represents the algorithm, than an analog version, in which the circuit itself represents the algorithm. This is particularly true if you are trying to build a system that is robust and reliable over the wide temperature and voltage ranges needed in commercial products. Analog design is an uncommon capability, and it is becoming less common as people find that they can do more with digital circuitry. For example, digital signal processors now perform most of what was once the domain of analog circuitry. Another advantage of the digital emulation of neural networks is that it significantly eases the integration of the neural network portion of the design with the larger digital system to which it connects.

Flexibility

Another factor working in favor of digital is that analog designs are generally algorithms wired into silicon. Such designs are inflexible. Though there is an interesting class of designs that are programmable analog. Perhaps the most powerful and widely studied is the Cellular Neural Network (Chua and Roska, 2001).

Digital designs can be either hardwired or programmable. Their flexibility is a major benefit, since it allows software control as well as an arbitrary level of precision (low to high and fixed or floating point). The price of this flexibility is reduced performance/cost, but the result is a chip that can solve a larger part of a problem. It also leads to a device that has broader applicability and can track incremental algorithm improvements by changing the software, not by redesigning the circuitry.

The role flexibility plays in system performance/cost can be understood more clearly by examining *Amdahl's law* (Hennessy and Patterson, 1991) that describes the execution time benefits of parallelizing a computation. Briefly, a computing task has portions or *subtasks* that often can be executed in parallel. Other, sequential tasks cannot begin until a previous task has been completed, which forces a sequential ordering of these tasks.

Amdahl's law states that no matter how many processors are available to execute subtasks, the speed of a particular task is roughly proportional to the subtasks that cannot be executed in parallel. In other words, sequential computation dominates as parallelism increases. Amdahl quantifies the relationship as follows:

$$S = 1/(op_s + (op_p/p))$$

where S is the total speed-up, op_s is the number of operations in the serial portion of the computation, op_p is the number of operations in the parallel portion, and p is the number of processors. As p gets large, S approaches $1/op_p$.

For example, suppose we have two chips to choose from. The first can perform 80% of the computation with a $20\times$ speed-up on that 80%. The second can perform only 20% of the computation but executes that 20% with a $1000\times$ total system speed-up. Plugging into the equation, the first chip gives us a total speed up of over $4\times$, while the second—and “faster”—chip has only a $1.25\times$ total system speed-up. A programmable device that accelerates several phases of an application generally offers a much larger benefit than a dedicated device.

Below we discuss FPGAs, which are flexible to the point of allowing the arbitrary configuration of physical digital circuitry.

They are a promising approach to efficiently implementing the inherent parallelism of neural-like structures.

Signal Intercommunication

One difference between silicon and biological networks is that for silicon, internode communication is relatively more expensive than for biological systems. Although several levels of wire interconnect (8–10 today) are available in most silicon processes, each level is restricted to two-dimensional interconnection because wires on the same level cannot pass over or touch one another.

Two-dimensional layout and large expensive wires require us to modify our biologically derived computational models to more closely match the strengths and weaknesses of the implementation technology. To show the need for such modifications, Bailey and Hammerstrom (1988) modeled a hypothetical neural circuit. This circuit, modest by biological standards, had one million “neurons” with one thousand inputs each, or one billion connections total.

The first calculation assumed a direct implementation—that is, one connection per wire. This billion-connection ANN required tens of square meters of silicon for dedicated communication pathways. Since silicon averages tens of dollars per square centimeter, such a system is too costly to be practical. These costs result from a theorem showing that the metal area required by direct communication is proportional to the *cube* of the fan-in or “convergence” at each node.

Their second calculation assumed a multiplexed interconnect structure, one in which several connections shared a metal wire. Wire multiplexing adds complexity at each end. Likewise, an address must be sent with each data packet to identify the sender, and some decoding must be performed on that address. Bailey and Hammerstrom (1988) showed that with the proper communication architecture, a $100\times$ reduction in silicon area over the direct approach was possible with little impact on performance. Since only a few nodes will be active in any given time interval for these large networks, multiplexing interconnect makes even more sense.

Even analog designers of neuromorphic circuitry have recognized the need for multiplexed interconnect. However, analog voltages and currents are difficult to multiplex. One alternative is to represent analog values by using pulses. There are several ways in which pulses can be used to represent information, including pulse rate, phase, and interpulse interval. It is possible for different pulse streams to share a single wire by sending, at the time the pulse occurs, the address of the pulse stream. This approach is called address event representation, or AER (Boahen, 2000). Pulse or “spike” signal representation is also much more neurobiologically plausible.

Digital Neural Networks: Off-the-Shelf Processors

One successful approach to high-speed ANN simulation has been to use arrays of commercial microprocessors. This approach works because desktop machines, thanks to Moore’s law, have achieved a tremendous level of performance/cost. Moore’s law states that the number of transistors that can be manufactured economically on a single silicon die doubles every 24 months. Moore’s law has held constant for roughly 32 doublings, which is truly impressive. There are not many industries that can claim exponential growth over such a long period.

In addition to raw clock speed, another effect of Moore’s law is that more transistors are available to dedicate to specialized functionality. Today, the latest microprocessors offer on-board single instruction multiple data (SIMD) parallel coprocessors. For Intel, these coprocessors have evolved from MMx to SSE (Pentium III) and now to SSE2 (Pentium 4) (Intel, 2001). The Motorola/IBM PowerPC has the similar AltiVec system. Although these copro-

cessors have been designed primarily for basic image processing, video codecs, and graphics, they can also be used to emulate certain ANN models.

A problem these machines have, though, is limited memory bandwidth. Most applications have a fair amount of referencing locality, in which a collection of physically contiguous addresses are referenced multiple times. Reference locality allows the processor to use several layers of cache memory (the Pentium IV has three). However, neural network algorithms typically require that an entire network be accessed for each state update. Since this network can be very large, it generally does not fit in the caches. Consequently, there is a significant slowdown as the processor ends up waiting for data from memory.

Perhaps the best approach is to use a commercial multiprocessor computer that hides the memory bandwidth problems by providing large numbers of processors. For example, the NASA Ames Research Center has several large Silicon Graphics parallel machines (Shan et al., 2000). The largest currently has 1024-processors. These systems use very high-speed interconnects and are able to emulate large, complex neural network structures. Our research group at OGI has simulated simple association networks approaching one million nodes on this machine.

However, such computational power is typically not available to the average researcher. One popular alternative has been to build large computer clusters using relatively inexpensive PCs. Often known as Beowulf clusters (Reschke et al., 1996), these systems connect large numbers of simple processors and are typically built from off-the-shelf hardware (PCs and LAN switches). The software is usually free. Programming is done by using traditional languages and MPI (the message-passing interface) or PVM (parallel virtual machine). Unfortunately, the interprocessor communication tends to be fairly slow relative to the computation, which compromises the total speed-up to some degree. However, they can be fairly efficient if complex models of the neuron are used that require more computation than intercommunication.

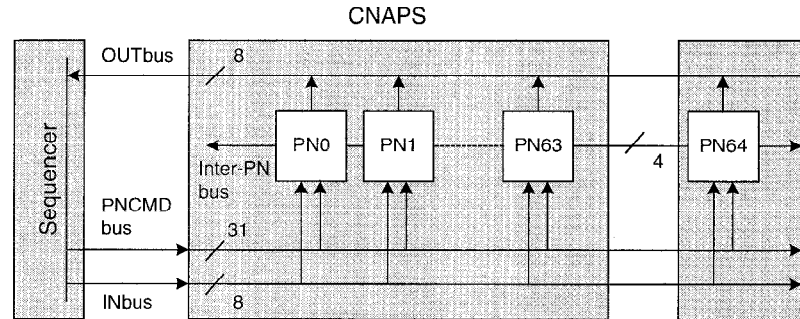
As neural network models become larger and more complex, the model connectivity issues become a major factor in the speed of emulation regardless of the hardware platform. Real neural structures demonstrate *sparseness*, that is, a small subset of all possible connections are actually made, and *locality*, that is, there is a higher probability of connections to neurons that are physically near each other. However, ANN models have typically not exhibited significant sparseness or locality, which is another reason for researchers to study more biologically plausible systems so that we can create structures that are computationally robust and have sparse, localized connections.

Digital Neural Networks: Full Custom Processors: CNAPS

Designing specialized architectures that are customized for ANN simulation permits significant improvements in performance/cost, since the processors and their interconnection architecture are optimized for the computations they perform. This section discusses the Adaptive Solutions CNAPS architecture, which was, for a time, a successful commercial product but is no longer produced. It represents the specialized functionality end of the design spectrum of digital chips.

The CNAPS architecture (Hammerstrom, 1995) had multiple processor nodes (PNs) connected in a one-dimensional structure, forming a SIMD array (Figure 1). SIMD architectures have one instruction storage and decode unit and many execution units, simplifying system design and reducing costs. Unlike a PC cluster, each CNAPS PN did not have program storage and sequencing hardware, and each executed the same instruction each clock. Node

Figure 1. CNAPS architecture. This is a single instruction multiple data (SIMD) architecture, in which all processor nodes (PNs) execute the same instruction during each clock. There is a single broadcast data bus that allows efficient one-to-many and many-to-many communication.



outputs were broadcast from each PN to all the others over a single broadcast bus.

Another major simplification of the CNAPS architecture, which is found in other digital ANN chips, was the use of limited-precision, fixed-point arithmetic. Many researchers have shown that floating point and high precision are unnecessary in ANN simulation (Fahlman and Hoehfeld, 1992). CNAPS supported 1-, 8-, and 16-bit precision in hardware. Consequently, the PNs were smaller and cheaper. This reduced precision was more than adequate for the applications implemented on CNAPS.

The CNAPS architecture had 64 PNs per chip. At the then frequency of 25 MHz, each chip executed at a rate of 1.6 billion connections computed per second. A single chip executed back-propagation learning at a rate of 300 million connection updates per second; each update consists of reading the weight associated with the connection, modifying it, and then writing it back. Each PN (Figure 2) had 4096 bytes of on-chip local memory, used to store synaptic weight data and other local values. Hence, a 64 PN

chip could store up to 256 KB of information. Multiple chips could be combined to create larger, more powerful systems. The general programmability of the device allowed it to execute a large range of functions, including many non-ANN algorithms such as the discrete Fourier transform, nearest neighbor classification, image processing, and dynamic time warping.

Figure 3 shows a simple two layer network mapped to a CNAPS array. The network nodes are labeled CN0–CN7; the processor nodes are labeled PN0–PN3. Multiple network nodes map to a single processor node; in this example, one node from each layer is mapped to a single PN. For feedforward calculation, assume that the outputs of nodes CN0–CN3 have been computed. To compute the inner product of nodes CN4–CN7, the output value of node CN0 is broadcast on the bus to all PNs in the first clock. Each PN then multiplies the CN0 output with the corresponding weight element, which is different for each PN. On the next clock, CN1's output is broadcast, and so on. After four clocks, all 16 products have been computed: $O(n^2)$ connections in $O(n)$ time.

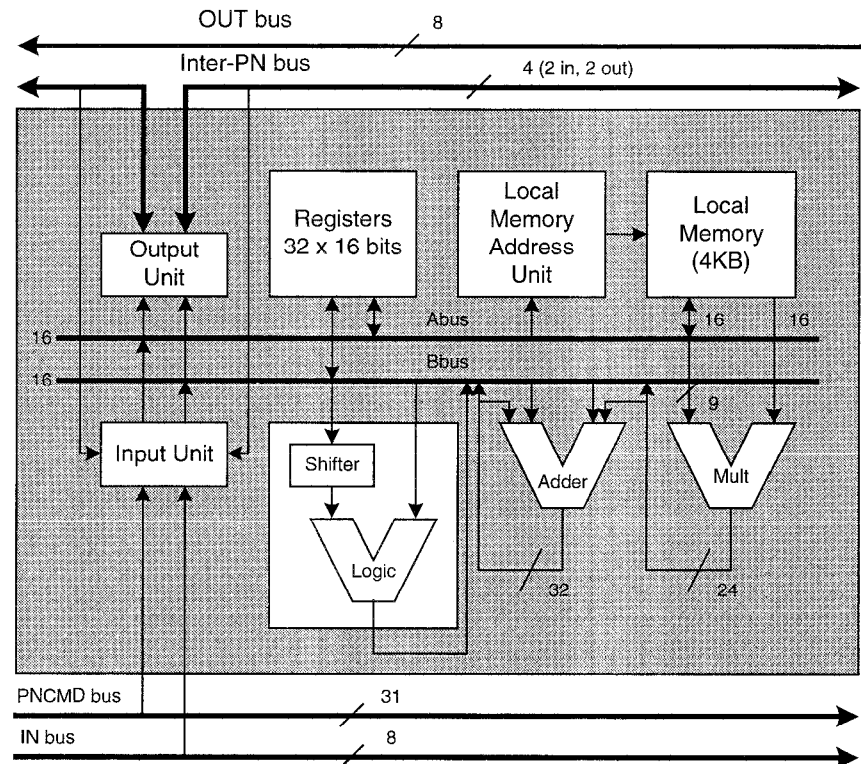


Figure 2. CNAPS PN architecture. A single PN has a multiplier, accumulator, logic/shifter unit, register file, and separate memory address adder. Each PN also has its own memory for storing weights, lookup tables, and other data. Each PN generates its own unique address to memory.

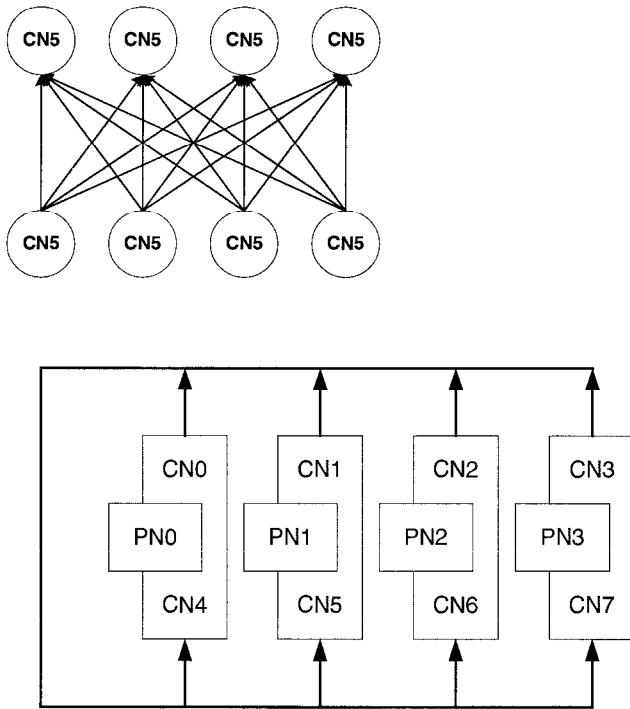


Figure 3. Mapping of a simple two-layer feedforward network to the CNAPS array. When emulating a feedforward network, each layer is spread across the PN array. The neuron outputs of one layer are broadcast sequentially to all PNs while they compute the multiply accumulations for the next layer of neurons.

Digital Neural Networks: Field Programmable Gate Arrays

Perhaps the most promising approach to emulating neural models digitally is the FPGA (Sharma, 1998). Briefly, an FPGA is a device with a large number of generic logic blocks and generic interconnect between those blocks. The functions that the logic blocks implement and how these blocks are connected to one another are determined by configuration bits that are loaded into the chip as one would load a program into a computer's memory. Because of Moore's law, it is now possible to buy FPGAs that are capable of emulating millions of logic gates at frequencies approaching several hundred megahertz. There are very sophisticated design tools that allow logic to be expressed in a high-level hardware description language and then be converted to FPGA configuration bits by an automated synthesis process. These devices can implement large neural structures in parallel (see, for example, Hatano et al., 1999).

Although FPGAs are appearing with larger on-chip memory, they still cannot approach the density of commercial DRAM. So for emulating very large networks, off-chip memory needs to be used to store the various parameters and state information associated with each neuron. However, unlike traditional processors, FPGAs are capable of supporting the access to several high-speed memory structures at once. Consequently, a board with several FPGAs could emulate networks at much higher speeds than a high-speed desktop PC. In addition, the inherent parallelism in each FPGA would allow parallel implementation of the various structures within the neuron, such as sophisticated spike-based computation.

Discussion

It is difficult to predict technology trends, but speculation is always possible. Today, most ANNs are used for pattern recognition. The final stage of most pattern recognition algorithms involves checking a series of classification results to see whether they fit in the larger context of the domain in question. Including this contextual knowledge can be as simple as spell checking, or it can be as complex as accessing high-order rules or schemas that reflect complex syntactical and semantic relationships. Since classification is imperfect, contextual processing, which makes knowledge of such higher-order relationships available to the classification process, is essential to guarantee the accuracy of the final result.

Although the results are still speculative, research (Ambros-Ingerson et al., 1990; Braitenberg and Schüz, 1998) has shown that scaling to large contexts requires networks with relatively sparse interconnect and sparse activation, in which only a few nodes are actively firing at a time. On the basis of research into VLSI connectivity (Bailey, 1993), digital-based systems can handle such networks more efficiently than analog. Therefore, at some point in the processing, the data will probably need to be converted from analog to digital representation. Today, the conversion is done at or just after the input transducer. On the basis of the state of analog technologies, systems of the future will probably take advantage of the computational density of analog VLSI to perform the feature extraction and some preliminary classification at the front end, with conversion to digital form for contextual processing and final classification by "higher-level brain regions."

Road Map: Implementation and Analysis

Related Reading: Analog VLSI Implementation of Neural Networks; Photonic Implementations of Neurobiologically Inspired Networks; Programmable Neurocomputing Systems

References

- Ambros-Ingerson, J., Granger, R., et al., 1990, Simulation of paleocortex performs hierarchical clustering, *Science*, 247:1344–1348.
- Bailey, J., 1993, A VLSI interconnect strategy for biologically inspired artificial neural networks, Ph.D. thesis, Department of Computer Science/Engineering, Oregon Graduate Institute, Beaverton, OR.
- Bailey, J., and Hammerstrom, D., 1988, Why VLSI implementations of associative VLCNs require connection multiplexing, *1988 International Conference on Neural Networks*, San Diego, CA, pp. 173–180.
- Boahen, K. A., 2000, Point-to-point connectivity between neuromorphic chips using address events, *IEEE Trans. Circuits and Systems II—Analog and Digital Signal Processing*, 47(5):416–434. ♦
- Braitenberg, V., and Schüz, A., 1998, *Cortex: Statistics and Geometry of Neuronal Connectivity*, New York: Springer-Verlag.
- Chua, L., and Roska, T., 2001, *Cellular Neural Networks and Visual Computing*, Cambridge, Engl.: Cambridge University Press.
- Fahlman, S. E., and Hoehfeld, M., 1992, Learning with limited numerical precision using the cascade-correlation algorithm, *IEEE Trans. Neural Networks*, 3(4):602–611.
- Hammerstrom, D., 1995, A digital VLSI architecture for real-world applications, in *An Introduction to Neural and Electronic Networks* (S. F. Zornetzer, J. L. Davis, C. Lau, and T. McKenna, Eds.), San Diego, CA: Academic Press, pp. 335–358. ♦
- Hatano, F., et al., 1999, Implementation of cell array neuro-processor by using FPGA, paper presented at IEEE International Joint Conference on Artificial Neural Networks, Washington, DC. ♦
- Hennessy, J. L., and Patterson, D. A., 1991, *Computer Architecture: A Quantitative Approach*, Palo Alto, CA: Morgan Kaufmann.
- Intel, 2001, *IA-32 Intel Architecture Software Developer's Manual*, vol. 1: *Basic Architecture*, Santa Clara, CA: Intel.
- Reschke, C., Sterling, T., et al., 1996, A design study of alternative network

topologies for the Beowulf parallel workstation, in *Proceedings of the IEEE International Symposium on High Performance Distributed Computing*, Piscataway, NJ: IEEE.

Shan, H., Singh, J. P., et al., 2000, A comparison of three programming

models for adaptive applications on the Origin2000, in *Proceedings of SC2000*, Dallas, TX.

Sharma, A. K., 1998, *Programmable Logic Handbook: PLDs, CPLDs & FPGAs*, New York: McGraw-Hill Handbooks.

Directional Selectivity

Norberto M. Grzywacz and David K. Merwine

Introduction

Directional selectivity refers to a neuron's ability to produce substantially different responses for stimulus motions of different direction. A directionally selective (DS) cell will fire many spikes in response to object motion in one direction (the preferred direction) while responding weakly, if at all, for motion in the opposite (null) direction. This directional "trigger feature" is often essentially independent of the contrast, contrast polarity, color, shape, or speed of the moving object. Cells displaying directional selectivity are found in the retinas and visual cortices of all the major vertebrate classes. These neurons support a host of visual tasks, ranging from motion perception, image segregation, and deblurring to the control of eye movements. The extraction of direction of motion is so crucial for vision that it is the first motion-related variable encoded in the visual pathway.

It is impossible, however, to determine the direction of a motion using an individual DS neuron. First, because these neurons have relatively small receptive fields, they can only report motion components that are perpendicular to the gradient of illumination. This phenomenon is known as the aperture problem. Second, motions orthogonal to the preferred-null axis will elicit intermediate, ambiguous responses. Moreover, nonmotion parameters, such as contrast, and motion parameters, such as speed, will affect the amplitude of the DS cell's response. However, by comparing over a population of DS cells, the true direction of a motion can be determined. One needs to compare the responses of DS neurons with different preferred directions. Regardless of object contrast or speed, the neuron whose preferred direction is closest to the actual direction of visual motion will have the largest response. Thus, a comparison of responses over a population of DS neurons can disambiguate the veridical motion direction.

Optic Flow

According to physics, the most fundamental variable of motion is velocity, a vector composed of direction and speed. Animals perceive three-dimensional (3D) velocities in the world through the world's two-dimensional (2D) projection onto these animals' retinas. Therefore, the true values of velocity in the world cannot be directly determined using the information from the eyes. A more useful velocity-related variable for the animal is *optic flow*. This is the spatial distribution of velocity vectors that is obtained by projecting the moving 3D world onto the 2D retina. An example of the utility of optic flow analysis occurs when an animal moves forward in a straight line. The resultant optic flow is that of an expansion, and the animal can maintain its heading by keeping the focus of expansion constant (see also MOTION PERCEPTION: NAVIGATION). To find this focus, information about directional selectivity must be combined across the image by "higher" cortical areas. One area that may contribute to this computation is the middle superior temporal cortex (MST), which appears to contain neurons

tuned to expansion and contraction (for a review, see Andersen, 1997).

Mathematically, the most popular definition of optic flow uses the image constraint equation $dE/dt = 0$, where E is the brightness of the image (see Grzywacz, Harris, and Amthor, 1994, for details and references). This equation assumes that brightness varies slowly over time and defines the optic flow \vec{v} as

$$\nabla E \cdot \vec{v} + \frac{\partial E}{\partial t} = 0 \quad (1)$$

This brightness-related definition is not unique, as nonmotion parameters, such as reflectance, can influence the solutions of the equation. Other definitions emphasize different useful components of the motion. For instance, the directional components alone have proved sufficient for determining heading direction as well as for recovering structure from motion and performing image-segregation tasks.

Speed

Before we embark on a detailed discussion of directional selectivity, we would like to comment briefly on the measurement of speed. This variable is considerably more difficult to measure than direction, as a local determination of speed in the image requires precise spatiotemporal information. In contrast, a measure of direction of motion requires two fairly imprecise positional measurements separated in time. For this reason, the visual system computes local speed with relatively less precision than direction and does so at a later stage of processing.

Computational models have been proposed that use DS signals to obtain local speed (reviewed in Grzywacz et al., 1994). For instance, Grzywacz and Yuille developed such a model by using model DS cells with receptive field profiles based on Gabor functions. They organized these cells in a 3D space whose coordinates were the optimal temporal frequency and the two components of optimal spatial frequency of the cells. They found that any arbitrary translation with constant velocity would yield maximal responses that fell on a plane in this space. Local speed and direction could then be determined by measuring the slant and tilt of this plane. Grzywacz and Yuille proposed a plausible neural architecture for performing this measurement.

Theory of Directional Selectivity

Reichardt, Poggio, and colleagues (Poggio and Reichardt, 1976) described the theoretical requirements for any model of directional selectivity (see also MOTION PERCEPTION, ELEMENTARY MECHANISMS). The first requirement is spatial asymmetry. If a neuron responds better to a motion coming from the left than to a motion coming from the right, then there must be some difference in the inputs from the left and right sides of the cell's receptive field.

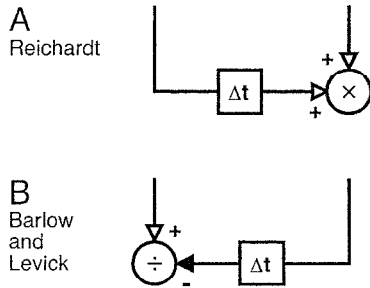


Figure 1. Models for directional selectivity. In each case, preferred direction is from left to right. The lines with arrows are inputs to the nonlinear interaction sites (circles). These inputs originate at different spatial locations as indicated by the nonarrow ends of the lines. Boxes with “ Δt ” symbols indicate that their corresponding lines are slow. Hence, movement going from the slow line to the fast line will generate signals that arrive together at the interaction sites but the opposite movement will not. In the Reichardt model, a multiplication exploits this difference to create directional selectivity. In the Barlow and Levick model, the difference is exploited by an inhibitory (possibly division-like) interaction.

Models for this asymmetry have always included a temporally asymmetric component, that is, some difference in time course between the left- and right-side inputs. However, this need not be the case. For example, the left-side input could “gate” the right-side input (Grzywacz et al., 1994). In this case, the cell will fire only when the motion comes from the left and opens the gate, before reaching the right-side input.

The second requirement for directional selectivity is a nonlinear mechanism. A spatiotemporal asymmetry alone will yield a directional difference in the responses (i.e., differing waveforms, depending on the direction of stimulus motion). However, such an asymmetry alone is not sufficient to produce two different single-number responses for preferred- and null-direction motions, a requirement to decide what the estimated direction of motion is. Poggio and Reichardt’s work proves that, without a nonlinear mechanism, any numbers obtained from the waveforms of the responses will be equal for all directions of motion. The nonlinearity can be as simple as a threshold or the gating mechanism mentioned above, but it must be present.

Figure 1A illustrates the simplest model proposed by Reichardt and colleagues for the insect’s retinal directional selectivity (RDS). For simplicity, the model uses inputs from only two locations. The proposed spatial asymmetry in this model is temporal. Inputs from the preferred side (the side first encountered by an object moving in the preferred direction, left in the figure) are propagated to the interaction site on a slow time scale compared with those from the null side. The proposed nonlinearity for the interaction is a multiplication. Thus, if an object moves in the preferred direction at an appropriate speed, the slowness of the left-side pathway is compensated for by the earlier arrival time of the stimulus, causing both inputs to arrive at the decision site at the same time, yielding a positive multiplication. For null-direction motions, the inputs will arrive at the decision site separately, and the result of the multiplication will be zero. The multiplicative nonlinearity therefore acts as a coincidence detector.

This multiplicative nonlinearity is one of many quadratic nonlinearity models supported by insect data. Poggio and Reichardt used the Volterra series formulation to examine the predictions common to all quadratic models of directional selectivity. For such a formulation of a smooth, time-invariant, nonlinear interaction between the responses to stimuli in spatial locations $a(z_a)$ and $b(z_b)$, the output is

$$y(t) = h_{0,0} + \sum_{m=1}^{\infty} \sum_{j=0}^m h_{j,m-j} *^m z_a^{(j)} z_b^{(m-j)} \quad (2)$$

where $*^m$ is the m th-order convolution and where $h_{j,m-j}$ are the m th-order kernels of the interaction. The m th-order kernel describes the nonlinear interaction between the responses to stimuli at m different instants in time. A quadratic nonlinearity is one for which if $m \geq 3$, $h_{j,m-j} = 0$. A quadratic nonlinearity thus describes multiplicative interactions between pairs of stimulus responses.

Two predictions of quadratic nonlinearity models are frequency doubling and superposition of nonlinearities. Frequency doubling is the appearance in the Fourier spectrum of the response to moving sinusoidal gratings of energy at a frequency of twice the fundamental but not at frequencies higher than that. In superposition of nonlinearities, the average of the nonlinear response to a grating composed of two sinusoidal gratings of different frequencies (whose ratio is a rational number) is equal to the sum of the responses to the two individual gratings. Both frequency doubling and superposition of nonlinearities can be used to test whether a system computes directional selectivity solely through quadratic nonlinearities.

Retinal Directional Selectivity

Although DS retinal neurons have been described in a number of vertebrates, the vast majority of work has been performed in the rabbit. In this animal, two types of DS cells have been described. The first, the On-Off DS ganglion cell, responds well to moving spots, bars, edges, or gratings of both contrast polarities over a broad range of speeds. Additionally, these cells can detect the direction of motion of long edges for displacements less than the spacing between photoreceptors, hence mediating directional hyperacuity. Four subtypes of On-Off DS cell exist, one each for temporal, nasal, superior, and inferior motions. Each subtype independently tiles the retinal surface and thus can independently sample the visual world. The second DS cell type (the On DS cell) responds only to bright edges and to considerably slower speeds than those preferred by the On-Off type. The On DS cells exist in three subtypes whose preferred directions are aligned with the animal’s semicircular canals. These cells have receptive fields whose areas are more than three times larger at any given eccentricity than the On-Off type, and their tiling is correspondingly less dense (Vaney et al., 2001). The On DS type supports optokinetic nystagmus through projections to subcortical areas. As they are infrequently encountered, the mechanisms responsible for these cells’ directional selectivity have not been well investigated, and they will not be considered further here.

Spatial Asymmetry

Although a spatial asymmetry is required for generating a DS circuit, no obvious asymmetry exists in the anatomical structure of the On-Off DS cell. The dendritic trees of these ganglion cells have a unique, looping morphology. But there is no relationship between asymmetries in the tree and the cell’s preferred direction of motion (Vaney et al., 2001). Therefore, the spatial asymmetry has been conjectured to exist in the connectivity between DS cells and their inputs. No spatially asymmetric connections have been conclusively identified to date. However, two strong candidates exist. First, cross-correlational and anatomical studies of the cholinergic amacrine cell and the On-Off DS cell suggest that the former makes spatially asymmetric connections to the latter. Support for this suggestion comes from the elimination of DS responses to moving gratings by cholinergic antagonists (Grzywacz, Amthor, and Mer-

wine, 1998). Second, asymmetric connections may also exist from some GABAergic amacrine cells to the DS cells. GABA antagonists strongly reduce directional selectivity.

Nonlinearities

In their seminal work on rabbit retinas, Barlow and Levick (1965) performed two-slit apparent-motion experiments on DS cells. They discovered that when two stimuli were presented in null sequence (as if an object were moving in the null direction), then the number of spikes elicited was far less than the sum of the spikes for each stimulus in isolation. From this, they concluded that RDS arises from a nonlinear, *inhibitory* mechanism that “vetoes” responses to null sequences (equivalent to the logical AND-NOT operation). As shown in Figure 1B, their proposed spatial asymmetry has two components. A central component is excitatory and is conducted to the interaction site quickly. A second, inhibitory component is offset to the null side and is conducted with a delay. Thus, an asymmetry exists in both the sign and time course of the two spatially separated components. This asymmetry causes motions in the preferred direction to yield responses, while responses to null-direction motions are vetoed. However, despite Barlow and Levick’s proposal, this veto mechanism turns out not to be a perfect veto. Studies reviewed by Grzywacz et al. (1994) show that a better description for the inhibitory interaction is a division-like nonlinearity, as shown in Figure 1B.

Torre and Poggio proposed a biophysical implementation of Barlow and Levick’s inhibition (see Grzywacz et al., 1994, for a review). Because RDS can be elicited by motions spanning remarkably short distances almost anywhere within the DS cell’s receptive field (Barlow and Levick, 1965), Torre and Poggio suggested that the inhibition acted separately within each branch of the cell’s dendritic tree. To constrain the computation spatially, they suggested that the inhibition works through a synapse that causes local changes of membrane conductance (shunting inhibition) and little hyperpolarization. To understand such a synapse, consider a patch of membrane receiving excitatory (g_e) and shunting inhibitory (g_i) synaptic conductances. Setting without loss of generality the resting and inhibitory reversal potentials to zero, the voltage V obeys

$$C \frac{dV(t)}{dt} + (g_e(t) + g_i(t) + g_{\text{leak}})V(t) = g_e(t)E_e + g_{\text{leak}}E_{\text{leak}} \quad (3)$$

where C is membrane capacitance, g_{leak} is the membrane’s leak conductance, and E_e and E_{leak} are reversal potentials of g_e and g_{leak} , respectively. When $g_i \gg g_e$, then V falls toward the following equilibrium value:

$$V(t) \rightarrow \frac{g_e E_e + g_{\text{leak}} E_{\text{leak}}}{g_i} \quad (4)$$

which is small, because g_i is large. Therefore, this inhibition is division-like rather than subtraction-like. Torre and Poggio argued that a shunting-inhibition mechanism might also be consistent with the insect’s quadratic nonlinearity, because, for sufficiently low contrasts, one can ignore the higher-order nonlinearities, as in a Taylor series approximation. However, experimentally a quadratic approximation is not valid for rabbit DS cells, as they fail both the frequency-doubling and superposition-of-nonlinearities tests even at near-threshold contrasts.

Although a shunting-inhibition mechanism can theoretically produce the localized interactions necessary to explain many DS properties, it has not been possible to record intracellularly within dendrites to test this mechanism. Recently, two additional nonlinearities have been proposed that could support dendritically lo-

calized directional selectivity. First, it has been suggested that this nonlinearity could be due to excitatory voltage-dependent conductances at the dendrites. There is evidence of dendritic spikes in rabbit’s ganglion cells. Second, it has been proposed that the unusual predominance of NMDA glutamatergic receptors on DS cells may have functional significance. The NMDA channel has a nonlinear behavior, due to channel blockade by magnesium ions at hyperpolarized potentials. Therefore, glutamatergic binding must occur *during depolarization* for the channel to operate. Even weak, null-direction inhibition could cause the NMDA channel to hyperpolarize and close. It could therefore act as the veto site hypothesized by Barlow and Levick (1965). When tested in magnesium-free medium, RDS is severely reduced, suggesting a critical role for NMDA receptors. The NMDA, spiking, and shunting nonlinearities need not be mutually exclusive, and each could play an important role in supporting robust RDS.

Pre- or Postsynaptic Nonlinearities?

In addition to the null-direction inhibition just described, it is known that preferred-direction motions facilitate DS cell responses (Barlow and Levick, 1965). If the spatiotemporal parameters of the stimulus are appropriate, then preferred-direction facilitation can be as strong as null-direction inhibition (reviewed in Grzywacz et al., 1994). Facilitation is believed to come to the DS cell from the cholinergic amacrine cells. A spatial asymmetry has been shown to exist in the input-output relationship of these cells’ dendrites. These dendrites receive excitatory inputs along their length, but they release excitatory transmitter (ACh) and may receive inhibitory inputs (through GABA) only at their tips (Figure 2A). If the GABA-inhibition acts in a division-like manner, then each dendrite contains a spatial asymmetry and a nonlinearity, and thus can act as an autonomous DS unit. Hence, it has been proposed that DS signals are at least partially generated presynaptically and flow from the cholinergic dendrites to the DS cell. DS cells would then

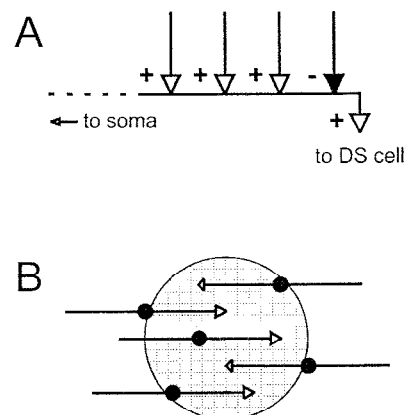


Figure 2. Schemes for presynaptic retinal directional selectivity. *A*, Model of a cholinergic-dendritic DS subunit. The bent line represents a dendrite of the cholinergic amacrine cell making a synapse onto a DS cell. This dendrite receives excitatory inputs throughout its length (open arrows) and inhibitory inputs (filled arrow) near the synaptic ending. Therefore, the output synapse is directionally selective, as the dendrite behaves like a Barlow-and-Levick model (Figure 1B). *B*, Spatial relationship between cholinergic dendrites and the DS cell’s receptive field (circle). Arrowhead size indicates synaptic strength, and black discs mark the cholinergic somas. Although all dendrites are directionally selective, the DS cell receives more inputs from the left than from the right. Consequently, its preferred direction is to the right.

preferentially sample from dendrites with the same preferred direction (Figure 2B).

There is accumulating evidence, however, that both pre- and postsynaptic asymmetries may be involved in RDS. Complete blockade of cholinergic synapses does not fully eliminate RDS to moving bars (Grzywacz et al., 1998). The residual direction selectivity is GABAergic and appears to be postsynaptic. In turn, GABA blockade not only does not always eliminate RDS, it occasionally reverses its preferred and null directions. Computer simulations of the cholinergic-dendritic RDS can account for these reversals as a result of synaptic saturation (reviewed in Grzywacz et al., 1994). Thus, asymmetric postsynaptic inhibition and asymmetric presynaptic facilitation may act cooperatively to produce robust RDS for a broad range of visual stimuli.

Development

The development of RDS is not well understood. In turtles, RDS emerges late in development, that is, after the establishment of concentric receptive fields, inhibitory surrounds, and orientation selectivity (selectivity to spatial orientation of anisotropic stimuli, such as lines) (Sernagor and Grzywacz, 1995). Evidence suggests that turtle RDS may emerge at the expense of orientationally selective (OS) cells. It has thus been proposed that turtle DS cells are modified OS cells. This late emergence of RDS suggests two hypotheses for its development: (1) it requires light exposure, and/or (2) it requires the late emergence of an inhibitory drive onto the network mediating orientation selectivity.

In rabbits, however, RDS emerges relatively early. The percentage of DS cells and DS cell gap-junctional coupling patterns both appear adult-like within a few days of eye opening. Directionally selective responses have been recorded from rabbit retinas before eye opening. Therefore, some authors have suggested that rabbit RDS is initially generated in the presynaptic cholinergic circuit. The primary excitatory stimulus for this mechanism would be the bursting spontaneous waves of activity known to occur during development (see Sernagor and Grzywacz, 1995, for references on developmental spontaneous activity). In this case, the first challenge for the DS cells would be to connect selectively to cholinergic dendrites of similar orientation. (Recall that the amacrine cell dendrites can each act as a DS subunit.) A Hebbian correlational process that would reinforce statistical biases in the initial contacts from the cholinergic dendrites to proto-DS cells could produce this selectivity. GABAergic null-direction inhibition would then be added to the cell's established directional preference.

Cortical Directional Selectivity

DS cells are found in multiple locations in the cortices of mammals, beginning with simple DS cells in the lateral geniculate nucleus (LGN) input layers of the primary visual cortex (V1). Nearly every simple and complex cell in mammalian V1 has some degree of directional selectivity. However, the strength of directionality varies widely. For example, only about 20% of macaque V1 DS cells achieve preferred-versus-null response ratios greater than 3:1 (De Valois et al., 2000). In comparison, rabbit retinal DS cells typically have preferred-null ratios around 10:1. As with the retinal DS cells, the preferred direction of motion cannot be accounted for by any spatial asymmetries in the dendritic trees of the cortical DS cells. However, unlike retinal DS cells, cortical DS cells are highly selective for the orientation and spatial frequency of the moving stimulus. Moreover, it is often possible to predict the preferred direction of motion for a simple DS cell from its spot-mapped receptive field. This has led to the development of quasilinear models of cortical directional selectivity, which we discuss below.

There is abundant evidence that cortical directional selectivity supports motion perception. Perceptual decisions in motion tasks correlate with the performance of cortical DS neurons (Andersen, 1997). Furthermore, lesions and/or current injections in MT affect motion integration tasks, biasing motion perception in specific and replicable ways (Salzman et al., 1992). In addition to the contribution of cortical DS cells to perception, they probably also assist in the control of eye movements. For instance, neuroanatomical data demonstrate that both MT and MST send large projections to the dorsolateral pons, an area known to be involved in smooth-pursuit eye movements.

Hierarchy

As a first-order approximation, motion is computed hierarchically in the cortex (Andersen, 1997). The hierarchy begins with the simple and complex DS cells in layers IV and VI of cortical area V1, as just described. (In animals phylogenetically close to primates, retinal DS cells project primarily to subcortical centers. Evidence suggests that directional selectivity in the primary visual cortex is computed independently from RDS.) The DS cells in V1 project to the MT cortical area (MT or V5) and to V2, which also projects to MT. Directional selectivity becomes more complex in MT; that is, cells there typically have very large, orientation-independent receptive fields, and many will respond best to the composite motion of a plaid, as opposed to its individual components. From MT, the motion pathway projects to MST, wherein directional selectivity information is further combined to produce neurons sensitive to complex motions, such as rotation, expansion, and contraction (Andersen, 1997).

Psychophysical Models

One class of models for the first stage of cortical directional selectivity is based on human psychophysics and is similar to the Reichardt model in Figure 1A. In these models a slow, laterally displaced input and a fast central input are multiplied. As described above, the two signals from these inputs will arrive simultaneously, yielding a response, for only one direction of object motion. Another class, called motion-energy models (Adelson and Bergen, 1985), proposes a distributed spatiotemporal asymmetry and a squaring nonlinearity (Figure 3A). The distributed spatial asymmetry occurs because different locations in the receptive field have different impulse responses. This property is known as space-time (S-T) inseparability and is illustrated in Figure 3B. For this type of space-time arrangement, it is only for preferred-direction motions that the responses of all areas occur simultaneously. Simple linear summation then results in differential responses to preferred- and null-direction motions, and the squaring nonlinearity converts this directional difference into directional selectivity.

Physiological Models

There is substantial physiological evidence for S-T inseparability in simple DS cells in visual cortex. Figure 3C shows an idealized simple-cell receptive field map as would be obtained from reverse-correlation experiments (De Valois et al., 2000). The cell contains on and off subregions with different time courses in different portions of space, resulting in an oriented spatiotemporal receptive field. Assorted variations of linear energy-motion models with static nonlinearities have been proposed that can account for approximately 50%–80% of the response of simple DS cells (Reid, Soodak, and Shapley, 1991; De Valois et al., 2000). However, the correlation between simple-cell S-T profile and direction selectivity varies widely in V1. Cells in layer IVB show very high correlation, those in IVA show only moderate correlation, and those in layer

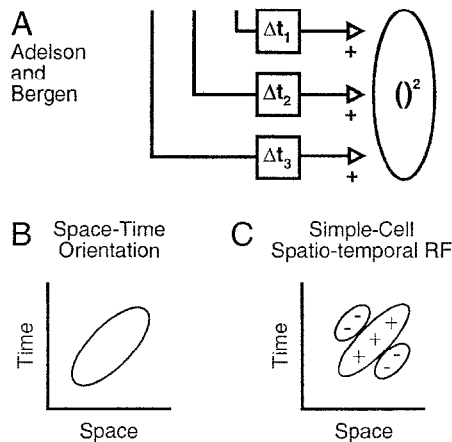


Figure 3. Motion-energy models of cortical directional selectivity. *A*, The Adelson and Bergen model. The symbols here are as in Figure 1, with “ $()^2$ ” indicating a squaring operation. *B*, Space-time orientation. If in *A*, $\Delta t_3 < \Delta t_2 < \Delta t_1$, then a plot of the response of the model to stimulation at different positions in space looks like this figure; that is, it is oriented in space-time. Hence, a motion going from left to right will have a positive slope in this plot and cause much response, whereas the opposite motion will not. *C*, Idealized representation of a simple-cell receptive field. The inhibitory flanks help to inhibit motions in the null (nonpreferred) direction (lines with negative slopes).

VI show very low correlation, despite equivalent directional tuning (Murthy et al., 1998). Thus, S-T structure alone cannot fully account for simple DS cell responses. In addition, quasilinear models of simple cells generally overestimate nonpreferred responses and sometimes underestimate preferred responses (Reid et al., 1991). And linear feedforward models also do not predict onset transients, which are commonly observed. Therefore, inhibitory (Reid et al., 1991; Heeger, 1993) or excitatory (Douglas and Martin, 1992) nonlinear feedback interactions between cortical cells have been proposed to account for these discrepancies. It has also been suggested that nonpreferred direction suppression might be due to a cortical inhibitory network devoted to response normalization (Heeger, 1993).

Although it has been accepted for many years that magnocellular cells from the LGN provide the input to the motion system, motion-energy models require inputs to simple DS cells that differ in latency (or temporal phase). Magno cells all have essentially identical response timings and therefore could not provide the range of latencies needed without an intracortical mechanism for creating delays (De Valois et al., 2000). It has been shown that blockade of cortical GABA_A inhibition reduces but does not eliminate S-T inseparability. Thus, intracortical inhibition may contribute to, but cannot be solely responsible for, input latency differences. Humphrey, Saul, and Fiedler (1998) note that about 40% of parvocellular geniculate cells display absolute phase delays and long latencies (lagged cells) relative to the remaining parvo cells (nonlagged cells). Because lagged and nonlagged timing signatures are identifiable in simple-cell receptive fields, these authors attribute S-T inseparability to converging lagged and nonlagged parvo cell inputs with spatially shifted receptive fields. De Valois et al. (2000), however, suggest that the requisite latency differences arise from two classes of nondirectional (preferred-null ratios $<3:1$) cortical cells, which receive inputs from magno and parvo cells. Both nondirectional types can be found in a single cortical column and have appropriate differences in response waveforms. Additionally, these two nondirectional cell types are often shifted 90° in spatial phase, though centered on the same spatial location. Thus, they have ex-

actly the spatiotemporal profile required to detect direction of motion using a linear energy-motion model.

Complex DS cells are generally found in the upper and lower layers of V1. Unlike most simple DS cells, complex DS cells lack first-order (linear) S-T oriented receptive fields. However, these cells show second-order (quadratic) S-T structure. In other words, it is the interactions between two sequentially stimulated locations in the cell's receptive field that are S-T inseparable. Thus, dynamic nonlinearities have been proposed to account for complex-cell directional selectivity. These nonlinearities would facilitate or inhibit, respectively, the responses to preferred- or null-direction motions. Similarly, S-T separable simple cells have also been shown to display some second-order, that is, nonlinear S-T structure. Because there is evidence for complex-to-simple-cell interactions, it has been proposed that directional selectivity and second-order S-T inseparability in simple cells arise from complex-cell inputs.

Development

At least 5% of the cells in areas V1 and V2 of the kitten are directionally selective at eye opening. Thus, some of the cortical directional selectivity is either genetically coded or epigenetically derived through developmental spontaneous activity. Interestingly, selective biases in the distribution of preferred directions can be produced in kittens by exposing them to stripes moving in a particular direction after eye opening. In addition, one can nearly eliminate cortical directional selectivity by raising animals in an environment that is illuminated only by brief, low-frequency, stroboscopic flashes of light. Under these rearing conditions, only about 10% of V1 cells develop directional selectivity, and their directional selectivity is considerably weaker than normal. Not surprisingly, elimination of S-T inseparability accompanies the loss of directional selectivity (Humphrey et al., 1998). Perceptually, strobe-reared cats require contrasts at least ten times higher than normal to determine the direction of a moving grating. Furthermore, these elevations in contrast threshold can be permanent. No recovery of S-T structure or improvement in contrast threshold was found in two cats that had received 12 years of training following strobe rearing. Thus, it appears that some critical period exists during development during which directional selectivity must be established or it is forever compromised.

Based on the loss of S-T inseparability during strobe rearing, some authors have modeled the development of cortical directional selectivity as due to inputs with different response timings forming connections with a common cortical cell. These connections would self-organize through a Hebbian process that would strengthen the synaptic connections of well-correlated inputs (Humphrey et al., 1998). Strobe rearing would restrict the range of timings that could be associated, and therefore could eliminate cortical directional selectivity.

Discussion

In this article, we have discussed the first step in the perception of motion—the determination of motion direction. This information is encoded by ensembles of directionally selective neurons both in the retina and within multiple visuocortical nuclei. Elucidating the cellular mechanisms subserving directional selectivity has been a major goal of neurobiologists for nearly 40 years.

Road Map: Vision

Related Reading: Feature Analysis; Motion Perception: Elementary Mechanisms; Retina; Visual Cortex: Anatomical Structure and Models of Function

References

- Adelson, E. H., and Bergen, J. R., 1985, Spatio-temporal energy models for the perception of motion, *J. Opt. Soc. Am. A*, 2:284–299.
- Andersen, R. A., 1997, Neural mechanisms of visual motion perception in primates, *Neuron*, 18:865–872. ♦
- Barlow, H. B., and Levick, W. R., 1965, The mechanism of directionally selective units in rabbit's retina, *J. Physiol.*, 178:477–504.
- De Valois, R. L., Cottaris, N. P., Mahon, L. E., Elfar, S. D., Wilson, J. A., 2000, Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity, *Vision Res.*, 40:3685–3702.
- Douglas, R. J., and Martin, K. A. C., 1992, Exploring cortical microcircuits: A combined anatomical, physiological, and computational approach, in *Single Neuron Computation* (T. McKenna, J. Davis, and S. F. Zornetzer, Eds.), Orlando, FL: Academic Press, pp. 381–412. ♦
- Grzywacz, N. M., Amthor, F. R., and Merwine, D. K., 1998, Necessity of acetylcholine for retinal directionally selective responses to drifting gratings in rabbit, *J. Physiol.*, 512:575–581.
- Grzywacz, N. M., Harris, J. M., and Amthor, F. R., 1994, Computational and neural constraints for the measurement of local visual motion, in *Visual Detection of Motion* (A. T. Smith and R. J. Snowden, Eds.), London: Academic Press, pp. 19–50. ♦
- Heeger, D. J., 1993, Modeling simple cell direction selectivity with normalized, half-squared, linear operators, *J. Neurophysiol.*, 70:1885–1898.
- Humphrey, A. L., Saul, A. B., and Fiedler, J. C., 1998, Strobe rearing prevents the convergence of inputs with different response timings onto area 17 simple cells, *J. Neurophysiol.*, 80:3005–3020.
- Murthy, A., Humphrey, A. L., Saul, A. B., and Fiedler, J. C., 1998, Laminar differences in the spatiotemporal structure of simple cell receptive fields in cat area 17, *Vision Neurosci.*, 15:239–256.
- Poggio, T., and Reichardt, W. T., 1976, Visual control of orientation behaviour in the fly: Part II. Towards the underlying neural interactions, *Q. Rev. Biophys.*, 9:377–438. ♦
- Reid, R. C., Soodak, R. E., and Shapley, R. M., 1991, Directional selectivity and spatiotemporal structure of receptive fields of simple cells in cat striate cortex, *J. Neurophysiol.*, 66:505–529.
- Salzman, C. D., Murasugi, C. M., Britten, K. H., and Newsome, W. T., 1992, Microstimulation in visual area MT: Effects on direction discrimination performance, *J. Neurosci.*, 12:2331–2355.
- Sernagor, E., and Grzywacz, N. M., 1995, Emergence of complex receptive field properties of ganglion cells in the developing turtle retina, *J. Neurophysiol.*, 73:1355–1364.
- Vaney, D. I., He, S., Taylor, W. R., and Levick, W. R., 2001, Direction-selective ganglion cells in the retina, in *Motion Vision: Computational, Neural, and Ecological Constraints* (J. M. Zanker and J. Zeil, Eds.), Berlin: Springer-Verlag, pp. 13–56. ♦

Dissociations Between Visual Processing Modes

Bruce Bridgeman

Introduction

The visual system has two kinds of jobs to do. One is to support visual cognition or perception—knowledge about the identities and locations of objects and surfaces in the world. Another, sensorimotor, function is to control visually guided behavior. The two functions require different kinds of visual information.

The cognitive function is concerned with pattern recognition and with the positions of objects relative to one another. Executing this function requires extensive interaction between bottom-up image data and top-down information about objects, faces, etc. Qualitative location information may be adequate for this function; humans are poor at quantitatively estimating distances, directions, etc. if the measures of these abilities are perceptual judgments rather than motor behaviors.

The sensorimotor function, in contrast, needs quantitative egocentrically calibrated spatial information to guide motor acts. It does not need the minute-of-arc acuity of the cognitive function, however: calibration is more important than resolution. As a result, the brain's sensorimotor representations can be much smaller than those supporting cognition.

It is an empirical question whether these two functions, the cognitive and the sensorimotor, should be modeled as a single visual representation with two readouts or as separate maps of visual space. There is now extensive evidence for two distinct maps or sets of maps of visual space in the brain, one set handling perception and the other supporting visually guided behavior (Figure 1). Evidence comes from physiological recordings from the separate maps, from neurological patients in which one system or the other is damaged, from fMRI and PET scans of humans doing cognitive or sensorimotor tasks, and from psychophysical work in which different spatial values are inserted into the two systems simultaneously.

Some of the earliest evidence for the two-visual-systems distinction came from experiments in hamsters, where lesions of the

midbrain's superior colliculus led to the inability to orient appropriately in a T-maze, combined with preserved abilities in pattern discrimination. In other animals, visual cortex lesions disturbed pattern discrimination without interfering with maze orienting (Schneider, 1969). This forebrain-midbrain distinction changed over the course of evolution, as both spatial orientation and pattern recognition became corticalized in primates.

Neurophysiology

The visual pathways begin as a unified system, from the retinas through the lateral geniculate nucleus of the thalamus to the primary visual cortex of the occipital lobe. From here, visual signals are relayed to approximately 27 topographic maps in other visual areas (VISUAL SCENE PERCEPTION). This characteristic of visual systems raises a question: do all of these maps work together in a

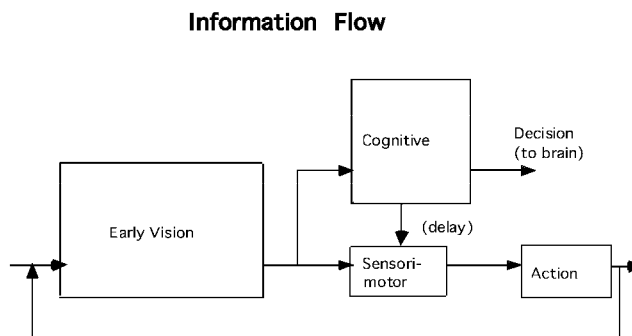


Figure 1. Information flow in cognitive and sensorimotor visual pathways. Visual image input is from the left, and cognitive or motor output to the right.

single visual representation, or are they functionally distinct? If they are distinct, how many functional maps are there and how do they interact?

The evidence reveals that the multiple maps support at least two functionally distinct representations of the visual world. The representations do not always function independently, but communicate with one another in some ways. Each representation uses several of the retinotopic maps; cognitive and sensorimotor representations correspond roughly to pathways into temporal and parietal cortex, respectively (Mishkin, Ungerleider, and Macko, 1983).

The temporal pathway consists of a number of regions in which neurons generally respond to stimuli in larger and larger regions of the visual field, but require increasing amounts of specific features or properties to excite them. This processing pathway culminates in the inferotemporal cortex, which specializes in pattern recognition problems involving choice and discrimination. Neurons in this region typically respond to very large areas of the visual field, usually including the fovea, and their responses are highly modified by visual experience and by the nature of the visual task currently being executed.

The parietal pathway, in contrast, specializes in physical features of the visual world, such as motion and location. One area contains a map of the cortex that specializes in motion of objects, whereas others contain neurons that respond both to characteristics of the visual world and to intended movements (GRASPING MOVEMENTS: VISUOMOTOR TRANSFORMATIONS). The areas serving the sensorimotor functions of vision are only a small part of the parietal lobe; dividing the cognitive and sensorimotor functions into temporal and parietal areas is a convenient first approximation, but several areas anatomically superior to the primary visual cortex (in occipito-parietal regions) are involved in the "temporal" stream. Some areas are involved in both functions.

A key task of the parietal stream is transformation of visual information into motor coordinates. As a first approximation, parietal visual cortex can be divided into five regions: (1) the lateral intraparietal area activated by both saccades and attention; (2) the parietal reach region that is activated by pointing and is also modulated by eye position; (3) the anterior intraparietal area active during visual grasp and also tactile manipulation; (4) the caudal intraparietal sulcus involved in object matching and grasping; and (5) the ventral intraparietal area, responding to visual motion toward the face (Culham and Kanwisher, 2001). Parietal regions are central to attention, eye movements, and orienting.

The spatial function can be further subdivided into two pathways that reflect different modes of spatial coding. Receptive fields of neurons in the lateral intraparietal area are spatially corrected before each rapid eye movement, so that they respond to stimuli that will be in their retinotopic receptive fields (i.e., in retinally based coordinates) following a planned eye movement (Duhamel, Colby, and Goldberg, 1992). The changes in these receptive fields can also be conceived as activity that signals candidates for planned eye movements.

A second coding scheme is seen in parietal area 7a. Neurons in this area provide information sufficient to reconstruct spatial position in head-centered coordinates (Andersen, Essick, and Siegel, 1985). These neurons respond strongly only if the eyes are in a particular position in the orbit and a target occupies a particular retinal location, in a multiplicative interaction. Simulations showed that such a network of cells codes information sufficient to derive spatiotopic output (Zipser and Andersen, 1988). A parallel distributed processing network was trained to respond to targets at particular locations in a visual field. After training, the response properties of the nodes in the model resembled the receptive fields of neurons in area 7a.

Spatial processing may be too basic a function to be limited to the parietal cortex, a relatively high-order structure that is well differentiated only in primates. Some features of the midbrain superior colliculus, such as broad intermodal integration and close connection to oculomotor control, suggest a role in sensorimotor vision (COLLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL), reflecting the earlier midbrain/forebrain distinction.

Clinical Evidence

Damage to part of the primary visual cortex in human patients results in functional blindness in the affected field—a scotoma. Patients have no visual experience in the scotoma, but when forced to point to targets located there, which they insist they cannot see, they point fairly accurately by using sensorimotor information unavailable to their perception. Visual information coexisting with a lack of visual experience is called blindsight (Weiskrantz et al., 1974); it is an example of visually guided behavior without the experience of perception. It may be made possible by an alternative pathway from the retina to nonstriate visual cortex through the superior colliculus or other subcortical structures. Recent work has shown surprisingly sophisticated processing without awareness, including orientation and color discriminations.

Another example of dissociation of sensorimotor and cognitive function has been found in a patient with damage to lateral occipital and occipitoparietal cortex (Goodale and Milner, 1992). When asked to match line orientations, the patient could not reliably distinguish horizontal from vertical, though she had no scotoma in which visual experience was altogether absent. Asked to put a card through a slot at varying orientations, however, she oriented the card correctly even as she raised her hand from the start position. There was a dissociation between her inability to perceive object orientation and her ability to direct accurate reaching movements toward objects of varying orientations. Her cognitive representation of space was unavailable, while sensorimotor representation remained intact. The complementary dissociation, with retention of perception accompanied by loss of ability to use visual information to control movements of the limb and hand, is seen clinically under the label of optic ataxia.

Such examples demonstrate separate cognitive and sensorimotor representations, but only in patients with brain damage. When a human suffers brain damage with partial loss of visual function, there is always the possibility that the visual system will reorganize itself, isolating fragments of the machinery that normally function as a unit. The clinical examples, then, leave open the possibility that the system may function differently in intact humans. Any rigorous proof that normal visual function shows the cognitive/sensorimotor distinction must include psychophysical measures in intact humans.

Psychophysics of Space

Some of the earliest evidence for a cognitive/sensorimotor distinction in normal subjects came from studies of rapid (saccadic) eye movements. Subjects are normally unaware of sizable displacements of the visual world if the displacements occur during saccadic eye movements. This implies that information about spatial location is degraded during saccades. There is a seeming paradox to this degradation, however, for people do not become disoriented after saccades, implying that spatial information is maintained. Experimental evidence supports this conclusion. For instance, the eyes can saccade accurately to a target that is flashed (and mislocalized) during an earlier saccade, and hand-eye coordination remains fairly accurate following saccades. How can perceptual information be lost while visually guided behavior is preserved?

To resolve this paradox, it should be noted that the conflicting observations use different response measures. The experiments on perception of displacement during saccades require a symbolic response, such as a nonspatial verbal report or a button press, with an arbitrary spatial relationship to the target. Orienting of the eye or hand, in contrast, requires quantitative spatial information, defined as requiring a 1:1 correspondence between a target position and a motor behavior, such as directing the hand or the eyes to the target. The conflict might be resolved if the two types of measure, which can be labeled as cognitive and sensorimotor, could be combined in a single experiment. If two visual pathways process different kinds of information, spatially oriented motor activities might have access to accurate position information even when that information is unavailable at a cognitive level.

The two conflicting observations (perceptual suppression on one hand and accurate motor behavior on the other) were combined by asking subjects to jab the position of a target that had been displaced and then extinguished (reviewed in Bridgeman, Peery, and Anand, 1997). On some trials the target jump was detected, while on others the jump went undetected due to a simultaneous eye movement (monitored photoelectrically). As one would expect, subjects could point accurately to the position of the now-extinguished target following a detected displacement. Pointing was equally good, however, following an undetected displacement. It appeared that updating information was available to the motor system but not to perception.

This result implied that quantitative control of motor activity was unaffected by the perceptual detectability of target position. One can also interpret the result in terms of signal detection theory as a high response criterion for the report of displacement. The first control for this possibility was a two-alternative forced-choice measure of saccadic suppression of displacement. This criterion-free measure showed an inability to perceive displacements under conditions where pointing was accurate even when the target had been displaced (Bridgeman and Stark, 1979). Information was available to a sensorimotor system controlling pointing, but not to a cognitive system informing visual perception.

Dissociation of cognitive and sensorimotor function has also been demonstrated by giving cognitive and sensorimotor systems opposite signals at the same time. This is a more rigorous way to separate cognitive and sensorimotor systems. A signal was inserted selectively into the cognitive system with stroboscopic induced motion. In this illusion a surrounding frame was displaced, creating the illusion that a target jumps although it remains fixed relative to the subject. We know that induced motion affects the cognitive system, because we experience the effect and subjects can make verbal judgments of it. But the above saccadic suppression experiments (Bridgeman and Stark, 1979) implied that the information used for motor behavior might come from sources unavailable to perception.

In the experiment, a target spot jumped in the same direction as a frame, but not far enough to cancel the induced motion. The spot still appeared to jump in the direction opposite the frame. Saccadic eye movements followed the actual jump direction, even though subjects perceived stroboscopic motion in the opposite direction (Wong and Mack, 1981). If a delay in responding was required, however, eye movements followed the perceptual illusion. This implies that the sensorimotor system has no memory, but must rely on information from the cognitive system when responding to what was previously present rather than what is currently present.

All of these techniques involve motion or displacement, leaving open the possibility that the dissociations are associated in some way with motion systems, rather than with representation of visual location *per se*. Motion and location may be confounded in some kinds of visual coding schemes. A newer design allows the examination of visual context in a situation where there is no motion

or displacement at any time (Bridgeman et al., 2000). The design is based on the Roelofs effect, a perceptual illusion seen when a static frame is offset to the left or the right of a subject's centerline. Objects that lie within the frame tend to be mislocalized in the direction opposite the offset of the frame. For example, in an otherwise featureless field, a rectangle is presented to the subject's left. Both the rectangle and stimuli within it tend to be localized too far to the right.

A Roelofs effect is seen with a perceptual measure, but subjects point without error. If a delay in responding causes subjects to switch from using sensorimotor information directly to using information imported from the cognitive representation, delaying the response should force them to switch to using cognitive information. By delaying the response cue long enough, all subjects showed a Roelofs effect in both pointing and judging. Thus, this design showed a switch from motor to cognitive information in directing the motor response; the cognitive illusion appears after a delay of about 2 s (Bridgeman et al., 2000; Rossetti, 2000). In the delay case, information flowed from the cognitive system, with the illusion, to the sensorimotor system and thence to behavior. However, the perception of the Roelofs effect shows that accurate sensorimotor information never corrects perception. Thus the arrow from cognitive to sensorimotor in Figure 1 extends in only one direction.

An even longer motor memory is seen for the slopes of hills, where veridical information endures for several minutes despite exaggerated perceptions of the slopes (Creem and Proffitt, 1998). The sensorimotor branch of the system seems to hold spatial information just long enough to direct current motor activity, but no longer. These authors have also shown that the sensorimotor system does not have access to top-down information about the conventional uses of an object. A group of objects is arrayed before a subject with their handles facing away. They will be grasped by their handles if the cognitive system is available to help organize the movement, but if the cognitive system is distracted by another, non-motoric task, the objects will be grasped skillfully but inappropriately by their closest parts. This result raises another issue, the extent of object coding in the sensorimotor system.

Psychophysics of Objects

The cognitive system needs detailed information to identify top-down information about the meanings and uses of objects, while the sensorimotor system needs only information about size, location, and graspability.

Research on object properties has followed several methods. One method is based on the Ebbinghaus illusion, also called the Titchner circles illusion. A circle looks larger if it is surrounded by smaller circles than if it is surrounded by larger circles. Haffenden and Goodale (1998) measured the illusion by asking subjects either to indicate the apparent size of a circle or to pick it up. In both cases neither hand nor target could be seen during the movement. The illusion was larger for the size estimations than for the grasp, indicating that the sensorimotor system was relatively insensitive to the illusion.

A challenge to this line of research came from Franz et al. (2000), who analyzed the Ebbinghaus illusion into two half-illusions: comparing a circle surrounded by larger circles to a circle alone, and comparing a circle surrounded by smaller circles to a circle alone. When the small-circle context and the large-circle context are compared directly, a super-additivity occurs: the illusion is larger than the sum of the two half-illusions. Because grasp necessarily involves a circle alone, the differences between grasp and perception were explained. Goodale has recently clarified this issue by showing that it is the physical distance between the inducing circles and the target circle, not the size of the circles, that influences grasp

aperture. Perception, in contrast, is based on the size difference between the inducing circles and the target circle.

Discussion

Information about egocentric spatial location and some object properties is available at a motor level despite cognitive illusions of location. Egocentric localization information is available to the sensorimotor system even while the cognitive system, relying on relative motion and relative position information, holds unreliable information about location. Spatial information can flow from the cognitive to the sensorimotor representation if the sensorimotor information has degraded due to delay, but information cannot flow in the other direction. The two-visual-systems conception applies not only to the localization of objects, but also to their properties.

Road Map: Vision

Related Reading: Object Recognition; Face Recognition; Neurophysiology and Neural Technology; Visual Attention; Visual Scene Perception

References

- Andersen, R. M., Essick, G., and Siegel, R., 1985, The encoding of spatial location by posterior parietal neurons, *Science*, 230:456–458.
- Bridgeman, B., Gemmer, A., Forsman, T., and Huemer, V., 2000, Processing spatial information in the sensorimotor branch of the visual system, *Vision Research*, 40:3539–3552.
- Bridgeman, B., Peery, S., and Anand, S., 1997, Interaction of cognitive and sensorimotor maps of visual space, *Perception and Psychophysics*, 59:456–469.
- Bridgeman, B., and Stark, L., 1979, Omnidirectional increase in threshold for image shifts during saccadic eye movements, *Perception and Psychophysics*, 25:241–243.
- Creem, S., and Proffitt, D., 1998, Two memories for geographical slant: Separation and interdependence of action and awareness, *Psychonomic Bull. Rev.*, 5:22–36. ♦
- Culham, J. C., and Kanwisher, N. G., 2001, Neuroimaging of cognitive functions in human parietal cortex, *Curr. Opin. Neurobiol.*, 2001, 11:157–163.
- Duhamel, J., Colby, C., and Goldberg, M. E., 1992, The updating of the representation of visual space in parietal cortex by intended eye movements, *Science*, 255:90–92.
- Franz, V., Gegenfurtner, K., Bühlhoff, H., and Fahle, M., 2000, Grasping visual illusions: No evidence for a dissociation between perception and action, *Psychol. Sci.*, 11:20–25.
- Goodale, M. A., and Milner, A. D., 1992, Separate visual pathways for perception and action, *Trends Neurosci.*, 15:20–25. ♦
- Haffenden, A. M., and Goodale, M. A., 1998, The effect of pictorial illusion on prehension and perception, *J. Cognit. Neurosci.*, 10:122–136.
- Mishkin, M., Ungerleider, L., and Macko, K., 1983, Object vision and spatial vision: Two cortical pathways, *Trends Neurosci.*, 6:414–417. ♦
- Rossetti, Y., 2000, Implicit perception in action: Short lived motor representations of space, in *Consciousness and Brain Circuitry* (P. Grossenbacher, Ed.), Amsterdam: John Benjamins Publishers, pp. 131–179.
- Schneider, G. E., 1969, Two visual systems, *Science*, 163:895–902.
- Weiskrantz, L., Warrington, E., Sanders, M., and Marshall, J., 1974, Visual capacity in the hemianopic field following a restricted occipital ablation, *Brain*, 97:709–728.
- Wong, E., and Mack, A., 1981, Saccadic programming and perceived location, *Acta Psychologica*, 48:123–131.
- Zipser, J., and Andersen, R. A., 1988, A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons, *Nature*, 33:679–684.

Dopamine, Roles of

Jean-Marc Fellous and Roland E. Suri

Introduction

Dopamine (DA) is a neuromodulator (see NEUROMODULATION IN INVERTEBRATE NERVOUS SYSTEMS) that originates from small groups of neurons in the mesencephalon [the ventral tegmental area (A10), the substantia nigra (A9) and A8], and the diencephalon (area A13, A14 and A15). Dopaminergic projections are in general very diffuse and reach large portions of the brain. The time scales of dopamine actions are diverse, from a few hundred milliseconds to several hours. This chapter will focus on the mesencephalic dopamine centers because they are the most studied, and because they are thought to be involved in diseases such as Tourette's syndrome, schizophrenia, Parkinson's disease, Huntington's disease, drug addiction, and depression (see Tzschentke, 2001). These centers are also involved in normal brain functions, such as working memory, reinforcement learning, and attention. This chapter briefly summarizes the main roles of dopamine with respect to recent modeling approaches.

Biophysical Effects of Dopamine

The effects of dopamine on membrane currents and synaptic transmission are complex and depend on the nature and distribution of the postsynaptic receptors. At the single-cell level, in the in vitro rat preparation, DA has been found to either increase or decrease the excitability of neurons, through the modulation of specific sets

of sodium, potassium, and calcium currents (see Gullledge and Jaffe, 1998, and Nicola, Surmeier, and Malenka, 2000, for reviews). Although the exact nature of the modulation is still debated, it is likely to depend on the opposing contributions of the D1/D5 and D2/D3 family of dopamine receptors that are respectively positively and negatively coupled with adenylate cyclase. Studies in monkey cortical tissue showed that the D1/D5 family of receptor was 20-fold more abundant than the D2/D3 family, and that these receptors were present distally in both pyramidal and non-pyramidal cells (Goldman-Rakic, Muly, and Williams, 2000).

Dopamine modulates excitatory and inhibitory synaptic transmission. Although the nature of neuromodulation of inhibitory transmission is still debated, it appears that in both the cortex and the striatum, D1 receptor activation selectively enhances NMDA but not AMPA synaptic transmission. Because of their voltage dependence, NMDA currents are smaller at rest than in a depolarized state when the postsynaptic cell is firing. Experimental and theoretical evidence suggest that the dopamine enhancement of NMDA currents may be used to induce working memory-like (see later discussion) bistable states in large networks of pyramidal neurons (Lisman, Fellous, and Wang, 1998).

In rats in vivo, stimulation of the ventral tegmental area or local application of dopamine decreases the spontaneous firing of the prefrontal cortex (Thierry et al., 1994), striatum, and nucleus accumbens (Nicola et al., 2000), suggesting that dopamine may be able to control the levels of noise, and hence signal-to-noise ratios.

Given that dopamine modulation strongly depends on the particular distribution of D1/D5 and D2/D3 receptors and on the particular pattern of incoming synaptic transmission, the biophysical effects of dopamine on the intrinsic and synaptic properties is likely to differ from one neuron to the next, raising the intriguing possibility of the existence of several subclasses of neurons that differ mainly by their responses to this neuromodulator.

Dopamine Levels Influence Working Memory

Working memory refers to the ability to hold a few items in mind, with the explicit purpose of working with them to yield a behavior (SHORT-TERM MEMORY). Typically, working memory tasks, such as spatial delayed match-to-sample tasks, consist of the brief presentation of a cue-stimulus (bright dot flashing once) in one of the four quadrants of a screen, followed by a delay period of several seconds, and by a test in which the subject has to respond only if the test stimulus appears the same quadrant as the cue-stimulus. Single-cell studies in monkeys revealed that some prefrontal cortical cells increased their firing rate during the delay period, when the stimulus is no longer present but when the animal has to remember its location in order to later perform the correct action. Both pyramidal cells and interneurons may present this property. The activity of these cells is stimulus dependent, so that only the cells that encode for the spatial location where the cue-stimulus occurred remain active during the delay period.

Local iontophoretic administrations of DA in the prefrontal cortex of monkeys performing a working memory task increase the cells' firing rate during the delay period, without increasing background noise, essentially increasing the signal-to-noise ratio during the task. There is, however, an optimal level of dopamine concentration above and below which working memory becomes impaired. Current theories propose that this effect is due to the enhancement by dopamine of excitatory inputs on pyramidal cells and interneurons. Because DA is more effective in facilitating excitatory transmission on pyramidal cells than on interneurons, intermediate levels of DA improve performance, while higher levels of DA recruit feedforward inhibition and decrease pyramidal cell output, thereby resulting in impairments in the task. Low levels of DA would not be sufficient to induce excitatory facilitation, yielding a poor pyramidal cell output, and hence an impairment (Figure 1 and Goldman-Rakic et al., 2000). There have been a few attempts at modeling the neural substrate of working memory, but very little has yet been done to account for the role of dopamine (Tanaka, 2001).

Dopamine Responses Resemble Reward Prediction Signal of TD Model

A large body of experimental evidence led to the hypothesis that Pavlovian learning depends on the degree of the unpredictability of the reinforcer (Dickinson, 1980). According to this hypothesis, reinforcers become progressively less efficient for behavioral adaptation as their predictability grows during the course of learning. The difference between the actual occurrence and the prediction of the reinforcer is usually referred to as the "error" in the reinforcer prediction. This concept has been used in the temporal-difference model (TD model) of Pavlovian learning (REINFORCEMENT LEARNING in MOTOR CONTROL). If the reinforcer is a reward, the TD model uses a reward prediction error signal to learn a reward prediction signal. The error signal progressively decreases and shifts to the time of earlier stimuli that predict the reinforcer. The characteristics of the reward prediction signal are comparable to those of anticipatory responses such as salivation in Pavlov's experiment.

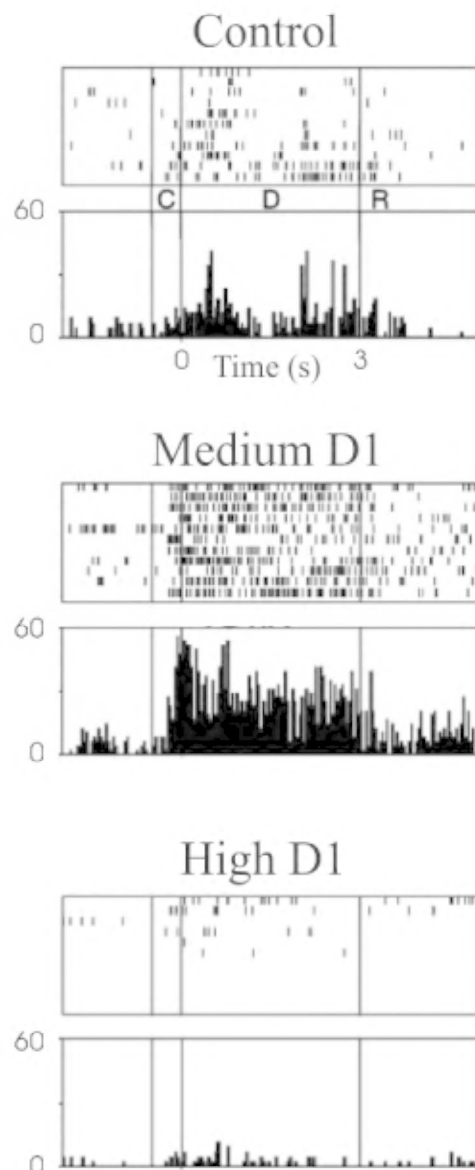
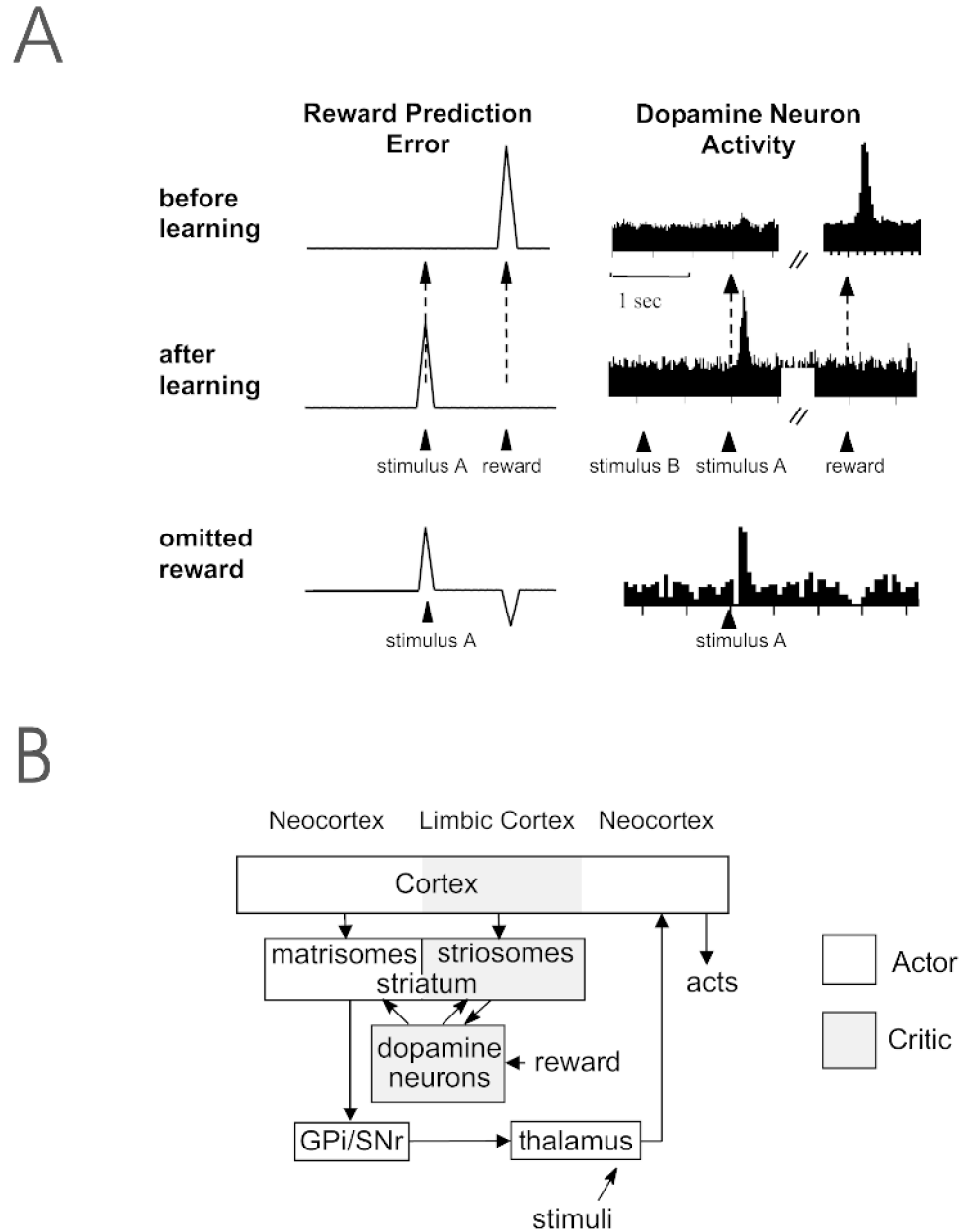


Figure 1. Biphasic effects of dopamine during a working memory task. The task consisted of the brief presentation of a cue (C), a delay of 3 seconds (D), and a response (R). Moderate levels of local application of SCH39166 (25 nA), a D1 receptor agonist, dramatically enhanced the activity of this cell, without significantly increasing its background activity (before cue). Higher levels of SCH39166 (75 nA) decreased the activity of this cell throughout the task. Histogram units are spikes/s. (Figure adapted from Goldman-Rakic, Muly, and Williams, 2000.)

The reward prediction error signal of the TD model remained a purely hypothetical signal until researchers discovered that the activity of midbrain dopamine neurons is strikingly similar to the reward prediction error of the TD model (Figure 2A) (Montague, Dayan, and Sejnowski, 1996; Schultz, 1998). Advances in reinforcement learning theories, as well as evidence for the involvement of dopamine in sensorimotor learning and in cognitive functions, led to the development of the Extended TD model. The reward prediction error signal of the TD model by Suri and Schultz (1999) reproduces dopamine neuron activity in several situations: (1) upon presentation of unpredicted rewards, (2) before, during,

Figure 2. A, Prediction error signal of the TD model (left) similar to dopamine neuron activity (right). If a neutral stimulus A is paired with reward, prediction error signal and dopamine activity respond to the reward (before learning). After repeated pairings, the prediction error signal and dopamine activity are already increased by stimulus A and on baseline levels at the time of the reward (after learning). If the stimulus A is conditioned to a reward but is occasionally presented without reward, the prediction error signal and dopamine activity are decreased below baseline levels at the predicted time of reward (omitted reward). B, Interactions between cortex, basal ganglia, and midbrain dopamine neurons according to the Actor-Critic models. The limbic areas correspond to the Critic and the sensorimotor areas to the Actor. The striatum is divided into matrisomes (sensorimotor) and striosomes (limbic). Limbic cortical areas project to striosomes, whereas neocortical areas chiefly project to matrisomes. Midbrain dopamine neurons are contacted by medium spiny neurons in striosomes and project to both striatal compartments. They are proposed to influence sensorimotor learning in the matrisomes (instrumental learning) and learning of reward predictions in the striosomes (Pavlovian learning). Striatal matrisomes inhibit the basal ganglia output nuclei Gpi/SNr and can elicit actions due to their projections via thalamic nuclei to motor cortical areas. Several additional functions of this architecture were proposed in Suri, Bargas, and Arbib, 2001. (Figure adapted from Suri and Schultz, 1998.)



and after learning that a stimulus precedes a reward, (3) when two stimuli precede a reward with fixed time intervals, (4) when the interval between the two stimuli are varied, (5) when a reward is unexpectedly omitted, (6) when a reward is delayed, (7) when a reward occurs earlier than expected, (8) when a reward-predictive stimulus is unexpectedly omitted, (9) when there is a novel, physically salient stimulus that has never been associated with reward (see allocation of attention, discussed later in this chapter), (10) and when a blocking paradigm is used. To reach this close correspondence, three constants of the TD model were tuned to characteristics of dopamine neuron activity (learning rate, decay of eligibility trace, and temporal discount factor), some weights were initialized with positive values to achieve (9), and some ad hoc changes of the TD algorithm were introduced to reproduce (7) (see later discussion).

In Pavlov's experiment, the salivation response of the dog does not influence food delivery. The TD model is a model of Pavlovian

learning and therefore computes predictive signals, corresponding to the salivation response, but does not select optimal actions. In contrast, instrumental learning paradigms, such as learning to press a lever for food delivery, demonstrate that animals are able to learn to perform actions that optimize reward. To model sensorimotor learning in such paradigms, a model component called the Actor is taught by the reward prediction error signal of the TD model. In such architectures, the TD model is also called the Critic. This approach is consistent with animal learning theory and was successfully applied to machine learning studies (REINFORCEMENT LEARNING IN MOTOR CONTROL). Midbrain dopamine neurons project to the striatum and cortex and are characterized by rather uniform responses throughout the whole neuron population. Computational modeling studies with Actor-Critic models show that such a dopamine-like reward prediction error can serve as a powerful teaching signal for learning with delayed reward and for learning of motor sequences (Suri and Schultz, 1999). These models are

also consistent with the role of dopamine in drug addiction and electrical self-stimulation (see later discussion). Comparison of the Actor-Critic architecture to biological structures suggests that the Critic may correspond to pathways from limbic cortex via limbic striatum (or striosomes) to dopamine neurons, whereas the Actor may correspond to pathways from neocortex via sensorimotor striatum (or matrisomes) to basal ganglia output nuclei (BASAL GANGLIA) (Figure 2B). Although this standard Actor-Critic model mimics learning of sensorimotor associations or habits, it does not imply that dopamine is involved in anhedonia.

Allocation of Attention

Several lines of evidence suggest that dopamine is also involved in attention processes. Although the firing rates of dopamine neurons can be increased or decreased for aversive stimuli, dopamine concentration in striatal and cortical target areas are often increased (Schultz, 1998). Both findings are not necessarily inconsistent, since small differences in firing rates of dopamine neurons are hard to detect with single neuron recordings, and measurement methods for dopamine concentration usually have less temporal resolution than methods used to measure spiking activity. Furthermore, dopamine concentration is not only influenced by dopamine neuron activity but also by local regulatory processes. Slow changes in cortical or striatal dopamine concentration may signal information completely unrelated to reward. Also, relief following aversive situations may influence dopamine neuron activity as if it were a reward, which would be consistent with opponent processing theories (CONDITIONING). Allocation of attentional resources seems to determine dopamine neuron activity in situations when a reward is delivered earlier than usual. In contrast to any linear model, including the standard TD model, dopamine neuron activity is on baseline levels at the time of the expected reward in this situation. This suggests that delivery of the reward earlier than usual seems to reallocate attentional resources through competitive mechanisms (Suri and Schultz, 1999).

Dopamine neurons respond to a novel, physically salient stimulus even if the stimulus has never been associated with a reward (Schultz, 1998). In contrast to reward-predictive responses, for stimuli of equal physical salience, the increase due to novelty responses seems to be smaller and is followed by a pronounced decrease of neural activity below baseline levels. (Brief and less pronounced decreases of dopamine neuron activity sometimes also occur after a response to a reward.) In contrast to responses to conditioned stimuli, novelty responses extinguish for repeated stimulus presentations. The characteristics of this novelty response are consistent with the TD model if certain associative weights are initialized with positive values instead of using initial values of zero (Suri and Schultz, 1999). Such weights initialization with positive values was proposed in machine learning studies to stimulate exploration of novel actions. Simulation studies demonstrated that such a novelty bonus hardly influences slow movements of more than 100 msec duration because the effects of the two phases in the firing of dopamine neurons cancel out and the movement starts after the biphasic response. However, dopamine novelty responses may stimulate exploration for very brief actions, which may include saccades or allocation of attentional resources (Suri and Schultz, 1999).

Redgrave and collaborators (Redgrave, Prescott, and Gurney, 1999) argued that the latency of dopamine responses is too short to be consistent with the hypothesis that dopamine is a reward prediction signal. Onsets of dopamine novelty responses as well as reward responses seem to occur just before the start of the saccade or during the saccade. The dopamine response will likely occur after the superior colliculus has detected a visual target but prior to the triggering (by collicular neurons) of the saccadic movement

required to bring the target to the fovea. If it is assumed that the animal must execute a saccade to a visually presented stimulus before it can adequately assess its predictive value, the latency of dopamine response would be too short to signal reward. We argue against this view. Neural activities in cortical and subcortical areas reflect the anticipated future visual image before a saccade is elicited (Ross et al., 2001). Therefore, the representations of future visual images may influence dopamine neuron activity as if the saccade has already been executed, and thus the dopamine response may start slightly before the saccade. The Extended TD model computes such predictive signals and uses them to select goal-directed actions in a cognitive task (Suri, Vargas, and Arbib, 2001). According to this Actor-Critic model, the interactions between dopamine neuron activities (computed by Critic) and activities that reflect the preparation for intended actions (in Actor) select the actions that maximize reward predictions. The model evaluates the expected values of future actions, without necessarily executing them, in order to select the action with the optimal predicted outcome. The model selects the optimal action from such “action ideas” or “imagined actions.” This optimal action is selected by assuming that dopamine neuron activity increases the signal-to-noise-ratio in target neurons. According to this advanced Actor-Critic model, dopamine improves focusing of attention to intended actions and selects actions. Since some neural activities anticipate the retinal images that result in saccades before these saccades are executed (Ross et al., 2001), animals may indeed use such predictive mechanisms for the selection of intentional saccades. Furthermore, similar internal mechanisms may bias intentional switching capabilities of the basal ganglia to facilitate the allocation of behavioral and cognitive processing capacity toward unexpected events (see BASAL GANGLIA and Redgrave et al., 1999). If we assume similar functions of dopamine for short-term memory, this model suggests that dopamine may select the items that should be kept in short-term memory and may also help to sustain their representation over time.

Conclusions

In vitro studies of the biophysical effects of dopamine demonstrate a wide range of dopamine effects on the intrinsic and synaptic properties of individual cells. In vivo studies suggest, however, that the main overall effect of dopamine may be to control noise levels and to selectively enhance the signal-to-noise-ratio of neural processing. This action may behaviorally lead to an improvement of working memory and to better selection of goal-directed actions. The TD model reproduces dopamine neuron activity in many behavioral situations and suggests that dopamine neuron activity codes for an error in reward prediction. This chapter described a TD model that solves cognitive tasks including goal-directed actions (also called planning or intentional actions) and that attempts to reproduce the function of dopamine in attention and preparation processes.

Road Map: Biological Networks

Related Reading: Basal Ganglia; Emotional Circuits; Neuromodulation in Invertebrate Nervous Systems; Neuromodulation in Mammalian Nervous Systems

References

- Dickinson, A., 1980, *Contemporary animal learning theory*, Cambridge, UK: Cambridge University Press.
- Goldman-Rakic, P. S., Muly III, E. C., and Williams, G. V., 2000, D(1) receptors in prefrontal cells and circuits, *Brain Res. Rev.*, 31:295–301.
- Gulledge, A. T., and Jaffe, D. B., 1998, Dopamine decreases the excitability of layer V pyramidal cells in the rat prefrontal cortex, *J. Neurosci.*, 18:9139–9151.

- Lisman, J. E., Fellous, J.-M., and Wang, X.-J., 1998, A role for NMDA-receptor channels in working memory, *Nature Neurosci.*, 1:273–275. ♦
- Montague, P. R., Dayan, P., and Sejnowski, T. J., 1996, A framework for mesencephalic dopamine systems based on predictive Hebbian learning, *J. Neurosci.*, 16:1936–1947.
- Nicola, S. M., Surmeier, J., and Malenka, R. C., 2000, Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens, *Annu. Rev. Neurosci.*, 23:185–215.
- Redgrave, P., Prescott, T. J., and Gurney, K., 1999, Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.*, 22:146–151.
- Ross, J., Morrone, M. C., Goldberg, M. E., and Burr, D. C., 2001, Changes in visual perception at the time of saccades, *Trends Neurosci.*, 24:113–121.
- Schultz, W., 1998, Predictive reward signal of dopamine neurons, *J. Neurophysiol.*, 80:1–27.
- Suri, R. E., and Schultz, W., 1998, Learning of sequential movements by neural network model with dopamine-like reinforcement signal, *Exp. Brain Res.*, 121:350–354.
- Suri, R. E., and Schultz, W., 1999, A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task, *Neurosci.*, 91:871–890. ♦
- Suri, R. E., Vargas, J., and Arbib, M. A., 2001, Modeling functions of striatal dopamine modulation in learning and planning, *Neurosci.*, 103:65–85. ♦
- Tanaka, S., 2001, Computational approaches to the architecture and operations of the prefrontal cortical circuit for working memory, *Prog. Neuropsychopharmacol. Biol. Psychiat.*, 25:259–281. ♦
- Thierry, A. M., Jay, T. M., Pirot, S., Mantz, J., Godbout, R., and Glowinski, J., 1994, Influence of afferent systems on the activity of the rat prefrontal cortex: Electrophysiological and pharmacological characterization, in *Motor and Cognitive Functions of the Prefrontal Cortex* (A. M. J. Thierry, P. S. Goldman-Rakic, and Y. Christen, Eds.), New York: Springer-Verlag. pp. 35–50.
- Tzschentke, T. M., 2001, Pharmacology and behavioral pharmacology of the mesocortical dopamine system, *Prog. Neurobiol.*, 63:241–320.

Dynamic Link Architecture

Christoph von der Malsburg

Introduction: The Architecture

The dynamic laws governing the brain's physical elements and their interaction enable it to fall into functionally useful states. The term neural architecture is taken here as referring to the shape of these dynamic laws. This article presents and discusses *dynamic link architecture* (DLA; von der Malsburg, 1981, 1985, 1986). There are various ways in which DLA has been couched in terms of equations (von der Malsburg, 1985; von der Malsburg and Schneider, 1986; von der Malsburg and Bienenstock, 1987; Bienenstock and von der Malsburg, 1987; Wiskott and von der Malsburg, 1996; Zhu and von der Malsburg, 2001). Because DLA has not yet received a canonical mathematical description, it is described here in abstract verbal terms. DLA is a construction site, and this article is an invitation to work at it.

According to DLA, the brain's data structure has the form of graphs composed of nodes (called units) connected by links. The graphs of DLA are dynamic: both units and links bear activity variables changing on the rapid functional time scale of fractions of a second. Graphs form a very versatile data format that is probably able to render the structure of any mental object. A particularly important feature is the ability of graphs to compose more complex data structures from simpler ones, an important requirement for the expression of cognitive structures (see COMPOSITIONALITY IN NEURAL SYSTEMS).

The units of DLA play the role of symbolic elements. This follows the tradition of associating neurons with elementary meaning (the identification of units with neurons is, however, not taken for granted here; see below). Units are endowed with structured signals changing in time. These signals can be evaluated under two aspects, intensity and correlation. Intensity measures the degree to which a unit is active in a given time interval, signifying the degree to which the meaning of the unit is alive in the mind of the animal. Correlations, on the other hand, quantify the degree to which the signal of one unit is related to that of others. The general idea is that identical signal patterns are strongly correlated, whereas statistically independent signal patterns have zero correlation. A correlation can be a binary relation, characterizing two units, or an n -ary relation, to be evaluated for n units at a time.

The strength of links can change on two time scales, represented by two variables called *temporal weight* and *permanent weight*.

The permanent weight corresponds to the usual synaptic weight, can change on the slow time scale of learning, and represents permanent memory. The temporary weight is constrained to the interval between zero and the permanent weight and can change on the same time scale as the unit activity (hence the name dynamic links).

Dynamic links constitute the glue by which higher data structures are built up from more elementary ones. Conversely, the absence of links (temporary or permanent) keeps mental objects separate from each other and prevents their direct interaction. In the simplest case, a link binds a descriptor to an object. For example, a link may bind a unit representing a color to another unit that stands for a specific object. More generally, mental objects are formed by binding together units representing constituent parts. The infinite richness and flexibility of the mind is thus made possible as a combinatorial game. The mental activity of familiar objects (like my grandmother, or a yellow Volkswagen) may be reliably correlated with the activity of specialized units, but these objects still acquire their substance—their imagined visual appearance, and so on—by the dynamical binding of appropriately structured arrays of other units. Units can be part of different functional contexts. They are integrated into a specific one by the activation of appropriate links. Dynamic links are the means by which the brain specializes its circuit diagram to the needs of the particular situation at hand.

Graph Dynamics

Under the influence of signal exchange, graphs and their units and links are subject to dynamic change, constituting a game of network self-organization (see SELF-ORGANIZATION AND THE BRAIN). The dynamic links have a resting strength near the value of the permanent weight. When the units connected by a permanent link become active, there is rapid feedback between the units' signal correlations and the link's strength, with a strong link tending to increase signal correlation and a strong correlation controlling the link to grow in strength toward the maximum set by the permanent weight. This feedback can also lead to a downward spiral, with a weak correlation reducing a link's strength and a weak link losing its grip on signals, which, under the influence of other links, drift apart toward lower correlation. Thus, links between active units

tend to be driven toward one of their extreme values, zero or the maximum set by the permanent weight.

Links are subject to divergent and convergent competition: links converging on one unit compete with each other for strength, as do links diverging from one unit. This competition drives graphs to sparsity. Links are also subject to cooperation. Several links carrying correlated signal structure cooperate in imposing that signal structure on a common target unit, helping them all to grow. Because the ultimate cause for all signal structure is random, correlations can only be generated on the basis of common origin of pathways. Thus, cooperation runs between pathways that start at one point and converge to another point. The common origin of converging pathways may, of course, be an event or a pattern in the environment.

Cooperation and competition conspire to favor certain graph structures. These are distinguished by being sparse (that is, activating relatively few of the permanent links in or out of units) and by having a large number of cooperative meshes—arrangements of alternative pathways from one source to one target unit. Beyond these statements, a general characterization of graph attractor states is an open issue. However, there are certain known graph structures that have been shown in simulations to be attractor states and that prove to be very useful (see the section on applications). All of these graph structures may be characterized as “topological graphs”: if their units are mapped appropriately into a low-dimensional display space (one- or two-dimensional in the known examples), the links of those graphs all run between units that are neighbors in the display space.

Slow Plasticity

In classical neural architectures, learning is modeled by *synaptic plasticity*, or the change of permanent synaptic weights under the control of neural signals. This general idea is also part of the dynamic link architecture. However, DLA imposes a further refinement in that a permanent weight grows only when the corresponding dynamic link has converged to its maximum strength, which happens only in the context of an organized graph structure. For a permanent link to grow, it is thus not sufficient for the two connected units to have high intensity in the same brain state; in addition, their signals must be correlated and their link must be active. This puts the extra condition on the growth of permanent connection weights that they be validated by indirect evidence, in the form of active indirect pathways between the units connected, and in the form of relative freedom from competition, the two conditions characterizing a well-structured dynamic graph. Thus, only the very few connections that are significant in this sense can grow.

Neural Implementation of Dynamic Links

How can the units, links, and dynamical rules of DLA be identified with known neural structures? This is possible in several ways. It will turn out that to some extent, DLA can be seen as a fair interpretation of known structures, whereas some experimental predictions also flow from it.

Units Are Individual Neurons

At the most fundamental level, units are to be identified with neurons, links with axons and synapses, signals with neural spike trains, and permanent weights with conventional synaptic strengths. Signal intensity is evaluated as firing rate, averaged over intervals of length Δ , whereas the stochastic signal fine structure within that interval is evaluated in terms of correlations with a resolution time τ , two spikes arriving within τ of each other being counted as simultaneous. The smallest reasonable choice for Δ is

probably 100 ms or a little less; the smallest choice for τ may be 3 ms, as proposed in von der Malsburg (1981). Neural signals in the cerebral cortex have a very rich stochastic structure on all time scales, much of which is not correlated strongly with external stimuli in neurophysiological experiments (and is usually suppressed by averaging in a post-stimulus time histogram).

A point of contention at the present time is the precision with which nervous tissue can process temporal signal structure. Some authors (e.g., Shadlen and Movshon, 1999) believe that fine temporal structure cannot be transmitted by neurons, and that meaningful signal correlations cannot be extracted. The proposed argument is, however, circular, as it was assumed that neural input signals are random and independent. If this assumption is violated in the brain, the argument falls flat. Indeed, it has been shown (Mainen and Sejnowski, 1995) that spike timing of cortical neurons can be reliable with 1-ms precision if neural input is sufficiently structured, and similarly precise spike timing was found in response to temporally structured visual input seven synaptic generations behind the retina (Bair and Koch, 1995). The latter and other studies would encourage the assumption of a τ of 1 ms. A heated discussion has also sprung up around the status of the interpretation of signal correlations in terms of dynamic binding (SYNCHRONIZATION, BINDING AND EXPECTANCY) as proposed in von der Malsburg (1981; see also Shadlen and Movshon, 1999; Gray, 1999; Singer, 1999; von der Malsburg, 1999, and other articles in the same issue of *Neuron*). At the present time, the issue is the subject of intensive experimental study in many laboratories.

Dynamic links are realized at the single-neuron level as rapid reversible synaptic plasticity (RRP). Starting from a resting value, the temporary weight of a synapse is increased by correlations between the pre- and postsynaptic signals and is decreased if both signals are active in a given time period but are not correlated. The resting weight of a synapse is probably not too far from the maximum set by the permanent weight (so that RRP will manifest itself mainly in the form of rapid weight reduction). The interactions between temporary synaptic strength and signals is such as to constitute a positive feedback loop. Changes in temporary synaptic weights must take place on a fast time scale to be of functional significance, possibly as quickly as within 10 ms. In the prolonged absence of presynaptic or postsynaptic activity, the temporary weight rises or falls back toward its resting value, with a time scale that corresponds to short-term memory (perhaps a few dozen seconds), or it is reset by an active mechanism (for example, in the visual cortex during saccades). Convergent synaptic competition (competition between synapses at the same postsynaptic neuron) could be implemented by the signals arriving on one synapse or one set of synapses spoiling the postsynaptic activity for others. Divergent competition could be implemented with the help of inhibition between the target cells, making it difficult or impossible to synchronize them all with the same presynaptic signal.

The existence of rapid reversible changes in synaptic strength in cortex is a broadly documented experimental fact (see Hempel et al., 2000, for examples and a review of the experimental literature; see also TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES). What has not been investigated experimentally in any detail is the dependence of rapid synaptic change on postsynaptic signals, and without such study the type of control postulated in RRP cannot be ascertained. There are many open details that must be determined experimentally. Among them is the identity of the relevant postsynaptic signal (membrane potential, some second messenger, e.g., Ca^{2+} , action potential, or other) and the precise definition of the dynamics of synaptic strength (which could require a delay between the presynaptic and postsynaptic signals, as described for long-term potentiation in Senn, Markram, and Tsodyks, 2001, and experimental work reviewed therein).

As was pointed out, implementation of DLA at the single-neuron level can be realized on a hierarchy of time scales Δ and the concomitant resolution time τ . So far I have discussed the faster end of the hierarchy. If Δ is taken to be a large fraction of a second or longer, we are in the domain of overt attention and the well-studied phenomenon of the mind shifting context sequentially on smaller and larger time scales. There is no doubt that a very important function of attention is keeping topics separate if their simultaneous activation would lead to confusion, and thus to provide temporal binding. A proper understanding of the mechanisms of attention will have to provide an answer to the question of how the focus of attention is formed. Part of the answer will be, of course, that it must unite elements that have something to do with each other (as recorded by the links between them), and not to activate simultaneously what would lead to confusion. The conceptual framework of DLA and its network self-organization are therefore appropriate for the description of attention dynamics.

Multicellular Units

Just as the DLA interpretation of neural dynamics can be applied at different temporal scales, it can also be applied at different spatial scales, either by identifying units with single neurons, as above, or by identifying units with groups of neurons. In this perspective, all individual neurons in a group, called a *multicellular unit* (MCU), are interpreted to have the same meaning. They differ, however, in the synaptic connections they have to neurons in other MCUs. The signal intensity of an MCU is the combined neural activity of all of its neurons. Signal correlations, however, are calculated by paying attention to the distribution of activity in MCUs and determining the combined synaptic weights of all connections between currently active cells in a pair of MCUs. By changing the distribution of activity over its neurons, an MCU can control the connectivity pattern it has to other MCUs. An important example of MCUs is constituted by the hypercolumns in visual cortex. All neurons in a hypercolumn have the same theme, subserving, by definition, one point in visual space. The neurons differ, however, in how they are connected with afferent neurons (which gives them different meaning on a more fine-grained level) and how they are connected to neurons in other hypercolumns.

MCU implementation of DLA differs in important points from the single-neuron implementation. The variables that constitute dynamic links are not temporary synaptic weights but neural spike activity, and correlations are not computed by time-consuming temporal integration over pairs of neural signals but by the instantaneous and parallel evaluation of the signals and connection weights of all active neurons in the MCUs involved. In consequence, with MCUs there is a much greater capacity to express highly structured graphs than in the single-cell implementation. The price for this greater power is much reduced flexibility, because appropriately specialized connectivity patterns within and between MCUs must first be installed.

Implementations of DLA on different temporal and spatial scales are not mutually exclusive and are probably realized concurrently in our brain. The single-cell version is indispensable because of its great flexibility and the absence of any need for specialized pre-existing connectivity structures, but it is limited in its capacity to distinguish detailed link structures in limited time. The MCU version is very powerful and may be seen as just an unconventional view of networks in classic architecture, but it requires highly specialized connectivity structures and appropriately tuned activity dynamics.

Applications

The aim of DLA is to serve as a framework for understanding brain functions. Conventional neural network architecture, lacking the

equivalent of dynamic binding, may be a universal medium for realizing individual functions when they are defined ahead of time (such that appropriate combination-coding neurons and connectivity patterns can be defined and binding ambiguities avoided), but in decades of modeling attempts, this architecture has shown itself to be too narrow to go beyond elementary functions. DLA has the full functional repertoire of conventional neural network architecture but goes beyond it in being able to build up structured objects, have them interact in a structured way, or keep them from interfering. The full potential of DLA is far from realized, but some applications have already been modeled, as briefly reviewed in von der Malsburg (1999).

Figure-ground segmentation in visual scenes or other modalities is most naturally modeled by DLA (see VISUAL SCENE SEGMENTATION). So far, most concrete models have employed temporal signal correlations to bind all elements of a figure together and to keep them separate from elements belonging to the ground. For this type of model there is experimental evidence, as reviewed in Gray (1999) and Singer (1999). Also, MCU implementations have been realized in which each unit is subdivided into two subpopulations, one for figure, one for ground, and in a final state all units belonging to the figure restrict their activity to the "figure" neurons, all units in the ground just activate their "ground" neurons.

Many mental objects are met first as sensory arrays of local features. They are most naturally handled, stored, and recognized if the neighborhood relations between features are expressed as bindings and stored and retrieved as dynamic links. This has been realized as *dynamic link matching* for the purpose of invariant visual object recognition (reviewed in von der Malsburg, 1999, and FACE RECOGNITION: NEUROPHYSIOLOGY AND NEURAL TECHNOLOGY). Dynamic link matching, implemented in terms of temporal binding, has rightly been criticized as too slow to account for object recognition in adults. However, a recent implementation employing direct interaction between links (to be implemented with the help of MCUs, for instance) was shown to be very fast, requiring only one or a few iterations (Zhu and von der Malsburg, 2001).

As to the potential of DLA for modeling brain function and cognitive processes, the cited applications are but the tip of the iceberg. Processing and learning the syntactical structure of natural language on the basis of conventional neural architecture has proved very difficult to impossible. The reason is that the flexibility to analyze or to form novel sentences requires dynamic binding. It is particularly important here to realize the general process of instantiation, in which an abstract syntactical structure is applied to a concrete set of elements. Instantiation requires the manipulation of dynamic links between abstract roles and concrete role fillers, and requires the recognition of structural relations between abstract structures and concrete instances. Both of these functions are not part of the repertoire of conventional neural networks.

Road Maps: Artificial Intelligence; Neural Plasticity; Vision

Related Reading: Structured Connectionist Models; Synchronization, Binding and Expectancy; Visual Scene Segmentation

References

- Bair, W., and Koch, C., 1995, Precision and reliability of neocortical spike trains in the behaving monkey, in *Computation and Neural Systems* (J. Bower, Ed.), Norwell, MA: Kluwer, pp. 53–58.
- Bienenstock, E., and von der Malsburg, C., 1987, A neural network for invariant pattern recognition, *Europhys. Lett.*, 4:121–126.
- Gray, C. M., 1999, The temporal correlation hypothesis of visual feature integration: Still alive and well, *Neuron*, 24:31–47.
- Hempel, C. M., Hartman, K. H., Wang, X.-J., Turigiano, G. G., and Nelson, S. B., 2000, Multiple forms of short-term plasticity at excitatory

- synapses in rat medial prefrontal cortex, *J. Neurophysiol.*, 83:3031–3041.
- Mainen, Z. F., and Sejnowski, T. J., 1995, Reliability of spike timing in neocortical neurons, *Science*, 268:1503–1506.
- Senn, W., Markram, H., and Tsodyks, M., 2001, An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing, *Neural. Computation*, 13:35–67.
- Shadlen, M. N., and Movshon, J. A., 1999, Synchrony unbound: A critical evaluation of the temporal binding hypothesis, *Neuron*, 24:67–77.
- Singer, W., 1999, Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24:49–65.
- von der Malsburg, C., 1981, *The Correlation Theory of Brain Function*, MPI Biophysical Chemistry, Internal Report 81–2, reprinted in *Models of Neural Networks II* (E. Domany, J. L. van Hemmen, and K. Schulten, Eds.), Berlin: Springer-Verlag, 1994, chap. 2, pp. 95–119. ♦
- von der Malsburg, C., 1985, Nervous structures with dynamical links, *Ber. Bunsenges. Phys. Chem.*, 89:703–710.
- von der Malsburg, C., 1986, Am I thinking assemblies? in *Proceedings of the Trieste Meeting on Brain Theory, October 1984* (G. Palm and A. Aertsen, Eds.), Berlin: Springer-Verlag, pp. 161–176.
- von der Malsburg, C., 1999, The what and why of binding: The modeler's perspective, *Neuron*, 24:95–104.
- von der Malsburg, C., and Bienenstock, E., 1987, A neural network for the retrieval of superimposed connection patterns, *Europhys. Lett.*, 3:1243–1249.
- von der Malsburg, C., and Schneider, W., 1986, A neural cocktail-party processor, *Biol. Cybern.*, 54:29–40.
- Wiskott, L., and von der Malsburg, C., 1996, Face recognition by dynamic link matching, in *Lateral Interactions in the Cortex: Structure and Function* (J. Sirosh, R. Miikkulainen, and Y. Choe, Eds.), electronic book, available: <http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96/>.
- Zhu, J., and von der Malsburg, C., 2001, Synapto-synaptic interactions speed up dynamic link matching, presented at the Computational Neuroscience Meeting (CNS*01), San Francisco, CA, June 30–July 5, 2001.

Dynamic Remapping

Alexandre Pouget and Terrence J. Sejnowski

Introduction

The term *dynamic remapping* has been used in many different ways, but one of the clearest formulations of this concept comes from the mental rotation studies by Georgopoulos et al. (1989) (see also MOTOR CORTEX: CODING AND DECODING OF DIRECTIONAL OPERATIONS). In these experiments monkeys were trained to move a joystick in the direction of a visual stimulus or 90° counterclockwise from it. The brightness of the stimulus indicated which movement was required on a particular trial; a dim light corresponded to a 90° movement and a bright light to a direct movement. An analysis of reaction time suggested that, by default, the initial motor command always pointed straight at the target and then continuously rotated if the cue indicated a 90° rotation, an interpretation that was subsequently confirmed by single unit recordings.

The term *remapping* is also commonly used whenever a sensory input in one modality is transformed to a sensory representation in another modality. The best-known example in primates is the remapping of auditory space, from head-centered in the early stages of auditory processing to the retinotopic coordinates used in the superior colliculus (Jay and Sparks, 1987). This type of remapping, equivalent to a change of coordinates, is closely related to sensorimotor transformations. It does not have to be performed over time but could be accomplished by the neuronal circuitry connecting different representations.

This review is divided into three parts. In the first part, we briefly describe the types of cortical representations typically encountered in dynamic remapping. We then summarize the results from several physiological studies where it has been possible to characterize the responses of neurons involved in temporal and spatial remappings. Finally, in the third part, we review modeling efforts to account for these processes.

Neural Representation of Vectors

A saccadic eye movement toward an object in space can be represented as a vector \mathbf{S} whose components S_x and S_y correspond to the horizontal and vertical displacement of the eyes. Any sensory, or motor, variable can be represented by a similar vector. There are two major ways of representing a vector in a neural population—by a topographic map and by a nontopographic vectorial representation.

The encoding of saccadic eye movements in the superior colliculus is an example of a topographic map representation. A saccade is specified by the activity of a two-dimensional layer of neurons organized as a Euclidean manifold (see COLLICULAR VISUOMOTOR TRANSFORMATIONS FOR GAZE CONTROL). Before a saccade, a bump of activity appears on the map at a location corresponding to the horizontal and vertical displacement of the saccade.

Another example of a vectorial code is the code for the direction of hand movements in the primate motor cortex. Neurons in the primary motor cortex respond maximally for a particular direction of hand movement with a cosine tuning curve around this preferred direction (Georgopoulos et al., 1989). This suggests that each cell encodes the projection of the vector along its preferred direction. [Todorov (2000) questions this interpretation, but the precise identity of the vector being encoded in motor cortex is not critical to the issue of remapping.]

In both cases, the original vector can be recovered from the population activity pattern using statistical estimators. Various examples of such estimators are described in POPULATION CODES.

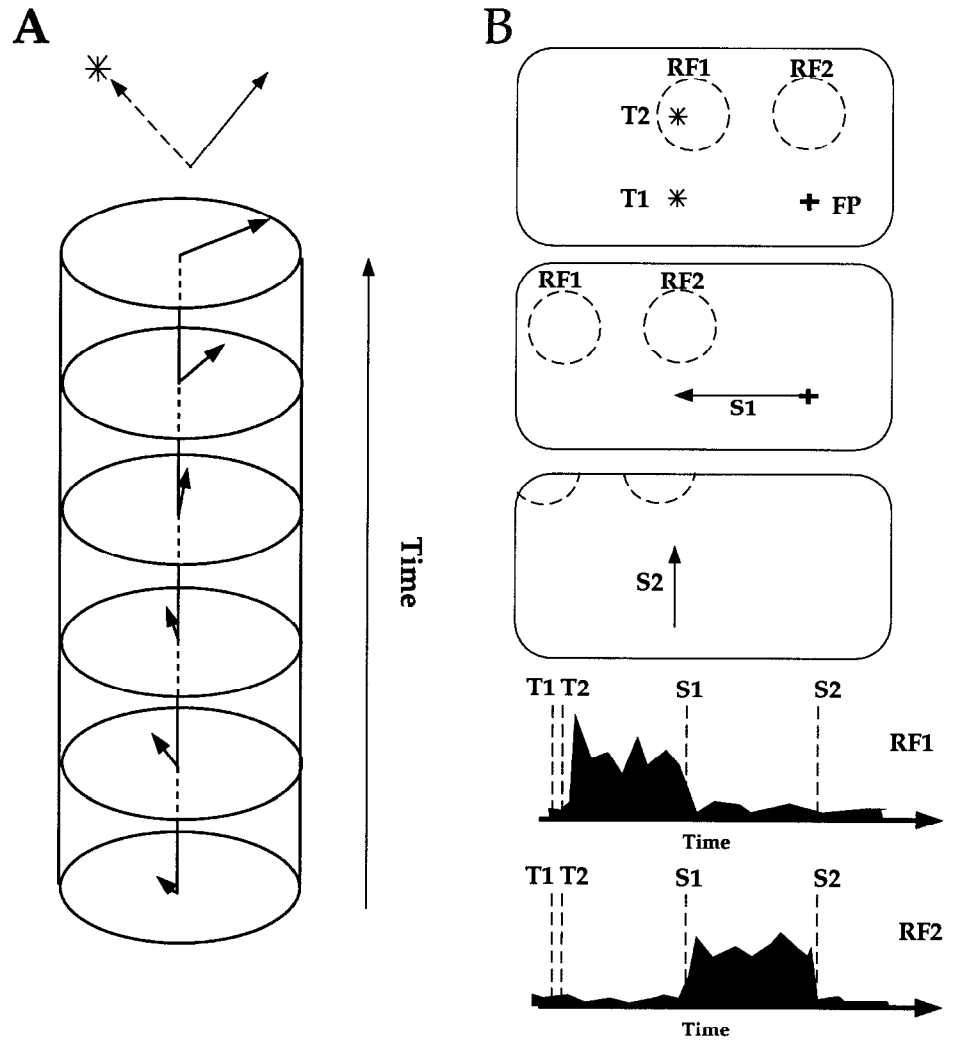
Neurophysiological Correlates of Remapping

Continuous Remappings

Georgopoulos et al. (1989) studied how the population vector varies over time in the mental rotation experiment described in the introduction. They found that for movement 90° counterclockwise from the target, the vector encoded in M1 initially pointed in the target direction and then continuously rotated 90° counterclockwise, at which point the monkey initiated a hand movement (Figure 1A). This is consistent with the interpretation of the reaction time experiments: The monkey had initially planned to move toward the stimulus, and then updated this command according to the task requirement.

Similar continuous remapping occurs in the postsubiculum of the rat, one of the cortical structures involved in navigation of space. Neurons in the postsubiculum provide an internal compass that encodes the direction of the head with respect to remembered visual landmarks. The neurons have bell-shaped tuning curves around their best direction, similar to the code for hand direction in the primary motor cortex. Electrophysiological recordings have revealed that this vector is continuously updated as the head of the

Figure 1. A, Rotation of population vector in the primary motor cortex when the brightness of the target (star) indicates a 90° clockwise movement. (Adapted from Georgopoulos et al., 1989.) B, Saccade remapping. The monkey makes a double saccade (S1 and S2) to the remembered positions of T1 and T2. C, Post-stimulus-time histograms showing the responses of two cells with receptive fields RF1 and RF2 illustrated in Figure 1B. The second cell (RF2) responds only after the first eye movement, encoding the new retinal location of T2, even though it is no longer present on the screen.



rat moves in space, even in complete darkness, suggesting that vestibular inputs are used for this updating (see RODENT HEAD DIRECTION SYSTEM).

Another example of continuous remappings has been reported in a double saccade task. In these experiments, two targets are briefly flashed in succession on the screen and the monkey makes successive saccades to their remembered locations (Figure 1B). Monkeys can perform this task with great accuracy, demonstrating that they do not simply keep a trace of the retinotopic location of the second target, since after the first eye movement this signal no longer corresponds to where the target was in space. Single unit recordings in the superior colliculus, frontal eye field, and parietal cortex have shown that the brain encodes the retinotopic location of the second target before the first saccade occurs. Then while the first eye movement is executed, this information is updated to represent where the second target would appear on the retina after the first saccade (Figure 1C; Mays and Sparks, 1980). In certain cases, this update is predictive; i.e., it starts prior to the eye movement (Duhamel, Colby, and Goldberg, 1992).

Graziano, Hu, and Gross (1997) have reported that the same mechanism appears to be at work in the premotor cortex. Bimodal, visuotactile neurons with receptive fields on the face remap the position of remembered visual stimuli after head movements. It is therefore becoming increasingly clear that continuous remappings

are widespread throughout the brain and play a critical role in sensorimotor transformations.

Although all these examples clearly involve vector remappings, it is not entirely clear that the remappings are continuous. Hence, in the Georgopoulos et al. (1989) experiment, the population vector rotation could be a consequence of the simultaneous decay and growth of the initial planned hand direction and the final one, respectively, without ever activating intermediate directions. This is an example of one-shot remapping considered in the next section. Moreover, it is often difficult to determine whether a remapping in one particular area is computed in that area or is simply the reflection of a remapping in an upstream area.

One-Shot Sensory Remapping

In the inferior colliculus and primary auditory cortex, neurons have bell-shaped auditory receptive fields in space whose positions are fixed with respect to the head. In contrast, in the multisensory layer of the superior colliculus, the positions of the auditory receptive fields are fixed in retinotopic coordinates, which implies that the auditory map must be combined with eye position (Jay and Sparks, 1987). Therefore, the auditory space is remapped in visual coordinates, presumably for the purpose of allowing auditory targets to

be foveated by saccadic eye movements, a function mediated by the superior colliculus.

A similar transformation has been found in the striatum and the premotor cortex, where some of the cells have visual receptive fields in somatosensory coordinates (skin-centered; Graziano et al., 1997). In all cases, these remappings are thought to reflect an intermediate stage of processing in sensorimotor transformations.

These remappings can be considered as a change of coordinates, which correspond to a translation operation. For example, the auditory remapping in the superior colliculus requires the retinal location of the auditory stimulus, \mathbf{R} , which, to a first approximation, can be computed by subtracting its head-centered location, \mathbf{A} , from the current eye position, \mathbf{E} :

$$\mathbf{R} = \mathbf{A} - \mathbf{E} \quad (1)$$

Remapping Models

The remappings we have described so far fall into two categories: vector rotation with a vectorial code (e.g., mental rotation) and vector translation within a topographic map (e.g., auditory remapping in the superior colliculus). These transformations are similar, since rotating a vector within a vectorial representation consists of translating a pattern of activity around a circle. Therefore, in both cases the remapping involves translating a bell-shaped pattern of activity across a map. Most models perform this operation either dynamically through time or in one shot through the hidden layer of a feedforward network (Figure 2).

Dynamical Models

Two kinds of mechanisms have been used in models of continuous remapping: the integration of a velocity signal or the relaxation of a recurrent network.

Integrative model for remapping. In the double saccade paradigm described above, the retinal coordinates of the second target were

updated during the first saccade, a process that might involve moving a hill of activity within the parietal cortex. A model by Droulez and Berthoz (1991) shows how this bump of activity could be moved continuously across the map by integrating the eye velocities during the first saccade (Figure 1A). Their model is essentially a *forward* model of motion: Given a velocity signal, it generates the corresponding moving image. Interestingly, the equations are similar to those used for *inverse* models of motion processing. In both cases, the analysis relies on the assumption that the temporal derivative of a moving image is zero. In other words, the overall gray level profile in the image is unchanged; only the positions of the image features change. It is possible to design a recurrent network to implement this constraint (Droulez and Berthoz, 1991), and the resulting network moves arbitrary patterns of activity in response to an instantaneous velocity signal.

Several variations of this idea have been developed. Dominey and Arbib have shown that an approximation of eye velocity, obtained from the eye position modulated neurons found in the parietal cortex is sufficient for this architecture to work (Dominey and Arbib, 1992). Their simulations show patterns of activation very similar to the ones shown in Figure 1B in the part of their model corresponding to the parietal cortex, FEF, and superior colliculus. Zhang (1996) has used line attractor networks to model head direction cells in the postsubiculum of the rat. In this model, the hill is moved by using the velocity signal—in this case a head velocity signal—to temporarily modify the efficacy of the lateral connections.

Recurrent networks. Mental rotation of a population vector can be reproduced by training a neural network to follow a circular trajectory over time. In this case, the population vector rotates as a consequence of the network dynamics in the absence of any input signals. This approach has been used by Lukashin and Georgopou-

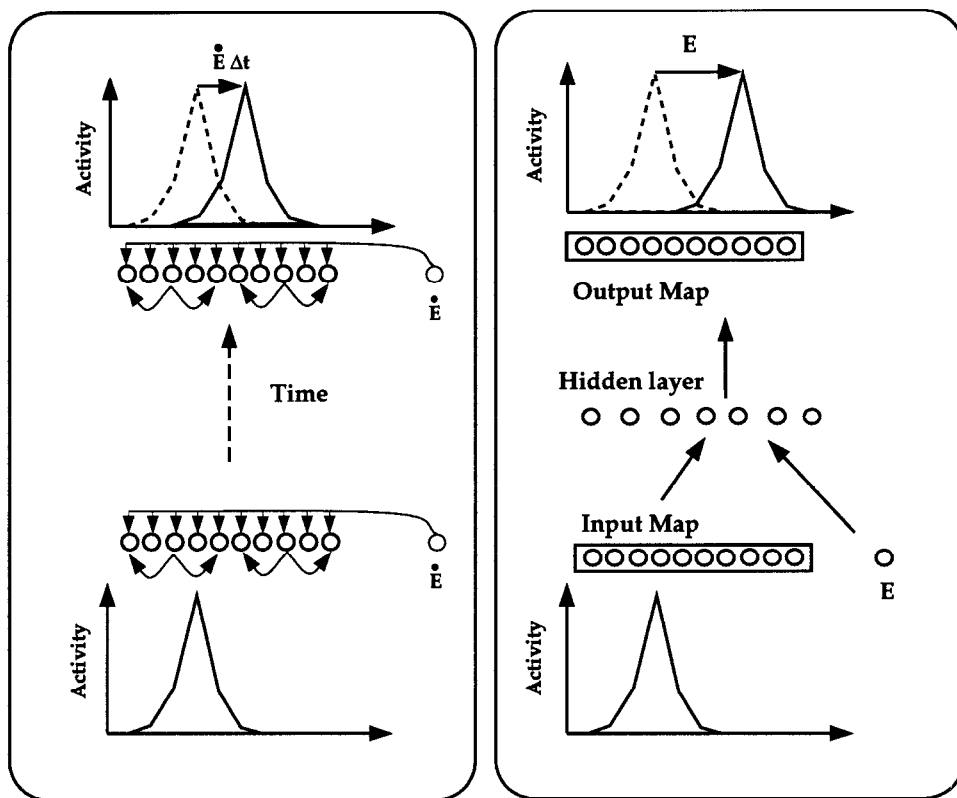


Figure 2. In a map representation, remappings involve moving hills of activity. These hills can be moved continuously in a recurrent network (A), or in one shot in a feedforward network (B). A, The recurrent network dynamically moves the hill of activity according to a velocity signal, \dot{E} . As described in the text, there are several ways to achieve this result. Droulez and Berthoz (1991) integrate the eye velocity signals through the lateral connections while Zhang (1996) uses the eye velocity signals to temporally bias the lateral connections. B, In feedforward remapping, the hill is moved in one shot by the full amount of the current displacement, E , via an intermediate stage of processing in the hidden layer. The weights can be adjusted with a learning algorithm such as backpropagation. Alternatively, one can use basis function units in the hidden layer and train the weights to the output units with a simple learning algorithm such as the delta rule.

los (1994) to model the generation of hand trajectories, but when the trajectory is a circle, mental rotation and a circular hand trajectory are equivalent. Although the model generates a rotating vector, additional mechanisms must be specified to stop the rotation.

Single-Shot Models

Feedforward models have been used for vectorial as well as map representations. They are used whenever the amplitude of the shift is available to the brain beforehand, such as auditory remapping in the superior colliculus in which the shift is directly proportional to the current eye position (Equation 1). In contrast, for mental rotation, the amplitude of the shift is specified by an external stimulus.

Shifter models. As demonstrated by Touretzky, Redish, and Wan (1993), rotation within a vectorial representation can be performed by using a shifter circuit (for more details on shifter circuits, see ROUTING NETWORKS IN VISUAL CORTEX in the First Edition). Their architecture uses N independent circuits, each implementing a rotation through a particular angle. This mechanism is limited in resolution since it rotates only by multiples of $360/N$ degrees. Whether such shifter circuits actually exist in the brain remains to be demonstrated.

Feedforward network models. There are many examples of three-layer networks, and variations thereof, that have been trained or handcrafted to perform sensory remappings. Since these remappings perform vector addition, it might appear unnecessary to deploy a fully nonlinear network for such a task. However, with a map representation, vector addition requires moving a hill of activity in a map as illustrated in Figure 2B, an operation that is highly nonlinear.

Special-purpose nonlinear circuits can be designed to perform this operation (Groh and Sparks, 1992), but more biologically realistic solutions have been found with networks of sigmoidal units trained with backpropagation. Hence, the model of Zipser and Andersen (see GAZE CODING IN THE POSTERIOR PARIETAL CORTEX in the First Edition), which was trained to compute a head-centered map from a retinotopic input, uses hidden units with retinotopic receptive fields modulated by eye position, as in parietal neurons (see also Krommenhoek et al., 1993).

However, backpropagation networks are generally quite difficult to analyze, providing realistic models but little insight into the algorithm used by the network. Pouget and Sejnowski (2001) have explored a way to analyze such networks using the theory of basis functions.

Basis functions. The process of moving a hill of activity in a single shot can be better understood when considered within the larger framework of nonlinear function approximation. For example, consider the feedforward network shown in Figure 2B, applied to a remapping from retinotopic, R_x , to head-centered coordinates, A_x . Because of the map format used in the output later, the responses of the output units are nonlinear in the input variables, namely, the retinal position, R_x , and eye position, E_x .

Therefore, the actual goal of the network is to find an appropriate intermediate representation to approximate this output function. One possibility is to use basis functions of R_x and E_x in the hidden layer (Pouget and Sejnowski, 2001; Salinas and Abbot, 1995).

Perhaps the best-known set of basis functions is the set of cosine and sine functions used in the Fourier transform. Another example is the set of Gaussian or radially symmetric functions with local support (see RADIAL BASIS FUNCTION NETWORKS). A good model

of the response of parietal neurons, which are believed to be involved in remapping, is a set of Gaussian functions of retinal position multiplied by sigmoid functions of eye position. The resulting response function is very similar to that of gain-modulated neurons in the posterior parietal cortex [see GAZE CODING IN THE POSTERIOR PARIETAL CORTEX in the First Edition, and Pouget and Snyder (2000) for a review].

Conclusions

Remappings can be continuous and dynamic or a single shot through several layers of neurons. In both cases, the problem amounts to moving a hill of activity in neuronal maps. Whether some models are better than others is often difficult to establish simply because the neurophysiological data available are relatively sparse. Models can be further constrained by considering deficits that accompany localized lesions in humans (see Pouget and Sejnowski, 2001). These data not only provide valuable insights into the nature of remappings but also might help bridge the gap between behavior and single-cell responses.

Road Map: Vision

Related Reading: Collicular Visuomotor Transformations for Gaze Control; Motion Perception: Elementary Mechanisms; Pursuit Eye Movements; Visual Attention; Visual Scene Perception

References

- Dominey, P., and Arbib, M., 1992, A cortico-subcortical model for the generation of spatially accurate sequential saccades, *Cerebral Cortex*, 2:153–175. ◆
- Droulez, J., and Berthoz, A., 1991, A neural model of sensoritopic maps with predictive short-term memory properties, *Proc. Natl. Acad. Sci. USA*, 88:9653–9657. ◆
- Duhamel, J. R., Colby, C. L., and Goldberg, M. E., 1992, The updating of the representation of visual space in parietal cortex by intended eye movements, *Science*, 255(5040):90–92.
- Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., and Massey, J. T., 1989, Mental rotation of the neuronal population vector, *Science*, 243:234–236.
- Graziano, M., Hu, X., and Gross, C., 1997, Coding the locations of objects in the dark, *Science*, 277:239–241.
- Groh, J., and Sparks, D., 1992, Two models for transforming auditory signals from head-centered to eye-centered coordinates, *Biol. Cybernetics*, 67:291–302.
- Jay, M. F., and Sparks, D. L., 1987, Sensorimotor integration in the primate superior colliculus: I. Motor convergence, *J. Neurophysiol.*, 57:22–34.
- Krommenhoek, K. P., Van Opstal, A. J., Gielen, C. C. A., and Van Gisbergen, J. A. M., 1993, Remapping of neural activity in the motor colliculus: A neural network study, *Vision Res.*, 33:1287–1298.
- Mays, L. E., and Sparks, D. L., 1980, Dissociation of visual and saccade-related responses in superior colliculus neurons, *J. Neurophysiol.*, 43:207–232.
- Pouget, A., and Sejnowski, T. J., 2001, Simulating a lesion in a basis function model of spatial representations: Comparison with hemineglect, *Psychol. Rev.*, 108:653–673. ◆
- Pouget, A., and Snyder, L., 2000, Computational approaches to sensorimotor transformations, *Nature Neurosci.*, 3:1192–1198.
- Salinas, E., and Abbot, L., 1995, Transfer of coded information from sensory to motor networks, *J. Neurosci.*, 15:6461–6474. ◆
- Todorov, E., 2000, Direct cortical control of muscle activation in voluntary arm movements: A model, *Nature Neurosci.*, 3:391–398.
- Touretzky, D., Redish, A., and Wan, H., 1993, Neural representation of space using sinusoidal arrays, *Neural Computation*, 5:869–884.
- Zhang, K., 1996, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory, *J. Neurosci.*, 16:2112–2126. ◆

Dynamics and Bifurcation in Neural Nets

Bard Ermentrout

Introduction

A *recurrent neural net*, whether it is continuous or discrete in space and time, defines a dynamical system. Thus, it is possible to apply the powerful qualitative and geometric tools of dynamical systems theory to understand the behavior of neural networks. These techniques are most useful when the behavior of interest is stationary in the sense that the inputs are at most time- or space-periodic. Thus, we can ask what kind of behavior we can expect over the long run for a given neural network. Such information is important both in artificial neural networks and biological neural nets. In the former, the final state of the neural network may represent the recognition of an input pattern, the segmentation of an image, or any number of machine computations. The stationary states of biological neural networks may correspond to cognitive decisions (e.g., binding via synchronous oscillations) or to pathological behavior such as seizures and hallucinations.

Another important issue that is addressed by dynamical systems theory is how the qualitative dynamics depends on parameters. The qualitative change in a dynamical system as a parameter is changed is the subject of *bifurcation theory*. The word *bifurcation* is derived from the Greek word for branching; we are concerned with the appearance and disappearance of branches of solutions to a given set of equations as some parameters vary. There are now a large number of very good general books on the mathematical theory behind dynamical systems and bifurcation. In this article we show how to use these techniques to understand the behavior of neural nets. A fundamental problem for both artificial and biologically motivated neural nets is to understand how the solutions depend on the parameters and the initial states of the network.

For excellent introductions to nonlinear dynamics and bifurcation theory, see Ermentrout (1998), Guckenheimer and Holmes (1983), Kuznetsov (1998), and Wiggins (1990).

Some Basic Definitions

A *dynamical system* consists of a phase space, X , a time domain, T , and a function that describes the evolution of the phase space, $\phi(x, t)$. The function ϕ gives the value of an element in the phase space at time t , given that at $t = 0$ it was x . The two main motivating examples are differential equations and maps. In the former case, the time domain is the real line (continuous time); in the latter the time domain is the integers (discrete time). Consider the ordinary differential equation:

$$\frac{dx}{dt} = F(x) \quad x \in X, t \in R \quad (1)$$

Then we define $\phi(x_0, t)$ to be the solution to Equation 1 with initial condition x_0 . For example, if

$$\frac{dx}{dt} = x \quad x(0) = x_0$$

then $x(t) = \phi(x_0, t) \equiv x_0 e^t$. (We have restricted our attention to *autonomous* systems in which there is no explicit time dependence.) Consider next the iteration:

$$x(n+1) = F(x(n)) \quad x \in X, n \in Z \quad (2)$$

Then $\phi(x_0, n)$ is defined as the solution to Equation 2 with initial conditions $x = x_0$. For example, if

$$x(n+1) = 2x(n) \quad x(0) = x_0$$

then $x(n) = \phi(x_0, n) = x_0 2^n$.

There is nothing that prevents us from considering infinite dimensional dynamical systems such as partial differential equations (see PATTERN FORMATION, BIOLOGICAL) or neural networks distributed in space. The set of states $\Gamma(x_0) = \{\phi(x_0, t) : t \in T, \phi(x_0, t) \text{ defined}\}$ is called the *orbit* or *trajectory* through x_0 . The orbit is a curve in state space for continuous systems and a sequence of points for discrete systems. If $\Gamma(x_0)$ consists of a single point in phase space, then we say that x_0 is a *fixed point* or *equilibrium* for the system. Fixed points are easily found by solving $F(x_0) = 0$ or $F(x_0) = x_0$ for continuous and discrete dynamical systems, respectively. If $\phi(x_0, t+P) = \phi(x_0, t)$ for some non-zero value P , then the orbit is called *periodic*, with period P . A set S is *invariant* with respect to the dynamical system if $y \in S$ implies that $\phi(y, t) \in S$ for all $t \in T$. Thus, any orbit is an invariant set, as is any fixed point or periodic solution. The partitioning of the state space into orbits is called the *phase portrait* of the dynamical system and is one of the goals of dynamical systems.

Another key question in dynamical systems is the issue of stability. We say that an invariant set S_0 is *stable* if for any y close to S_0 , $\phi(y, t)$ stays close to S_0 for all $t \in T$. An invariant set is *asymptotically stable* if for any y near S_0 , the distance between $\phi(y, t)$ and S_0 tends to zero as $t \rightarrow \infty$. The stability of a fixed point in a discrete or continuous dynamical system is easily determined by studying the eigenvalues of an associated linear operator or matrix. Suppose that x_0 is a fixed point. Let $A = DF(x_0)$ be the matrix obtained by taking the partial derivatives of F with respect to the state variables and evaluating it at the fixed point. The dynamical system obtained by replacing $F(x)$ with Ax is called the *linearized system*. For Equation 1, if all of the eigenvalues of A have strictly negative real parts, then the fixed point is asymptotically stable (and all solutions to the linearized system decay to 0). If any eigenvalue has a positive real part, then the fixed point is unstable. For Equation 2, if all of the eigenvalues of A lie inside the unit circle, then the fixed point is asymptotically stable. If any eigenvalue lies outside the unit circle, then the fixed point is unstable. As long as none of the eigenvalues have zero real part (respectively, lie on the unit circle), we say the fixed point of the differential Equation 1 (respectively map, Equation 2) is *hyperbolic*. Eigenvalues that have negative real parts (lie in the unit circle) are called *stable eigenvalues*, those with positive real parts (lie outside the unit circle) are called *unstable eigenvalues*, and those that have zero real parts (lie on the unit circle) are called *neutral eigenvalues* for the continuous-time (discrete-time) fixed point. The invariant set $W^s(x_0) = \{y \in X : \phi(y, t) \rightarrow x_0 \text{ as } t \rightarrow \infty\}$ (respectively $W^u(x_0) = \{y \in X : \phi(y, t) \rightarrow x_0 \text{ as } t \rightarrow -\infty\}$) is called the *stable manifold* (respectively *unstable manifold*) of the fixed point x_0 . The stable (unstable) manifold is just the set of all points that tend to the fixed point as time increases (decreases) to infinity (negative infinity). The dimension of the stable (respectively, unstable) manifold is the number of stable (respectively unstable) eigenvalues of the linearized system. A fixed point that has both unstable and stable eigenvalues is called a *saddle point*. If the fixed point is asymptotically stable, then the stable manifold has the same dimension as the phase space and is then often called the *basin of attraction* for the fixed point. For neutral eigenvalues there is a *center manifold* that is invariant and has the same dimension as the number of neutral eigenvalues. The center manifold is extremely useful and important, as we shall see, since it allows one to study the behavior of high-dimensional systems in a lower-dimensional setting.

So far, the discussion of asymptotic behavior has been restricted to fixed points. A continuous dynamical system can often be re-

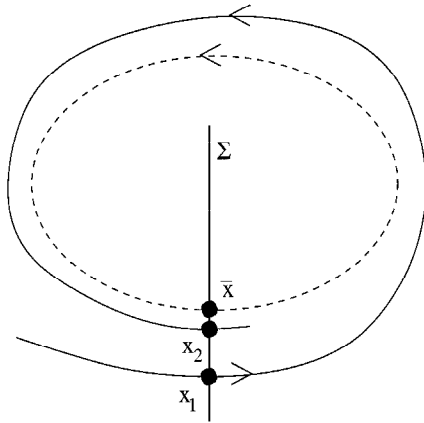


Figure 1. Construction of the Poincaré map for a two-dimensional system.

duced to a discrete one by introducing a *Poincaré map*. Suppose the system has a periodic orbit. A *cross-section* for an n -dimensional continuous dynamical system is an $n - 1$ -dimensional hypersurface that is orthogonal to the tangent of the periodic orbit (Figure 1). A point on the surface that starts near the periodic orbit will be brought back to the surface at a later time. This produces a locally defined map from the surface back to itself, which is then a discrete dynamical system. From Figure 1, it is clear that a fixed point of the Poincaré map corresponds to a periodic solution of the original system. Thus, we can determine stability, stable, unstable, and center manifolds for periodic solutions just by studying the behavior of a Poincaré map defined in some local neighborhood of it. An isolated periodic solution is called a *limit cycle* and its stability is determined by studying the stability of the fixed point for the associated Poincaré map. A limit cycle solution is hyperbolic if the fixed point of the Poincaré map is hyperbolic.

Invariant sets are not all as simple as periodic solutions and fixed points. In fact, they can be quite complicated. A stable invariant set that has irregular behavior (e.g., it is not a simple curve or point) is called a *strange attractor*. Similarly, the behavior of an invariant set can be quite complex as well. We say that an invariant set is *chaotic* if it displays sensitive dependence on initial conditions; that is, the orbits through two arbitrarily close points on the set diverge from each other exponentially. (For examples of chaotic behavior in neurons, see CHAOS IN NEURAL SYSTEMS; SYNAPTIC NOISE AND CHAOS IN VERTEBRATE NEURONS; and CHAOS IN BIOLOGICAL SYSTEMS).

A dynamical system that has only hyperbolic fixed points and periodic orbits will maintain the same qualitative behavior if the parameters are varied slightly. Thus, qualitative changes are seen when the fixed points and periodic orbits become nonhyperbolic. This happens when eigenvalues cross the critical axis and thus a fixed point loses stability. This is one of the key ideas behind *bifurcation theory*. Roughly, we expect to see qualitative changes as a parameter varies when some of the fixed points or periodic orbits become nonhyperbolic. Bifurcation theory gives us a method of studying arbitrary dynamical systems near these critical values of parameters. The idea is that near a critical parameter value, there will be some neutral eigenvalues. This will imply that there is a nontrivial center manifold and in fact the local dynamics of the full system can be completely understood by studying the dynamics restricted to this center manifold. Bifurcation methods give a recipe for computing the form of the equations on this low-dimensional system. This means that one can study a possibly infinite-dimensional system by looking at the dynamics of a possibly one-dimensional system!

The simplified dynamical systems that one obtains near critical values of the parameters are called *normal forms*. Thus, the understanding of the local behavior of the full system comes from studying the behavior of the relevant normal form.

Local Bifurcations

Local bifurcation theory allows one to study the behavior of discrete and continuous dynamical systems near fixed points. Since the behavior of periodic solutions to continuous systems reduces to the analysis of fixed points of the Poincaré map, local bifurcation of maps enables us to analyze bifurcations of limit cycles in continuous systems.

Continuous-Time Systems

There are two ways in which a fixed point of a continuous-time dynamical system can become nonhyperbolic as a parameter varies: (1) an eigenvalue crosses zero or (2) a pair of complex eigenvalues crosses the imaginary axis. In the case of a zero eigenvalue, this signifies the appearance of new fixed points near the original one. In the most general setting, with no symmetries, a zero eigenvalue implies a *fold* or *turning point* bifurcation. In this bifurcation, the deviation from the fixed point obeys one-dimensional dynamics:

$$r' = ar^2 + c(\mu - \mu^*) \quad (3)$$

where a, c are problem-dependent parameters, $r \in \mathbb{R}$, and μ is the parameter that is varied. The fixed points of this simple system (called the normal form) correspond to fixed points of the full system even if it is infinite dimensional. In Figure 2A, steady-state solutions are shown for $a, c > 0$. As μ increases past μ^* , two fixed points, a stable one and an unstable one, coalesce and disappear, leaving no nearby fixed point. (The picture is essentially the same

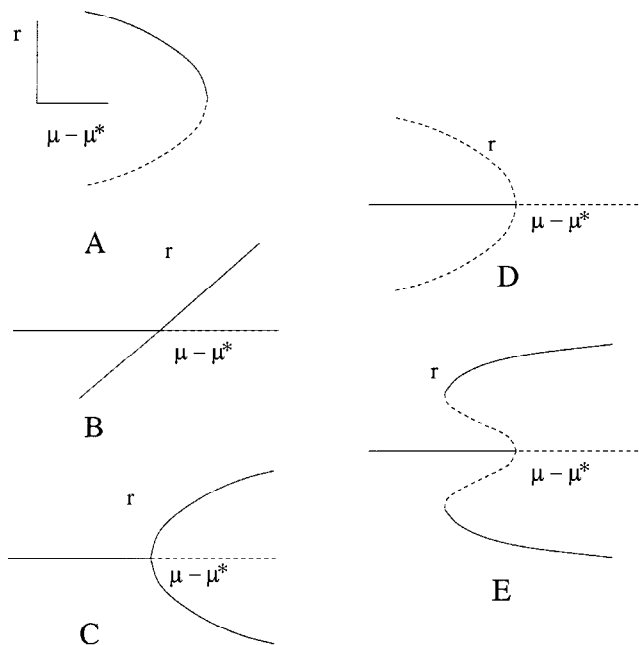


Figure 2. Bifurcation diagrams for fixed points of continuous-time dynamical systems. Solid lines are stable, dashed unstable. A, Steady-state solutions for a fold or turning point bifurcation. B, A transcritical or exchange-of-stability bifurcation. C, A supercritical bifurcation. D, A subcritical bifurcation. E, Subcritical pitchfork bifurcations turning around at a pair of fold points and restabilizing, leading to bistability.

for other choices of c , a .) In cases with additional symmetry (such as the requirement that there always be at least one fixed point or that the system have some symmetry), there are two additional common normal forms:

$$r' = ar^2 + cr(\mu - \mu^*) \quad (4)$$

$$r' = ar^3 + cr(\mu - \mu^*) \quad (5)$$

Equation 4 leads to the *transcritical* bifurcation shown in Figure 2B. This is also called an *exchange-of-stability* bifurcation since the stability of the two fixed points switches at the point of bifurcation. As in the fold, the signs of a , c are irrelevant to the picture. With additional symmetry, Equation 5 occurs, and this is called the *pitchfork* bifurcation. In this case, the sign of a is important. If $a < 0$, then the bifurcation is *supercritical*, and two new *stable* fixed points arise (Figure 2C). If $a > 0$, then two new *unstable* fixed points occur, and the bifurcation is called *subcritical* (Figure 2D). In many biological and physical systems, subcritical pitchfork bifurcations “turn around” at a pair of fold points and restabilize (as in Figure 2E). This leads to *bistability* between several fixed points and to hysteresis.

Suppose that stability is lost when a pair of complex eigenvalues crosses the imaginary axis. Then the system can undergo what is called a *Hopf* or *Andronov* bifurcation. The dynamics are locally determined by the simple complex differential equation:

$$z' = az^2\bar{z} + cz(\mu - \mu^*) + iwz \quad a, c, z \in \mathbb{C} \quad (6)$$

If the real part of a is negative, then a stable periodic orbit will bifurcate from the fixed point with amplitude that is proportional to $\sqrt{\mu - \mu^*}$. The bifurcation is called *supercritical*. If the real part of a is positive, then an *unstable* periodic orbit bifurcates from the fixed point and the bifurcation is called *subcritical*. The bifurcation diagrams look like those of the pitchfork bifurcation (Figures 2C and 2D), where $r = |z|$. As with the pitchfork bifurcation, physical systems that have subcritical Hopf bifurcations often restabilize, as in Figure 2E. Thus, there will be a range of the parameter for which there is a stable limit cycle and a stable fixed point (analogous to Figure 2E.)

Discrete Dynamics

For a discrete dynamical system to undergo a local bifurcation, eigenvalues must cross the unit circle. There are three ways in which this can happen: (1) an eigenvalue is $+1$, (2) an eigenvalue is -1 , or (3) a complex pair of eigenvalues lies on the unit circle. The first case, an eigenvalue of 1 , is completely analogous to the case of a zero eigenvalue for continuous systems. The third case is similar to the Hopf bifurcation for continuous systems and is called a *Neimark-Sacker* bifurcation. However, the dynamics of the bifurcating solution can be complicated, and all one can conclude is that there is a small invariant circle. The second case, a -1 eigenvalue, is called a *flip* or *period-doubling* bifurcation. The dynamics is determined by the behavior of the simple one-dimensional map:

$$r_{n+1} = -r_n + c(\mu - \mu^*)r_n + ar_n^3 \quad (7)$$

If $a > 0$, then a stable period 2 fixed point appears; that is, every other iterate of the map is the same. If $a < 0$, then an unstable period 2 fixed point occurs. Period-doubling bifurcations are very important, as they often signal the onset of chaotic behavior. Indeed, often a period 2 point itself will become unstable through another period-doubling bifurcation to a period 4 point. This continues as the parameter is changed, and a whole *cascade* of period doublings occurs, terminating in chaotic behavior.

Periodic Orbits in Continuous Systems

Periodic orbits undergo bifurcations similar to those of discrete dynamical systems since their local behavior is reducible to a discrete system. Folds of the Poincaré map correspond to the annihilation of a stable and unstable limit cycle; flips correspond to period doubling of the limit cycle; Neimark-Sacker bifurcations correspond to the appearance of an invariant 2-torus.

In addition to these local bifurcations, there are two common *global* bifurcations in which the period of the orbit tends to infinity. A *heteroclinic* orbit, $\gamma(t)$, is a nonconstant orbit satisfying

$$\lim_{t \rightarrow \pm\infty} \gamma(t) = x^\pm$$

where x^\pm are fixed points. If $x^+ = x^-$, we call $\gamma(t)$ a *homoclinic* orbit. In general, one cannot expect to find a homoclinic orbit; rather, as some parameter μ changes, the homoclinic orbit appears at one value of that parameter, μ^* (Figure 3A). Thus, the appearance of a homoclinic is a bifurcation. If the fixed point $x^+ = x^-$ is a hyperbolic saddle, then, as the parameter moves away from criticality, a periodic orbit arises, and the period of this orbit goes to infinity as the homoclinic orbit is approached. The period is given by

$$T_{hom} \sim K \log \frac{1}{|\mu - \mu^*|}$$

If instead of a hyperbolic saddle point, the fixed point is a fold point, then a *saddle node on a limit cycle* occurs (Figure 3B). Limit cycles occur as the parameter is moved from criticality with a period

$$T_{sn} \sim \frac{K}{\sqrt{|\mu - \mu^*|}}$$

Applications to Continuous-Time Neural Nets

So far, we have introduced definitions but not used them in the context of neural networks. Numerous papers have used bifurcation

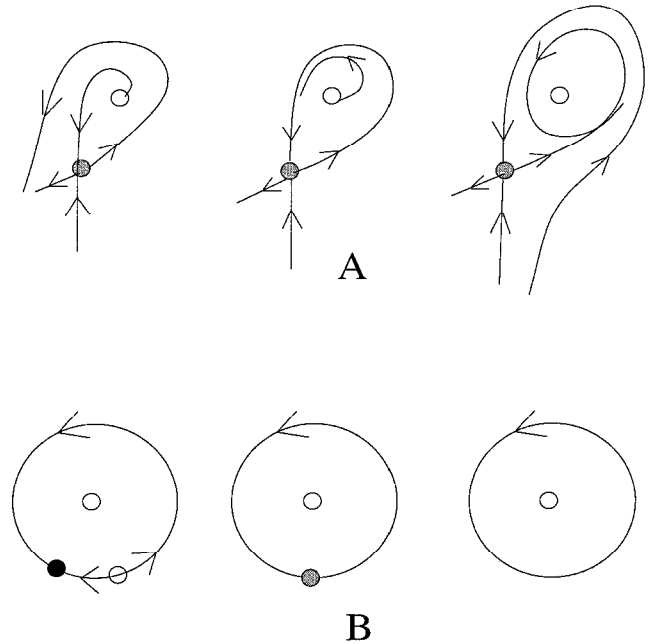


Figure 3. Two distinct types of homoclinic bifurcation. Bifurcation of (A) a homoclinic orbit from a hyperbolic saddle or (B) a saddle node on a limit cycle, leading to periodic orbits.

methods to analyze neural networks (see, e.g., PHASE-PLANE ANALYSIS OF NEURAL NETS; COOPERATIVE PHENOMENA; OSCILLATORY AND BURSTING PROPERTIES OF NEURONS; and PATTERN FORMATION, NEURAL; see also Izhikevich, 2000, for a long and well-illustrated review). The types of neural net models to which the theory has been applied generally have one of the following two forms:

$$\tau_j \frac{du_j}{dt} = -u_j + F_j \left(\sum_k W_{jk} u_k + I_j \right) \quad (8)$$

$$u_j(n+1) = \mu_j u_j + F_j \left(\sum_k W_{jk} u_k(n) + I_j \right) \quad (9)$$

The functions F_j are generally monotone and increasing, W_{jk} are the weights, I_j are the inputs, and $1/\tau_j$, $1 - \mu_j$ are decay rates. With proper limits, the sums in Equations 8 and 9 can go to integrals over space in order to represent a continuum model for neural nets. In this case the interaction takes the form

$$\int_{\Omega} W(x, y) u(y) dy$$

where Ω is the spatial domain, a one- or two-dimensional region in space. The behavior of these models is generally impossible to analyze completely except in certain simple cases. For example, Hopfield (1984) shows that if the weights W_{jk} are symmetric, then all solutions to Equation 8 tend to be fixed points. In the case of one or two dimensions, Equation 8 can be completely analyzed (see Hoppensteadt and Izhikevich, 1997, chap. 2; PHASE-PLANE ANALYSIS OF NEURAL NETS). The discrete system in Equation 9 can be completely characterized only in one dimension.

Local Bifurcations of Fixed Points

To see how bifurcation methods can be used in neural nets, consider a simple continuous neural network with no inputs and satisfying $\tau_j = 1$, $F_j(u) = F(u)$ with $F(0) = 0$, $F'(0) = \mu$, a gain parameter. Thus the trivial state of the network, $u_j = 0$, is a fixed point. To study stability and bifurcation, we linearize about $u_j = 0$ and obtain the matrix

$$M_{jk} = -\delta_{jk} + \mu W_{jk}$$

where δ_{jk} is the Kronecker delta function. If λ is an eigenvalue of W_{jk} , then $-1 + \mu\lambda$ is an eigenvalue for M ; thus, if the gain is sufficiently small, the trivial fixed point is stable. If some of the eigenvalues of the weight matrix W have positive real parts, then, for sufficient gain, some of the eigenvalues of M will cross the imaginary axis and the rest state will lose stability. In particular, let λ_0 be the eigenvalue with maximal real part and let Φ_0 be the corresponding eigenvector. Since this is a continuous dynamical system, there are only two cases of interest. Suppose that λ_0 is real. Then if $\mu = \mu^* = 1/\lambda_0$, there will be a zero eigenvalue and a bifurcation to new fixed points. Since 0 is always a fixed point, the bifurcation will be transcritical, or a pitchfork. In any case, the new solution will have a non-zero amplitude and be proportional to Φ_0 . This is the essence of pattern formation. The trivial state $u_j = 0$ loses stability and a new state that is coded in the weight matrix bifurcates from rest. Analysis of pattern formation in neural networks (and for that matter, any pattern formation models) all comes down to this calculation (see Murray, 1989). If the weight matrix is symmetric, then the eigenvalues are real and the initial bifurcation from the rest state will always be through a zero eigenvalue. If the weight matrix is non-negative, then the eigenvalue with largest modulus will be real and positive and the first bifurcating mode will be proportional to the principal component of the weight matrix.

Suppose the weight matrix is not symmetric and not all the same sign; that is, there are “inhibitory” and “excitatory” interactions. Then the eigenvalue with maximal real part could be a complex eigenvalue. This means that the first instability is through a Hopf bifurcation and an oscillatory pattern of activities can bifurcate. In situations in which there is symmetry or in which the network has a particular geometry, this type of bifurcation can lead to periodic wave trains. Recently, Hoppensteadt and Izhikevich (1997) proposed a general theory of neural networks that exploits the kind of local analysis that we have only sketched here.

Beyond Local Bifurcations

To go beyond the simple bifurcation analysis described above, it is necessary to turn to numerical methods. There are several numerical packages for the analysis of bifurcations in nonlinear dynamics. AUTO (Doedel et al., 1997) is among the best of them and works on many operating systems. To illustrate the concepts discussed in the previous sections, I present a global numerical diagram for a six-neuron model whose weight matrix was chosen randomly from a uniform distribution with mean 0 and a standard deviation of 0.5. (Note that this particular example was picked because of its rich behavior.) The function $F(x) = \tanh(\mu x)$ was chosen for simplicity. Once the weight matrix is chosen, there is only one parameter, the gain, μ . Figure 4 shows the norm of the solutions that are computed numerically as the gain is increased. The eigenvalues for the weight matrix are approximately -0.83 , $0.25 \pm 0.2i$, -0.16 , 0.12 , -0.47 . Thus, for positive gains, there will be a Hopf bifurcation at approximately $\mu = 4$ and a pitchfork bifurcation at approximately $\mu = 8.33$. Both of these branches are shown in the diagram (labeled H1 and P, respectively). The pitchfork is subcritical and undergoes a fold bifurcation (labeled F1) and stabilizes. A subcritical Hopf bifurcation (H3) occurs on this branch,

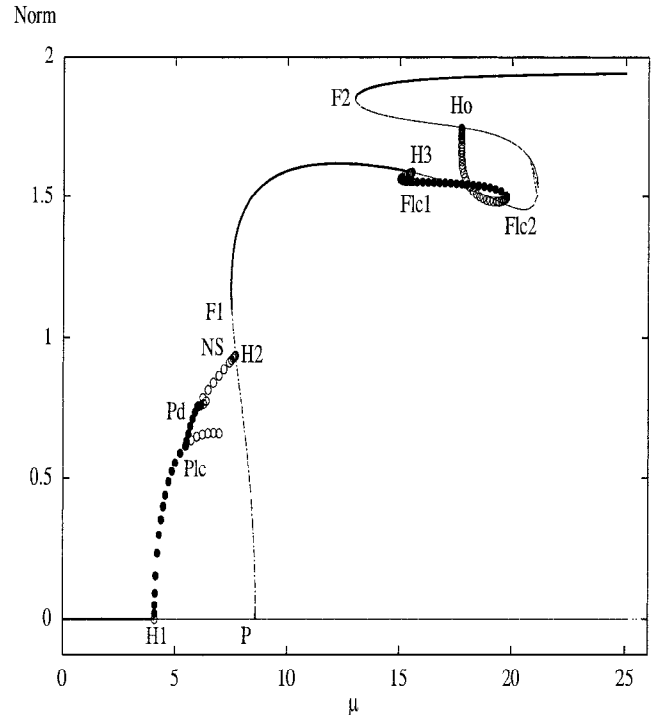


Figure 4. Numerically computed bifurcation diagram for a six-neuron network with random weights chosen between -0.5 and 0.5 .

which turns around at the limit-cycle fold bifurcation (Flc1), giving rise to a stable periodic orbit. This orbit turns around again at Flc2, and the resulting unstable limit cycle terminates at a homoclinic bifurcation (Ho). The curve of fixed points turns around at F2, leaving a stable fixed point that persists for all higher values of μ .

The fate of the periodic orbit that arises at H1 is more interesting. This orbit is supercritical and leads to small-amplitude, stable periodic solutions. As μ increases, the periodic orbit (and hence the fixed point to the Poincaré map) loses stability as an eigenvalue crosses $+1$. This results in a pitchfork bifurcation at Plc. The unstable periodic orbit appears to persist for all values of μ beyond the bifurcation point, but never restabilizes. The pitchfork bifurcation of the Poincaré map is supercritical and results in a stable periodic orbit. This orbit becomes unstable at a flip or period-doubling bifurcation (Pd); the branch eventually terminates at a Hopf bifurcation (H2) on the branch of fixed points that bifurcated at P. The periodic branch also undergoes a Neimark-Sacker bifurcation (NS), resulting in an unstable torus. The flip bifurcation is supercritical and leads to a new branch of periodic solutions with twice the period. This branch in turn undergoes a flip bifurcation, and so on, producing a period-doubling cascade. The regimen between $\mu = 5.5$ and $\mu = 7.5$ is very complicated. There are many bifurcations, and the appearance of many stable and unstable periodic orbits as well as chaotic behavior, that are not described on this plot. In order to depict this behavior, we plot the Poincaré map whose cross-section is defined as the hyperplane where $u'_4 = 0$. The u_2 component of this map is shown in Figure 5 as a function of μ . In this diagram, each dot represents an iteration of the Poincaré map, so that for a fixed value of μ , a single dot implies a periodic solution, a pair is a period 2 orbit, and so on. The period-doubling cascades are clear. A period 3 orbit at about $\mu = 6.85$ is seen to undergo a period-doubling cascade as μ is decreased. Chaotic behavior terminates near a saddle point at $\mu \approx 7.57$ and dis-

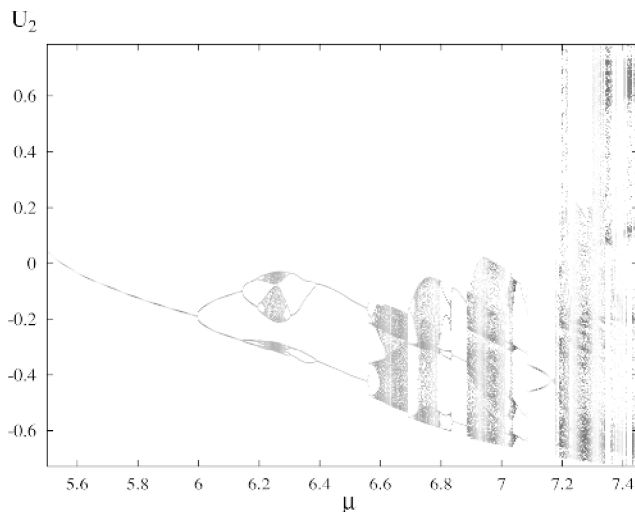


Figure 5. Poincaré map of a region from Figure 3 showing complex dynamics.

appears “instantly.” This global bifurcation is called a *crisis* by Yorke.

Conclusion

The general behavior of recurrent neural networks as parameters vary remains an open problem. Dynamical systems methods and bifurcation theory provide a general approach to analyzing these interesting systems. Pattern formation, spatiotemporal behavior, and complex dynamics are all aspects of recurrent neural networks that can be understood by these useful mathematical tools. Analytical methods along with the careful use of numerical methods allow one to globally characterize complex biological and artificial neural networks.

Appendix

The neural network used in Figures 3 and 4 has the form:

$$U'_j = -U_j + \tanh \left(\mu \sum_{k=1}^6 W_{jk} U_k \right)$$

where

$$W = \begin{pmatrix} -0.42473 & 0.243325 & -0.267939 & 0.308063 & -0.0370201 & 0.394969 \\ -0.166832 & 0.474204 & -0.0151443 & 0.476774 & 0.211162 & 0.401305 \\ -0.427914 & 0.370044 & -0.0675567 & 0.276535 & 0.45988 & -0.457168 \\ -0.362131 & 0.033334 & -0.196538 & -0.037606 & -0.125548 & 0.143851 \\ 0.429334 & -0.306886 & 0.402954 & -0.166799 & -0.45518 & -0.0304156 \\ 0.294663 & -0.346348 & -0.138444 & 0.334973 & 0.13884 & -0.364227 \end{pmatrix}$$

Road Map: Dynamic Systems

Related Reading: Chaos in Neural Systems; Computing with Attractors; Oscillatory and Bursting Properties of Neurons; Pattern Formation, Biological; Phase-Plane Analysis of Neural Nets

References

- Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., and Wang, X. J., 1997, AUTO97: Continuation and bifurcation software for ordinary differential equations (with HomCont), Montreal: Computer Science Department, Concordia University available: <ftp://ftp.cs.concordia.ca/pub/doedel/auto>.
- Ermentrout, B., 1998, Neural networks as spatio-temporal pattern-forming systems, *Rep. Prog. Phys.*, 61:353–430. ♦
- Guckenheimer, J., and Holmes P. J., 1983, *Nonlinear oscillations, Dynamical Systems, and Bifurcations of Vector Fields.*, Heidelberg: Springer-Verlag.
- Hopfield, J., 1984, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA*, 81:3088–3092.
- Hoppensteadt, F., and Izhikevich, E., 1997, *Weakly Connected Neural Networks*, New York: Springer-Verlag. ♦
- Izhikevich, E. M., 2000, Neural excitability, spiking, and bursting, *Int. J. Bifurcat. Chaos*, 10:1171–1266.
- Kuznetsov, Y. A., 1998, *Elements of Applied Bifurcation Theory*, New York: Springer-Verlag. ♦
- Murray, J. D., 1989, *Mathematical Biology*, Heidelberg: Springer-Verlag.
- Wiggins, S., 1990, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, New York: Springer-Verlag.

Dynamics of Association and Recall

Ton Coolen

Introduction

The concept and relevance of associative memory in neural networks are discussed elsewhere in this *Handbook* (see, e.g., ASSOCIATIVE NETWORKS, STATISTICAL MECHANICS OF NEURAL NETWORKS, and CORTICAL HEBBIAN MODULES). Associative memory networks are usually (but not always) recurrent, which implies that one cannot simply write down the values of successive neuron states (as with layered networks). The latter must be solved from coupled dynamic equations. Dynamical studies shed light on the pattern recall process and its relation to the choice of initial state, the properties of the stored patterns, the noise level, and the network architecture. In addition, for nonsymmetric networks (where the equilibrium statistics are not known), dynamical techniques are the *only* tools available. Since our interest is usually in large networks and in global recall processes, the common strategy of the theorist is to move away from the *microscopic* equations (i.e., equations at the level of individual neurons) and to derive dynamical laws at a *macroscopic* level (i.e., in terms of quantities that depend on many neuron states). One then faces the following questions:

- Which are the appropriate macroscopic quantities we should attempt to calculate?
- How do we extract the corresponding macroscopic equations from the microscopic ones?

All research in this field basically involves launching or elaborating proposals for dealing with these two issues, which are related in that the “natural” set of macroscopic quantities can be defined as the *smallest* set that will obey *closed deterministic* equations in the limit of an infinitely large network. Since the mid-1970s, many theoretical approaches to recall dynamics have been developed in parallel, while their relations were only understood later. To give an overview of the field I will therefore not always discuss papers in chronological order, but use the benefit of hindsight. In the absence of recent textbooks or reviews that deal with recall dynamics in a satisfactory manner (at least from the viewpoint of what we know today), most of the references will indeed be to research papers. In the interest of transparency, in this article I restrict myself mainly to simple Hopfield-type models, and discuss some variations and generalizations in a final section.

The Core of the Problem

In Hopfield-type models one has N binary neurons, with states $S_i \in \{-1, 1\}$. I will write $\mathbf{S} = (S_1, \dots, S_N)$. The S_i evolve stochastically, driven by postsynaptic potentials (or “local fields”) h_i :

$$S_i(t + \Delta) = \text{sgn}[h_i(\mathbf{S}(t)) + \eta_i(t)] \quad h_i(\mathbf{S}) = \sum_{j \neq i} J_{ij} S_j \quad (1)$$

The $\eta_i(t)$ are independent random variables, modeling threshold noise. The updates in Equation 1 can be executed in parallel or sequentially. Comparable time units result when choosing $\Delta = 1$ for parallel dynamics and $\Delta = 1/N$ for sequential dynamics. Hopfield networks are equipped with Hebbian-type synapses:

$$J_{ij} = \frac{c_{ij}}{c} \sum_{\mu=1}^p \xi_i^\mu \xi_j^\mu \quad (2)$$

These involve p patterns $\xi^\mu = (\xi_1^\mu, \dots, \xi_N^\mu) \in \{-1, 1\}^N$. The $c_{ij} \in \{0, 1\}$ specify the network connectivity ($c_{ij} = 1$ if a connection $j \rightarrow i$ is present, $c_{ij} = 0$ if not), whereas the factor $c = \langle 1/$

$N \sum_{ij} c_{ij} \rangle$ (the average number of neurons contributing to a local field) ensures that the local fields remain finite as $N \rightarrow \infty$. One usually defines also the relative load factor $\alpha = p/c$ (the number of patterns stored per synapse). A fully connected network thus corresponds to $c_{ij} = 1$ for all (i, j) and $c = N$. A so-called randomly extremely diluted network, on the other hand, would have $\lim_{N \rightarrow \infty} c^{-1} = 0$ and $\lim_{N \rightarrow \infty} c/N = 0$ (i.e., the number of synapses attached to a typical neuron diverges but remains small when compared to the total number of neurons). Theorists often choose to study these two connectivity types because they are found to simplify the analysis of the dynamics (e.g., one need not worry about spatial aspects).

The process of interest is that where, triggered by correlation between the initial state $\mathbf{S}(0)$ and a stored pattern ξ^1 , the state vector \mathbf{S} evolves toward ξ^1 . If this happens, pattern ξ^1 is said to be recalled. Numerical simulations of a simple fully connected network (with randomly drawn patterns) already clearly illustrate the main features and complications of recall dynamics. Here the correlation between a state vector and the stored patterns is measured by so-called overlaps (or “directional cosines”):

$$m_\mu(\mathbf{S}) = \frac{1}{N} \sum_i \xi_i^\mu S_i \quad (3)$$

For large N we distinguish structural overlaps, where $m_\mu(\mathbf{S}) = \mathcal{O}(1)$, from accidental ones, where $m_\mu(\mathbf{S}) = \mathcal{O}(N^{-1/2})$ (as for a randomly drawn \mathbf{S}). Figure 1 shows the result of measuring the quantities

$$m = m_1(\mathbf{S}) \quad r = \alpha^{-1} \sum_{\mu > 1} m_\mu^2(\mathbf{S}) \quad (4)$$

following initial states that are correlated with pattern ξ^1 only. Overlaps with non-nominated patterns are seen to remain $\mathcal{O}(N^{-1/2})$, i.e., $r(t) = \mathcal{O}(1)$. We immediately observe competition between pattern recall ($m \rightarrow 1$) and interference of non-nominated patterns ($m \rightarrow 0$, with r increasing). The initial overlap (the “cue”) needed to trigger recall is found to increase with increasing α (the loading) and increasing T (the noise), until beyond a critical value $\alpha_c(T)$ recall is no longer possible and all trajectories will ultimately lead to the $m = 0$ state (irrespective of the initial state). The competing forces are easily recognized when working out the local fields described by Equation 1, using Equation 2 with $c = N$:

$$h_i(\mathbf{S}) = \xi_i^1 m_1(\mathbf{S}) + \frac{1}{N} \sum_{\mu > 1} \xi_i^\mu \sum_{j \neq i} \xi_j^\mu S_j + \mathcal{O}(N^{-1}) \quad (5)$$

The first term in Equation 5 drives \mathbf{S} toward pattern ξ^1 as soon as $m_1(\mathbf{S}) > 0$. The other terms represent interference caused by correlations between the state vector \mathbf{S} and non-nominated patterns. One easily shows that for $N \rightarrow \infty$, the fluctuations in the values of the recall overlap m must vanish at any time, and that for the present types of initial states and threshold noise, the overlap m will obey

$$m(t + 1) = \int dz P_t(z) \tanh[\beta(m(t) + z)] \quad (6)$$

$$P_t(z) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \left\langle \delta \left[z - \frac{1}{N} \sum_{\mu > 1} \xi_i^1 \xi_i^\mu \sum_{j \neq i} \xi_j^\mu S_j(t) \right] \right\rangle \quad (7)$$

(with $\beta = T^{-1}$). For general (symmetric) threshold noise distributions, one just substitutes $\tanh[\beta x] \rightarrow 2 \int_0^1 d\eta w(\eta)$. Note that Equation 6 can be interpreted as describing the dynamics of the average firing state $m(t) = \langle S(t) \rangle$ of a single “effective neuron” with an effective local field $h(t) = m(t) + z(t)$, where $z(t)$ (distrib-

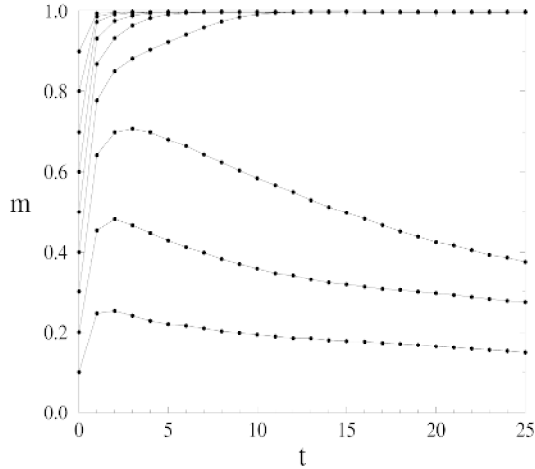


Figure 1. Simulations of a fully connected Hopfield model with $N = 30,000$, threshold noise distribution $w(\eta) = (1/2T)[1 - \tanh^2(\eta/T)]$, $\alpha = T = 0.1$, parallel dynamics, and randomly drawn pattern bits. *Left*, Overlap $m = m_1(S)$ with pattern 1 as functions of time, following initial states $S(0)$

uted according to $P_t(z)$) acts as an extra contribution to the threshold noise. If all $S_i(0)$ are drawn independently, $\text{Prob}[S_i(0) = \pm \xi_i^1] = (1/2)[1 \pm m(0)]$, the central limit theorem tells us that $P_0(z)$ is Gaussian. One easily derives $\langle z \rangle_0 = 0$ and $\langle z^2 \rangle_0 = \alpha$, so that at $t = 0$, Equation 6 reduces to

$$m(1) = \int \frac{dz}{\sqrt{2\pi}} e^{-(1/2)z^2} \tanh[\beta(m(0) + z\sqrt{\alpha})] \quad (8)$$

The above ideas, and Equation 8 in particular, go back to Amari (1977). For times $t > 0$, however, the independence of the states S_i need no longer hold. Solving the recall dynamics thus boils down to calculating the (often nontrivial) interference noise distribution $P_t(z)$ in Equation 6 for non-zero times.

Small Numbers of Patterns

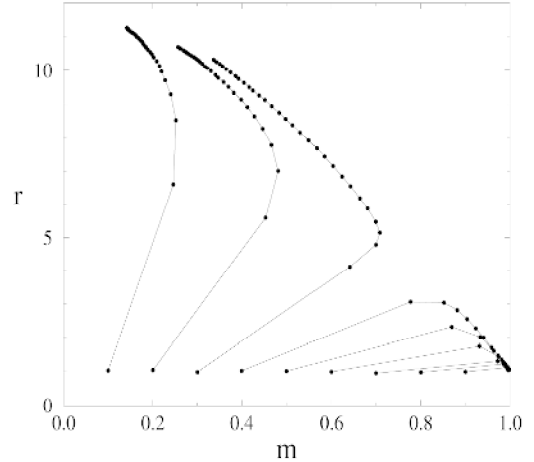
The simplest way to avoid trouble is to study situations where the interference noise vanishes in the $N \rightarrow \infty$ limit, which happens, for example, in fully connected networks when $\alpha = \lim_{N \rightarrow \infty} p/N = 0$ (as with finite p). Equation 6 and its sequential dynamics counterpart now reduce to (with the integrated threshold noise distribution $f[x] = 2f_0^+ d\eta w(\eta)$)

$$\begin{aligned} \text{parallel: } m(t+1) &= f[m(t)] \\ \text{sequential: } \frac{d}{dt} m &= f[m] - m \end{aligned}$$

This situation was exploited by several authors (e.g., Buhmann and Schulten, 1987; Coolen and Ruijgrok, 1988; Riedel, Kühn, and van Hemmen, 1988), who, in view of the simplicity of the finite p regime, were able to consider more complicated choices than Equation 2 for the synapses and arbitrary initial states $S(0)$. If, for instance, we choose the more general form $J_{ij} = (1/N) \sum_{\mu=1}^p \xi_i^\mu A_{\mu\nu} \xi_j^\nu$ (with p finite) and if we put $\mathbf{m} = (m_1, \dots, m_p)$, we find the above equations for a single overlap generalizing to coupled equations involving all p overlaps:

$$\begin{aligned} \text{parallel: } \mathbf{m}(t+1) &= \langle \xi f[\xi \cdot \mathbf{A} \mathbf{m}(t)] \rangle_\xi \\ \text{sequential: } \frac{d}{dt} \mathbf{m} &= \langle \xi f[\xi \cdot \mathbf{A} \mathbf{m}] \rangle_\xi - \mathbf{m} \end{aligned}$$

with $\langle g[\xi] \rangle_\xi = 2^{-p} \sum_{\xi \in \{-1,1\}^p} g[\xi]$ and $\mathbf{A} = \{A_{\mu\nu}\}$. These equations, and more elaborate versions (with, e.g., neuronal transmis-



correlated with pattern 1 only, with $m_1(S(0)) \in \{0.1, \dots, 0.9\}$. *Right*, Corresponding trajectories in the (m, r) plane, where the observable $r = \alpha^{-1} \sum_{\mu>1} m_\mu^2(S)$ measures the overlaps with non-nominated patterns.

sion delays) have since been used extensively to study recall properties of models storing static patterns and of models storing pattern sequences (with nonsymmetric J_{ij}). It is even possible to generalize the allowed synapse types further to all matrices of the form $J_{ij} = (1/N)Q[\xi_i; \xi_j]$, with $\xi_i = (\xi_i^1, \dots, \xi_i^p)$, such as the “clipped Hebbian” synapses $J_{ij} = (1/N)\text{sgn}[\sum_\mu \xi_i^\mu \xi_j^\mu]$, by using a different (and larger) set of macroscopic quantities than the p pattern overlaps.

Gaussian Approximations

Let us now return to the nontrivial regime, where $\alpha > 0$ and the interference noise problem must be confronted. As a simple approximation one could just assume that the S_i remain uncorrelated at all times, i.e., $\text{Prob}[S_i(t) = \pm \xi_i^1] = (1/2)[1 \pm m(t)]$ for all $t \geq 0$, such that the argument given earlier for $t = 0$ (leading to a Gaussian interference noise distribution $P(z)$) would hold generally, and where the mapping described by Equation 7 would describe the overlap evolution at all times:

$$m(t+1) = \int \frac{dz}{\sqrt{2\pi}} e^{-(1/2)z^2} f[m(t) + z\sqrt{\alpha}] \quad (9)$$

(again with $f[x] = 2f_0^+ d\eta w(\eta)$). This equation, however, must be generally incorrect: Figure 1 already shows that, at least for fully connected networks, knowledge of $m(t)$ only does not yet permit prediction of $m(t+1)$. However, for extremely and asymmetrically diluted networks, which are constructed by drawing the synaptic wiring variables c_{ij} in Equation 2 independently at random from $p(c_{ij}) = (c/N)\delta_{c_{ij},1} + (1 - (c/N))\delta_{c_{ij},0}$, with $\lim_{N \rightarrow \infty} c/N = 0$ and $c \rightarrow \infty$, Equation 8 is indeed found to be correct on finite times (Derrida, Gardner, and Zippelius, 1987). In these networks the time it takes for correlations between neuron states to build up diverges with N , so that correlations are simply not yet noticeable on finite times. For the common choice $f[x] = \tanh[x/T]$ (i.e., $w(\eta) = (1/2T)[1 - \tanh^2(\eta/T)]$), Equation 9 predicts a critical noise level (at $\alpha = 0$) of $T_c = 1$, and a storage capacity (at $T = 0$) of $\alpha_c = 2/\pi \approx 0.637$.

Rather than taking all S_i to be independent, a weaker assumption would be only to assume the interference noise distribution $P_t(z)$ to be a zero-average Gaussian one at any time (with statistically independent noise variables z at different times): $P_t(z) =$

$[\sigma(t)\sqrt{2\pi}]^{-1}e^{-(1/2)z^2/\sigma^2(t)}$. One can then derive (for $N \rightarrow \infty$ and fully connected networks) an evolution equation for the width $\sigma(t)$, giving (Amari and Maginu, 1988; Nishimori and Ozeki, 1993):

$$m(t+1) = \int \frac{dz}{\sqrt{2\pi}} e^{-(1/2)z^2} f[m(t) + z\sigma(t)]$$

$$\sigma^2(t+1) = \alpha + 2\alpha m(t+1)m(t)h[m(t), \sigma(t)] + \sigma^2(t)h^2[m(t), \sigma(t)]$$

$$h[m, \sigma] = \int \frac{dz}{\sqrt{2\pi}} e^{-(1/2)z^2} f'[m + z\sigma]$$

These equations describe correctly the qualitative features of recall dynamics, and are found to work quite well when retrieval actually occurs. For nonretrieval trajectories, however, they appear to underestimate the impact of interference noise, and for $f[x] = \tanh[x/T]$, they predict $T_c = 1$ (at $\alpha = 0$) and a storage capacity (at $T = 0$) of $\alpha_c \approx 0.1597$ (whereas this should have been roughly 0.139). A final refinement of the Gaussian approach by Okada (1995) consisted in allowing for correlations between the noise variables z at different times (while still describing them by Gaussian distributions). This then results in a hierarchy of macroscopic equations that improve upon the previous Gaussian theories and even predict the correct stationary state and phase diagrams, but still fail to be correct at intermediate times.

There is a fundamental problem with all Gaussian theories, however sophisticated: apart from special cases, the interference noise distribution is generally *not* of a Gaussian shape. This becomes clear when we follow Nishimori and Ozeki (1993) and return to the simulation experiments of Figure 1 to measure the distribution $P_z(z)$ of Equation 7, resulting in Figure 2. $P_z(z)$ is only (approximately) Gaussian when pattern recall occurs. Hence the successes of Gaussian theories in describing recall trajectories, and their perpetual problems in describing the nonrecall ones.

Exact Results

The only exact procedure known at present is based on a philosophy different from those described so far. Rather than using the probability $p_i(S)$ of finding a microscopic state S at time t in order to calculate the statistics of a macroscopic observable $m(S)$ at time t , one here turns to the probability $\text{Prob}[S(0), \dots, S(t_m)]$ of finding

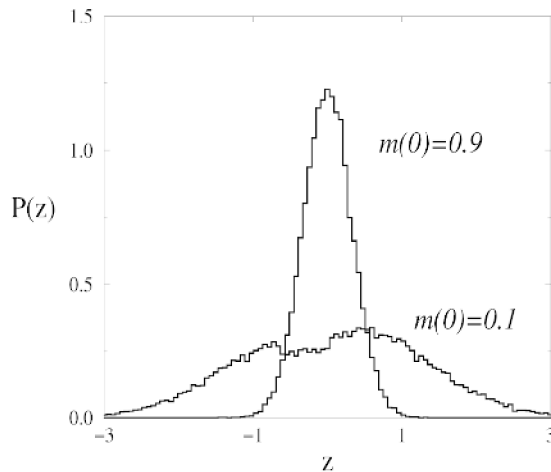


Figure 2. Distributions of interference noise variables $z_i = (1/N)\sum_{\mu>1} \xi_i^\mu \xi_j^\mu \sum_{j \neq i} \xi_j^\mu S_j$, as measured in the simulations of Figure 1, at $t = 10$. Unimodal histogram: noise distribution following $m(0) = 0.9$ (leading to recall). Bimodal histogram: noise distribution following $m(0) = 0.1$ (not leading to recall).

a microscopic *path* $S(0) \rightarrow S(1) \rightarrow \dots \rightarrow S(t_m)$. One also adds time-dependent external sources (similar to injected currents) to the local fields, $h_i(S) \rightarrow h_i(S) + \theta_i(t)$, to probe the networks via small perturbations. The statistics of paths are fully captured by the following moment-generating function:

$$Z[\psi] = \langle e^{-i\sum_i \xi_i^m \psi_i \phi(S_i(t))} \rangle \quad (10)$$

It generates expectation values of most relevant observable quantities, including those involving neuron states at different times, such as correlation functions $C_{ij}(t, t') = \langle S_i(t)S_j(t') \rangle$ and response functions $G_{ij}(t, t') = \partial \langle S_i(t) \rangle / \partial \theta_j(t')$:

$$\langle S_i(t) \rangle = i \lim_{\psi \rightarrow 0} \frac{\partial Z[\psi]}{\partial \psi_i(t)} \quad C_{ij}(t, t') = - \lim_{\psi \rightarrow 0} \frac{\partial^2 Z[\psi]}{\partial \psi_i(t) \partial \psi_j(t')}$$

$$G_{ij}(t, t') = i \lim_{\psi \rightarrow 0} \frac{\partial^2 Z[\psi]}{\partial \psi_i(t) \partial \theta_j(t')}$$

The idea is next to assume (correctly) that for $N \rightarrow \infty$, only the statistical properties of the stored patterns will influence the above macroscopic quantities, so that the generating function $Z[\psi]$ can be averaged over all pattern realizations, i.e., $Z[\psi] \rightarrow \bar{Z}[\psi]$. After a certain amount of algebra to evaluate $\bar{Z}[\psi]$ one finds that the recall dynamics can again be described in terms of a single “effective neuron” $S(t)$ with an effective local field. However, this effective local field will generally depend on past states of the neuron and on zero-average but temporally correlated Gaussian noise contributions $\phi(t)$:

$$h(t|\{S\}, \{\phi\}) = m(t) + \theta(t) + \alpha \sum_{t' < t} K(t, t')S(t') + \sqrt{\alpha}\phi(t) \quad (11)$$

The first comprehensive neural network studies along these lines, dealing with fully connected networks, were carried out by Rieger, Schreckenberg, and Zittartz (1988) and Horner et al. (1989), followed by applications to asymmetrically and symmetrically extremely diluted networks (Kree and Zippelius, 1991; Watkin and Sherrington, 1991) (in symmetrically diluted networks the dilution is subject to the constraint $c_{ij} = c_{ji}$ for all (i, j)). More recent applications include sequence processing networks (Düring, Coolen, and Sherrington, 1998), where the overlap m is defined with respect to the “moving” target, i.e., $m(t) = (1/N)\sum_i S_i(t)\xi_i^t$. The differences between the results obtained for different models are mainly in the actual form taken by the effective local field in Equation 11, i.e., in the dependence of the “retarded self-interaction” kernel $\mathbf{K} = \{K(t, t')\}$ and the covariance matrix $\Phi = \{\Phi(t, t') = \langle \phi(t)\phi(t') \rangle\}$ of the interference-induced Gaussian noise on the macroscopic quantities $\mathbf{C} = \{C(s, s') = \lim_{N \rightarrow \infty} (1/N)\sum_i C_{ii}(s, s')\}$ and $\mathbf{G} = \{G(s, s') = \lim_{N \rightarrow \infty} (1/N)\sum_i G_{ii}(s, s')\}$. For instance:

Model	Synapses J_{ij}	$K(t, t')$	$\langle \phi(t)\phi(t') \rangle$
Fully connected, static patterns	$\frac{1}{N} \sum_{\mu=1}^N \xi_i^\mu \xi_j^\mu$	$(\mathbf{I} - \mathbf{G})^{-1} \mathbf{G}$	$(\mathbf{I} - \mathbf{G})^{-1} \mathbf{C} (\mathbf{I} - \mathbf{G})^{-1}$
Fully connected, pattern sequence	$\frac{1}{N} \sum_{\mu=1}^N \xi_i^{\mu+1} \xi_j^\mu$	0	$\sum_{n \geq 0} (\mathbf{G}^\dagger)^n \mathbf{C} \mathbf{G}^n$
Symmetric extremely diluted, static patterns	$\frac{c_{ij}}{c} \sum_{\mu=1}^{ac} \xi_i^\mu \xi_j^\mu$	\mathbf{G}	\mathbf{C}
Asymmetric extremely diluted, static patterns	$\frac{c_{ij}}{c} \sum_{\mu=1}^{ac} \xi_i^\mu \xi_j^\mu$	0	\mathbf{C}

The correlation and response functions are to be solved self-consistently from the following (closed) equations, involving the statistics of the single effective neuron experiencing the field described by Equation 10:

$$C(t, t') = \langle S(t)S(t') \rangle \quad G(t, t') = \frac{\partial}{\partial \theta(t')} \langle S(t) \rangle \quad (12)$$

In the case of sequential dynamics the picture is found to be very similar to the one above; instead of discrete time labels $t \in \{0, 1, \dots, t_m\}$, path summations, and matrices, there one has a real-time variable $t \in [0, t_m]$, path integrals, and integral operators.

It is now clear what happens with Gaussian theories: they can at most produce exact results for asymmetric networks. Any degree of symmetry in the synapses is found to induce, via the kernel $K(t, t')$ a non-zero retarded self-interaction that constitutes a non-Gaussian contribution to the local fields. One also sees that it is not extreme dilution that is responsible for a drastic simplification of the dynamics, but synaptic asymmetry. This is underlined by Figure 3, which shows the phase diagrams of the asymmetrically and the symmetrically extremely diluted Hopfield model (both with $\lim_{N \rightarrow \infty} c/N = 0$ and $c \rightarrow \infty$), as derived from the above equations. The indirect effect on the dynamics of (even partial) synaptic symmetry is fundamental and can be understood as follows: such (partial) symmetry implies that an excitatory synapse $i \rightarrow j$ is more likely to be accompanied by an excitatory synapse $j \rightarrow i$ than an inhibitory one; conversely, an inhibitory synapse $i \rightarrow j$ is more likely to be accompanied by an inhibitory synapse $j \rightarrow i$. In both cases the net effect for an active (inactive) neuron i is an effective self-excitation (self-inhibition); any degree of synaptic symmetry thus acts as a retarded “mirror,” which complicates (and slows down) the dynamics.

Non-Gaussian Approximations

Owing to the generally complicated nature of a rigorous treatment of recall dynamics, there is still a market for approximate theories. In view of the non-Gaussian shape of the interference noise distribution, several attempts have been made to construct non-Gaussian

approximations. In all cases the aim is to arrive at a theory involving only macroscopic quantities with a *single* time argument. Henkel and Oppen (1990) approximated the theory of a fully connected network with parallel dynamics by replacing the field in Equation 10 with the simpler expression

$$h(t) = m(t) + \theta(t) + d(t)S(t-1) + \sqrt{\alpha}\phi(t)$$

$$\langle \phi(t)\phi(t') \rangle = \sigma^2(t)\delta_{tt'}$$

(with independent zero-average Gaussian fields $\phi(t)$), followed by a self-consistent calculation of $d(t)$ (representing the retarded self-interaction) and of the width $\sigma(t)$ of the distribution of $\phi(t)$. This results in an interference noise distribution $P_t(z)$ (Equation 6), which is the sum of *two* Gaussians, which appears sensible in view of Figure 2, and a nice (but not perfect) agreement with numerical simulations.

A different philosophy was followed by Coolen and Sherrington (1994). Here (as yet exact) equations are derived for the evolution of the macroscopic quantities m and r in Equation 4 (for sequential dynamics), which both involve $P_t(z)$:

$$\begin{aligned} \frac{d}{dt} m &= \int dz P_t(z) \tanh[\beta(m + z)] \\ \frac{d}{dt} r &= \frac{1}{\alpha} \int dz P_t(z) z \tanh[\beta(m + z)] + 1 - r \end{aligned}$$

Next one closes these equations *by hand*, using a maximum-entropy (or “Occam’s razor”) argument: Instead of calculating $P_t(z)$ with the (unknown) distribution $p_t(S)$, it is calculated upon assigning equal probabilities to all states S with $m(S) = m$ and $r(S) = r$, followed by averaging over all realizations of the stored patterns. This results in an explicit (non-Gaussian) expression for the noise distribution in terms of (m, r) only, a theory that is exact for short times and in equilibrium, accurate predictions of the macroscopic flow in the (m, r) plane, but (again) deviations in predicted time dependences at intermediate times. This theory, and its performance, was later improved by applying the same ideas to a derivation of a dynamic equation for the function $P_t(z)$ itself (rather than for m and r only).

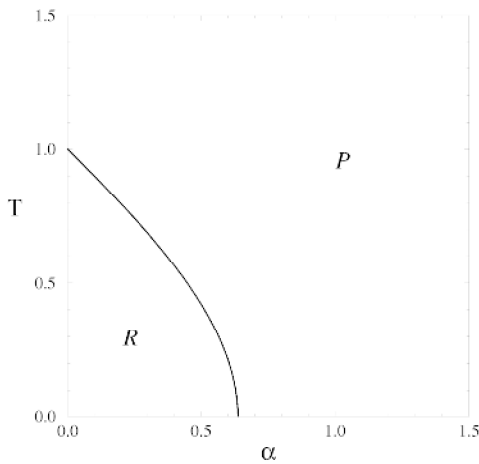
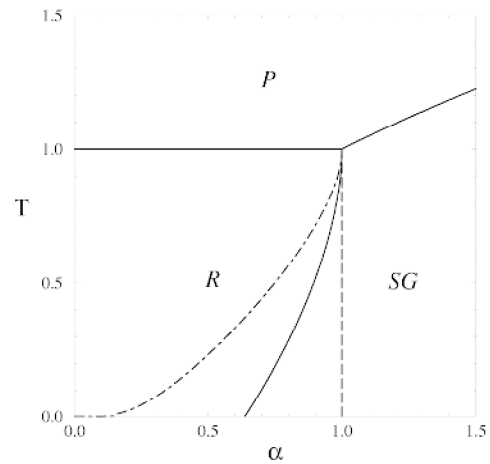


Figure 3. *Left*, Phase diagram of *asymmetrically* extremely diluted networks. Solid line traces continuous transition, separating a paramagnetic (P) region (with pseudo-random motion) from a recall (R) region, reaching the $T = 0$ axis at $\alpha_c = 2/\pi \approx 0.637$. *Right*, Phase diagram of *symmetrically* extremely diluted networks. Solid lines indicate continuous transitions, separating a paramagnetic region from a recall region ($\alpha < 1$) and from a spin-glass region (SG) ($\alpha > 1$) (SG describes the unwanted state where m



$= 0$ and $r > 0$, in the language of Figure 1), and separating the recall region from a spin-glass region (calculated using the so-called replica symmetry assumption). Dashed-dotted line indicates the AT instability (where this assumption breaks down), reaching the $T = 0$ axis at $\alpha_c^{RS} = 2/\pi \approx 0.637$; replica symmetry breaking displaces the solid line to a new (dashed) line, with a new storage capacity $\alpha_c^{RSB} = 1$.

Discussion

In this review I have attempted to explain the various theories that have been developed since the mid-1970s, in often disjunct communities, to understand the dynamics of pattern recall in recurrent neural networks. The methods and ideas described in this review can be, and indeed have been, extended and generalized in many ways. This includes networks with spatial structure, axonal or neuronal delays, nonbinary neurons such as graded-response or integrate-and-fire ones, correlated patterns, networks of coupled oscillators, and so on. Increased biological realism of a model is, as always, paid for by a proportionate reduction in how far one can push the mathematical analysis. However, remarkable progress has been made over the last two decades, with a wave of dynamical studies especially around 1990 (unfortunately, many popular textbooks on neural network theory were written in about 1989, and thus contain very little on recall dynamics), but with new ground being covered up to the present day. Recall dynamics is interesting and relevant in its own right (and a nice challenge for the theorist), yet one can hardly underestimate the importance, especially to those interested in understanding recurrent modules in real nervous tissue, of the other deliverable of dynamical studies: they open up for mathematical analysis the area of recurrent networks with *non-symmetric* (i.e., biologically more realistic) synapses, which is not accessible with the more common equilibrium tools.

Road Maps: Dynamic Systems; Learning in Artificial Networks

Related Reading: Computing with Attractors; Statistical Mechanics of Neural Networks

References

- Amari, S. I., 1977, Neural theory of association and concept-formation, *Biol. Cybern.*, 26:175–185. ♦
- Amari, S. I., and Maginu, K., 1988, Statistical neurodynamics of associative memory, *Neural Netw.*, 1:63–73.
- Buhmann, J., and Schulten, K., 1987, Noise driven temporal association in neural networks, *Europhys. Lett.*, 4:1205–1209.
- Coolen, A. C. C., and Ruijgrok, T. W., 1988, Image evolution in Hopfield networks, *Phys. Rev. A*, 38:4253–4255. ♦
- Coolen, A. C. C., and Sherrington, D., 1994, Order parameter flow in the fully connected Hopfield model near saturation, *Phys. Rev. E*, 49:1921–1934; *Phys. Rev. E*, 49:5906.
- Derrida, B., Gardner, E., and Zippelius, A., 1987, An exactly solvable asymmetric neural network model, *Europhys. Lett.*, 4:167–173.
- Düring, A., Coolen, A. C. C., and Sherrington, D., 1998, Phase diagram and storage capacity of sequence processing neural networks, *J. Phys. A Math. Gen.*, 31:8607–8621.
- Henkel, R. D., and Oppen, M., 1990, Distribution of internal fields and dynamics of neural networks, *Europhys. Lett.*, 11:403–408.
- Horner, H., Bormann, D., Frick, M., Kinzelbach, H., and Schmidt, A., 1989, Transients and basins of attraction in neural network models, *Z. Phys. B*, 76:383–398.
- Kree, R., and Zippelius, A., 1991, Asymmetrically diluted neural networks, in *Models of Neural Networks* (R. Domany, J. L. van Hemmen, and K. Schulten, Eds.), Berlin: Springer-Verlag, pp. 193–212.
- Nishimori, H., and Ozeki, T., 1993, Retrieval dynamics of associative memory of the Hopfield type, *J. Phys. A Math. Gen.*, 26:859–871.
- Okada, M., 1995, A hierarchy of macrodynamical equations for associative memory, *Neural Netw.*, 8:833–838.
- Riedel, U., Kühn, R., and van Hemmen, J. L., 1988, Temporal sequences and chaos in neural networks, *Phys. Rev. A*, 38:1105–1108.
- Rieger, H., Schreckenberg, M., and Zittartz, J., 1988, Glauber dynamics of the Little-Hopfield model, *Z. Phys. B*, 72:523–533.
- Watkin, T. L. H., and Sherrington, D., 1991, The parallel dynamics of a dilute symmetric Hebb-rule network, *J. Phys. A Math. Gen.*, 24:5427–5433.

Echolocation: Cochleotopic and Computational Maps

Nobuo Suga

Introduction

The order *Chiroptera* (bats) accounts for one-fifth of mammalian species and has two suborders, Megachiroptera, with 154 species, and Microchiroptera, with about 800 species. All microchiropterans thus far studied echolocate, but only one megachiropteran, the genus *Rousettus*, does. The morphology and ecology of bats are so diverse that echolocation behavior and orientation sounds (biosonar pulses, or simply pulses) are quite different among different species of bats. Accordingly, the auditory system differs among species. Studies in the neuroscience of echolocation have been mainly performed with four different species of microchiropterans: *Pteronotus parnellii* (mustached bat), *Rhinolophus ferrumequinum* (horseshoe bat), *Myotis lucifugus* (little brown bat), and *Eptesicus fuscus* (big brown bat). Among them, the parallel/hierarchical processing of biosonar information was best explored in *Pteronotus*. Thus, the neurophysiology of the auditory system of *Pteronotus* is mainly described in this article.

Properties of Biosonar Signals

For insect capture and orientation, certain microchiropterans emit constant-frequency (CF) and/or frequency-modulated (FM) sounds. The biosonar pulses of *Pteronotus* always consist of a long CF component followed by a short FM component. Since each biosonar pulse contains four harmonics (H_{1-4}), there are potentially

eight major components (CF_{1-4} and FM_{1-4} in Figure 1A). The second harmonic (H_2) is always predominant, with CF_2 at about 61 kHz and FM_2 sweeping from 61 kHz to about 49 kHz. The CF_2 “resting” frequency differs among individual bats and is sexually dimorphic. In target-directed flight, a CF-FM pulse varies in duration and emission rate, but its spectrum changes little. Target echoes usually overlap with the emitted pulses (Figure 1B). *Rhinolophus* also emits CF-FM sounds with H_2 at about 83 kHz. In comparison, *Myotis* and *Eptesicus* emit FM pulses that sweep downward about one octave within a range between 100 and 15 kHz. The properties of FM pulses vary, depending on the species and situations in echolocation. Target echoes usually do not overlap with the emitted pulses (Figure 1C and 1D).

Different components or parameters of echoes carry different types of target information (Table 1). The long CF component is an ideal signal for target detection and measurement of target relative velocity (2.84 m/s/kHz Doppler shift for a 61-kHz carrier) and the velocity of an insect’s wing beat. *Pteronotus* optimizes the acquisition of velocity information by an acoustic behavior called Doppler-shift compensation, by which the frequency of the echo CF_2 is stabilized at approximately 61 kHz. The short FM component is more appropriate for ranging (17.3 cm/ms), localization, and characterization of a target. Since bats emit acoustic pulses at rates of 5–200/s, their auditory system creates stroboscopic acoustic images through enormous parallel processing of pulse-echo pairs in all aspects listed in Table 1.

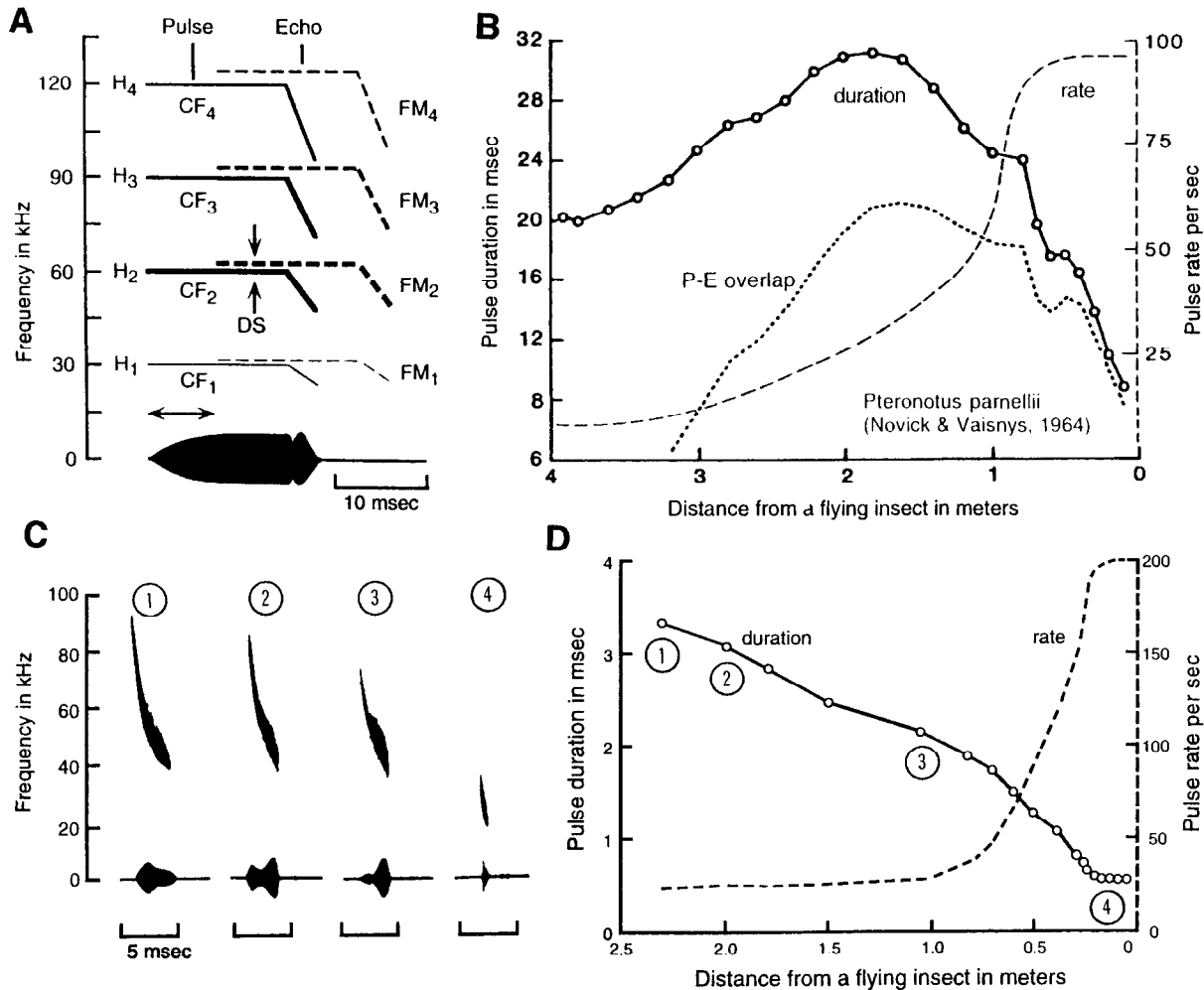


Figure 1. Biosonar “pulses” of the mustached bat, *Pteronotus parnellii* (A, B) and the little brown bat, *Myotis lucifugus* (C, D). A, The four harmonics (H_{1–4}) of a *Pteronotus* biosonar pulse each consist of a long CF component terminating in a short FM sweep, so that there are eight components, CF_{1–4} and FM_{1–4}. Most of the sound energy is in the second harmonic (H₂), with CF₂ at approximately 61 kHz and FM₂ sweeping from approximately 61 kHz to approximately 49 kHz. As the bat approaches a target, a reflected echo is Doppler-shifted (DS) in frequency, and its delay from the pulse becomes shorter. B, During target-directed flight, *Pteronotus* increases the

rate of its pulse emissions. The bat initially lengthens the duration of the pulse and then systematically shortens it. Pulse-echo overlap is maximal when the pulses emitted are longest in duration. C, *Myotis* emits a short, downward-sweeping FM pulse. D, During target-directed flight, *Myotis* increases the rate of pulse emissions, shortens pulse duration as to avoid pulse-echo overlap, and lowers the frequency sweep range of the pulse. Circled numbers 1–4 correspond to those in C. (B adapted from Novick, A., and Vaisnys, J. R., 1964, Echolocation of flying insects by the bat, *Chilonycteris parnellii*. *Biol. Bull.*, 127:478–488.)

Table 1. Different acoustic parameters of an echo relative to the biosonar pulse at the ear carry different types of target information

Echo	Target
Doppler shift	Velocity
Steady component	Relative velocity
Periodic component	Insect wingbeat
Amplitude	Subtended angle
Steady component	Relative size
Periodic component	Insect wing beat
Delay	Range
Amplitude + delay	Absolute size
Amplitude spectrum	Fine characteristics
Envelope	Fine characteristics
Binaural cues	Azimuth
Pinna-tragus cue	Elevation

The Auditory System

The gross structure of the auditory system of *Pteronotus* is basically the same as that of other mammals, but it shows a unique functional organization reflecting the properties of its biosonar pulse and echolocation behavior. One of the most striking features is the extremely sharp tuning and the large population of neurons at the auditory periphery for fine frequency analysis of the CF₂ component at approximately 61 kHz. Because of this extremely sharp tuning, they can easily code the small frequency modulation of echoes that would be evoked by the wings of flying insects. The CF₂ processing channel is disproportionately large, from the cochlea through the auditory cortex. The other striking feature of the auditory system of *Pteronotus* is the parallel/hierarchical organization to extract certain types of biosonar information by “combination-sensitive” neurons, as described below.

All the CF and FM components of the biosonar pulse differ in frequency (Figure 1A), and hence they are separately analyzed in the cochlea and are sent in parallel into the brain by different auditory nerve fibers. At the auditory periphery, frequency is represented by the location along the basilar membrane in the cochlea and phase (or stimulus)-locked discharges of auditory nerve fibers. There are no anatomical axes for the representation of stimulus amplitude (intensity) and time (i.e., duration of stimuli and interval between the stimuli). In the central auditory system, divergent/convergent interactions repeatedly take place, and multiple frequency maps and neuronal response properties are formed that differ from those at the periphery. Inhibition (in particular, lateral inhibition), coincidence detectors (AND gates), and multipliers play important roles in the processing of auditory information in the frequency and amplitude domains, whereas delay lines, in addition to the above three, play key roles in time domain processing. For example, the response properties of neurons tuned to echo delays, durations, repetition rates, sequences of sounds, or interaural time differences are created with these four neural mechanisms. The length of delay lines is very short for the processing of interaural time differences, but it can be very long for the processing of other acoustic parameters. Long delay lines are created by inhibition that evokes rebound off-response.

In the CF_{1-3} processing channels, the frequency-tuning curves of peripheral neurons are much sharper than those in other channels. However, they are still triangular in shape, so that the ambiguity in coding frequency by a single neuron is large at high-stimulus amplitudes. All these peripheral neurons respond not only to tone bursts, but also to FM sounds and noise bursts. They are

not "combination sensitive." At higher levels, however, many neurons show sharp "level-tolerant" frequency tuning for further analysis of CF signals. A level-tolerant tuning curve has a very narrow bandwidth regardless of stimulus level (amplitude) and plays a role in level-tolerant fine-frequency analysis. Level-tolerant tuning is created by lateral inhibition, which occurs at different levels of the central auditory system (Figure 2). Sharpening of tuning by lateral inhibition is most dramatic in the CF_2 processing channel, and it is practically completed in the medial geniculate body. Inhibition also creates amplitude selectivity, so that some central auditory neurons are tuned not only in frequency but also in amplitude (Figure 2). The CF_2 processing channel mostly projects to the Doppler-shifted CF (DSCF) area of the primary auditory cortex (Figure 3). For the extraction of velocity information from biosonar pulse-echo pairs, the parts of the CF_{1-3} processing channels are integrated in the inferior colliculus (IC), and perhaps in the medial geniculate body (MGB). (It is likely that this integration greatly depends on corticofugal feedback.) As a result, CF/CF combination-sensitive neurons are produced (Figure 4). These neurons project to the CF/CF and DIF areas in the auditory cortex (AC) (Figure 3).

In the FM_{1-4} processing channels, some neurons at higher levels respond selectively to FM sounds because of neural circuits incorporating *disinhibition* and/or *facilitation*. The IC has frequency-versus-latency coordinates. Part of the latency axis is perhaps used as delay lines, as explained later. The parts of the FM_{1-4} processing channels are integrated in the inferior colliculus to create FM-FM neurons, which are tuned to particular echo delays for the extraction of target-distance information from biosonar pulse-echo pairs. The neural mechanisms for creating FM-FM neurons consist of at least

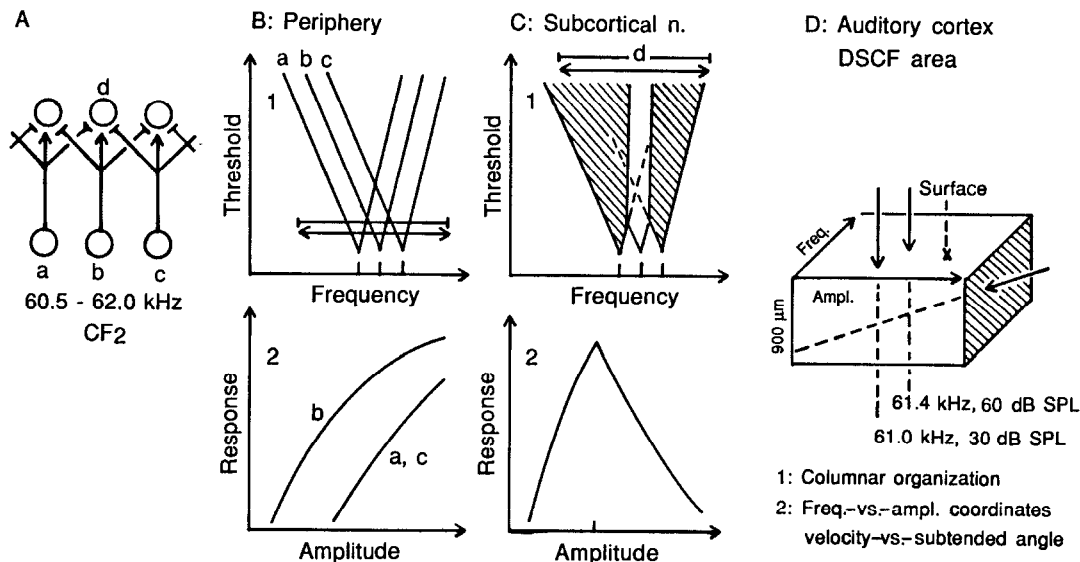


Figure 2. Sharpening of frequency-tuning curves and creation of amplitude tuning by inhibition. *A*, Neural circuit for lateral inhibition. The arrowheads and bar heads represent excitatory and inhibitory synapses, respectively. *B*, Frequency-tuning curves (*B1*) and impulse-count functions (*B2*) of three peripheral neurons: *a*, *b*, and *c* in *A*. The impulse-count functions are measured with a tone burst at a best frequency of neuron *b*. The double-headed arrow in *B1* indicates that the neurons respond to upward as well as downward-sweeping FM sounds and that the threshold of the response is slightly higher than that to a CF tone at a neuron's best frequency. The horizontal bar in *B1* indicates that these neurons respond to a noise burst and that the threshold of the response is slightly higher than those to the FM sounds. *C*, Subcortical neuron *d* in *A* has a sharp, level-tolerant frequency-tuning curve sandwiched between inhibitory tuning curves

(shaded) as a result of lateral inhibition (*C1*). This subcortical neuron is tuned to a weak tone burst because of the inhibition. The impulse-count function measured with a tone burst at a best frequency of neuron *d* is highly nonmonotonic (*C2*). The double-headed arrow and horizontal bar in *C1* indicate that neuron *d* does not respond to FM sounds and noise bursts. *D*, The functional organization of the DSCF area in the auditory cortex. The DSCF area shows columnar organization characterized by a particular combination of best frequency and best amplitude. It has the frequency-versus-amplitude coordinates. (Based on Suga, N., 1994, Processing of auditory information carried by complex species-specific sounds, in *Cognitive Neuroscience* (M. S. Gazzaniga, Ed.), Cambridge, MA: MIT Press, pp. 295–318; Suga and Manabe, 1982.)

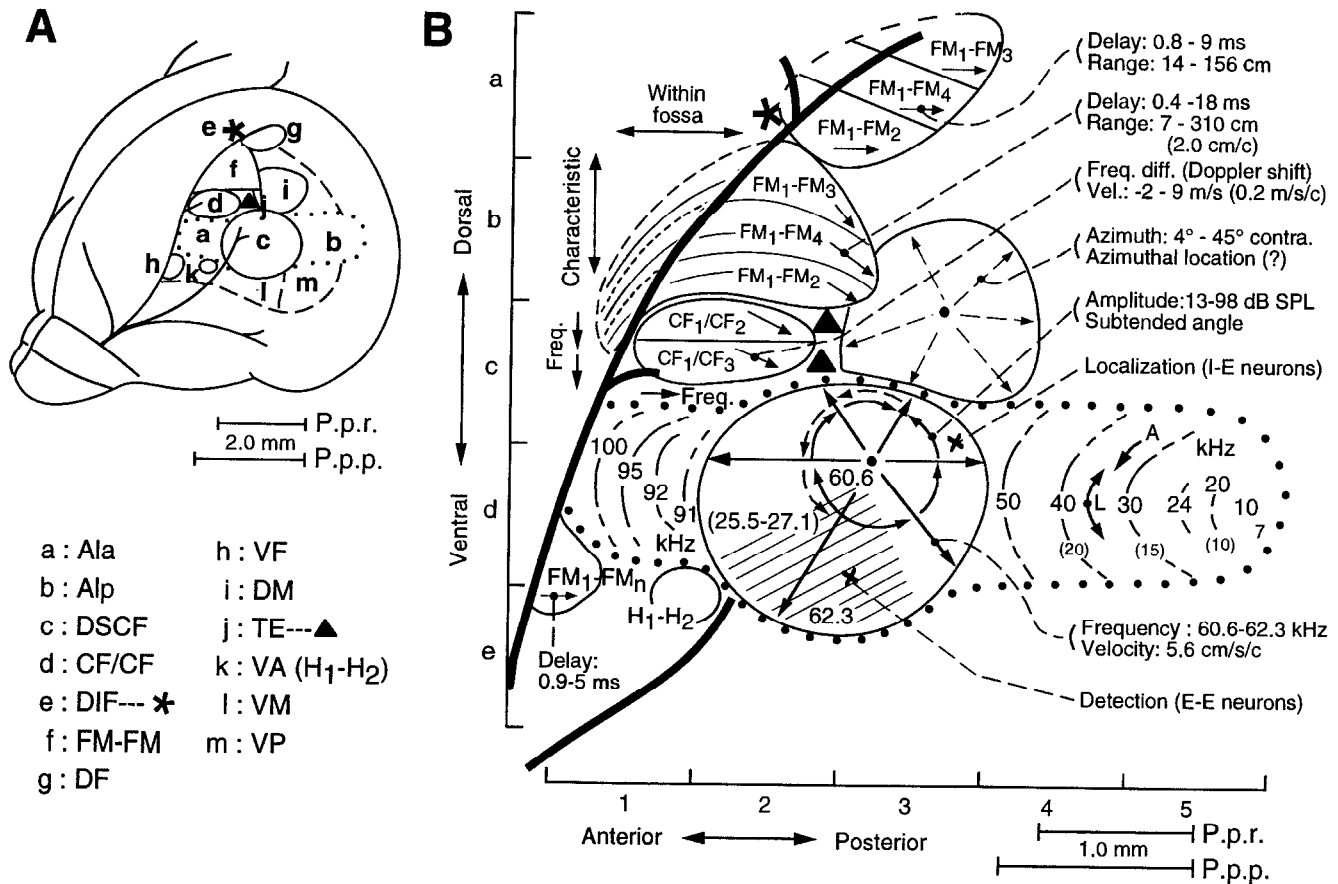


Figure 3. The auditory cortex of the mustached bat. **A**, Dorsolateral view of the left cerebral hemisphere and the branches of the median cerebral artery. The long branch is on the fossa. The auditory cortex consists of several areas (a-m), are called AIa, Alp, etc. The functional organization of certain areas has been electrophysiologically explored and is graphically summarized in **B**. **B**, The tonotopic representation of AI (AIa, DSCF, and Alp) and the functional organization of the other areas are indicated by lines and arrows. The DSCF area has axes representing target-velocity information (echo frequency: 60.6–62.3 kHz) or subtended target-angle information (echo amplitude: 13–98 dB SPL). It consists of two subdivisions suited for either target detection (shaded) or target localization (unshaded). These subdivisions are occupied mainly by E-E or I-E binaural neurons. The anterior and posterior halves of the DSCF area are hypothesized to be adapted for processing echoes from either fluttering or stationary targets. Neurons in the DSCF area are sensitive to an FM₁-CF₂ combination. The FM-FM area consists of three major types of FM-FM neurons (FM₁-FM₂, FM₁-FM₃, and FM₁-FM₄), which form separate clusters. Each cluster has

an axis representing target ranges from 7 to 310 cm (echo delay: 0.4–18 ms). The dorsoventral axis of the FM-FM area probably represents fine target characteristics. The CF/CF area consists of two major types of CF/CF neurons (CF₁/CF₂ and CF₁/CF₃), which are also found in separate clusters. Each cluster has two frequency axes and represents a target velocity from -2 to +9 m/s (echo Doppler shift: -0.7 to +3.2 kHz for CF₂, and -1.1 to +4.8 kHz for CF₃). The DF area consists of the three clusters of FM-FM neurons and has an axis representing target ranges from 7 to 140 cm. The VF area also contains FM-FM neurons and represents target ranges up to 80 cm. The DM area has an axis representing the azimuthal location of the target on the contralateral side in front of the animal. This azimuthal representation is incorporated with tonotopic representation. In the VP area, azimuthal motion-sensitive neurons have been found. The functional organization of the VA and VP areas remains to be studied further. (Adapted from Suga, N., 1994, Processing of auditory information carried by complex species-specific sounds, in *Cognitive Neurosciences* (M. S. Gazzaniga, Ed.), Cambridge, MA: MIT Press, pp. 295–318.)

four physiological components: phasic/constant latency responding neurons, delay lines, coincidence detectors, and amplifiers. An FM sound sequentially stimulates an array of neurons tuned to different frequencies by sweeping across their frequency tuning curves. Phasic/constant latency responders are suited for coding that moment of the stimulus. They code the exact timing of the pulse and its echo, i.e., the echo delay. They are mostly created in the subcollicular nuclei. The other three components are now considered to be in the IC, although part of the delay lines are perhaps created in subcollicular nuclei. Delay lines shift the response to the pulse FM₁ in time. A coincidence detector (FM-FM neuron) has two inputs: one input carries activity evoked by the pulse FM₁ through a delay line, and the other input carries activity evoked by the echo FM_n without delay lines. An echo always delays acoustically from the

pulse emitted by the bat according to the distance to an echo source. At a coincidence detector whose neural delay is equal to an echo delay, the excitatory response to the echo FM_n arrives at the same time as the delayed excitatory response evoked by the pulse FM₁. Then, the coincidence detector shows a strong facilitative response (Figure 5). This facilitation greatly depends on corticofugal feedback.

Collicular FM-FM combination-sensitive neurons produced in this way are tuned to particular echo delays, i.e., target distances. They project to thalamic FM-FM neurons, which show stronger facilitative responses to pulse-echo pairs and sharper delay tuning than the collicular FM-FM neurons. These thalamic FM-FM neurons project to the FM-FM, DF, and VF areas in the AC. The response of these thalamic neurons greatly depends on corticofugal

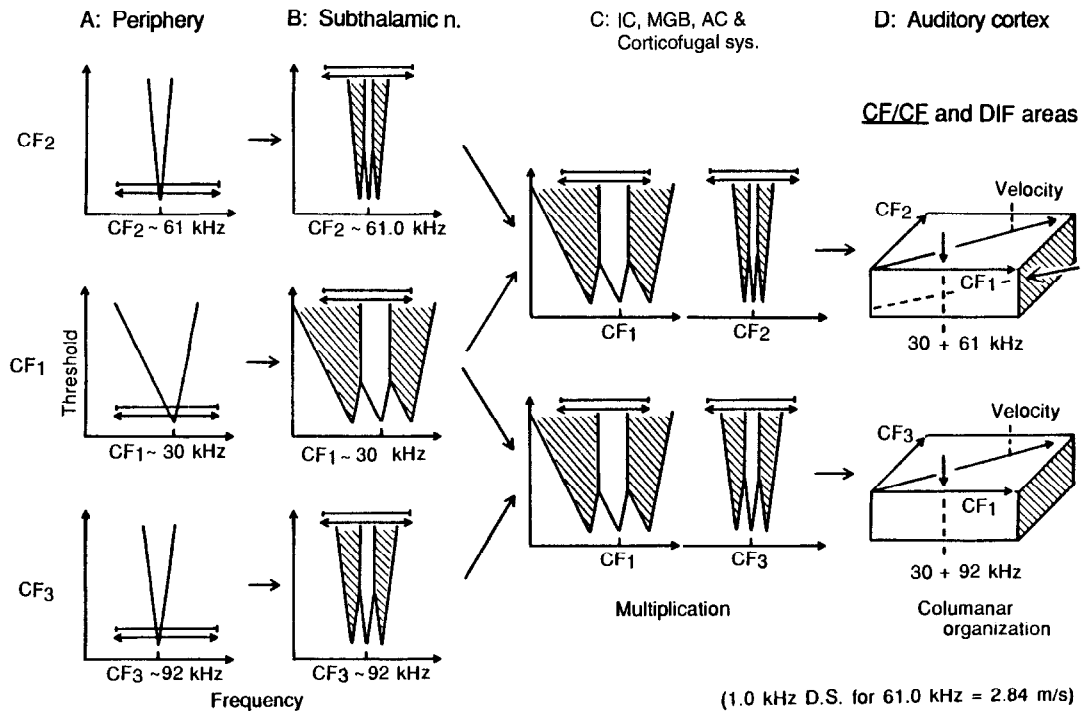


Figure 4. Neural mechanisms for creating frequency-versus-frequency coordinates. Triangular frequency-tuning curves of three channels ($CF_{1,2}$ and CF_3) at the periphery (A) are changed into level-tolerant frequency-tuning curves in the subthalamic auditory nuclei by lateral inhibition (B). Then, the CF_1 and CF_2 or CF_3 channels are integrated in the inferior colliculus (IC) and, perhaps, in the medial geniculate body (MGB) to create two types of velocity-tuned, CF/CF combination-sensitive neurons (C). These neu-

rons project to the CF/CF area in the auditory cortex and form the CF_1/CF_2 and CF_1/CF_3 subdivisions. In each subdivision, a cortical minicolumn is characterized by a particular combination of two frequencies. Each subdivision has frequency-versus-frequency coordinates, in which relative velocities of targets are systematically mapped. CF/CF neurons also cluster in the DIF area. (Based on Suga and Tsuzuki, 1985.)

feedback. As a consequence of such parallel-hierarchical processing of complex sounds, there are many functional divisions in the auditory cortex (Figure 3).

Cochleotopic (Frequency) Map in the Auditory Cortex

The central auditory system of *Pteronotus* processes different types of biosonar information and communication calls in a parallel/hierarchical way. As a result, its AC consists of many areas containing different types of combination-sensitive neurons. Therefore, each of these areas has more than one frequency axis. Accordingly, frequency maps based on best frequencies (BFs) of combination-sensitive neurons are much more complex than those based on BFs measured with single-tone stimuli, and indicate the following four important facts. (1) The frequency axes in the AC are not exact copies of the frequency axis along the basilar membrane in the cochlea. Certain portions of the peripheral frequency axis are reduced or enlarged in a cortical frequency axis. (2) Different portions of the peripheral frequency axis are superimposed in parallel or orthogonally across certain cortical areas. (3) An area where the frequency axis can hardly be demonstrated with single-tone stimuli has distinct frequency axes when studied with combination tones. (4) The complex multiple frequency representations in the AC are directly related to the representations of different types of biosonar information.

Computational Maps in the Auditory Cortex

The response properties of certain types of combination-sensitive neurons can be easily related to the processing of particular types

of biosonar information, because the CF and FM components of the biosonar signals are suited for the measurements of target velocities and distances, respectively.

Neurons in the DSCF processing area are tuned to particular frequencies and amplitudes of the CF_2 component and show a facilitative response to the echo CF_2 combined with the pulse CF_1 and/or FM_1 . Most of them are sensitive to periodic frequency modulation. In the DSCF area, individual minicolumns, often called "cortical modules," are characterized by a particular combination of frequency and amplitude. They form frequency-versus-amplitude coordinates for the fine spatiotemporal representation of periodic changes of echoes in frequency and amplitude that would be evoked by insect wing beats (Figure 3B). Inactivation of the DSCF area disrupts frequency but not echo-delay discrimination.

Another cortical area (CF/CF) consists of two subdivisions: CF_1/CF_2 and CF_1/CF_3 (Figure 3B). The responses of CF_1/CF_2 and CF_1/CF_3 neurons to an echo CF_2 or CF_3 are facilitated by the pulse CF_1 emitted by the bat when the echo CF_2 or CF_3 returns from a target with a particular relative velocity and/or with beating wings. Their frequency-tuning curves are extremely sharp. They act as coincidence detectors tuned to particular combinations of CF sounds in the frequency (Doppler shift) domain. Since emitted pulses and Doppler-shifted echoes both vary independently in frequency from each other, the amount of a Doppler shift should be expressed with the frequency-versus-frequency coordinate system. In the CF/CF area, individual minicolumns are characterized by a particular combination of two frequencies. They form the frequency-versus-frequency coordinates by which the relationships between CF_1 and CF_2 or CF_3 (i.e., relative target velocities) are systematically mapped (Figures 3B and 4). Some CF/CF neurons are sensitive to

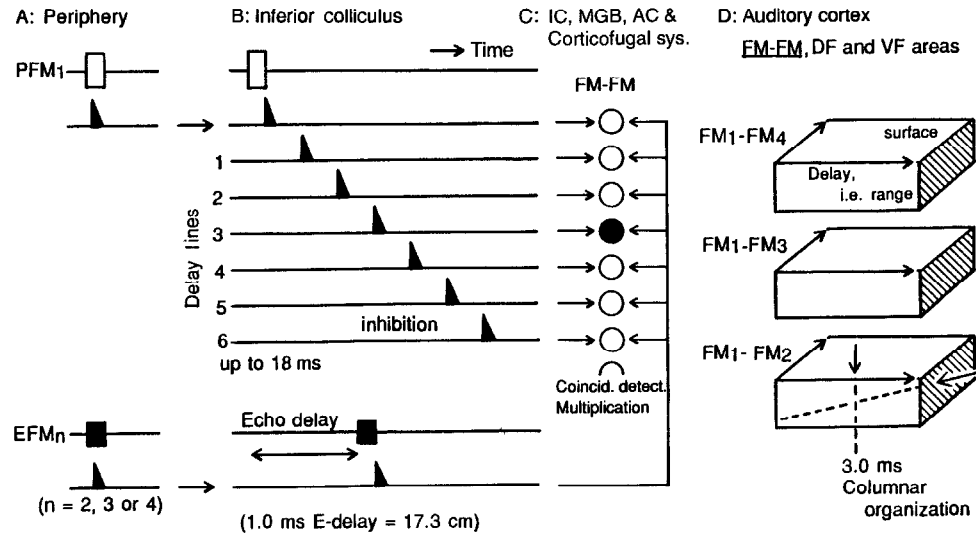


Figure 5. Neural mechanisms for creating delay-tuned neurons and a delay (range) axis. Delay-tuned neurons utilize delay lines (0.4–18 ms long) that are mostly created by inferior collicular neurons tuned to frequencies swept by the FM_1 (top portion of A and B). However, the delay-tuned neurons do not use delay lines created by neurons tuned to frequencies swept by FM_n ($n = 2, 3$, or 4 ; bottom portion of A and B). An array of delay-tuned neurons in the inferior colliculus (IC) receives signals from both the FM_1 and FM_n channels (C). In a delay-tuned neuron (filled circle), where an echo FM_n delay from the pulse FM_1 is equal to the neural delay line associated with it, both the signals arrive at the same time. The amount of facilitative ex-

citation of the neuron depends on the degree of coincidence. Three types of thalamic delay-tuned neurons separately project to the FM-FM area of the auditory cortex, forming three subdivisions within the FM-FM area (D). Each subdivision of the FM-FM area shows columnar organization characterized by a particular value of echo delay. It has an echo delay axis, i.e., target-range axis. Delay-tuned neurons also cluster in the DF and VF areas. For simplicity, phasic on-response (onset detector), amplitude tuning, and FM sensitivity are eliminated from the model. (Adapted from Suga, 1990, based on O'Neill and Suga, 1982.)

a combination of three CF signals: CF_1 , CF_2 , and CF_3 . CF/CF neurons are also found in another cortical area (dorsal intrafossa, or DIF) (Figure 3B).

Another specialized area (FM-FM) consists of three subdivisions: FM_1 - FM_2 , FM_1 - FM_3 , and FM_1 - FM_4 (Figure 3B). The responses of FM-FM neurons to an echo FM_n ($n = 2, 3$, or 4) are facilitated by the pulse FM_1 emitted by the bat when the echo FM_n returns from a target at a particular distance. They act as coincidence detectors tuned to particular combinations of FM sounds in the time (echo-delay) domain. In the FM-FM area, individual minicolumns are characterized by a particular echo delay. They form an echo-delay (range) map in each subdivision of the FM-FM area (Figures 3B and 5). Some FM-FM neurons are sensitive to a combination of three or even four FM signals. Inactivation of the FM-FM area disrupts delay but not frequency discrimination. FM-FM neurons are clustered not only in the FM-FM area, but also in the dorsal fringe (DF) and ventral fringe (VF) areas, which also have an echo delay map (Figure 3B). It is also important to note that most FM-FM neurons are more sensitive to the combinations of FM sounds than to the combinations of CF tones.

As explained above, velocity and distance information is extracted by combination-sensitive neurons comparing the fundamental of the emitted pulse (CF_1 or FM_1) with the higher harmonics of echo components (CF_n or FM_n). This heteroharmonic sensitivity is one of seven possible mechanisms used to protect the echolocation system from the jamming effect of biosonar pulses produced by conspecifics.

There are many important topics and findings of echolocation that are not described in this short article. One of the important topics to be discussed is sound localization. As in other mammalian and avian species, the bat's auditory system creates binaural neurons tuned to particular interaural time or intensity differences in the subcollicular auditory nuclei and the auditory space map in the

superior colliculus, which is an important nucleus for sensorimotor integration. In this nucleus, some neurons are tuned to a sound source at a combination of a particular azimuth, elevation, and depth. In the AC, however, the auditory space map has not yet been found in any animal, although it has been found that two types of binaural neurons (I-E and E-E) are separately clustered in the DSCF area and that the best azimuth to excite neurons varies systematically along the frequency axis of the AC.

Since the amplitude, velocity, and range maps do not exist in the periphery but are created centrally, these maps are called *computational maps*. The findings made in the AC of the mustached bat and in the visual cortex of the macaque monkey indicate that the auditory and visual systems share an identical principle for processing acoustic or visual scenes: different types of information-bearing elements and parameters are represented/processed in separate cortical areas in parallel. As visual information, auditory information is expressed by spatiotemporal patterns of neural activity in different cortical areas, and cortical neurons contributing to these patterns are quite different from peripheral neurons in response properties. (The cortical areas highly specialized for the processing of biosonar information are also involved in processing communication calls. Each cortical area apparently has multiple functions in auditory information processing.)

Ascending and Descending (Corticofugal) Auditory Systems

Auditory information sent into the brain from the cochlea by the auditory nerve is sent up to the AC in the cerebrum through the brainstem auditory nuclei, the IC in the midbrain, and the MGB in the thalamus. This ascending system is incorporated with the descending (corticofugal) system. In the corticofugal system, neural signals originating from the AC are sent down to the MGB and IC,

and then further down to the brainstem auditory nuclei and eventually to the cochlea. The corticofugal system forms feedback loops and plays an important role in adjusting and improving signal processing according to auditory experiences.

Adjustment and Improvement of Cochleotopic and Computational Maps

Corticofugal feedback amplifies excitatory responses to single tones by 1.5 times in the IC and 2.5 times in the MGB, whereas it amplifies facilitative responses to paired sounds by 2.9 times in the IC and 5.6 times in the MGB. Therefore, the auditory responses of subcortical neurons would be very weak without corticofugal feedback. Detailed studies on the corticofugal modulation of subcortical signal processing indicate the following important facts. (1) Neurons in a cortical minicolumn increase the auditory responses of “physiologically matched” subcortical neurons and sharpen, but do not shift, their tuning curves. (2) Neurons in a cortical minicolumn decrease the auditory responses of “physiologically unmatched” subcortical neurons, and sharpen and shift their tuning curves away from the tuning curves of the cortical neurons. (3) As the tuning curves of adjacent neurons along a frequency or echo delay axis overlap each other, the direction of the shift in tuning curve depends on the relative contribution of positive feedback and lateral inhibition. (These cortical functions, mediated by a highly focused positive feedback associated with widespread lateral inhibition, are referred to as *egocentric selection*.) (4) Lateral inhibition spreads only in a given cortical functional area. (5) The effect of egocentric selection is larger on thalamic neurons than on collicular neurons, and is larger on the facilitative responses of combination-sensitive neurons than on the excitatory responses of neurons primarily responding to single tones. (6) Without egocentric selection, the auditory responses of subcortical neurons are significantly weaker than normal. (7) Egocentric selection evoked overrepresentation of a particular value of an acoustic parameter and underrepresentation of adjacent values. In other words, it increases the contrast in neural representation of acoustic signals. (8) Short tone bursts at moderate intensity delivered for 30 minutes can evoke a change (overrepresentation of these tone bursts) in the IC, MGB, and AC. (9) The change becomes larger when the tone bursts become behaviorally relevant. (10) Egocentric selection plays a role in self-organizing the central auditory system according to auditory experiences.

Both the cortical cochleotopic and computational maps are the result not only of divergent and convergent projections in the ascending auditory system, but also of corticofugal feedback. For echolocation, these maps and response properties of central auditory neurons are continuously adjusted and improved by corticofugal feedback according to auditory experience.

Acknowledgments. This work was supported by NIDCD research grant No. DC00175.

Road Map: Other Sensory Systems

Related Reading: Auditory Cortex; Electrollocation; Sound Localization and Binaural Processing

References

- Fay, R. R., and Popper, A. N., 1995, *Hearing by Bats*, New York: Springer-Verlag.
- Gao, E., and Suga, N., 1998, Experience-dependent corticofugal adjustment of midbrain frequency map in bat auditory system, *Proc. Natl. Acad. Sci. USA*, 95:12663–12670.
- Huffman, R. F., and Henson, O. W., Jr., 1990, The descending auditory pathway and acousticomotor systems: Connections with the inferior colliculus, *Brain Res.*, 15:295–232.
- Nachtigall, P. E., and Moore, P. W. B., 1988, *Animal Sonar: Process and Performance*, New York: Plenum Press.
- Novick, A., and Vainsys, J. R., 1964, Echolocation of flying insects by the bat, *Chilonycteris*, *Biol. Bull.*, 127:478–488.
- O'Neill, W. E., and Suga, N., 1982, Encoding of target-range information and its representation in the auditory cortex of the mustached bat, *J. Neurosci.*, 2:17–31.
- Simmons, J. A., 1989, A view of the world through the bat's ear: The formation of acoustic image in echolocation, *Cognition*, 33:155–199. ♦
- Suga, N., 1990, Cortical computational maps for auditory imaging, *Neural Netw.*, 3:3–21.
- Suga, N., 1994, Processing of auditory information carried by complex species-specific sounds, in *Cognitive Neurosciences* (M. S. Gazzaniga, Ed.), Cambridge, MA: MIT Press, pp. 295–318. ♦
- Suga, N., Gao, E., Zhang, Y., Ma, X., and Olsen, J. F., 2000, The corticofugal system for hearing: Recent progress, *Proc. Natl. Acad. Sci. USA*, 97:11807–11814.
- Suga, N., Manabe, T., 1982, Neural basis of amplitude-spectrum representation in auditory cortex of the mustached bat, *J. Neurophysiol.*, 47:225–255.
- Suga, N., O'Neill, W. E., Kujirai, K., and Manabe, T., 1983, Specialization of “combination-sensitive,” neurons for processing of complex biosonar signals in the auditory cortex of the mustached bat, *J. Neurophysiol.*, 49:573–1626.
- Suga, N., and Tsuzuki, K., 1985, Inhibition and level-tolerant frequency tuning in the auditory cortex of the mustached bat, *J. Neurophysiol.*, 53:1109–1145.
- Thomas, J., Moss, C., and Vater, M., 1999, *Advances in the Study of Echolocation*, Chicago: University of Chicago Press.
- Valentine, D. E., and Moss, C. F., 1997, Spatially selective auditory responses in the superior colliculus of the echolocating bat, *J. Neurosci.*, 17:1720–1733. ♦
- Yan, J., and Suga, N., 1996, Corticofugal modulation of time-domain processing of biosonar information in bats, *Science*, 273:1100–1103.
- Zhang, Y., Suga, N., and Yan, J., 1997, Corticofugal modulation of frequency processing in bat auditory system, *Nature*, 387:900–903.

EEG and MEG Analysis

Fernando H. Lopes da Silva and Jan Pieter Pijn

Introduction

The electroencephalogram, or EEG, consists of the electrical activity of relatively large neuronal populations that can be recorded from the scalp. Along with the EEG, the magnetic fields generated by these populations can also be recorded as the magnetoencephalogram or MEG using very sensitive transducers. In this

article we discuss these two types of activity jointly, since the same methods of analysis apply to both.

Over the course of time, EEG and MEG have become valuable tools in the diagnosis of functional brain disorders, and indispensable in sleep and epilepsy research. The related study of EVENT-RELATED POTENTIALS (q.v.) became essential for studies of brain function in neurology and pathopsychology. This research field re-

ceived a strong impetus with the development of whole-head MEG systems, since the latter produce data that are less ambiguous in their interpretation. In contrast to EEG, MEG does not need a reference point, and MEG data are more readily modeled in terms of localization of brain sources. These sources of EEG and MEG reflect the dynamics of populations of neurons that have the capacity to work in synchrony. Current understanding emphasizes that synchronous activity in neuronal populations is coupled to the emergence of “local field potentials” (LFPs), which may display oscillations over a wide range of frequencies (Salinas and Sejnowski, 2001).

In general, the same brain sources account for EEG and MEG, with the reservation that the orientation of the active neuronal populations with respect to the cortical surface affects these two modalities differently. MEG signals reflect magnetic fields perpendicular to the skull that are caused by tangential current dipolar fields, whereas EEG signals reflect both radial and tangential fields. This property can be used advantageously to disentangle radial sources lying in the convexity of cortical gyri from tangential sources lying in the sulci.

Over the past few decades, the tools available for functional studies of the brain have been enriched with brain imaging modalities that measure changes in hemodynamics and/or in brain metabolism (e.g., positron emission tomography [PET], functional magnetic resonance imaging [fMRI]). Although these imaging methods have a higher spatial resolution than EEG or MEG, the former have an insurmountable time resolution problem. By contrast, EEG and MEG signals can follow the dynamics of brain activities on a time scale of milliseconds, which corresponds well to the time span for cognitive processing.

A constant preoccupation of EEG research has been to develop techniques to extract information from EEG/MEG signals, recorded at the scalp, that may be relevant for the study of brain functional states and disorders. To this end, a large number of quantitative analytic methods applicable to EEG/MEG have been developed. In general, one can analyze EEG/MEG in time, in space, or both together (spatiotemporal analysis). Recently, the mathematical theory of dynamical nonlinear systems has started to influence the field of brain sciences, in particular by providing a framework that can lead to a better understanding of the dynamics of EEG/MEG signals in relation to brain functions.

In this article we briefly discuss the main aspects of EEG/MEG analysis, considering first, EEG/MEG as a spatiotemporal time series, second, EEG/MEG in terms of the corresponding brain sources, and third, EEG/MEG as a signal that provides information about the state of complex neuronal networks considered as nonlinear dynamical systems (reviewed in Nunez, 1995).

Analysis of EEG/MEG Signals as Spatiotemporal Signals

EEG/MEG signals are complex *spatiotemporal signals*, the statistical properties of which depend on the state of the subject and on external factors. Even when the subject’s behavioral state is almost constant, the duration of epochs that have the same statistical properties (i.e., that are *stationary*) is limited. Therefore, EEG/MEG signals present essential *nonstationary* properties. According to the interest of the researcher, emphasis may be placed on analysis of EEG/MEG signals during steady states or on the detection and characterization of transients, such as the paroxysmal patterns that commonly occur in epileptic patients, or alterations of the basic rhythmic activity that occur during changes of the state of alertness (Figure 1). A special type of EEG/MEG transients is formed by event-related, or event-evoked, potentials (see EVENT-RELATED POTENTIALS).

In an analysis of ongoing EEG/MEG activity, it is customary to subdivide EEG/MEG signals into quasi-stationary epochs and to characterize them by a number of statistical parameters, such as probability distributions, correlation functions and frequency, or power spectra. EEG/MEG time series often present a certain degree of *interdependence*. *Correlation functions* have commonly been used to analyze this property. Here we must distinguish two main questions: the determination of whether within one EEG/MEG signal, a dependency between successive time samples exists, such as in the case of brain rhythmic activity, and the determination of the degree of relationship between two or more EEG/MEG signals. The former can be approached by computing the time average of the product of the signal and a replica of itself shifted by a given time interval—i.e., the *autocorrelation function* (for a more extensive and formal description, see Lopes da Silva, 1999). An important property of the autocorrelation function is that its Fourier transform is the power density spectrum, or simply the *power spectrum*. It gives the distribution of the (squared) amplitude of different frequency components. In general, power spectra represent steady-state variables, but there are also interesting applications to dynamic changes that may reveal how neuronal processes are correlated with specific behaviors. In the latter case, changes in ongoing EEG/MEG activity within given frequency bands, induced by some event (e.g., a sensory stimulus, a cognitive task, or a movement), can be adduced in evidence. Such changes may be characterized by a reduction in EEG/MEG power within a frequency band, usually called event-related desynchronization (ERD), or, in the opposite case, by an increase in power, or event-related synchronization (ERS) (Pfurtscheller and Lopes da Silva, 1999). This form of analysis has been applied to the assessment of cortical areas involved in different behaviors, with interesting results. An example is the planning of specific movements (Figure 2). This example shows that in preparation for a finger movement, the rhythmic activity at 10–12 Hz (i.e., the mu rhythm of the central region) desynchronizes, while a burst of gamma (36–40 Hz) oscillations emerges. The former reflects the focused arousal preparation of the movement, while the latter corresponds to the increase in coupling between clusters of neurons, oscillating within the gamma frequency range, that likely mediates the coordination of the movement. This may be the LFP that corresponds to the neuronal population vector encoding for particular finger movements (Georgopoulos et al., 1999). The postmovement ERS in the beta frequency range probably reflects a reset phenomenon of the dynamical state of the cortical networks engaged in this behavior.

The degree of relationship between two EEG/MEG signals can be estimated using the *cross-correlation function*. Similar to autocorrelation, cross-correlation is the time average of the product of two signals as a function of the time delay between both. The Fourier transform of this function yields the *cross-power spectrum*. The latter is a complex quantity that has magnitude and phase. To quantify the degree of relationship between pairs of EEG/MEG signals as a function of frequency, the magnitude of the cross-power spectrum is usually normalized by dividing it by the value of the autospectra at that frequency of the corresponding signals. This yields a normalized quantity called the *coherence function*. Coherence functions can be used to estimate the degree of the relationship, in the frequency domain, between pairs of EEG/MEG signals.

The counterpart of the coherence function is the *phase function*, which provides information about the time relationship between two signals as a function of frequency. The computation of phase functions has been used to estimate time delays between EEG/MEG signals, in order to obtain evidence for the propagation of EEG/MEG signals in the brain.

As indicated earlier, the correlated activity in neuronal populations (which may display oscillations over a wide range of fre-

Awake: low voltage-random, fast



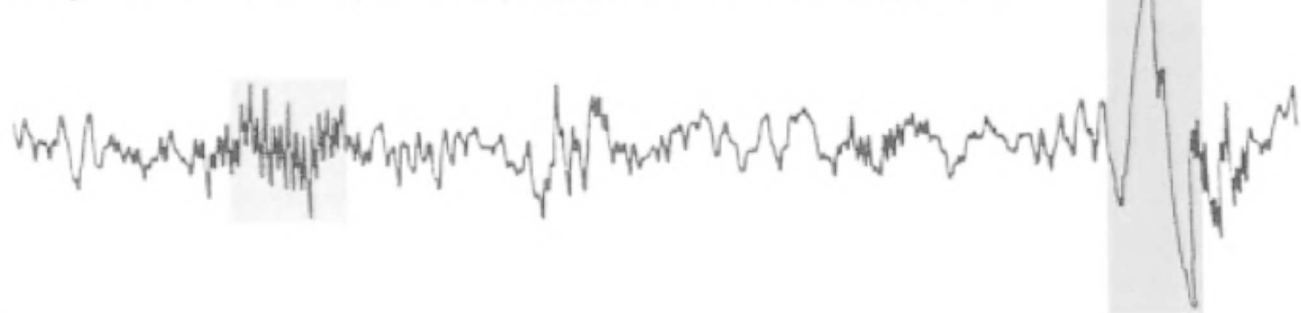
Drowsy: 8 to 12 cps- alpha waves



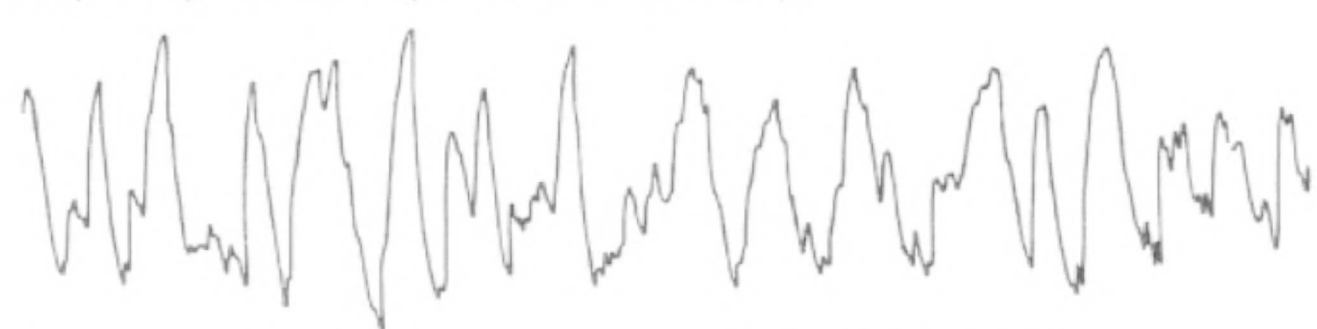
Stage 1: 3 to 7 cps- theta waves



Stage 2: 12 to 14 cps- sleep spindles and K complexes



Deep sleep: 1/2 to 2 cps- delta waves >75 μV



REM sleep: low voltage-random, fast with sawtooth waves



Figure 1. EEG/MEGs recorded during different behavioral stages: awake, drowsy, and sleep stages from superficial sleep (stages 1 and 2) to deep or slow-wave sleep. The last trace shows the EEG during rapid-eye-movement sleep (REM). Note that the awake state is characterized by a mixed pattern, dominated by low-voltage, fast-amplitude (beta/gamma) activities. In the relaxed, drowsy state with the eyes closed, the trace is dominated by oscillations within the alpha frequency range (8–12 Hz). As the subject falls asleep, systematic changes occur: deepening of drowsiness is associated with an increase in slow activity, with occasional bursts of 3–7 Hz (theta)

waves (shaded area). As sleep deepens, sleep spindles, i.e., bursts of 11 or 12–14 Hz oscillations, and K-complexes (shaded areas), which are evoked responses to afferent stimuli, are the dominant features. During REM sleep the EEG is difficult to distinguish from that of the awake state: it is characterized by low-voltage polyrhythmic activity with occasional “sawtoothed waves” in the 2–6 Hz range (shaded area) that may occur in conjunction with ocular movements (Modified with permission from Zigmond et al., *Fundamental Neuroscience*, San Diego: Academic Press, 1999.)

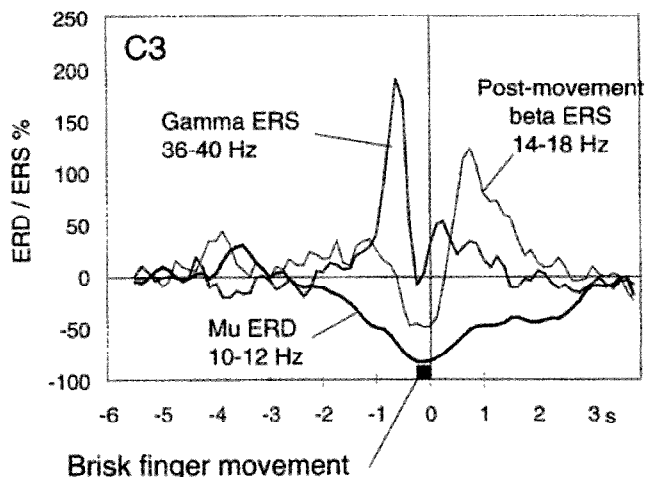


Figure 2. Running power spectra computed for three frequency bands of an EEG recording from electrode site C3, referenced to the left mastoid, and covering the rolandic cortical area, during a brisk right index finger movement. The latter lasted about 240 ms (black bar). The analysis is triggered with respect to movement offset and shows the average of 40 trials. ERD means event-related desynchronization: it consists of a significant power decrease with respect to baseline; ERS indicates the inverse: increase of power or of synchrony. Note that the rhythm in the alpha frequency range (10–12 Hz), called the mu rhythm of the central region, displays ERD starting about 2 s before the movement, while a burst of ERS of gamma frequency components (35–40 Hz) reaches a maximum just before the movement. After movement offset, a rebound ERS in the beta range (14–18 Hz) appears. (Modified with permission from Pfurtscheller, G., and Lopes da Silva, F. H., 1999, Event-related EEG/EMG synchronization and desynchronization: Basic principles, *Clin. Neurophysiol.*, 110:1842–1857.)

quencies as revealed in the EEG and MEG) forms the substrate for a variety of brain functions and associated cognitive processes. Namely transient periods of phase synchrony of oscillating neuronal discharges, particularly in the frequency range of 30–80 Hz (gamma oscillations), appear to act as an integrative mechanism that brings a widely distributed set of neurons together into a coherent ensemble that underlies a cognitive process or the preparation of a movement, as shown in the example of Figure 2. This implies that mechanisms of synchronization at various levels of brain organization, from individual pairs of neurons to LFPs to the much larger scale of EEG and MEG signals recorded from the scalp, are necessary for the coordination of neural activities distributed over distinct brain areas (Varela et al., 2001). Because phase synchrony in EEG/MEG signals must be detected and measured in order to investigate neurocognitive processes, ways of measuring synchrony are of increasing interest. This has led to the development of analytical tools that allow the phase component to be obtained separately from the amplitude component for a given frequency range (Le Van Quyen et al., 2001). With this methodology it was shown, for example, that the scalp EEG/MEG of subjects performing the perceptive task of recognizing human faces induces a long-distance pattern of phase synchronization that represents active coupling of the underlying neural populations. This coupling appears to be necessary for the realization of this cognitive task (Rodriguez et al., 1999).

Estimation of Brain Sources from Scalp EEG/MEG Recordings

A most fundamental aspect of EEG/MEG analysis is to be able to estimate from multiple scalp EEG/MEG recordings the distribution

of the corresponding sources within the brain. This implies solving the so-called *inverse problem* of volume conduction theory, i.e., determining the location within the brain tissue of the sources of electrical activity, taking into consideration the properties of both the brain and the conductive media surrounding the brain. This problem has no unique solution: it is not possible to determine a unique current source distribution in a volume conductor from measurements taken at the conductor surface. However, it is possible to solve this problem by putting constraints on the current source distributions, i.e., by defining a specific model of the source. A commonly assumed source model is the equivalent current dipole that represents an active patch of cortex. Therefore, two models are required, one of the source and another one of the volume conductor. Of course, it is necessary to have a sufficient number of measurement points at the scalp in order to obtain a satisfactory solution.

We consider, briefly, the main problems posed by the source and the volume conductor models. The problem of estimating the source is a nonlinear problem that has to be solved iteratively. De Munck (1990) has proposed a general approach that takes into account both the time functions of the activity of the sources and the corresponding spatial properties (positions and orientations). Regarding the volume conductor models, the most commonly used model is a set of concentric spheres that represent, from inside to outside, the brain, the cerebrospinal fluid (in some cases), the skull, and the scalp. However, the head deviates appreciably from a sphere. With the increasing availability of MRI, it is now possible to reconstruct the different head compartments in a more realistic way. This is important, since several simulation studies have shown that deviations from a realistic shape of the head can significantly influence the magnetic fields and the potential distributions (Hämäläinen et al., 1993).

In most applications of EEG/MEG spatial analysis, either in neurology or in psychophysiology, one does not attempt an estimation of the brain sources using the inverse approach, because of the inherent difficulties of this method and the uncertainties of the estimated localizations. Most researchers are satisfied with representing the sets of multiple EEG/MEG signals projected as a map at the surface of the scalp. This is called electric or magnetic *brain mapping*.

EEG/MEG Signals as Expressions of Dynamical Nonlinear Systems

An overview of the basic concepts of nonlinear dynamics can be found in CHAOS IN NEURAL SYSTEMS. Here we briefly state that EEG/MEG signals may correspond to multiple kinds of dynamical states that are characterized by the corresponding attractors, depending on the network's parameters and input conditions. Thus, bifurcations between different modes of dynamics of the same network can take place. This nonlinear dynamical behavior can account for the fact that an epileptic seizure, with a typical paroxysmal EEG/MEG pattern, can emerge suddenly from an apparently normal state characterized by resting EEG/MEG activity.

Quantitative measures for identifying complex nonlinear dynamics have been applied to several kinds of EEG/MEG signals. The *correlation dimension* (D_2) (Grassberger and Procaccia, 1983) has been the most widely used. To interpret estimates of D_2 , it is important to use *surrogate* signals (controls) obtained by randomizing the phase of the original EEG/MEG signals. The latter should yield a large D_2 value, since the transformed signal is not distinguishable from Gaussian noise (Theiler, 1990; Pijn et al., 1991). In most cases of ongoing EEG/MEGs, including those recorded during sleep, the difference between the real signals and the surrogate signals is very small, such that the hypothesis that the EEG/MEG signals are generated by a deterministic chaotic process cannot be supported in

these cases. However, in the case of EEG/MEG signals recorded during epileptic seizures, the value of D_2 was shown to be much smaller than that of the corresponding surrogate signals. This finding led several groups to explore whether this kind of methodology could be used to unravel changes in EEG/MEG dynamics that may take place *before* an epileptic seizure becomes manifest (Elger and Lehnertz, 1998; Martinerie et al., 1998). Similar results were also obtained from the spatiotemporal evolution of other nonlinear parameters (reviewed by Lehnertz et al., 2001). Taken together, these results indicate that the dynamical properties of the interictal, preictal, ictal, and postictal states are clearly different and have different attractors. This opens the exciting possibility of using these nonlinear analysis methods to predict the occurrence of impending epileptic seizures in clinical practice, and of possibly avoiding their occurrence.

In conclusion, EEG/MEG recordings are complex signals that may provide valuable information about the underlying brain systems, since they have unsurpassed resolution in time, although their spatial resolution is rather limited. Therefore, mapping cortical activity using EEG/MEG signals combined with realistic models of the brain, extracted from MRI scans, may yield new possibilities for functional imaging of dynamical brain states.

Road Map: Cognitive Neuroscience

Related Reading: Brain Signal Analysis; Event-Related Potentials; Hippocampal Rhythm Generation

References

- De Munck, J., 1990, The estimation of time-varying dipoles on the basis of evoked potentials, *Electroencephalogr. Clin. Neurophysiol.*, 77:156–160.
- Elger, C. E., and Lehnertz, K., 1998, Seizure prediction by non-linear time series analysis of brain electrical activity, *Eur. J. Neurosci.*, 10:786–789.
- Georgopoulos, A. P., Pellizzer, G., Poliakov, A. V., and Schieber, M. H., 1999, Neural coding of finger and wrist movements, *J. Comput. Neurosci.*, 6:279–288.
- Grassberger, P., and Procaccia, I., 1983, Measuring the strangeness of strange attractors, *Physica*, 9:183–208.
- Hämäläinen, M., Hari, R., Ilmoniemi, R., Knuutila, J., and Lounasmaa, O. V., 1993, Magnetoencephalography: Theory, instrumentation, and applications to noninvasive studies of the working human brain, *Rev. Mod. Phys.*, 65:413–497.
- Lehnertz, K., Andrzejak, R. G., Arnhold, J., Kreuz, T., Mormann, F., Rieke, C., Widman, G., and Elger, C. E., 2001, Nonlinear EEG analysis in epilepsy: Its possible use for interictal focus localization, seizure anticipation, and prevention, *J. Clin. Neurophysiol.*, 18:209–222.
- Le Van Quyen, M., Foucher, J., Lachaux, J., Rodriguez, E., Lutz, A., Martinerie, J., and Varela, F. J., 2001, Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony, *J. Neurosci. Methods*, 111:83–98.
- Lopes da Silva, F. H., 1999, EEG analysis: Theory and practice, in *Electroencephalography: Basic Principles, Clinical Applications and Related Fields* (E. Niedermeyer and F. H. Lopes da Silva, Eds.), 4th ed., Baltimore: Williams and Wilkins, pp. 1135–1163. ♦
- Martinierie, J., Adam, C., Le Van Quyen, M., Baulac, M., Clemenceau, S., Renault, B., and Varela, F., 1998, Epileptic seizures can be anticipated by non-linear analysis, *Nature Med.*, 4:1173–1176.
- Nunez, P. L., 1995, *Neocortical Dynamics and Human EEG Rhythms*, New York: Oxford University Press. ♦
- Pfurtscheller, G., and Lopes da Silva, F. H., 1999, Event-related EEG/MEG synchronization and desynchronization: Basic principles, *Clin. Neurophysiol.*, 110:1842–1857.
- Pijn, J. P. M., van Nerveen, J., Noest, A., and Lopes da Silva, F. H., 1991, Chaos or noise in EEG signals: Dependence on state and brain site, *Electroencephalogr. Clin. Neurophysiol.*, 79:371–381.
- Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B., and Varela, F. J., 1999, Perception's shadow: Long-distance synchronization of human brain activity, *Nature*, 397:430–433.
- Salinas, E., and Sejnowski, T. J., 2001, Correlated neuronal activity and the flow of neural information, *Nature Rev. Neurosci.*, 2:539–550. ♦
- Theiler, J., 1990, Estimating fractal dimension, *J. Opt. Soc. Am. A*, 7:1055–1073.
- Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J., 2001, The brainweb: Phase synchronization and large-scale integration, *Nature Rev. Neurosci.*, 2:229–239.

Electrolocation

Joseph Bastian

Introduction

The electrosensory systems of weakly electric fishes are widely recognized as attractive model systems for studies of the neural bases of behavior. These animals are specialists, relying heavily on information acquired via this unique sensory system, and the importance of electroreception in the animals' life history is reflected in the hypertrophy of brain regions devoted to processing electrosensory information. Animals with active electrosensory systems generate an electric field around their body by means of an electric organ located in the trunk and tail, and measure this field via electroreceptors embedded in the skin. Distortions of the electric field due to animate or inanimate targets in the environment or signals generated by other fishes provide inputs to the system, and several distinct behaviors can be linked to simple patterns of electrosensory input. The ease with which behaviorally relevant stimuli can be identified and simulated, and the wealth of background anatomical and physiological information, are major advantages to studying this system.

Since publication of the first edition of the *Handbook*, advances have been made in understanding the electric organ discharge field of weakly electric fishes and its interaction with targets in the en-

vironment (Assad, Rasnow, and Stoddard, 1999). Studies of feeding behavior revealed the motor strategies used during prey capture and have determined detection limits for the system (Nelson and MacIver, 1999). A novel algorithm for determining fish-target distance (depth perception), using information acquired from a single two-dimensional (2D) receptor array, has also been discovered (von der Emde, 1999). Properties of the epidermal electroreceptors have been further defined; numerical models that accurately predict their responses are now available (Nelson, Xu, and Payne, 1997), and analyses of their information-encoding capabilities (see RATE CODING AND SIGNAL PROCESSING) have been completed (Wessel, Koch, and Gabbiani, 1996; Gabbiani and Metzner, 1999). Studies of the primary electrosensory processing region, the electrosensory lateral line lobe (ELL), have revealed how the interaction between ascending sensory information and that descending from higher centers contributes to gain control, receptive field organization, and attentional mechanisms (Berman and Maler, 1999). In addition, these descending inputs participate in an adaptive filtering mechanism, enabling the system to reject predictable patterns of input while preserving sensitivity to novel stimuli (Bastian, 1999; Bell et al., 1997, 1999; Bodznick, Montgomery, and Carey, 1999).

This review focuses on recent progress made in understanding electrolocation behavior, and on the neural implementation of an adaptive filter that attenuates electrosensory stimuli resulting from the fish's own movements. Movement-caused signals can exceed those caused by small prey by more than two orders of magnitude; hence, without a reduction in this "noise," the animal may be unable to detect food efficiently. Reviews of recent progress made in areas ranging from electric fish taxonomy and behavior through the physiology and molecular biology of important neural circuits can be found in Turner, Maler, and Burrows (1999).

Overview of Electoreception

Figure 1A illustrates a South American electric fish, *Apteronotus leptorhynchus*. *Apteronotus* produces a quasi-sinusoidal electric organ discharge, or EOD, that ranges in frequency from about 600 to 1,000 Hz, depending on the individual. The bold arrows in Figure 1A show the pattern of current flow due to the EOD, which results in a pattern of transdermal potential changes. Electrorceptors specialized to measure the amplitude and timing of this potential are found at high densities within the skin. Objects in the environment having an impedance different from that of the water distort the pattern of EOD current, resulting in a localized change in the transdermal potential or electric image. As the animal moves relative to an object, an EOD amplitude modulation (EOD AM) is produced that is encoded as a change in the firing pattern of electrorceptor afferents (Figure 1B). The electrorceptor afferents convey this information to the first processing region in the brain, the ELL. The pyramidal cells (Figure 1C) are the principal efferent neurons of the ELL; they receive massive descending or feedback inputs as well as receptor afferent input. Insofar as many of the higher-order neurons providing these feedback signals receive inputs from the pyramidal cells themselves, the ELL is positioned within an electrosensory processing loop, with the result that the functional characteristics of the pyramidal cells are subject to continuous feedback control.

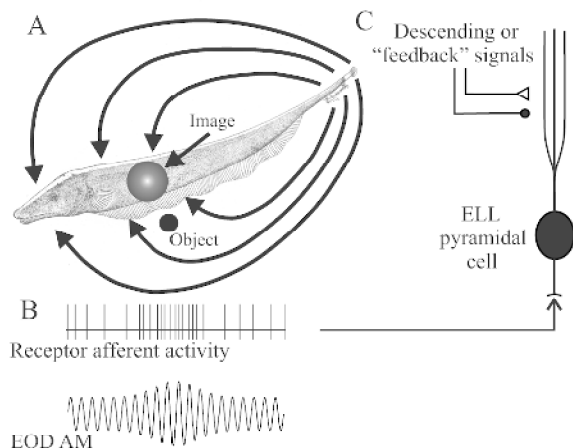


Figure 1. A, Pattern of current flow due to the electric organ discharge in *Apteronotus leptorhynchus*. The circle labeled "image" illustrates the pattern of electric organ discharge amplitude modulation (EOD AM) at the animal's body surface as a result of the presence of an object near the fish. B, The pattern of EOD AM expected to occur as the animal moves relative to the object, and the resulting change in electrorceptor afferent activity. C, Electrosensory lateral line lobe (ELL) pyramidal cell. These "principal cells" of the first-order central electrosensory processing station receive receptor afferent inputs along with feedback signals from higher centers.

Electric Field Measurements and Electrolocation Behavior

Most of the early studies of electrolocation focused on the physiological properties of the receptor afferents and ELL pyramidal cells. Of necessity, these physiological studies used immobilized animals and relatively simple stimulus patterns. Although such studies provided important descriptions of electrosensory system function under well-defined conditions, they were also limited. Natural stimulus patterns such as living prey were not used, and, more important, the animals were not able to engage in normal exploratory behaviors. Recently, high-resolution measurements of the EODs of several species have been completed, and techniques for accurate 3D modeling of the EOD field and various object-caused EOD distortions have become available. These advances have allowed accurate predictions to be made of the spatiotemporal patterns of electrosensory input that occur as a fish encounters targets in its environment. Such studies have also suggested sensory processing algorithms that the animals might use to accomplish critical tasks (Rasnow, 1996; Assad et al., 1999). For example, prey capture based on electrosensory cues requires that the 3D location of the prey relative to the fish be determined. Because the prey or electrolocation target "casts" an electric image on the skin surface (see Figure 1A), the 2D position of a target relative to the animal's body surface can be determined from the pattern of activity over the 2D array of electrorceptors. However, estimation of the distance of an object lateral to the fish, or depth perception, is much more difficult. Unlike visual systems, in which depth information can be extracted from paired sensors that view the target from slightly different positions, electrosensory systems probably gauge distance using information from a single receptor array. The peak change in amplitude of the transdermal potential caused by a given object decreases rapidly with increasing lateral distance; hence, this cue could provide information related to target distance. However, the peak amplitude is also a function of object size, resulting in size-distance ambiguities: larger objects at greater distances can generate the same peak potential change as smaller objects closer to the fish. Empirical and modeling studies suggest that distance estimation could be achieved by simultaneous evaluation of multiple image characteristics. It was specifically proposed that measuring the *relative* dimensions of electrical images (e.g., the ratio of electric image width to its peak amplitude) would allow unambiguous distance determination (Rasnow, 1996; Assad et al., 1999).

A series of elegant behavioral experiments demonstrated that another weakly electric fish, *Gnathonemus petersii*, does determine electrolocation target distance using information received over a single 2D receptor array by simultaneously evaluating two electric image characteristics (von der Emde, 1999). Fishes were trained to swim through one of two openings in a partition dividing an experimental tank contingent on the relative distances of electrolocation targets positioned behind the openings. Fishes had to choose the opening associated with the object further inside the partition. Not only could the fishes accurately determine the relative distances to identical objects, but, once trained, they could generalize, correctly discriminating relative distance even when objects of different sizes were presented. However, for this species, the maximum slope of the electrical image's spatial profile in combination with peak image amplitude was found to be the best combination of parameters. Von der Emde also found that, at a given distance, spherical objects always produced significantly smaller slopes relative to amplitude than did cubes of the same size. This observation led to the prediction that the animals could be fooled and should misjudge distances when spheres and cubes were compared. This prediction was fulfilled, and the fishes interpreted spherical objects as being farther away than cubes, even when the spheres were as much as 1.5 cm closer to the animal. Such responses to the pre-

dicted illusory position of spheres provide compelling evidence in support of the proposed mechanism for distance determination. Furthermore, identification of the sensory processing algorithm used for a given perceptual task provides physiologists with important clues that should greatly facilitate the search for neural correlates of this behavior.

Although it has long been assumed that weakly electric fishes rely heavily on the electrosensory system for prey capture, detailed descriptions of feeding behavior have only recently appeared. Nelson and MacIver (1999) used a video tracking and reconstruction technique to analyze the behavior of two related fishes, *Apteronotus albifrons* and *A. leptorhynchus*, as they fed on small aquatic crustaceans. Stereotyped changes in swimming direction and velocity were found to be associated with prey detection, and, based on the timing of these behavioral landmarks, prey distance at the time of detection was determined. Additionally, given the relationships between electrolocation target size and distance from a fish and the resulting electric image characteristics provided by Rasnow (1996), Nelson and MacIver reconstructed the temporal sequence of electrical images experienced by the fish during prey capture. Finally, given the sequence of electric images as input to a numerical model of electroreceptor afferent responsiveness, Nelson and MacIver arrived at predictions of the spatiotemporal patterns of electroreceptor afferent activity during prey capture. At a typical detection distance of 1 cm, the fishes would experience a peak electric image amplitude of approximately 3 μV , or about 0.3% of the transdermal potential. This translates into a change in receptor afferent firing of about 2.5%, or 7 spikes/s. More recent results also showed that maximum detection distance is dependent on water conductivity, which lends additional support to the idea that the animal's behavior is guided by electrosensory information instead of other cues, such as water turbulence caused by the prey (MacIver, Sharabash, and Nelson, 2001).

These fishes also often change posture during swimming and prey capture; the body is frequently bent into an arc-like posture like that shown in Figure 2. Because the electric organ is located in the trunk and tail, such changes in posture result in large transdermal potential modulations that cause changes in electroreceptor afferent activity 10- to 100-fold greater than changes due to a small prey. These large electrosensory signals resulting from locomotor activity are obviously problematic; they could easily mask small signals due to the presence of prey. However, an adaptive filter operating at the level of the ELL removes the predictable signals related to locomotion while preserving sensitivity to novel signals such as those generated by prey.

Adaptive Plasticity in the Electrosensory Systems

Movements of the trunk and tail of *A. leptorhynchus* through an arc of less than ± 20 degrees, which mimic changes in posture that commonly occur during swimming, cause transdermal potential changes in excess of 100 μV . It has long been recognized that mechanisms must exist that enable the animals to differentiate between such movement-related or reafferent inputs and prey-related signals. Studies in several lower vertebrate species have identified a general mechanism by which reafferent or other predictable and potentially disruptive stimulus patterns can be canceled without compromising sensitivity to relevant stimuli. Importantly, this cancellation process is adaptive; that is, the network mediating the cancellation learns, enabling the system to filter out reafferent inputs that change or evolve over time.

The operation of the ELL pyramidal cell filter is shown diagrammatically in Figure 2. During electrolocation, the fishes often make bending movements of the trunk and tail as they approach and orient to targets (Figure 2A). Consequently, the input to the receptor

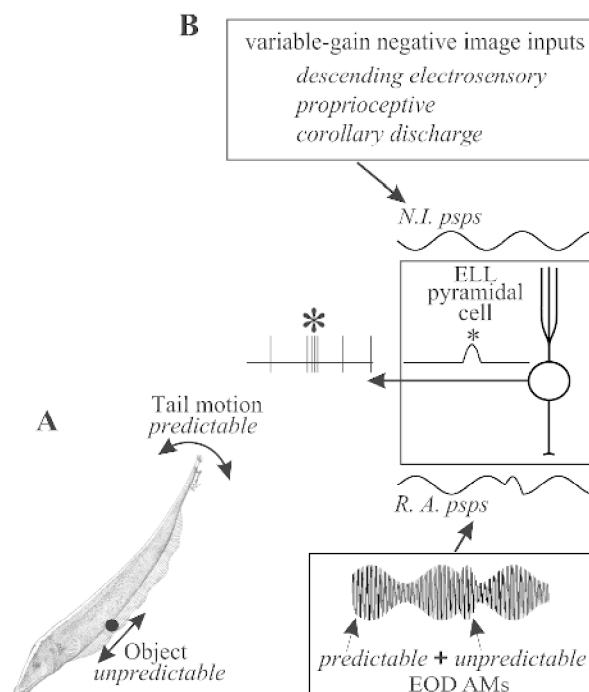


Figure 2. Adaptive filtering by ELL pyramidal cells. *A*, Changes in the fish's posture result in reafferent patterns of afferent input that are *predictable* and can be differentiated from *unpredictable* inputs due to the presence of an object. *B*, ELL pyramidal cells integrate receptor afferent psp's (*R.A. psp's*) with negative image inputs (*N.I. psp's*), with the result that predictable afferent inputs are attenuated while unpredictable inputs are encoded without interference (*).

afferents consists of predictable, often cyclical EOD AMs due to changes in posture as well as unpredictable AMs due to the object. This AM pattern, shown by the lowest waveform in Figure 2B, leads to modulation of receptor afferent firing, with the possibility that responses to the cyclical postural changes mask responses to the object. The pattern of receptor afferent synaptic input received by the pyramidal cell is illustrated by the lower waveform, labeled *R.A. psp's*. Descending electrosensory, proprioceptive, and corollary discharges of motor commands are received at the pyramidal cell's apical dendrites, and this constellation of excitatory and inhibitory inputs provides a signal that is approximately the inverse of the expected or predictable component of the receptor afferent input. Pyramidal cells sum this negative image input, *N.I. psp's*, with the *R.A. psp's*, canceling the predictable component of the afference while preserving sensitivity to the unpredictable stimulus (Figure 2B*).

Very similar mechanisms have been described for four different groups of fishes; elasmobranchs, mormyrid weakly electric fish, gymnotid weakly electric fish, and a nonelectric teleost (Bastian, 1999; Bell et al., 1997, 1999; Bodznick et al., 1999). In all cases the neural networks underlying the process are components of the octavolateral system; they process either electroreceptor or normal lateral line inputs and have a cerebellum-like organization. In each case a population of principal cells (e.g., the gymnotid ELL pyramidal cells) receives receptor afferent inputs as well as inputs from large numbers of cerebellar-like parallel fibers. The parallel fibers provide the predictive or negative image inputs. The adaptive characteristic of the cancellation is due to an anti-Hebbian form of

synaptic plasticity (see HEBBIAN SYNAPTIC PLASTICITY) at the parallel fiber to principal cell synapse, and a simple set of learning rules can account for the circuit's behavior (reviewed in Bodznick et al., 1999). First, coincident activity of parallel fibers and principal cells leads to a reduction in the strength of active excitatory dendritic synapses. Second, parallel fiber inputs active at times when the postsynaptic cell is inactive lead to increased excitatory synaptic strength. These rules are anti-Hebbian: rather than coincident pre- and postsynaptic activity leading to increased synaptic strength, the opposite occurs. Although these two rules governing plasticity at excitatory synapses are sufficient to account for the adjustment of negative image inputs as described in Figure 2, it is very likely that the strength of inhibitory synapses can also be adjusted, and a complementary pair of rules should govern inhibitory plasticity (Bodznick et al., 1999).

Parallel studies of the cancellation mechanism in these different species have verified that plasticity is associated with parallel fiber to apical dendritic synapses (reviewed in Bastian, 1999; Bell et al., 1999; Bodznick et al., 1999). The plasticity can be blocked by glutamate antagonists applied at these synapses, and it can be evoked by pairing direct electrical stimulation of the parallel fibers with depolarization of the principal cells via intracellular current injection. The depression of excitatory synaptic strength that results is similar to the long-term depression (see CEREBELLUM: NEURAL PLASTICITY) that occurs at cerebellar parallel fiber to Purkinje cell synapses. As in the cerebellum, the depression requires a postsynaptic Ca^{2+} influx; however, unlike in the cerebellum, the depression can be blocked by NMDA receptor antagonists, suggesting that the Ca^{2+} influx may occur via activation of these channels (see NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK MODELS). Second-messenger systems operating within the ELL pyramidal cells have also been implicated, and it has been suggested that protein kinase A (PKA) and Ca^{2+} /calmodulin-dependent kinase (CaMK2 β) may be involved in pre- and postsynaptic mechanisms underlying the anti-Hebbian plasticity (see Berman and Maler, 1999). Most recently, Han, Grant, and Bell (2000), using an in vitro preparation of the mormyrid ELL, demonstrated that the anti-Hebbian depression at excitatory synapses occurs at the same locus as a nonassociative potentiation. Having both depression and potentiation operative at the same locus has long been recognized as critical to ensuring true reversibility of plastic changes and preventing saturation of potentiation or depression mechanisms. This is the first demonstration of potentiation and depression occurring at the same locus in a cerebellum-like structure.

Discussion

Although electrosensory organisms are highly specialized, they face the same problems and constraints as most other animals do. They must find prey, avoid predators, communicate with conspecifics, and reproduce. Although many of these critical behaviors rely heavily on information acquired via this single sensory system, the properties of the neural circuits involved are likely to reflect general principles operating in many other systems. For example, a widespread but poorly understood characteristic of sensory processing circuits is the presence of massive descending or feedback connections by which higher centers presumably modulate the operation of lower centers. Studies of the ELL of

these fishes have shown not only that fundamental properties such as response gain and receptive field organization are controlled by these descending connections, but also that elegant adaptive filtering mechanisms exist that enable the rejection of stimuli that otherwise might mask critical functions. Understanding the cellular mechanisms underlying the synaptic plasticity that forms the basis for this filter should contribute to our understanding of closely related neural circuits such as those found in the cochlear nuclei and the cerebellum. In addition, defining general principles of operation, such as the use of stored sensory expectations for the cancellation or perhaps the identification of specific input patterns, may lead to increased understanding of more diverse neural circuits.

Road Maps: Neuroethology and Evolution; Other Sensory Systems

Related Reading: Auditory Periphery and Cochlear Nucleus; Echolocation: Cochleotopic and Computational Maps; Sound Localization and Binaural Processing

References

- Assad, C., Rasnow, B., and Stoddard, P. K., 1999, Electric organ discharges and electric images during electrolocation, *J. Exp. Biol.*, 202:1185–1193.
- Bastian, J., 1999, Plasticity of feedback inputs in the apteronotid electrosensory system, *J. Exp. Biol.*, 202:1327–1337. ♦
- Bell, C. C., Bodznick, D., Montgomery, J., and Bastian, J., 1997, The generation and subtraction of sensory expectations within cerebellum-like structures, *Brain Behav. Evol.*, 50(suppl. 1):171–178. ♦
- Bell, C. C., Han, V. Z., Sugawara, Y., and Grant, K., 1999, Synaptic plasticity in the mormyrid electrosensory lobe, *J. Exp. Biol.*, 202:1339–1347.
- Berman, N. J., and Maler, L., 1999, Neural architecture of the electrosensory lateral line lobe: Adaptations for coincidence detection, a sensory searchlight and frequency-dependent adaptive filtering, *J. Exp. Biol.*, 202:1243–1253.
- Bodznick, D., Montgomery, J. C., and Carey, M., 1999, Adaptive mechanisms in the elasmobranch hindbrain, *J. Exp. Biol.*, 202:1357–1364. ♦
- Gabbiani, F., and Metzner, W., 1999, Encoding and processing of sensory information in neuronal spike trains, *J. Exp. Biol.*, 202:1267–1279. ♦
- Han, V. Z., Grant, K., and Bell, C. C., 2000, Reversible associative depression and nonassociative potentiation at a parallel fiber synapse, *Neuron*, 27:611–622.
- MacIver, M. A., Sharabash, N. M., and Nelson, M. E., 2001, Prey-capture behavior in gymnotid electric fish: Motion analysis and effects of water conductivity, *J. Exp. Biol.*, 204:543–557.
- Nelson, M. E., and MacIver, M. A., 1999, Prey capture in the weakly electric fish *Apteronotus albifrons*: Sensory acquisition strategies and electrosensory consequences, *J. Exp. Biol.*, 202:1195–1203. ♦
- Nelson, M. E., Xu, Z., and Payne, J. R., 1997, Characterization and modeling of P-type electrosensory afferent responses to amplitude modulations in a wave-type electric fish, *J. Comp. Physiol. A*, 181:532–544.
- Rasnow, B., 1996, The effects of simple objects on the electric field of *Apteronotus leptorhynchus*, *J. Comp. Physiol. A*, 178:397–411.
- Turner, R. W., Maler, L., and Burrows, M., 1999, *Electroreception and Electrocommunication*, Cambridge, Engl.: Company of Biologists. ♦
- von der Emde, G., 1999, Active electrolocation of objects in weakly electric fish, *J. Exp. Biol.*, 202:1205–1215.
- Wessel, R., Koch, C., and Gabbiani, F., 1996, Coding of time-varying electric field amplitude modulations in a wave-type electric fish, *J. Neurophysiol.*, 75:2280–2293.

Embodied Cognition

Olaf Sporns

Introduction

The central tenet of embodied cognition is that cognitive processes emerge from the interactions between neural, bodily, and environmental factors (Varela, Thompson, and Rosch, 1991). Brain, body, and environment are seen as reciprocally and dynamically coupled, with neural and behavioral processes exerting specific effects across the boundaries of brain and body and over different time scales. Clark (1997) has called this complex interplay “continuous reciprocal causation,” the coupling of distinct subsystems leading to the emergence of qualitatively new structures. Embodied cognition places strong emphasis on perception-action loops, in which internal and external processes are intricately and cyclically interwoven. The distributed nature of neural, bodily, and environmental interactions has led many authors to deemphasize or, in radical formulations, even abandon some fundamental concepts of cognitive science such as internal representations and the computational nature of mind. They claim that embodied systems generate their cognitive power through real-world interaction, not by manipulating an internal world model, organized in a predominantly sequential (sense-think-act) processing scheme. Whether all internal models (symbolic or neurally based) should be abandoned is a matter of much debate. More balanced theories of embodied cognition would place emphasis on the dynamic coupling between brain and body while allowing for the existence and use of internal models in motor control, planning, and linguistic and symbolic behavior.

Not only has embodied cognition sparked much philosophical and foundational discussion within cognitive science, it has also inspired numerous attempts to build embodied systems as working models of cognitive processes. This article briefly examines embodied models that focus on real-time perception-action coupling, developmental processes, the role of sensorimotor activity in category formation, and value systems. Before we discuss examples of embodied models, we briefly examine some of the underlying design principles.

Design Principles of Embodied Models

While embodied cognitive models show great diversity across different task domains and physical instantiations, most of them fall into distinct classes using different sets of design principles. Roughly, these classes comprise connectionist, dynamic, and neurocomputational (synthetic) models.

Most formal symbolic models rely heavily on computation and representation and do not address neural implementations, learning mechanisms, or the coupling between brain and body. Not surprisingly, embodied cognition eschews such models as appropriate means for implementation. Connectionist models seem more appropriate because they attempt to embed cognitive processes in a neural context. Typical connectionist architectures operate by converting encoded inputs into outputs via intermediate layers, with an emphasis on learning as the search for optimal configurations of synaptic weights. At first glance, however, some basic design principles of connectionist networks appear inconsistent with the philosophy of embodied cognition. Connectionist models employ internal representations in the form of distributed activity patterns that encode higher-order statistical properties of inputs. Motor action is usually not an explicit part of the model and thus rarely contributes to the generation or selection of input patterns. Learning is limited to the optimization of synaptic weights and does not include active exploration or real-world interaction. Although these design principles apply to many classical connectionist architec-

tures (consisting of three layers with hidden units and backpropagation), a number of connectionist approaches have evolved that aim at constructing internal models while taking interactions with the real world into account (see *SENSORIMOTOR LEARNING*). In addition, numerous connectionist models make explicit reference to neural architectures and dynamics and employ realistic rules of synaptic change. Such neurocomputational models can be embedded in behaving autonomous systems and yield interesting behavior.

A distinct class of models of embodied cognition is based on concepts of dynamical systems theory (Thelen and Smith, 1994; Kelso, 1995). Here the emphasis is less on the mechanistic implementation of neural structures and more on the interplay between the internal dynamics of the agent and the external dynamics of the body and environment. Some central concepts of dynamical systems theory are those of *attractor* and *phase space*. Any point in phase space corresponds to a particular set of values of the system's state variables (which form the dimensions of the space). Over time, the system goes through a trajectory within this phase space, a set of points that are occupied as time progresses. The temporal evolution of the system's trajectory is described by a set of dynamic equations (usually nonlinear differential equations). An attractor is that portion of the state space that the system's trajectory converges upon over time; attractors can be points (stable steady state), limit cycles (periodic states), quasi-periodic, or chaotic. When this formal approach is applied to modeling embodied cognition, cognitive processes (perceiving, planning, deciding, remembering) are described using the language and tools of dynamical systems theory. This has the advantage of unifying the formal treatment of internal and external processes within a common (dynamical) framework. The basic building blocks of dynamical embodied models are state variables, realized as spatially continuous activation fields and usually defined in a behavioral or task-dependent context (e.g., “movement planning field,” “decision field”). The temporal evolution of these fields is governed by sets of dynamical equations that ultimately determine behavior.

Another class of models attempts to create autonomous and embodied systems by synthetically assembling such systems from simple components. This approach, also called “synthetic modeling,” incorporates explicit mechanisms at all levels and studies their emergent dynamic behavior (Edelman et al., 1992; Pfeifer and Scheier, 1999; Sporns, Almasy, and Edelman, 2000; see also Braatenberg, 1986). In synthetic models, the neural system is defined in terms of a specific physiology and anatomy. When combined with a matching body structure, this approach produces models that often closely resemble particular animal species; examples in the literature include robotic ants, bees, crickets, lobsters, frogs, and primates (see also *NEUROETHOLOGY, COMPUTATIONAL*). Real and simulated neural systems often attain specific (input- and state-dependent) dynamic states as a result of neuronal interactions. Thus, in synthetic models, large-scale dynamics are the emergent product of low-level components (neurons and connections) and their specific structural and functional properties (e.g., spike dynamics, neuronal morphology, excitatory and inhibitory effects, synaptic plasticity, and connectivity patterns). The synthetic approach allows us to use established principles of computational neuroscience and to relate results obtained with embodied models to the empirical findings of neurobiology. However, the computational expense of simulating realistic neural architectures still places serious limits on the size and complexity of such models, given that they must function in real time as part of a behaving creature.

The emphasis on body structure and environment as causal elements in the emergence of organized behavior requires their explicit inclusion in models of embodied cognition. Although some researchers use simulated environments, more often they use actual robots as physical analogues of real organisms. Robot sensors and effectors are interfaced with a computational model (a neural or dynamic simulation). The robot moves or acts within an environment, either an unconstrained real-world setting (office, lab) or an enclosure containing various kinds of objects. When designing embodied models, it is important to include the relevant physical, dynamic, and kinematic properties of the robot body (Beer et al., 1998), as well as to match the sensorimotor capabilities of the robot and the complexity of its task environment (the principle of “ecological balance”; Pfeifer and Scheier, 1999). Finally, embodied cognitive models must function without human intervention and without the use of supervised learning strategies. A truly autonomous system not only must be embodied, it should also be able to seek out and gather its own sensory inputs (“situatedness”) and generate its own experiential and behavioral history. Autonomous systems may exist in a social context. A full account of their behavior and development should include their social interactions and socially mediated processes such as observational or imitation learning.

Real-Time Coupling Between Brain and World

Embodied models of active vision and motor coordination show that several computationally hard problems of information extraction and control are more naturally addressed when the real-time coupling between neural and bodily structures is taken into account. Rather than relying on explicit internal representations or internal world models in computing appropriate outputs, such systems use “the world as its own model.”

For example, for many years researchers in machine vision believed that the purpose of vision is to generate an accurate and comprehensive internal representation of the surrounding three-dimensional (3D) world by extracting information from 2D images. According to this view, given an image, the visual system computes a solution, and ultimately issues appropriate motor commands. Movement implements the outcome of perceptual decision making but does not participate in the perceptual process as a source or generator of useful information. About 10 years ago, several authors, among them Dana Ballard and Ruzena Bajcsy (see review in Clark, 1997), proposed an alternative strategy called active or animate vision. According to this approach, vision is best understood in the context of visual behaviors. Organisms use vision to guide motor action in real time, and motor action serves to seek out sources of perceptual information in the environment, for example by orienting sensory surfaces during gaze control. For example, visuomotor behaviors greatly facilitate efficient sampling of sensory environments by autonomous sensorimotor agents. Active vision simplifies several problems of standard machine vision, such as invariant recognition, by using sensorimotor strategies such as foveation to reduce variance across multiple views of the same object. In addition, visual agents can utilize and continually reference objects in the outside world during the generation of behavior instead of relying exclusively on internal models and representations to guide action. Active vision strategies provide only one set of examples of how agent-environment interactions can be exploited in perception; other examples include the use of haptic exploration strategies in active touch.

Another set of examples demonstrating how coupling to the real world can simplify classical control problems comes from embodied models of coordinated movement (see *LOCOMOTION, INVERTEBRATE*). Several researchers (e.g., Brooks, 1991; Beer et al., 1998) have studied locomotion in insect-like hexapod robots. In these models, walking and other behaviors were not prepro-

grammed or controlled by a central processor; rather coordination emerged from the interactions between individual leg dynamics, layered control architectures, and the physics of the real world. Real-time coupling to an environment has also been exploited in robotic models of swimming in fish and flying in insects.

Motor Development and the Embodied Perspective

Many theories of human cognitive development primarily rely on the gradual maturation of an internal representation-based processing architecture as an explanatory basis for developmental change. Thelen and Smith (1994) proposed an alternative account focusing on the intimate linkage between brain, body, and environment. These authors claim that, in the course of development, structured action and perception result from dynamic interactions between all these domains, without the necessity for the prior construction of underlying internal representations.

Recently, Thelen and colleagues have investigated the development of Jean Piaget’s classic “A-not-B” error in infants. The basic phenomenon involves infants reaching for and retrieving an object from one of two identical containers (labeled A and B) after the object was hidden in full view of the infants. First, the object is repeatedly hidden at and retrieved from location A. If the object is then hidden in container B, the infants, after a brief delay, continue to reach for container A, even though they viewed the object being hidden in B. Numerous contextual effects on the A-not-B task have been demonstrated, including the visual appearance of the containers, timing effects, and infant posture. Several cognitive theories have been proposed to account for the A-not-B error, some suggesting an immature concept of object permanence as the principal cause, others focusing on a dissociation or modular segregation between “knowing” and “acting.” Thelen and colleagues (2001) designed a detailed dynamical model to account for the A-not-B error and its context dependency. Important components of the model include a “movement planning field” whose dynamics determine the goal location of the reaching movement. This field receives activation from sensory input fields and from a “memory field” that maintains a memory of recent reaching locations. The A-not-B error emerges dynamically if acute or specific sensory inputs cuing target location B decay (for example, during a brief temporal delay) while inputs from the memory field (with high activity at location A and a slow time course) continue to dominate. Thus, in the model, the A-not-B error is not due to immature or weak internal representations or to a separation between “knowing” and “acting” but is the result of the internal dynamics of the reaching system, specifically the interaction between the effects of memory (long time course) and those of acute sensory inputs (short time course). The model was tested in numerous other contexts and for different sets of task parameters and produced results that were consistent with empirical findings.

A major advance of the dynamical perspective has been to unify processes of neural, bodily, and environmental change within a common dynamical framework. According to this approach, for example, “knowing” and “acting” do not constitute neatly separable modules or domains within the human cognitive architecture. Rather, in the course of human cognitive development, “knowing” and “acting” are intricately coupled and manifest themselves in different contexts of embodied, situated action and individual experience.

Sensorimotor Processes in Category Formation

Perceptual categorization is one of the most fundamental cognitive processes. The formation of new categories is crucial for an organism’s ability to continually adapt within a changing and unpredictable environment (Edelman, 1987). A large number of computational and connectionist models of categorization have been

proposed (see Pfeifer and Scheier, 1999). The majority of these models work by constructing an optimal mapping between representations of the stimulus (input) and discrete category representations (output). Synaptic weights linking input and output representations are adjusted by using supervised (backpropagation) or unsupervised learning schemes.

Embodied cognition offers a different perspective on category learning. An embodied system is not passively exposed to sensory "data" or to coded feature vectors. Rather, embodied systems exploit movement and interactions with the environment to actively seek out sensory stimulation. In the process, they not only sample but also may generate "good" sensory information, for example by introducing temporal correlations due to bodily movement that help in constructing perceptual invariants. To test this hypothesis, Sporns and colleagues designed an embodied neural model of the development of translation invariance and object selectivity in the primate inferior temporal cortex (Almassy, Edelman, and Sporns, 1998). The model demonstrated that smooth lateral displacement of visual objects due to visual scanning movements are essential for constructing large homogeneous receptive fields of inferior temporal cortical neurons. When such movements were disrupted, translation-invariant and object-selective cells failed to emerge. Even after the initial developmental phase was completed, ongoing synaptic changes within the visual neural maps produced representational experience-dependent plasticity. Groups of neuronal units continued to compete for neural inputs and showed differential increases and decreases reflecting the behavioral history of the model. In another model of category learning, Pfeifer and Scheier (1999) showed that an embodied system can construct sensorimotor categories by generating consistent spatiotemporal correlations across sensor readings while interacting with objects. A simple circling behavior was sufficient to learn the distinction between large and small objects, encountered as a mobile robot navigates its environment.

Both of these models show how embodied systems can actively generate information about objects that is not contained in individual sensor "snapshots." The categorization of objects requires bodily movement and real-world interaction to generate temporally correlated sensory inputs. Correlations can be generated within one sensory modality across time, which may lead to a disambiguation of sensory input through a reduction of the dimensionality of sensory space. In addition, correlations across different sensory channels can be exploited to form cross-modal associations, which are a prerequisite for concept formation. In the context of category and concept formation, embodiment may be viewed as an essential principle that supports the ability of autonomous systems to extract statistical regularities ("knowledge") from an environment. In order to accomplish this function, the temporal dynamics of the nervous system, the morphology of the body (the spatial arrangement and characteristics of its sensors and effectors), and the movement repertoire act together in ways that are not encompassed by pure information-processing approaches.

Value and Embodied Cognition

Embodied cognition requires that neural, bodily, and environmental domains attain dynamically coupled states. Embodied systems need to function coherently within a given environment or task domain. If behavior is to be adaptive, mechanisms must exist by which global functional goals can shape the elementary (neural and bodily) components of the organism, enabling adaptive dynamic states consistently to emerge. In other words, the global performance or adaptiveness of the system must be able to influence its local (internal) structure such that the propensity of the system to behave adaptively is increased. For the system to be truly autonomous, the mechanisms that mold local structure to yield global function must reside wholly within the system itself.

Supervised learning is clearly inconsistent with this requirement. Reinforcement learning (see REINFORCEMENT LEARNING) and in particular temporal difference learning provide a promising set of computational principles, although generally they do not specify the neural mechanisms by which consequences of behavior are sensed by an organism. In mammals, several neuromodulatory systems (including the dopaminergic and noradrenergic systems) are known to project diffusely and widely throughout the cerebral cortex (see DOPAMINE, ROLES OF; NEUROMODULATION IN MAMMALIAN NERVOUS SYSTEMS). They are responsive to salient events in the environment (e.g., reward stimuli) and exert physiological effects on cortical neural activity and plasticity. Sporns and colleagues have implemented such "value systems" in embodied synthetic models (Almassy et al., 1998; Sporns et al., 2000) to study their role in adaptive behavior. Computationally, value acts to gate plasticity during brief episodes of high behavioral saliency. Through synaptic plasticity in sensory afferents to value systems, previously neutral sensory stimuli or modalities can acquire the ability to trigger value and thus become salient to the organism. Sporns et al. (2000) showed how innate and acquired response characteristics can shape the behavioral history of an autonomous robot and provide a neural basis for avoidance and secondary conditioning.

Discussion

Embodied cognition presents a stark contrast to classical views of cognition. The theoretical challenge of embodied cognition has provoked numerous philosophical discussions of core principles of cognition such as representation and the internal world model. It is unclear whether internal models, explicitly symbolic or based on neural representations, should be fully abandoned. Modern neuroscience provides abundant evidence for neural coding in all sensory and motor domains, as well as neural activity underlying internally generated states related to attention, memory, prediction, and planning. Embodied cognition needs to strive for a conceptual synthesis between such internal processes and coupled sensorimotor and behavioral interactions across brain and body. Perhaps it is fair to say that the ultimate success of embodied cognition will depend on whether empirical findings in developmental, cognitive, and neural science buttress its far-reaching theoretical claims and whether implementations of embodied models will compare favorably with those based on other approaches.

In the future we may expect to see a fruitful convergence between the methods and concepts from dynamical systems theory and neuroscience. For example, the large-scale dynamics of an extended network such as the cerebral cortex can be characterized by global interactions between locally specialized (segregated) areas, leading to the emergence of functionally integrated cognitive and perceptual states (Tononi, Edelman, and Sporns, 1998). Thus, in a sense, perceptual and cognitive states are characterized not only by the activity of certain brain regions, but also by their dynamic co- and interactivity. In the future, as we continue to explore the important connection between neural dynamics and cognition (Arbib, Érdi, and Szentágothai, 1998), more advanced synthetic approaches will begin to study neurodynamical processes of increasing complexity within the embodied systems of organisms and robots.

Road Map: Psychology

Related Reading: Neuroethology, Computational; Philosophical Issues in Brain Theory and Connectionism

References

- Almassy, N., Edelman, G. M., and Sporns, O., 1998, Behavioral constraints in the development of neuronal properties: A cortical model embedded in a real world device. *Cerebr. Cortex*, 8:346–361.

- Arbib, M. A., Érdi, P., and Szentágothai, J., 1998, *Neural Organization: Structure, Function, and Dynamics*, Cambridge, MA: MIT Press.
- Beer, R. D., Chiel, H. J., Quinn, R. D., and Ritzmann, R. E., 1998, Bio-robotic approaches to the study of motor systems, *Curr. Opin. Neurobiol.*, 8:777–782.
- Braitenberg, V., 1986, *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA: MIT Press.
- Brooks, R. A., 1991, New approaches to robotics, *Science*, 253:1227–1232.
- Clark, A., 1997, *Being There: Putting Brain, Body and World Together Again*, Cambridge, MA: MIT Press. ♦
- Edelman, G. M., 1987, *Neural Darwinism*, New York: Basic Books. ♦
- Edelman, G. M., Reeke, G. N., Gall, W. E., Tononi, G., Williams, D., and Sporns, O., 1992, Synthetic neural modeling applied to a real-world artifact, *Proc. Natl. Acad. Sci. USA*, 89:7267–7271.
- Kelso, J. A. S., 1995, *Dynamic Patterns*, Cambridge, MA: MIT Press. ♦
- Pfeifer, R., and Scheier, C., 1999, *Understanding Intelligence*, Cambridge, MA: MIT Press. ♦
- Sporns, O., Almassy, N., and Edelman, G. M., 2000, Plasticity in value systems and its role in adaptive behavior, *Adapt. Behav.*, 8:129–148.
- Thelen, E., and Smith, L. B., 1994, *A Dynamic Systems Approach to the Development of Cognition and Action*, Cambridge, MA: MIT Press. ♦
- Thelen, E., Schöner, G., Scheier, C., and Smith, L. B., 2001, The dynamics of embodiment: A field theory of infant perseverative reaching, *Brain Behav. Sci.*, 24:1–34.
- Tononi, G., Edelman, G. M., and Sporns, O., 1998, Complexity and coherency: Integrating information in the brain, *Trends Cognit. Sci.*, 2:474–484.
- Varela, F. J., Thompson, E., and Rosch, E., 1991, *The Embodied Mind*, Cambridge, MA: MIT Press.

Emotional Circuits

Jean-Marc Fellous, Jorge L. Armony, and Joseph E. LeDoux

Introduction

Emotion is clearly an important aspect of the mind; yet it has been largely ignored by the “brain and mind (cognitive) sciences” in modern times. However, there are signs that this is beginning to change. This chapter surveys some issues about the nature of emotion, describes what is known about the neural basis of emotion, and considers some efforts that have been made to develop computer-based models of different aspects of emotion.

What Is Emotion?

The nature of emotion has been debated within psychology for the past century. The formal debate goes back to William James’s famous question: Do we run from the bear because we are afraid, or are we afraid because we run? James suggested that we are afraid because we run. Subsequently, the psychological debate over emotion has centered on the question of what gives rise to the subjective states of awareness that we call feelings, or emotional experiences. Theories of emotional experience typically seek to account for how different emotional states come about, and can be grouped into several broad categories: feedback, central, arousal, and cognitive theories (for review, see LeDoux, 1996). Though very different in some ways, each of these theories proposes that emotional experiences are the result of prior emotional processes. Feedback and arousal theories require that the brain detect emotionally significant events and produce responses appropriate to the stimulus; these responses then serve as a signal that determines the content of emotional experience. Central and cognitive appraisal theories, which are in some ways different levels of description of similar processes, assume that emotional experience is based on prior evaluations of situations; these evaluations then determine the content of experience. Interestingly, the evaluative processes that constitute central and appraisal theories are also implicitly necessary for the elicitation of the peripheral responses and arousal states of feedback and arousal theories.

The disparate theories of emotional experience thus all point to a common mechanism—an evaluative system that determines whether a given situation is potentially harmful or beneficial to the individual. Since these evaluations are the precursors to conscious emotional experiences, they must, by definition, be unconscious processes. Such processes are the essence of the ignored half of James’s question. That is, we run from a bear because our brain determines that bears are dangerous. Many emotional reactions are

likely to be of this type: unconscious information processing of stimulus significance, with the experience of “emotion” (the subjective feeling of fear) coming after the fact.

Although the manner in which conscious experiences emerge from prior processing is poorly understood, progress has nevertheless been made in understanding how brain circuits process emotion. Just as vision researchers have achieved considerable understanding of the neural mechanisms underlying the processing of color while still knowing little about how color experience emerges from color processing (COLOR PERCEPTION), it is possible to study how the brain processes the emotional significance of situations without first solving the problem of how those situations are experienced as conscious content.

The Neural Basis of Emotional Processing

Traditionally, emotion has been ascribed to the brain’s limbic system, which is presumed to be an evolutionarily old part of the brain involved in the survival of the individual and species (LeDoux, 2000). Some of the areas usually included in the limbic system are the hippocampal formation, septum, cingulate cortex, anterior thalamus, mammillary bodies, orbital frontal cortex, amygdala, hypothalamus, and certain parts of the basal ganglia. However, the limbic system anatomical concept and the limbic system theory of emotion are both problematic (LeDoux, 2000). The survival of the limbic system theory of emotion is due in large part to the fact that the amygdala, a small region in the temporal lobe, was included in the concept.

The amygdala has been consistently implicated in emotional functions (LeDoux, 1996; Rolls, 1998; Damasio, 1999; various chapters in Aggleton, 2000). Lesions of this region interfere with both positive and negative emotional reactions. Moreover, unit-recording studies show that cells in the amygdala are sensitive to the rewarding and punishing features of stimuli and to the social implications of stimuli. Other limbic areas have been less consistently implicated in emotion, and when they have been implicated, it has been difficult to separate out the contribution of the region to emotion per se as opposed to some of the cognitive prerequisites of emotion. The amygdala therefore serves as an experimentally accessible entry point into the distributed network of brain regions that mediate complex emotional evaluations.

The contribution of the amygdala to emotion results in large part from its anatomical connectivity (reviewed in LeDoux, 2000). The

amygdala receives inputs from each of the major sensory systems and from higher-order association areas of the cortex. The sensory inputs arise from both the thalamic and cortical levels. These various inputs allow a variety of levels of information representation (from raw sensory features processed in the thalamus to whole objects processed in sensory cortex to complex scenes or contexts processed in the hippocampus) to impact on the amygdala and thereby activate emotional reactions. Most of these sensory inputs converge in the lateral nucleus of the amygdala, and the higher order information in the basal nucleus (Figure 1). These can be viewed as the sensory and cognitive gateways, respectively, into the amygdala's emotional functions. At the same time, the amygdala sends output projections to a variety of brainstem systems involved in controlling emotional responses, such as species-typical behavioral responses (including facial expressions and whole-body responses such as freezing), autonomic nervous system responses, and endocrine responses. Most of these outputs originate from the central nucleus of the amygdala. Recent anatomical and physiological work has, however, shown that the amygdaloid complex consists of several interacting subnuclei that may have specific individual contribution to the overall emotional computation performed (see the following discussion and Figure 1). If the amygdala is consistently found to contribute to the evaluation of the emotional significance of a stimulus, are there systems that control the processing of the amygdala? Recent work suggests that the amygdaloid complex can be modulated by neurochemical systems, such as serotonergic or dopaminergic, that are activated in relation to the overall behavioral state of the organism.

Much of the anatomical circuitry of emotion described previously has been elucidated through studies of fear conditioning, a procedure whereby an emotionally neutral stimulus, such as a tone or light, is associated with an aversive event, such as a mild shock to the foot (LeDoux and Phelps, in Lewis and Haviland-Jones, 2000; Davis, 1998). After such pairings, the tone or light comes to elicit emotional reactions that are characteristically expressed when members of the species in question are threatened. Although there are other procedures for studying emotion, none has been as successfully applied to the problem of identifying stimulus-response connections in emotion. The fear conditioning model is at this point

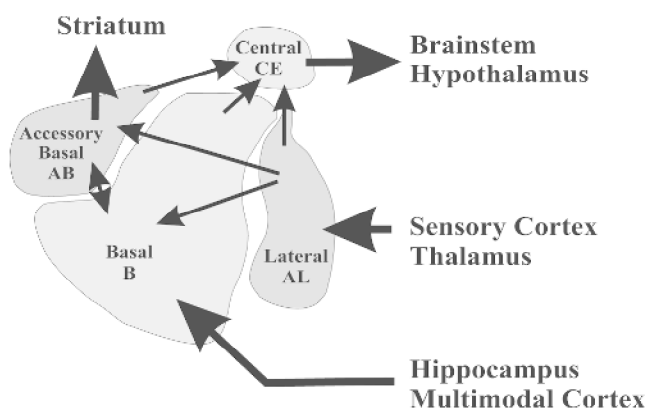


Figure 1. Simplified diagram of the amygdala intrinsic connections. The two main inputs to the amygdala are from the sensory/thalamic structures to AL, providing low level sensory information, and from polymodal and multimodal cortical association areas to B, providing more processed sensory information. The central nucleus receives convergent information from many other amygdaloid nuclei, and generate behavioral outputs (low level motor, autonomic, endocrine responses) that are a reflection of the intrinsic computations performed by the amygdala as a whole. Higher order motor control outputs are generated by the AB. See LeDoux and Pitkanen's chapter and Aggleton's chapter in Aggleton (2000) for more details.

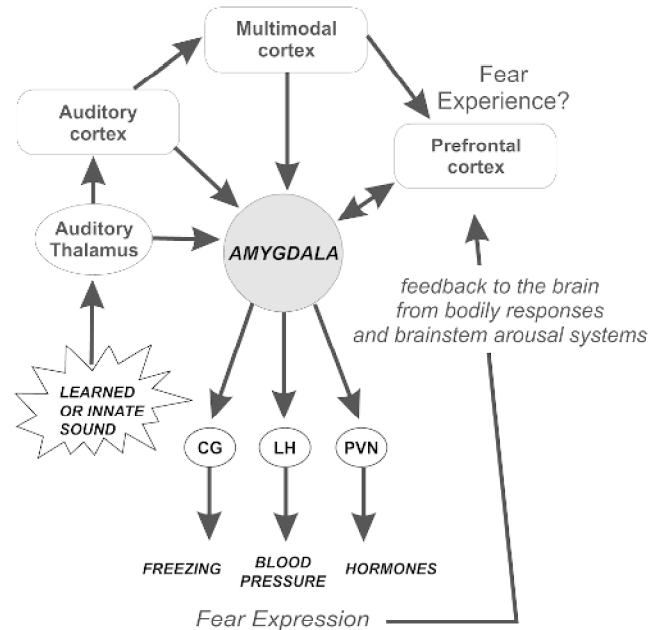


Figure 2. Emotional expression and emotional experience of auditory fear. Fearful stimuli (learned or innate) follow two main routes. The fast route involves the thalamo-amygdala pathway and responds best to simple stimulus features (such as a tone), the slow route involves the thalamo-cortical-amygdala pathway and carries more complex features (such as context). The expression of fear is mediated by the outputs of the amygdala to lower brain centers (brainstem, hypothalamus), while the experience of fear involves the prefrontal cortex circuitry.

particularly attractive since it has laid out pathways from the sensory input stage to the motor output stage of processing, showing how simple stimulus features, stimulus discriminations, and contexts control the expression of behavioral, autonomic, and endocrine responses in threatening situations (Figure 2).

Although many emotional response patterns are hardwired in the brain's circuitry, the particular stimulus conditions that activate these are mostly learned by association through classical conditioning. The amygdala appears to contribute significantly to this aspect of learning and memory and may be a crucial site of synaptic plasticity in emotional learning (LeDoux, 2000). This form of memory is quite different from what has come to be called *declarative memory*, the ability to consciously recall some experience from the past (CORTICAL MEMORY; SHORT-TERM MEMORY). Declarative memory, in contrast to *emotional memory*, crucially requires the hippocampus and related areas of the cortex. When we encounter some stimulus that in the past had aversive consequences, we recall the details of who we were with and where we were and even that it was a bad experience. However, in order to give the declarative memory an emotional flavor, it may be necessary for the stimulus, simultaneously and in parallel, to activate the emotional memory system of the amygdala. It is likely to be this dual activation of memory systems that gives our ongoing declarative memories their emotional coloration. Emotional memories are formed by the amygdala, in the same manner as declarative memories are formed in the hippocampus. The actual site of storage of emotional and declarative memories is still a matter of debate (Cahill et al., 1999), but may involve distant cortical and subcortical areas in addition to the amygdala and hippocampus (LeDoux, 2000).

In the last several years, the basic findings regarding fear conditioning in animals have been confirmed and extended by studies

of brain-damaged patients and functional imaging studies. This work has shown that the human amygdala is also involved in fear learning and other emotional processes (for reviews see Damasio, 1999; LeDoux, 2000; and Dolan's chapter in Aggleton, 2000).

At this point, we have mentioned "emotional experience" a number of times, and it may be worth speculating on just what an emotional experience is and how it might emerge. The emotion of fear will be used as an example. All animals, regardless of their stage of evolutionary development, must have the ability to detect and escape from or avoid danger. The widespread distribution of these behaviors in the animal kingdom makes it unlikely that the subjective experience of fear is at the heart of this ability. It may well be the case that subjective, consciously experienced fear is a mental state that occurs when the defense system of the brain (the system that detects threats and organizes appropriate responses) is activated, but only if that brain also has the capacity for consciousness. That is, by this reasoning, fear and other emotions reflect the representation of the activity of neural systems shaped by evolution and the responses they produce as conscious content. If this is true, then it is important that we focus our research efforts on these stimulus-detection and response-organizing systems, as these are the systems that generate the conscious content we call emotions. Although emotional behaviors may be triggered by sensory inputs that bypass or pass through the neocortex, the experience of emotion is likely to involve the cortical representation of the emotional episode. Although our understanding of the cortical representation of emotion episodes (or other conscious experiences) is poor at present, considerable evidence suggests that working memory circuits involving the frontal lobe may play a key role (LeDoux, 2000, and Figure 2). For a different view of the neural basis of emotional experience, see Damasio, 1999.

Computational Models of Emotion

Using computers to understand emotions has always been a challenge. Popular beliefs define computing devices as inherently incapable of exhibiting and experiencing any emotions and, at present, no definite claims have been made that computers may be suitable for such a task. Nevertheless, consistent with the notion put forth in the introduction, computers are used as tools for modeling certain aspects of emotional processing.

Models of Emotional Learning and Memory

As proposed by most central theories, many emotional responses are hardwired in brain circuitry. Nevertheless, in humans and animals, the environmental events that trigger these responses are often learned through experiences in which emotionally neutral stimuli come to be associated with emotionally charged stimuli. One important aspect of emotional processing, therefore, involves the manner in which the brain forms, stores, and uses associations between meaningless and meaningful stimuli.

Grossberg developed models of conditioned affective states based on the notion that conditioned reinforcement involves pairs of antagonistic neural processes, such as fear and relief. The model suggests a mechanism by which neutral events are charged with a reinforcing value (either positive or negative) depending on the previous activity of the model. The simulated neural circuits are suggestive of the role of brain structures involved in the processing of certain emotions, such as the hippocampo-amygdaloid system (described as a zone of convergence of conditioned (CS) and unconditioned (US) stimulus pathways), the septum (described as a zone in which the opposition of the processes is represented), the hypothalamus, the nucleus of the solitary tract, and the reticular formation (described as zones of visceral and somatosensory inputs). These models have been used to explain some aspects of the

dysfunctional behaviors seen in diseases such as schizophrenia (see Grossberg, 2000, for a review).

Armony and co-workers have implemented another connectionist model of emotional learning and memory that, like the previous model, also focuses on zones of convergence of US and CS pathways (Armony et al., 1997). This model is anatomically constrained by the known data of the fear conditioning circuitry. It examines processing in two parallel sensory (CS) transmission pathways to the amygdala from the auditory thalamus and the auditory cortex in a learning situation involving an auditory CS paired with a footshock US. The model is initially trained using a modified Hebb-type learning rule and, under testing conditions, reproduces data related to frequency-specific changes of the receptive fields known to exist in the auditory thalamus and amygdala. The model predicted that lesions of the cortical auditory route would not affect the specificity of the behavioral response to a range of frequencies centered on the training (aversively meaningful) frequency. This prediction has been verified experimentally. Because cortical representations are subject to attentional focus, this modeling study, like the previous one, suggests a close link between the amygdala and the attentional system of the midbrain. A separate connectionist-like model proposed that the amygdala might work in concert with the perirhinal cortex to generate conditioned responses to fear in cases in which the CS-US interval lasts several seconds (Tieu et al., 1999).

Recent anatomical studies coupled to physiological experiments *in vivo* and *in vitro* have provided invaluable data that can be used to build biophysically realistic computational models of amygdala circuits. Such models explore the interactions between converging thalamic and cortical inputs onto neurons in the lateral nucleus of the amygdala (Armony and LeDoux, 1997), as well as the role of local feedforward and feedback inhibition in stimulus processing (reviewed in Armony et al., 1997).

Computational Models of Cognitive-Emotion Interactions and Appraisal

Researchers in experimental psychology, artificial intelligence (AI), and cognitive science have long recognized the mutual influences of emotion and cognition. However, these interactions are still not clearly understood. We still do not have adequate theories defining each of these components of human mentation (emotion and cognition), much less a full understanding of how cognition and emotion might relate (EMOTION-COGNITION INTERACTIONS in the First Edition).

As described previously, most theories of emotion recognize the importance of evaluative or appraisal processes. Although there is considerable disagreement as to how these processes should best be viewed, most workers nevertheless see evaluative or appraisal processes as functioning by comparing sensed characteristics of the world to internal goals, standards, and attitude structures, deducing the emotional significance of the stimulus, guiding the expression of emotional behavior and other physiological responses, and influencing other modules pertaining to behavioral decisions.

In principle, it is possible to model appraisal processes using classical symbolic AI techniques (see Picard, 1997, for a review). It is possible, for example, using a vector space approach, to find a plausible mapping between appraisal features (e.g., novelty, urgency, intrinsic pleasantness) and emotion categories (e.g., fear, joy, pride). Relying on a posteriori verbal reports and a predefined set of emotions, one could then derive a limited set of appraisal criteria, sufficient for emotion prediction and differentiation. Other AI approaches, such as decision trees, pattern matching, and production rules (expert systems), are also possible, although each of these methods encounters theoretical difficulties. These types of systems, however, do not generally account for neurophysiological data.

One criticism often made of cognitive models of emotion is related to the complexity of processing involved and to the time they consequently require. From an AI point of view, the criticism has been addressed by introducing reactivity to “classical” cognitive models. Classical AI approaches assume that systems possess a well-defined representation of their environment, state, actions, and goals. In contrast, reactive systems do not make such assumptions; they are mostly based on real-time, incomplete evaluations, their performance being based more on the properties of the evaluative mechanisms than on the quality and quantity of their internal representations (REACTIVE ROBOTIC SYSTEMS).

It is interesting to note that, as we mentioned earlier, appraisal of sensory information might be one of the most prominent functions of the amygdala, placing this structure in a key position to actually perform the mapping of the emotional value of the stimuli. In this view, the relation between amygdala activity and emotion is a computational one (in the broad sense of the term) rather than a subjective one. The existence of multiple pathways to the amygdala from input processing systems of various levels of complexity (see previous discussion) provides a biological resolution to some of the concerns that have been raised about the importance of cognition in driving emotion. The involvement of cognition can be minimal or maximal, depending on the situation.

Models of Facial Expressions of Emotion

Of interest to feedback and arousal theories, the expression of emotion in the face is an important biological aspect of emotion that has significant implications for how emotion is communicated in social situations. Face recognition and analysis of facial expression has only recently been an active field of research in the computer vision community (for review, see Bartlett, 2001). Face analysis can be computationally divided into three subproblems: detecting the face in a scene; identifying the face; and analyzing its expression. At present, each of these tasks uses different features of the face, and different computational approaches. These approaches are based on psychophysical observations and are not yet explicitly based on neurophysiological data. However, a number of neurophysiological studies have been conducted (for review, see Rolls's chapter in Aggleton, 2000). These studies have shown cells selectively responsive to particular faces in areas of the temporal neocortex and in the amygdala. Functional imaging studies of humans have led to similar results (see Dolan's chapter in Aggleton, 2000). Other studies have shown that there might be an influence of facial expressions on the actual neural correlates of the emotional states experienced, through modifications of blood flow characteristics (for review, see Ekman, 1992). Other approaches are more physico-mathematical, relying on image processing techniques. These implementations address exclusively the problem of emotional expression (and, possibly, communication of emotions) without relying on any theory of emotional experience (Bartlett, 2001).

Conclusions

It is important to distinguish between emotional experiences and the underlying processes that lead to emotional experiences. One

of the stumbling blocks to an adequate scientific approach to emotion has been the focus of the field on constructing theories of the subjective aspects of emotion. Studies of the neural basis of emotion and emotional learning have instead focused on how the brain detects and evaluates emotional stimuli and how, on the basis of such evaluations, emotional responses are produced. The amygdala was found to play a major role in the evaluation process. It is likely that the processing that underlies the expression of emotional responses also underlies emotional experiences, and that progress can be made by treating emotion as a function that allows the organism to respond in an adaptive manner to challenges in the environment rather than to a subjective state. Although computational approaches to subjective experiences of the emotional or non-emotional kind are not likely to be easily achieved, computational approaches to emotional processing are both possible and practical. Although relatively few models currently exist, this situation is likely to change as researchers begin to realize the opportunities that are present in this long-neglected area.

Road Maps: Cognitive Neuroscience; Psychology

Related Reading: Cognitive Maps; Conditioning; Embodied Cognition; Motivation; Pain Networks; Sparse Coding in the Primate Cortex

References

- Aggleton, J. P., 2000, *The Amygdala: A Functional Analysis*, 2nd ed., Oxford: Oxford University Press. ♦
- Armony, J. L., and LeDoux, J. E., 1997, How the brain processes emotional information, *Ann. NY Acad. Sci.*, 821:259–270. ♦
- Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and LeDoux, J. E., 1997, Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning, *Trends Cog. Sci.*, 1:28–34.
- Bartlett, M., 2001, *Face Image Analysis by Unsupervised Learning*, Boston: Kluwer Academic Publishers.
- Cahill, L., Weinberger, N. M., Roozendaal, B., and McGaugh, J. L., 1999, Is the amygdala a locus of “conditioned fear”? Some questions and caveats, *Neuron*, 23:227–228.
- Damasio, A., 1999, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, New York: Harcourt Brace.
- Davis, M., 1998, Anatomic and physiologic substrates of emotion in an animal model, *J. Clin. Neurophysiol.*, 15:378–387.
- Ekman, P., 1992, Facial expressions of emotion: New findings, new questions, *Psychol. Sci.*, 3:34–38.
- Grossberg, S., 2000, The imbalanced brain: From normal behavior to schizophrenia, *Biol. Psychiat.*, 48:81–98.
- LeDoux, J., 1996, *The Emotional Brain*, New York: Simon and Schuster. ♦
- LeDoux, J. E., 2000, Emotion circuits in the brain, *Annu. Rev. Neurosci.*, 23:155–184. ♦
- Lewis, M., and Haviland-Jones, J. M., 2000, *Handbook of Emotions*, 2nd ed., New York: Guilford Press.
- Picard, R. W., 1997, *Affective Computing*, Boston: MIT Press. ♦
- Rolls, E. T., 1998, *The Brain and Emotion*, Oxford, UK: Oxford University Press.
- Tieu, K. H., Keidel, A. L., McGann, J. P., Faulkner, B., and Brown, T. H., 1999, Perirhinal-amygdala circuit-level computational model of temporal encoding of fear conditioning, *Psychobiology*, 27:1–25.

Energy Functionals for Neural Networks

Eric Goles

Introduction

In this article we survey some of the most common neural network models for which an *energy* can be defined, where an energy is a quantity $E(x(t))$, depending on the current configuration of the network, $x(t)$, that does not increase when the network is updated, i.e., $E(x(t + \tau)) \leq E(x(t))$, and that is constant in steady state. Classically, these quantities appeared in the dynamical study of ordinary differential equations as what is called a Lyapunov function. In fact, the Lyapunov function approach consists in determining a positive quantity that decreases when the differential system approaches the equilibrium points. In such a case, it is possible to study the stability of the solutions (see Grossberg, 1988, and Hirsch, 1989, in the neural networks context). From the physical point of view, the possibility of associating an energy with neural networks arose from the deep analogy between them and the spin-glass magnetic model (Little and Shaw, 1975; Hopfield, 1984; see also OPTIMIZATION, NEURAL). The interest in determining when such quantities exist arises from the fact that their existence allows study of the convergence rate, for specific update modes of the network, to stable or short-period configurations. Moreover, the attractors are local minima of the energy E , so this kind of network can be used to model associative memories (Hopfield, 1982) and hill-climbing optimization strategies.

Energies have been developed for discrete and continuous networks. There are three principal models: discrete transition–discrete time, continuous transition–discrete time, and continuous transition–continuous time. The first model was made famous by Hopfield (1982) for associative memories with Hebb interactions updated asynchronously. It was later extended to sequential and parallel update (Fogelman-Soulié, Goles, and Weisbuch, 1983; Goles, Fogelman-Soulié, and Pellegrin, 1985). The discrete time–continuous function appears in the context of the brain-state-in-a-box model (Golden, 1986). A survey of the energy approach can be found in Goles and Martinez (1990). Finally, the continuous time–continuous transition model has been studied first by Cohen and Grossberg (1983; see COMPUTING WITH ATTRACTORS) and in a particular case by Hopfield (1984). It is important to point out that Hopfield’s energy approach was based on a spin-glass analogy of symmetric neural networks. This analogy and some preliminary results were presented in Little (1974) and Little and Shaw (1975).

In this article, we present first the linear argument model for discrete time and continuous transition. We determine, under symmetric assumptions about interconnections, the associated energy. Moreover, we extend the approach to the related class of quasi-symmetric interconnections. In a similar way, we present the discrete time–continuous transition model by taking as a local function a real, bounded, increasing function. It is important to point out that in this case symmetry is also the key hypothesis in determining the energy. Further, we extend the analysis to any increasing function.

For continuous time–continuous transition, we present the general model studied by Grossberg (1988), and we illustrate the energy determination for the particular case developed in Hopfield (1984).

The Binary State–Discrete Time Model

Suppose the neurons take values in a binary set, usually $\{0, 1\}$ or $\{1, 1\}$. We present here the threshold case, i.e., $y_i = 1(\sum_{j=1}^n w_{ij}x_j - b_i)$, for $x \in \{0, 1\}^n$ and $1(u) = 1$ iff $u \geq 0$ (0 otherwise).

Throughout this paragraph we will assume, without loss of generality, that for any $x \in \{0, 1\}^n$ and $i \in \{1, \dots, n\}$, $\sum_{j=1}^n w_{ij}x_j - b_i \neq 0$. (Otherwise, it suffices to make a small change in threshold b_i without changing the dynamics; Goles and Martínez, 1990.) The threshold model is the classical one. Other binary models, such as states in the set $\{-1, 1\}$ with the sign transition function, can be reduced to the threshold model, with similar expressions for the energies.

Asynchronous Update

Let us consider the foregoing model with asynchronous update, i.e., the neurons are updated one by one in random order. In this context, we have the following result.

Theorem 1. (Hopfield, 1982; Fogelman-Soulié et al., 1983.) Let $W = (w_{ij})$ be a symmetric $n \times n$ matrix with non-negative diagonal entries (i.e., $\text{diag}(W) \geq 0$). Then the quantity $E(x) = -\frac{1}{2} \sum_{i,j=1}^n w_{ij}x_i x_j + \sum_{i=1}^n b_i x_i$ is an energy associated with the asynchronous iteration of the network.

Proof. Let $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ be the current configuration. Suppose we update the k th neuron, obtaining the new configuration $\tilde{x} = (x_1, \dots, x_{k-1}, \tilde{x}_k, \dots, x_n)$, where $\tilde{x}_k = 1(\sum_{j=1}^n w_{kj}x_j - b_k)$. Let $\Delta_k E = E(\tilde{x}) - E(x)$. Since W is symmetric, we get:

$$\Delta_k E = -(\tilde{x}_k - x_k) \left(\sum_{j=1}^n w_{kj}x_j - b_k \right) - \frac{1}{2} w_{kk}(\tilde{x}_k - x_k)^2 \quad (1)$$

By definition of the threshold function, the first term is negative when $\tilde{x}_k \neq x_k$. Since $w_{kk} \geq 0$, we conclude that $\Delta_k E \leq 0$, with $\Delta_k E < 0$ iff $\tilde{x}_k \neq x_k$. \square

The first determination of this energy for a particular symmetric threshold model was done by Hopfield (1982) for associative memories and interactions w_{ij} defined by Hebb’s rule, i.e., $w_{ij} = \sum_{k=1}^p x_i^{(k)} x_j^{(k)}$, ($w_{ij} = 0$), where $x = \{x^{(1)}, \dots, x^{(p)}\}$ is the set of prototypes to be memorized. This rule was used before Hopfield in a neural model proposed by Anderson (see ASSOCIATIVE NETWORKS).

As a corollary to the previous theorem, we can state that the only stable states of the network are fixed points, i.e., configurations remaining invariant by the application of the threshold rule. In fact, it suffices to remark that between two successive different configurations, the energy decreases. Furthermore, since the energy is bounded on the set $\{0, 1\}^n$, the network converges in a finite number of steps to a fixed point. We will come back to this aspect in the next section. On the other hand, when $\text{diag}(W) = 0$ (as in Hopfield, 1982), it is easy to verify that the fixed points of the network are local minima of E . In fact, consider a fixed point $x \in \{0, 1\}^n$ and its k th neighborhood in the hypercube, $\tilde{x} = (x_1, \dots, 1 - x_k, \dots, x_n)$. Since x is a fixed point and $\text{diag}(W) = 0$, one gets $x_k = 1(\sum_{j \neq k} w_{kj}x_j - b_k)$ and $\Delta_k E = E(\tilde{x}) - E(x) = -(1 - 2x_k)(\sum_{j \neq k} w_{kj}x_j - b_k) > 0$, so x is a local minimum of E . The previous aspect is important for modeling, by hill-climbing neural strategies, some hard combinatorial optimization problems (see OPTIMIZATION, NEURAL).

Periodic Update

Assume now that the neurons are updated one by one in a periodic order: $\{1 \rightarrow 2 \rightarrow \dots \rightarrow n \rightarrow 1\}$. Clearly, this update strategy is a

particular case of the asynchronous one, but the periodicity permits us to obtain bounds on the transient time. It is obvious that the network has the same energy, E , as the asynchronous iteration. Given a matrix W and a threshold vector b , we define $\tau(W, b)$ as the maximum number of steps taken by the network, for any initial condition, to reach a stable configuration.

Theorem 2. (Fogelman-Soulié et al., 1983.) Let W be an $n \times n$ symmetric matrix with non-negative diagonal. Then the transient time $\tau(W, b)$ for periodic update is bounded by

$$\tau(W, b) \leq \frac{\|W\|_1 + 2\|b\|_1}{2\left(e + \min_i w_{ii}\right)} \quad (2)$$

where $e = \min \{|\sum_{j=1}^n w_{ij}x_j - b_i| : i \in \{1, \dots, n\}, x \in \{0, 1\}^n\}$, $\|W\|_1 = \sum_{i,j=1}^n |w_{ij}|$, and $\|b\|_1 = \sum_{i=1}^n |b_i|$.

Proof. From the proof of Theorem 1, one gets, for any $k \in \{1, \dots, n\}$, $|\Delta_k E| \geq e + \frac{1}{2} \min_i w_{ii}$. On the other hand, $|E(x)| \leq \frac{1}{2} \sum_{i,j=1}^n |w_{ij}| + \sum_{i=1}^n |b_i|$. From previous inequalities we obtain the bound directly. \square

Better bounds can be obtained with a finer analysis of $|\Delta_k E|$ and $|E(x)|$. See, for instance, Kamp and Hasler (1990) and Goles and Martínez (1990).

Parallel Update

Suppose we update the network synchronously:

$$x_i(t+1) = 1 \left(\sum_{j=1}^n w_{ij}x_j(t) - b_i \right) \quad 1 \leq i \leq n, x(0) \in \{0, 1\}^n \quad (3)$$

In this context we have the following result.

Theorem 3. (Goles et al., 1985.) Let W be an $n \times n$ symmetric matrix. Then the expression

$$E(t) = - \sum_{i,j=1}^n w_{ij}x_i(t)x_j(t-1) + \sum_{i=1}^n b_i(x_i(t) + x_i(t-1))$$

is an energy associated with the parallel update.

Proof. Let $\Delta E = E(t) - E(t-1)$. Since W is symmetric, one gets

$$\Delta E = - \sum_{i=1}^n (x_i(t) - x_i(t-2)) \left(\sum_{j=1}^n w_{ij}x_j(t-1) - b_i \right) \quad (4)$$

By definition of the threshold function, $\Delta E \leq 0$ and it is strictly negative when $x(t) \neq x(t-2)$. \square

This energy can also be obtained in the framework of the statistical mechanics model proposed by Little (1974). In fact, the energy $E(t)$ with threshold $b = 0$ is the zero temperature limit of the Hamiltonian: $H(x) = -\beta^{-1} \sum_{i,j=1}^n \log_2 \cosh(\beta n^{-1} \sum_{j \neq i} w_{ij}x_j)$, where $\beta = 1/kT$, k being the Boltzmann constant (Peretto, 1984). More information about the physical approach can be found in STATISTICAL MECHANICS OF NEURAL NETWORKS (q.v.).

Corollary 1. For a symmetric matrix W , the parallel iteration converges to fixed points or two-periodic configurations.

Proof. Suppose that $\{x(0), \dots, x(T-1), x(T) = x(0)\}$ is a cycle of period T . From Theorem 3 it follows that $E(t)$ is necessarily

constant on the cycle. If $T > 2$, we have $x(0) \neq x(2)$, so $E(2) < E(0)$, which is a contradiction. \square

Corollary 2. For a positive-definite matrix W , the parallel update converges to fixed points.

Proof. From Corollary 1 it is enough to prove that there are no two-cycles. Suppose $\{x(0), x(1)\}$ is a two-cycle, i.e., $x(2) = x(0)$. Since W is positive definite, $\alpha = (x(1) - x(0))^T(Wx(1) - x(0)) \geq 0$, with equality only if $x(1) = x(0)$. Further, by Equation 4,

$$\alpha = (x(0) - x(1))^T(Wx(1) - b) - (x(1) - x(0))^T(Wx(0) - b) = \Delta_2 E + \Delta_1 E$$

From the proof of Theorem 3, we know that $\Delta_2 E \leq 0$ and $\Delta_1 E \leq 0$, so $\alpha = 0$. Hence, $x(1) = x(0)$, i.e., two-cycles do not exist. \square

The application of the foregoing result to associative memory models is straightforward. Consider the Hopfield model with generalized Hebb interconnections on the prototype set $\{x^{(1)}, \dots, x^{(p)}\} \subseteq \{0, 1\}^n$, with interconnection matrix $W = (1/p)X^T X$, where $X = (x^{(1)}, \dots, x^{(p)})$. Consider also the projection or pseudo-inverse interconnection model, i.e., $W = (X^T X)^{-1} X^T$. Since in both cases W is positive definite, the parallel updates of previous models converge only to fixed points (Kamp and Hasler, 1990).

From the energy given in Theorem 3, one may determine that the transient time for the parallel update is bounded by $\tau(W, b) \leq (1/e)(\|W\|_1 + 3\|2b - W\bar{1}\|_1 - 2\sum_{i=1}^n e_i)$ if $e > 0$ (0 if $e = 0$), where $e = \min \{-E^*(2) - E^*(1) : x(0) \neq x(2)\}$, $e_i = \min \{|\sum_{j=1}^n w_{ij}u_j - b_i| : u \in \{0, 1\}^n\}$, and $\bar{1} = (1, \dots, 1)$. Clearly, if all the vectors belong to a two-cycle, then $e = 0$ and $\tau(W, b) = 0$.

It is important to remark that this bound is not necessarily polynomial. In fact, it is possible to build symmetric neural networks of size n with exponential transient time (recall that n neurons corresponds to 2^n states). When the matrix W takes values on the integers, the quantities e_i, e can be controlled and an explicit bound on $\tau(W, b)$ can be given in terms of W, b . Further, there exist symmetric neural networks where the bound is attained. More information about these topics can be found in Goles and Martínez (1990).

Another model that can be studied by using the energy approach is the bidirectional associative memory (BAM) model proposed by Kosko (1988). Roughly, it consists of a two-layer bidirectional network that achieves heteroassociations with a smaller correlation matrix. Given its correlation matrix W and a pair of vectors $(x, y) \in \{0, 1\}^{2n}$, the energy is $E = -x^T W y$ (Wang, Cruz, and Mulligan, 1991).

Energies for Nonsymmetric Models

When the matrix W is no longer symmetric, it is difficult, for general classes of matrices with other nontrivial regularities, to determine energies that ensure convergence to fixed points (in the asynchronous and sequential update) and to fixed points and two-cycles (for the parallel update). That fact is not surprising, since if we permit arbitrary interconnections and enough neurons we can model arbitrary automata (see NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY).

Furthermore, the dynamics (sequential or parallel) is very dependent on symmetry. In fact, arbitrarily small variations in a symmetric matrix can generate long cycles. A similar situation occurs for the sequential iteration and the non-negativity diagonal hypothesis: a negative diagonal also generates long cycles (Goles and Martínez, 1990). So one may say that the existence of energies relies very much on the hypothesis of symmetric interconnections. Of course, one may find nonsymmetric matrices that accept an energy, but one could not find a nontrivial class, really different

from the symmetric one, with the energy property. However, minor variation on symmetry are possible.

One can generalize the foregoing results to the quasi-symmetric class of matrices. We say that a matrix $W = (w_{ij})$ is quasi-symmetric if there exists a positive vector (u_1, \dots, u_n) such that for any $i, j \in \{1, \dots, n\}$, $u_i w_{ij} = u_j w_{ji}$. We have the following result.

Theorem 4. (Goles and Martinez, 1990.) Let W be a quasi-symmetric matrix. Then the quantity $E(x) = -\frac{1}{2} \sum_{i,j} u_i w_{ij} x_i x_j + \sum_{i=1}^n u_i b_i x_i$, is an energy for the asynchronous and sequential update.

The quantity $E(t) = -\frac{1}{2} \sum_{i,j=1}^n u_i w_{ij} x_i(t) x_j(t-1) + \sum_{i=1}^n u_i b_i (x_i(t) + x_i(t-1))$ is an energy for the parallel update.

The Continuous Transition–Discrete Time Model

Let us now consider a real transition function, $f: \mathbb{R} \rightarrow \mathbb{R}$, continuous, strictly increasing on an interval $S = (-a, a)$, $f(0) = 0$, and constant outside S , i.e., $f(x) = f(-a)$ for $x \leq -a$ and $f(x) = f(a)$ for $x \geq a$. As an example, f could be a truncated “sigmoidal” similar to those used by Hopfield (1984) and classically employed in multilayered networks. Another example is the BSB model (Golden, 1986) where f is linear in S . In the previous context, the update function is as follows:

$$y_i = f(\arg_i(x)) \quad 1 \leq i \leq n, x \in S^n \quad (5)$$

where $\arg_i(x) = \sum_{j=1}^n w_{ij} x_j - b_i$. For this model and the iterations defined in previous paragraphs we have the following result.

Theorem 5. (Goles and Martínez, 1990.) Suppose W is an $n \times n$ symmetric matrix. Then when $\text{diag}(W) \geq 0$ the asynchronous or periodic update admits the energy

$$E(x) = -\frac{1}{2} \sum_{i,j=1}^n w_{ij} x_i x_j + \sum_{i=1}^n \left(\int_0^{x_i} f^{-1}(s) ds + b_i x_i \right) \quad (6)$$

Furthermore, the quantity

$$\begin{aligned} E(t) = & -\sum_{i,j=1}^n w_{ij} x_i(t) x_j(t-1) \\ & + \sum_{i=1}^n \left[\int_0^{x_i(t)} f^{-1}(s) ds + \int_0^{x_i(t-1)} f^{-1}(s) ds \right] \\ & + \sum_{i=1}^n b_i (x_i(t) + x_i(t-1)) \end{aligned} \quad (7)$$

is an energy associated with the parallel update.

Proof. We first give the proof for the asynchronous (analogous to the periodic) update. Suppose we update the k th neuron, so that $\tilde{x} = (x_1, \dots, \tilde{x}_k, \dots, x_n)$, where $\tilde{x}_k = f(\arg_k(x))$. Suppose also that each neuron has been updated at least one time, so $x_k = f(\arg_k(z))$, where $z \in S^n$. Since W is symmetric, one gets

$$\begin{aligned} \Delta_k E = & -(\tilde{x}_k - x_k) \arg_k(x) - \frac{1}{2} w_{kk} (\tilde{x}_k - x_k)^2 \\ & + \int_0^{\tilde{x}_k} f^{-1}(s) ds - \int_0^{x_k} f^{-1}(s) ds \end{aligned}$$

Since $w_{kk} \geq 0$, the quadratic term is clearly negative. For the other two terms, let $u = \arg_k(x)$ and $v = \arg_k(z)$. Then $\Delta_k E = -(f(u) - f(v))u + \int_0^{f(u)} f^{-1}(s) ds - \int_0^{f(v)} f^{-1}(s) ds$.

From the definition of f , we have easily

$$\int_0^{f(\alpha)} f^{-1}(s) ds = \alpha f(\alpha) - \int_0^\alpha f(s) ds \quad (8)$$

so $\Delta_k E = (u - v)f(v) + \int_0^v f(s) ds - \int_0^u f(s) ds$. Since f is an increasing function, one concludes that $\Delta_k E \leq 0$. For the parallel update the proof is similar. \square

One may also prove that the only finite orbits are fixed points for the asynchronous and periodic update, while fixed points and/or two-cycles are possible for the parallel update.

The generalization to a nonsymmetric interval, and functions such that $f(0) \neq 0$, can be studied in a similar way. A more general approach would suppose that local update functions act in a high-dimensional space. In this framework, by considering $f: \mathbb{R}^p \rightarrow \mathbb{R}^p$, it has been proved that when f is positive (i.e., $\langle f(x) - f(y), x \rangle \geq 0$ for all $x, y \in \mathbb{R}^p$, where $\langle \cdot, \cdot \rangle$ is a scalar product) the sequential and parallel update for symmetric interconnections admits an energy (Goles, 1985).

Further, if f is a subgradient of a convex function g (i.e., $g(y) \geq g(x) + \langle f(x), y - x \rangle$ for all $x, y \in \mathbb{R}^p$), it can be proved that, under the symmetry hypothesis, the parallel iteration also admits an energy (Goles and Martínez, 1990).

Continuous Transition–Continuous Time Models

Several authors have introduced differential models of neural networks such that the transition function and the time steps are continuous. In this context Cohen and Grossberg (1983) proposed the nonlinear update

$$\frac{dx_i}{dt} = a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^n w_{ij} d_j(x_j) \right] \quad 1 \leq i \leq n \quad (9)$$

To study the dynamics of Equation 9, the authors assume that $a_i(x) \geq 0$, $d'_i(x) \geq 0$, and the matrix W is symmetric. Then they determine the following energy function:

$$E(x) = -\sum_{i=1}^n \int^{x_i} b_i d'_i(x) dx + \frac{1}{2} \sum_{i,j=1}^n w_{ij} d_i(x_i) d_j(x_j) \quad (10)$$

A particular case of this model was studied by Hopfield (1984):

$$c_i \left(\frac{dy_i}{dt} \right) = \sum_{j=1}^n w_{ij} x_j - \frac{y_i}{R_i} + I_i \quad (11)$$

where $y_i = g_i^{-1}(x_i)$.

Theorem 6. (Hopfield, 1984.) Let W be a symmetric matrix and g_i a sigmoid. Then

$$E(x) = -\frac{1}{2} \sum_{i,j=1}^n w_{ij} x_i x_j + \sum_{i=1}^n \left(\frac{1}{R_i} \right) \int_0^{x_i} g_i^{-1}(s) ds + \sum_{i=1}^n I_i x_i \quad (12)$$

is an energy function associated with Equation 11.

Proof. Since W is symmetric, one gets

$$\frac{dE}{dt} = -\sum_{i=1}^n \frac{dx_i}{dt} \left(\sum_{j=1}^n w_{ij} x_j - \frac{y_i}{R_i} + I_i \right)$$

From Equation 11 we have

$$\frac{dE}{dt} = -\sum_{i=1}^n c_i \frac{dx_i}{dt} \frac{dy_i}{dt} = -\sum_{i=1}^n c_i (g_i^{-1})'(x_i) \left(\frac{dx_i}{dt} \right)^2$$

Since g_i^{-1} is increasing, one concludes that $dE/dt \leq 0$. \square

Good surveys of differential models of neural networks are provided by Grossberg (1988) and Hirsch (1989).

Discussion

In this article we have reviewed the principal neural models that accept an energy function. In all cases, the existence of such op-

erators depends strongly on regularities of interconnections: symmetry and quasi-symmetry. One may also determine an energy for a class of antisymmetric matrices (Goles and Martínez, 1990), but it seems difficult to find other nontrivial classes of matrices that accept energy functions. In fact, arbitrarily small perturbation of the matrices discussed in this article can induce long cycles in the network dynamics. Necessary and sufficient conditions for the existence of an energy function have been formally studied by Kobuchi (1991) under a plausible decomposition hypothesis of ΔE . But, in practice, the authors recover only the quasi-symmetry class.

Further, some of the results can be extended to high-order networks, i.e., neural networks with polynomial interactions. In this context, under a generalization of the symmetry hypothesis, one may find an energy function for the periodic update (Xu and Tsai, 1990; Kamp and Hasler, 1990).

[Reprinted from the First Edition]

Road Map: Dynamic Systems

Background: I.3. Dynamics and Adaptation in Neural Nets; Computing with Attractors

Related Reading: Dynamics and Bifurcation in Neural Nets

References

- Cohen, M., and Grossberg, S., 1983, Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Trans. Syst. Man Cybern.*, SMC-13:815–826.
- Fogelman-Soulié, F., Goles, E., and Weisbuch, G., 1983, Transient length in sequential iteration of threshold functions, *Discrete Appl. Math.*, 6:95–98.
- Golden, R. M., 1986, The brain-state-in-a-box neural model is a gradient descent algorithm, *J. Math. Psychol.*, 30:73–80.
- Goles, E., 1985, Dynamic of positive automata networks, *Theoret. Comput. Sci.*, 41:19–32.
- Goles, E., Fogelman-Soulié, F., and Pellegrin, D., 1985, Decreasing energy functions as a tool for studying threshold networks, *Discrete Appl. Math.*, 12:261–277.
- Goles, E., and Martínez, S., 1990, *Neural and Automata Networks*, Norwell, MA: Kluwer. ♦
- Grossberg, S., 1988, Nonlinear neural networks: Principles, mechanisms, and architectures, *Neural Netw.*, 1:17–61. ♦
- Hirsch, M., 1989, Convergent activation dynamics in continuous time networks, *Neural Netw.*, 2:331–349. ♦
- Hopfield, J. J., 1982, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.
- Hopfield, J. J., 1984, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA*, 81:3088–3092.
- Kamp, Y., and Hasler, M., 1990, *Réseaux de neurones récurrents pour mémoires associatives*, Lausanne: Presses Polytechniques et Universitaires Romandes. ♦
- Kobuchi, Y., 1991, State evaluation functions and Lyapunov functions for neural network, *Neural Netw.*, 4:505–510.
- Kosko, B., 1988, Adaptive bidirectional associative memories, *Appl. Opt.*, 26:4947–4960.
- Little, W. A., 1974, The existence of persistent states in the brain, *Math. Biosci.*, 19:101–120.
- Little, W. A., and Shaw, G. L., 1975, A statistical theory of short and long term memory, *Behav. Biol.*, 14:115.
- Peretto, P., 1984, Collective properties of neural networks: A statistical physics approach, *Biol. Cybern.*, 50:51–62.
- Wang, Y. F., Cruz, J. B., and Mulligan, J. H., 1991, Guaranteed recall of all training pairs for bidirectional associative memory, *IEEE Trans. Neural Netw.*, 2:559–567.
- Xu, X., and Tsai, W. T., 1990, Constructing associative memories using neural networks, *Neural Netw.*, 3:301–309.

Ensemble Learning

Thomas G. Dietterich

Introduction

Learning describes many different activities, ranging from CONCEPT LEARNING (q.v.) to REINFORCEMENT LEARNING (q.v.). The best understood form of statistical learning is known as *supervised learning* (see LEARNING AND STATISTICAL INFERENCE). In this setting, each data point consists of a vector of features (denoted \mathbf{x}) and a class label y , and it is assumed that there is some underlying function f such that $y = f(\mathbf{x})$ for each training data point (\mathbf{x}, y) . The goal of the learning algorithm is to find a good approximation h to f that can be applied to assign labels to new \mathbf{x} values. The function h is called a *classifier*, because it assigns class labels y to input data points \mathbf{x} . Supervised learning can be applied to many problems, including handwriting recognition, medical diagnosis, and part-of-speech tagging in language processing.

Ordinary machine learning algorithms work by searching through a space of possible functions, called *hypotheses*, to find the one function, h , that is the best approximation to the unknown function f . To determine which hypothesis h is best, a learning algorithm can measure how well h matches f on the training data points, and it can also assess how consistent h is with any available prior knowledge about the problem.

As an example, consider the problem of learning to pronounce the letter k in English. Consider the words *desk*, *think*, and *hook*, where the k is pronounced, and the words *back*, *quack*, and *knave*,

where the k is silent (in *back* and *quack*, we will suppose that the c is responsible for the k sound). Suppose we define a vector of features that consists of the two letters prior to the k and the two letters that follow the k . Then each of these words can be represented by the following data points:

x_1	x_2	x_3	x_4	y
e	s	—	—	+1
i	n	—	—	+1
o	o	—	—	+1
a	c	—	—	−1
a	c	—	—	−1
—	—	n	a	−1

where $y = +1$ if k is pronounced and -1 if k is silent, and where “—” denotes positions beyond the ends of the word.

One of the most efficient and widely applied learning algorithms searches the hypothesis space consisting of decision trees. Figure 1 shows a decision tree that explains the data points given above. This tree can be used to classify a new data point as follows. Starting at the so-called root (i.e., top) of the tree, we first check whether $x_2 = c$. If so, then we follow the left (“yes”) branch to the $y = -1$ “leaf,” which predicts that k will be silent. If not, we follow the right (“no”) branch to another test: Is $x_3 = n$? If so, we follow

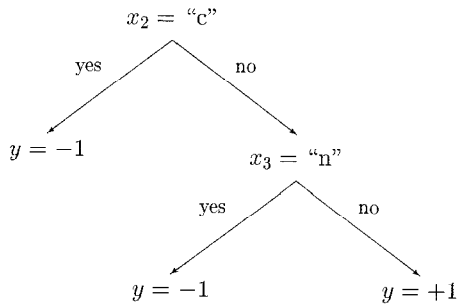


Figure 1. A decision tree for pronouncing the letter *k*. First, feature x_2 is tested to see if it is the letter *c*. If not, the feature x_3 is tested to see if it is the letter *n*. *K* is pronounced only if x_2 is not *c* and x_3 is not *n*.

the left branch to another $y = -1$ leaf. If not, we follow the right branch to the $y = +1$ leaf, where the tree indicates that *k* should be pronounced.

A decision tree learning algorithm searches the space of such trees by first considering trees that test only one feature (in this case x_2 was chosen) and making an immediate classification. Then they consider expanding the tree by replacing one of the leaves by a test of a second feature (in this case, the right leaf was replaced with a test of x_3). Various heuristics are applied to choose which test to include in each iteration and when to stop growing the tree. For a good discussion of decision trees, see the books by Quinlan (1993) and Breiman et al. (1984).

In addition to decision trees, there are many other representations for hypotheses that have been studied, including PERCEPTRONS, ADALINES, AND BACKPROPAGATION (q.v.), RADIAL BASIS FUNCTION NETWORKS (q.v.), GAUSSIAN PROCESSES (q.v.), graphical models, Helmholtz machines, and SUPPORT VECTOR MACHINES (q.v.). In all cases, these algorithms find one best hypothesis h and output it as the “solution” to the learning problem.

Ensemble learning algorithms take a different approach. Rather than finding one best hypothesis to explain the data, they construct a *set* of hypotheses (sometimes called a *committee* or *ensemble*) and then have those hypotheses “vote” in some fashion to predict the label of new data points. More precisely, an ensemble method constructs a set of hypotheses $\{h_1, \dots, h_K\}$, chooses a set of weights $\{w_1, \dots, w_K\}$, and constructs the “voted” classifier $H(\mathbf{x}) = w_1 h_1(\mathbf{x}) + \dots + w_K h_K(\mathbf{x})$. The classification decision of the combined classifier H is $+1$ if $H(\mathbf{x}) \geq 0$ and -1 otherwise.

Experimental evidence has shown that ensemble methods are often much more accurate than any single hypothesis. Freund and Schapire (1996) showed improved performance on 22 benchmark problems, equal performance on one problem, and worse performance on four problems. These and other studies are summarized in Dietterich (1997).

Why Ensemble Methods Work

Learning algorithms that output only a single hypothesis suffer from three problems that can be partly overcome by ensemble methods: the statistical problem, the computational problem, and the representation problem.

The statistical problem arises when the learning algorithm is searching a space of hypotheses that is too large for the amount of available training data. In such cases, there may be several different hypotheses that all give the same accuracy on the training data, and the learning algorithm must choose one of these to output. There is a risk that the chosen hypothesis will not predict future data points well. A simple vote of all of these equally good classifiers can reduce this risk.

The computational problem arises when the learning algorithm cannot guarantee finding the best hypothesis within the hypothesis space. In neural network and decision tree algorithms, for example, the task of finding the hypothesis that best fits the training data is computationally intractable, so heuristic methods must be employed. These heuristics (such as gradient descent) can get stuck in local minima and hence fail to find the best hypothesis. As with the statistical problem, a weighted combination of several different local minima can reduce the risk of choosing the wrong local minimum to output.

Finally, the representational problem arises when the hypothesis space does not contain any hypotheses that are good approximations to the true function f . In some cases, a weighted sum of hypotheses expands the space of functions that can be represented. Hence, by taking a weighted vote of hypotheses, the learning algorithm may be able to form a more accurate approximation to f .

A learning algorithm that suffers from the statistical problem is said to have high *variance*. An algorithm that exhibits the computational problem is sometimes described as having *computational variance*. And a learning algorithm that suffers from the representational problem is said to have high *bias*. Hence, ensemble methods can reduce both the bias and the variance of learning algorithms. Experimental measurements of bias and variance have confirmed this.

Review of Ensemble Algorithms

Ensemble learning algorithms work by running a base learning algorithm multiple times, and forming a vote out of the resulting hypotheses. There are two main approaches to designing ensemble learning algorithms.

The first approach is to construct each hypothesis independently in such a way that the resulting set of hypotheses is accurate and diverse, that is, each individual hypothesis has a reasonably low error rate for making new predictions and yet the hypotheses disagree with each other in many of their predictions. If such an ensemble of hypotheses can be constructed, it is easy to see that it will be more accurate than any of its component classifiers, because the disagreements will cancel out. Such ensembles can overcome both the statistical and computational problems discussed above.

The second approach to designing ensembles is to construct the hypotheses in a coupled fashion so that the weighted vote of the hypotheses gives a good fit to the data. This approach directly addresses the representational problem discussed above.

We will discuss each of these two approaches in turn.

Methods for Independently Constructing Ensembles

One way to force a learning algorithm to construct multiple hypotheses is to run the algorithm several times and provide it with somewhat different training data in each run. For example, Breiman (1996) introduced the bagging (*bootstrap aggregating*) method, which works as follows. Given a set of m training data points, bagging chooses in each iteration a set of data points of size m by sampling uniformly with replacement from the original data points. This creates a resampled data set in which some data points appear multiple times and other data points do not appear at all. If the learning algorithm is *unstable*—that is, if small changes in the training data lead to large changes in the resulting hypothesis—then bagging will produce a diverse ensemble of hypotheses.

A second way to force diversity is to provide a different subset of the input features in each call to the learning algorithm. For example, in a project to identify volcanoes on Venus, Cherkauer (1996) trained an ensemble of 32 neural networks. The 32 networks were based on eight different subsets of the 119 available input features and four different network sizes. The input feature subsets were selected (by hand) to group together features that were based

on different image processing operations (such as principal component analysis and the fast Fourier transform). The resulting ensemble classifier was significantly more accurate than any of the individual neural networks.

A third way to force diversity is to manipulate the output labels of the training data. Dietterich and Bakiri (1995) describe a technique called error-correcting output coding. Suppose that the number of classes, C , is large. Then new learning problems can be constructed by randomly partitioning the C classes into two subsets, A_k and B_k . The input data can then be relabeled so that any of the original classes in set A_k are given the derived label -1 and the original classes in set B_k are given the derived label $+1$. This relabeled data is then given to the learning algorithm, which constructs a classifier h_k . By repeating this process K times (generating different subsets A_k and B_k), an ensemble of K classifiers h_1, \dots, h_K is obtained.

Now, given a new data point \mathbf{x} , how should it be classified? The answer is to have each h_k classify \mathbf{x} . If $h_k(\mathbf{x}) = -1$, then each class in A_k receives a vote. If $h_k(\mathbf{x}) = +1$, then each class in B_k receives a vote. After each of the K classifiers has voted, the class with the highest number of votes is selected as the prediction of the ensemble.

An equivalent way of thinking about this method is that each class j is encoded as a K -bit codeword C_j , where bit k is 1 if $j \in B_k$ and 0 otherwise. The k th learned classifier attempts to predict bit k of these codewords (a prediction of -1 is treated as a binary value of 0). When the L classifiers are applied to classify a new point \mathbf{x} , their predictions are combined into a K -bit binary string. The ensemble's prediction is the class j whose codeword C_j is closest (measured by the number of bits that agree) to the K -bit output string. Methods for designing good error-correcting codes can be applied to choose the codewords C_j (or, equivalently, subsets A_k and B_k). Dietterich and Bakiri (1995) report that this technique improves the performance of both decision-tree and backpropagation learning algorithms on a variety of difficult classification problems.

A fourth way of generating accurate and diverse ensembles is to inject randomness into the learning algorithm. For example, the backpropagation algorithm can be run many times, starting each time from a different random setting of the weights. Decision tree algorithms can be randomized by adding randomness to the process of choosing which feature and threshold to split on. Dietterich (2000) showed that randomized trees gave significantly improved performance on 14 out of 33 benchmark tasks (and no change on the remaining 19 tasks).

Ho (1998) introduced the random subspace method for growing collections of decision trees ("decision forests"). This method chooses a random subset of the features at each node of the tree, and constrains the tree-growing algorithm to choose its splitting rule from among this subset. She reports improved performance on 16 benchmark data sets. Breiman (2001) combines bagging with the random subspace method to grow random decision forests that give excellent performance.

Methods for Coordinated Construction of Ensembles

In all of the methods described above, each hypothesis h_k in the ensemble is constructed independently of the others by manipulating the inputs, the outputs, or the features, or by injecting randomness. Then an unweighted vote of the hypotheses determines the final classification of a data point.

A contrasting view of an ensemble is that it is an *additive model*, that is, it predicts the class of a new data point by taking a weighted sum of a set of component models. This view suggests developing algorithms that choose the component models and the weights so that the weighted sum fits the data well. In this approach, the choice of one component hypothesis influences the choice of other hy-

potheses and of the weights assigned to them. In statistics, such ensembles are known as *generalized additive models* (Hastie and Tibshirani, 1990).

The Adaboost algorithm, introduced by Freund and Schapire (1996, 1997), is an extremely effective method for constructing an additive model. It works by incrementally adding one hypothesis at a time to an ensemble. Each new hypothesis is constructed by a learning algorithm that seeks to minimize the classification error on a *weighted* training data set. The goal is to construct a weighted sum of hypotheses such that $H(\mathbf{x}_i) = \sum_k w_k h_k(\mathbf{x}_i)$ has the same sign as y_i , the correct label of \mathbf{x}_i .

The algorithm operates as follows. Let $d_k(\mathbf{x}_i)$ be the weight on data point \mathbf{x}_i during iteration k of the algorithm. Initially, all training data points i are given a weight $d_1(\mathbf{x}_i) = 1/m$, where m is the number of data points. In iteration k , the underlying learning algorithm constructs hypothesis h_k to minimize the weighted training error. The resulting weighted error is $r = \sum_i d(\mathbf{x}_i) y_i h_k(\mathbf{x}_i)$, where $h_k(\mathbf{x}_i)$ is the label predicted by hypothesis h_k . The weight assigned to this hypothesis is computed by

$$w_k = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

To compute the weights for the next iteration, the weight of training data point i is set to

$$d_{k+1}(\mathbf{x}_i) = d_k(\mathbf{x}_i) \frac{\exp(-w_k y_i h_k(\mathbf{x}_i))}{Z_k}$$

where Z_k is chosen to make d_{k+1} sum to 1.

Breiman (1997) showed that this algorithm is a form of gradient optimization in function space with the goal of minimizing the objective function

$$J(H) = \sum_i \exp(-y_i H(\mathbf{x}_i))$$

The quantity $y_i H(\mathbf{x}_i)$ is called the *margin*, because it is the amount by which \mathbf{x}_i is correctly classified. If the margin is positive, then the sign of $H(\mathbf{x}_i)$ agrees with the sign of y_i . Minimizing J causes the margin to be maximized. Friedman, Hastie, and Tibshirani (2000) expand on Breiman's analysis from a statistical perspective.

In most experimental studies (Freund and Schapire, 1996; Bauer and Kohavi, 1999; Dietterich, 2000), Adaboost (and algorithms based on it) gives the best performance on the vast majority of data sets. The primary exception are data sets in which there is a high level of mislabeled training data points. In such cases, Adaboost will put very high weights on the noisy data points and learn very poor classifiers. Current research is focusing on methods for extending Adaboost to work in high noise settings.

The exact reasons for Adaboost's success are not fully understood. One line of explanation is based on the margin analysis developed by Vapnik (1995) and extended by Schapire et al. (1998). This work shows that the error of an ensemble on new data points is bounded by the fraction of training data points for which the margin is less than some quantity $\Theta > 0$ plus a term that grows as

$$\sqrt{\frac{d}{m}} \frac{\log(m/d)}{\Theta}$$

ignoring constant factors and some log terms. In this formula, m is the number of training data points, and d is a measure of the expressive power of the hypothesis space from which the individual classifiers are drawn, known as the VC-dimension. The value of Θ can be chosen to minimize the value of this expression.

Intuitively, this formula says that if the ensemble learning algorithm can achieve a large "margin of safety" on each training data point while using only a weighted sum of simple classifiers,

then the resulting voted classifier is likely to be very accurate. Experimentally, Adaboost has been shown to be very effective at increasing the margins on the training data points; this result suggests that Adaboost will make few errors on new data points.

There are three ways in which this analysis has been criticized. First, the bound is not tight, so it may be hiding the real explanation for Adaboost's success. Second, even when Adaboost is applied to large decision trees and neural networks, it is observed to work very well even though these representations have high VC-dimension. Third, it is possible to design algorithms that are more effective than Adaboost at increasing the margin on the training data, but these algorithms exhibit worse performance than Adaboost when applied to classify new data points.

Related Nonensemble Learning Methods

In addition to the ensemble methods described here, there are other nonensemble learning algorithms that are similar. For example, any method for constructing a classifier as a weighted sum of basis functions (see, e.g., RADIAL BASIS FUNCTION NETWORKS) can be viewed as an additive ensemble where each individual basis function forms one of the hypotheses.

Another closely related learning algorithm is the hierarchical mixture-of-experts method (see MODULAR AND HIERARCHICAL LEARNING SYSTEMS). In a hierarchical mixture, individual hypotheses are combined by a gating network that decides, based on the features of the data point, what weights should be employed. This differs from Adaboost and other additive ensembles, where the weights are determined once during training and then held constant thereafter.

Discussion

The majority of research into ensemble methods has focused on constructing ensembles of decision trees. Decision tree learning algorithms are known to suffer from high variance, because they make a cascade of choices (of which variable and value to test at each internal node in the decision tree) such that one incorrect choice has an impact on all subsequent decisions. In addition, because the internal nodes of the tree test only a single variable, this creates axis-parallel rectangular decision regions that can have high bias. Consequently, ensembles of decision tree classifiers perform much better than individual decision trees. Recent experiments suggest that Breiman's combination of bagging and the random subspace method is the method of choice for decision trees: it gives excellent accuracy and works well even when there is substantial noise in the training data.

If the base learning algorithm produces less expressive hypotheses than decision trees, then the Adaboost method is recommended. Many experiments have employed so-called decision stumps, which are decision trees with only one internal node. In order to learn complex functions with decision stumps, it is important to exploit Adaboost's ability to directly construct an additive model. This usually gives better results than bagging and other accuracy/diversity methods. Similar recommendations apply to ensembles constructed using the naive Bayes and Fisher's linear discriminant algorithms. Both of these learn a single linear discrimination rule. The algorithms are very stable, which means that even substantial (random) changes to the training data do not cause the learned discrimination rule to change very much. Hence, methods like bagging that rely on instability do not produce diverse ensembles.

Because the generalization ability of a single feedforward neural network is usually very good, neural networks benefit less from ensemble methods. Adaboost is probably the best method to apply, but favorable results have been obtained just by training several networks from different random starting weight values, and bagging is also quite effective.

For multiclass problems, the error-correcting output coding algorithm can produce good ensembles. However, because the output coding can create difficult two-class learning problems, it is important that the base learner be very expressive. The best experimental results have been obtained with very large decision trees and neural networks. In addition, the base learning algorithm must be sensitive to the encoding of the output values. The nearest neighbor algorithm does not satisfy this constraint, because it merely identifies the training data point \mathbf{x}_i nearest to the new point \mathbf{x} and outputs the corresponding value y_i as the prediction for $h(\mathbf{x})$, regardless of how y_i is encoded. Current research is exploring ways of integrating error-correcting output codes directly into the Adaboost algorithm.

Road Map: Learning in Artificial Networks

Related Reading: Modular and Hierarchical Learning Systems; Radial Basis Function Networks

References

- Bauer, E., and Kohavi, R., 1999, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learn.*, 36:105–139.
- Breiman, L., 1996, Bagging predictors, *Machine Learn.*, 24:123–140. ♦
- Breiman, L., 1997, *Arcing the Edge*, Technical Report 486, Department of Statistics, University of California, Berkeley. Available: <http://citeseer.nj.nec.com/breiman97arcing.html>.
- Breiman, L., 2001, Random forests, *Machine Learn.*, 45:5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., 1984, *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks. ♦
- Cherkauer, K. J., 1996, Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks, in *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models* (P. Chan, Ed.), Menlo Park, CA: AAAI Press, pp. 15–21.
- Dietterich, T. G., 2000, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, *Machine Learn.*, 40:139–158.
- Dietterich, T. G., 1997, Machine learning research: Four current directions, *AI Magazine*, 18:97–136. ♦
- Dietterich, T. G., and Bakiri, G., 1995, Solving multiclass learning problems via error-correcting output codes, *J. Artif. Intell. Res.*, 2:263–286.
- Freund, Y., and Schapire, R. E., 1996, Experiments with a new boosting algorithm, in *Proceedings of the 13th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, pp. 148–156.
- Freund, Y., and Schapire, R. E., 1997, A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.*, 55:119–139.
- Friedman, J. H., Hastie, T., and Tibshirani, R., 2000, Additive logistic regression: A statistical view of boosting, *Ann. Statist.*, 28:337–407. ♦
- Hastie, T. J., and Tibshirani, R. J., 1990, *Generalized Additive Models*, London: Chapman and Hall. ♦
- Ho, T. K., 1998, The random subspace method for constructing decision forests, *IEEE Trans. Pattern Anal. Machine Intell.*, 20:832–844.
- Quinlan, J. R., 1993, *C4.5: Programs for Empirical Learning*, San Francisco: Morgan Kaufmann. ♦
- Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S., 1998, Boosting the margin: A new explanation for the effectiveness of voting methods, *Ann. Statist.*, 26:1651–1686.
- Vapnik, V., 1995, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Equilibrium Point Hypothesis

Reza Shadmehr

Introduction

If one were to take a robot arm and replace each of its motors with a pair of opposing rubber bands, the arm would tend to settle to the same configuration, regardless of where it was released. That configuration is the *equilibrium point* of the system. If we now change the length-tension properties of the rubber bands, such as by changing the resting lengths or stiffnesses, the equilibrium point of the system will change. Our muscles share a property with rubber bands in that the static force they generate depends on length: the greater the length, the greater the force (see *MUSCLE MODELS*). The activations received by motor neurons, whether from direct descending commands from the brain or from the spinal reflex circuitry, can change the force-length relation for each muscle, resulting in a change in the equilibrium position of the system. When we reach for an object, is the smooth, stable motion a consequence of a simple trajectory of equilibrium points? Are our muscles and the associated spinal reflex circuitry designed in a way that makes control of motion particularly simple for the brain?

If the answer is yes, then it implies that many of the problems inherent to control of a multijoint limb, such as nonlinear state-dependent dynamics, might be simplified by a well-designed low-level control system. In this article I review the evidence regarding this hypothesis.

Mathematical Basis of the Hypothesis

Equilibrium refers to a state of a system in which the forces acting on it are zero. For example, if the dynamics of the system are

$$\dot{q} = h(q, u) \quad (1)$$

where q is the state of the system and $u(t)$ is a control input, then the equilibrium points q^* satisfy the following condition:

$$0 = h(q^*, u) \text{ for all } t \geq t_0$$

In short, if the system reaches an equilibrium position, it will remain there.

For a mechanical system, the state is an ordered pair $q = \{\theta, \dot{\theta}\}$, where θ and $\dot{\theta}$ are the position and velocity of the system. A change in the state occurs when there are forces acting on it. This can be written in the framework of Equation 1 as

$$\ddot{\theta} = I(\theta)^{-1}(f_c(\dot{\theta}, \theta, u(t)) - f_m(\dot{\theta}, \theta)) \quad (2)$$

where I is the system's inertia, f_c is the external force field imposed on the system due to the controller with control input $u(t)$, and f_m is the force field produced by the motion of the inertial coordinate frames (Coriolis and centripetal) and other forces. It follows that the system is at equilibrium at any state $\{\theta = 0, \dot{\theta} = 0\}$ where the force in the net field $f_c - f_m$ is zero. Any such position θ^* is an equilibrium point for the system.

We call each state where a field has zero force a *null point* of that field. The equilibrium points for the system, however, are a subset of these null points: the equilibrium point exists only at those null points of the force field $f_c - f_m$ where the state has zero velocity.

Let us consider how we could go about controlling the system of Equation 2. Our objective may be to select the input u in such a way that the system follows a desired trajectory $\theta_d(t)$. For this to occur, we might select u at any time t in such a way that if we were at state $\{\theta_d, \dot{\theta}_d\}$, our controller would produce a force $f_c = \hat{f}_m + \hat{I}\ddot{\theta}_d(t)$, where \hat{x} is the controller's estimate of x . Since there may be

uncertainties in the environment, it is a good idea to also have a mechanism to push us toward where we should be if the need arises:

$$f_c = \hat{f}_m + \hat{I}\ddot{\theta}_d - B(\dot{\theta} - \dot{\theta}_d) - K(\theta - \theta_d) \quad (3)$$

where B and K should be positive definite matrices. We can think of the estimates as a feedforward component of the controller and the remainder as the feedback component of the controller. If the estimates were perfect, substitution of Equation 3 into Equation 2 would give:

$$\ddot{e} + c_1\dot{e} + c_2e = 0$$

where $e = \theta - \theta_d$ is the error in tracking our desired trajectory, and c_1 and c_2 are positive definite (because the inertia matrix I is also positive definite for a mechanical system). Therefore, the tracking error would exponentially decline with time and the system would be stable about the desired trajectory.

Equation 3 makes plain the notion that the forces produced by the controller must take into account the system's mass if it is to move the system along the desired trajectory. The estimates are *internal models* that the brain would presumably have to know (see *MOTOR CONTROL, BIOLOGICAL AND THEORETICAL* for a discussion of how these models might be learned). However, the equilibrium point hypothesis suggests that the feedback system in Equation 3 is designed in a way that largely eliminates the need for the estimates of the dynamics of the limb. In this hypothesis, the muscles and the spinal reflexes function as the feedback system about the desired trajectory, i.e., the stiffness and viscosity of the system. The main question is the extent to which the mechanical behavior of muscles and the reflex system can compensate for the dynamics of the limb.

Biomechanical Behavior at Rest

In a seminal paper by Feldman (1966), it was observed that the spinal control system acting on the elbow joint of the human arm (composed of muscles and the local feedback circuitry) had static characteristics similar to those of a nonlinear spring. When the elbow was displaced from its equilibrium position, muscles produced monotonically increasing force (as measured at the hand):

$$f = a(\exp[b(x(t) - x_2(t))] - 1) \quad (4)$$

where f is muscle force, t is time, $x(t)$ is length of a muscle, and $x_2(t)$ is the threshold length beyond which the muscle will produce force. Feldman's thesis was that the signals sent from the brain to the spinal reflexes and muscles could be interpreted as setting the threshold length $x_2(t)$ for each muscle. Feldman and Orlovsky (1972) later showed that stimulation of a motor center in the brainstem (of cats), resembling what might happen in a voluntary change in the brain's input to the spinal cord, did result in force-length changes in the muscles. These changes appeared as changes in $x_2(t)$ in the above system.

For a constant input $x_2(t)$ in Equation 4, muscle force reflects both the mechanical properties of the isolated muscles (increased production of force when muscle is lengthened) and the effect of local neural feedback (recruitment of more motor neurons if length exceeds a set threshold). There is now independent support for the formulation in Equation 4. Hoffer and Andreassen (1981) measured the rate of change in stiffness with respect to force in muscles of a cat's hindlimb. They found the relation between force and stiffness to be independent of muscle length, and of the form:

$$\frac{df}{dx} = k(1 - \exp[-\alpha f]) \quad (5)$$

where df/dx is muscle stiffness. Shadmehr and Arbib (1992) noted that the solution to the above differential equation has the form:

$$f = \frac{1}{\alpha} \ln(\exp[\alpha k(x - \lambda)] + 1) \quad (6)$$

In the above, λ is the constant of integration and depends on the initial conditions for Equation 5. This result demonstrated that an intact muscle reflex system has a static behavior that resembles that of a nonlinear spring with an adjustable threshold.

If a single-joint limb is controlled by a pair of muscles, then setting λ for each muscle sets the equilibrium point of the system and describes a force field about this equilibrium. Hogan (1985) showed that in a multijoint system, this field will be conservative. This means that if the nervous system produces a force field f_c in Equation 3 through setting of threshold lengths for the muscles of the limb, then when $\dot{\theta}$ and $\ddot{\theta}$ are zero, curl of the field f_c should be zero. Mussa-Ivaldi, Hogan, and Bizzi (1985) measured the static component of f_c in humans. The procedure was to have subjects hold on to the handle of a robotic arm. The robot produced force perturbations at various directions and measured the steady-state force response of the subject's arm as a function of position. It was found that the resulting force field was essentially curl-free. Taken together, static behavior of muscles and the spinal control circuitry appeared to be well described as a nonlinear spring with an adjustable threshold length.

Movements as a Shift in Equilibrium Position

When threshold lengths are set for each muscle, the result is a corresponding equilibrium position θ^* for the limb. The major contribution of the equilibrium point hypothesis has been to suggest that motion is generated by the CNS through a gradual transition of equilibrium points along the desired trajectory without an explicit compensation for dynamics. The evidence for this initially came from a simulation study by Flash. She suggested that in the case of human reaching movements in the horizontal plane, it was possible to predict the hand's motion accurately by smoothly shifting the equilibrium point along a straight line from the start point to a target location. Interestingly, she showed that in the simulation, because the controller was not attempting to compensate for the limb's dynamics, the hand's trajectory slightly deviated from a straight line. However, it turns out that the trajectories recorded in human subjects also show similar deviations, matching her simulations. In this model, the controller was composed of a linear spring-dashpot system with adjustable threshold:

$$f_c = K(\theta - \theta^*(t)) + B\dot{\theta}$$

The field had the property that its static behavior about equilibrium was defined by a stiffness matrix K . This matrix was measured about the equilibrium position of a resting arm by Mussa-Ivaldi et al. (1985).

Taking a different approach, Shadmehr, Mussa-Ivaldi, and Bizzi (1993) suggested that if a movement was generated through a gradual shift of the equilibrium position toward the target, then from measurements of the force field about the hand at rest, one should be able to predict the direction and magnitude of forces that should be produced by the muscles during the initiation of the reaching movement (Figure 1). Because the field at rest is not isotropic and depends on the position of the hand, the forces measured during the initiation of a movement should not point toward the target and be position dependent. These movement initiation forces were measured, and it was found that the pattern of forces from measure-

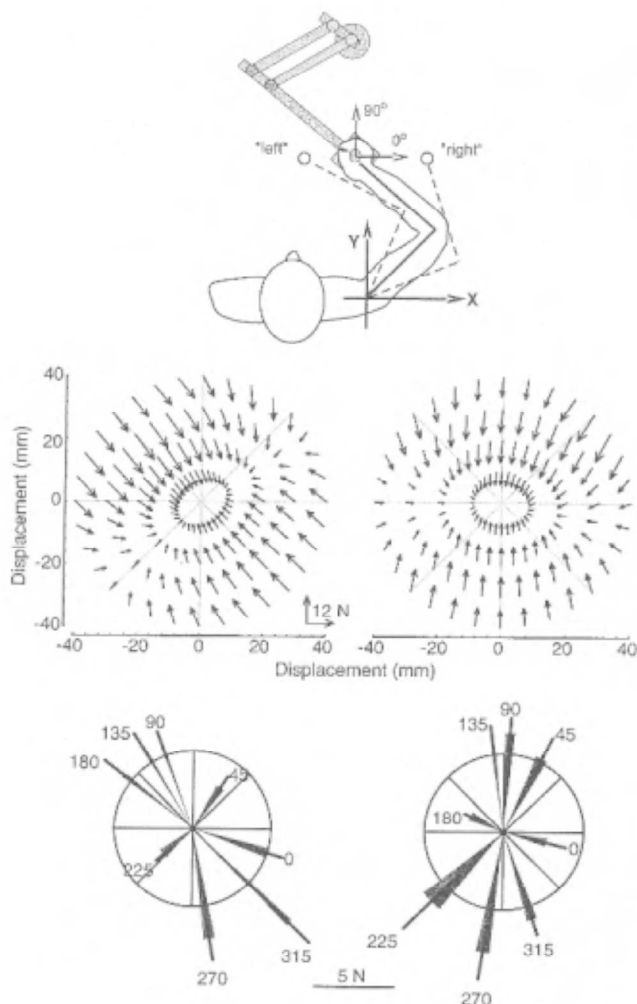


Figure 1. Subject was seated in front of a robotic arm and instructed to hold the handle at either the *right* or *left* configuration. Robot slowly displaced the hand from the origin and measured restoring forces. These forces represent the static component of the force field produced by the muscles, i.e., the postural field. Note the anisotropic shape. Now the subject is told to reach to a target. For randomly selected targets, the robot prevents initiation of the movement (applies a break) and measures the force that the subject is generating in order to make the movement. The magnitude and standard deviations of the movement-related forces are plotted for targets at 0°, 45°, . . . , 315°. The magnitude and direction of movement-related forces are in agreement with the hypothesis that movement is generated through a shift of the equilibrium position of the postural force field toward the target. (From Shadmehr, R., Mussa-Ivaldi, F. A., and Bizzi, R., 1993, Postural force fields of the human arm and their role in generating multi-joint movements, *J. Neurosci.*, 13:45–62. Reprinted with permission.)

ments at rest agreed with the measured forces during initiation of movement. In other words, during the start of movements, the equilibrium point of the field had shifted toward the target.

Won and Hogan (1995) went a step further and suggested that during the entire movement, the static component of the field f_c should be similar to that measured when the hand was at rest; i.e., it should converge to an equilibrium position. In their experiment, the hand was displaced from its intended trajectory via a rigid mechanical constraint. It was shown that as the arm was being displaced, it produced forces directed toward the intended trajectory.

The notion of stability about a trajectory (Equation 3) was clearly demonstrated in the data as the controller's output during movement was a force field with an equilibrium point moving roughly along the path connecting the start to the target position.

Dynamics of the Muscle-Reflex System During Movement

Katayama and Kawato (1993) noted that the simulations by Flash used a magnitude of stiffness K that was approximately three times that measured when the arm was at rest. This correctly highlighted the fact that a very stiff system has no need to take into account dynamics of the system in generating its motor output. Although the actual stiffness of the arm was a crucial factor in the simulations, its actual value during movement was unknown, and its estimation had been difficult. Bennett et al. (1992) had found that stiffness during a highly practiced movement was significantly less than that measured when the hand was at rest, while Milner (1993) had found a value that was near the rest levels. It seemed clear that accurate measures of the arm's stiffness during motion were required.

Gomi and Kawato (1996) designed a high-performance robotic manipulandum and measured the arm's stiffness during motion. They found that the stiffness of the arm was near those measured at rest but was temporally modulated about this level during motion. They used measures of local stiffness to estimate the point of convergence of the static component of field f_c by assuming that the static muscle forces were linearly related to distance from equilibrium. Gomi and Kawato concluded that motion of the arm could not be due to a simple shift of the equilibrium point along the desired trajectory. This suggested that ultimately, control of motion required explicit compensation for dynamics of the limb.

The crucial question in the work of Gomi and Kawato (1996) was how to estimate the null point of a force field from local measures of stiffness. Most if not all of the experimental data on intact muscle reflex systems describe only the static behavior, as in Equations 4 and 6. Gomi and Kawato showed that if the dynamic behavior of the muscle reflex system is dominated by its static properties, then it is unlikely that the brain can produce a desired movement via a simple shift of the equilibrium point of the system. But what about the dynamic properties of the muscle reflex system? How do they contribute to control?

Gribble et al. (1998) approached this question by modifying Equation 4 to include the effect of delayed sensory feedback on recruitment of motor neurons, and dependence of muscle force on velocity of contraction and temporal summation of activations. The result was a muscle reflex model that, as before, was controlled via a threshold muscle length, and had a static behavior that remained similar to Equation 4, but was now a complex dynamical system. Remarkably, it was found that if the threshold lengths of the muscles acting on a simulated two-joint arm were shifted along a smooth desired trajectory to the target, the resulting motion was also a smooth trajectory. Furthermore, the local stiffness of the system about the actual trajectory was very similar to that reported by Gomi and Kawato (1996). This suggested that the dynamical behavior of the muscle reflex system was a crucial element in compensating for the arm's dynamic, and that the input to the system might change rather simply from a starting location to a desired end point in order to produce a smooth hand trajectory.

Hodgson and Hogan (2000) performed an elegant experiment that appears to resolve this issue. Rather than estimating a limb's equilibrium position from local stiffness properties and extrapolating to its null point, they perturbed the limb until its equilibrium point was found. Their results (Figure 2) clearly demonstrated that for simple reaching movements, the attractor trajectory was not along the actual trajectory but led it considerably. Therefore, this

demonstrates that while the limb has simple reflex mechanisms that stabilize it in the case of perturbation, the system is not stiff enough that its input can simply ignore dynamics.

Learning and Modulation in Stiffness Properties of the Limb

The motor commands that act on muscles produce not only force but also an attractor that stabilizes the limb about a trajectory. This is a fundamental property that is crucial for control of our movements because there exists a significant delay in transmission of sensory information from the limbs to our brain. The delay results in multiple levels of feedback: muscles respond with almost no delay with increased force as they are stretched; muscle spindles sense this stretch and activate spinal reflex pathways that enhance this force production with a delay of approximately 30 ms (termed "short-loop" reflexes); the afferent information is conveyed to the thalamus and then to the motor cortex, and the brain responds to the muscle stretch by altering the descending commands, affecting the muscle in approximately 100 ms (termed "long-loop" reflexes). Therefore, the stiffness of the limb and the behavior of the attractor have a time-dependent component. The brain has the ability to modulate the attractor (force response to a displacement) at a time scale of about 100 ms.

Two recent experiments demonstrate that the brain can modulate the long-loop reflexes so as to match the properties of the task. Burdet et al. (2001) asked subjects to make reaching movements while holding a robot that produced a field that was zero along a straight line connecting the start point to end point, but pushed the hand away if the hand strayed from the straight line. Effectively, the hand was traveling along a knife's edge. They demonstrated that the stiffness of the arm increased only along the dimension where the forces were acting (perpendicular to the direction of motion). The results of Wang, Dordevic, and Shadmehr (2001) in subjects who also learned to reach in force fields suggest that this modulation is limited to the long-latency component of the reflex. Therefore, not only do stiffness properties of the limb produce an attractor trajectory that stabilizes the limb during movements, but the brain can also modulate the shape of the restoring forces about the attractor to match the properties of the task.

Summary

Biological muscles are spring-like systems. It was thought that because of their elastic behavior, the CNS might simply describe motor commands in terms of the resting lengths of these springs, effectively producing trajectories in terms of equilibrium points of the system. These commands would ignore the inertial dynamics

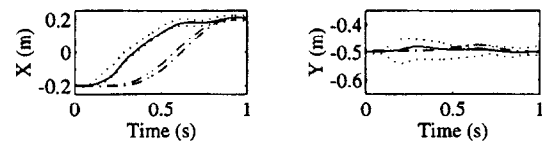


Figure 2. The hand is making a reaching movement while holding a robotic arm. In some cases, the movement is perturbed so that the robot takes the arm along the current estimate of the equilibrium trajectory. The process is repeated until the equilibrium trajectory is found. The bounds for the actual movement are shown by the dash-dot lines. The bounds for the equilibrium trajectory are shown by the dotted lines. The estimate for the equilibrium trajectory is shown by the solid line. Note the large distance by which the equilibrium trajectory leads the actual trajectory. (From Hodgson, A. J., and Hogan, N., 2000, A model-independent definition of attractor behavior applicable to interactive tasks, *IEEE Trans. Syst. Man Cybern.*, 30:105–118. Reprinted with permission.)

of the limb, simplifying the process of control. However, there is now convincing evidence that in programming motor commands to the muscles, the brain does take into account the dynamics of the task (see **SENSORIMOTOR LEARNING**). The motor commands result in an attractor trajectory that leads the hand in simple reaching movements. It would be expected that both the trajectory of the attractor and the shape of the restoring field about the attractor would change as the dynamics of the task change. Although it is quite possible that biological motor commands can be described in terms of changes in the equilibrium position of the limb, the hypothesis that control of movements by the brain is explicitly performed through this manipulation because it somehow simplifies the process of control appears to be inconsistent with the current data.

Road Maps: Mammalian Motor Control; Dynamic Systems

Related Reading: Arm and Hand Movement Control; Cerebellum and Motor Control; Geometrical Principles in Motor Control; Limb Geometry; Neural Control; Muscle Models; Optimization Principles in Motor Control

References

- Bennett, D. J., Hollerbach, J. M., Xu, Y., and Hunter, I. W., 1992, Time varying stiffness of the human elbow joint during cyclic voluntary movement, *Exp. Brain Res.*, 88:433–442.
- Burdet, E., Osu, R., Franklin, D. W., Milner, T. E., and Kawato, M., 2001, The central nervous system stabilizes unstable dynamics by learning optimal impedance, *Nature*, 414:446–449.
- Feldman, A. G., 1966, Functional tuning of the nervous system with control of movement or maintenance of a steady posture: II. Controllable parameters of the muscles, *Biophysics*, 11:565–578. ♦
- Feldman, A. G., and Orlovsky, G. N., 1972, The influence of different descending system on the tonic stretch reflex in the cat, *Exp. Neurol.*, 37:481–494.
- Gomi, H., and Kawato, M., 1996, Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement, *Science*, 272:117–120. ♦
- Gribble, P. L., Ostry, D. J., Sanguineti, V., and LaBoissiere, R., 1998, Are complex control signals required for human arm movement? *J. Neurophysiol.*, 79:1409–1424.
- Hodgson, A. J., and Hogan, N., 2000, A model-independent definition of attractor behavior applicable to interactive tasks, *IEEE Trans. Syst. Man Cybern.*, 30:105–118. ♦
- Hoffer, J. A., and Andreassen, S., 1981, Regulation of soleus muscle stiffness in premammillary cats: Intrinsic and reflex components, *J. Neurophysiol.*, 45:267–285.
- Hogan, N., 1985, The mechanics of multi-joint posture and movement control, *Biol. Cybern.*, 52:315–331.
- Katayama, M., and Kawato, M., 1993, Virtual trajectory and stiffness ellipse during multijoint arm movements predicted by neural inverse models, *Biol. Cybern.*, 68:353–362.
- Milner, T. E., 1993, Dependence of elbow viscoelastic behavior on speed and loading in voluntary movements, *Exp. Brain Res.*, 93:177–180.
- Mussa-Ivaldi, F. A., Hogan, N., and Bizzi, E., 1985, Neural, mechanical, and geometric factors subserving arm posture, *J. Neurosci.*, 5:2732–2743. ♦
- Shadmehr, R., and Arbib, M. A., 1992, A mathematical analysis of the force-stiffness characteristics of muscles in control of a single joint system, *Biol. Cybern.*, 66:463–477.
- Shadmehr, R., Mussa-Ivaldi, F. A., and Bizzi, E., 1993, Postural force fields of the human arm and their role in generating multi-joint movements, *J. Neurosci.*, 13:45–62.
- Wang, T., Dordevic, G. S., and Shadmehr, R., 2001, Learning dynamics of reaching movements results in the modification of arm impedance and long-latency perturbation responses, *Biol. Cybern.*, 85:437–448.
- Won, J., and Hogan, N., 1995, Stability properties of human reaching movements, *Exp. Brain Res.*, 107:125–136.

Event-Related Potentials

Steven L. Bressler

Introduction

It is commonly believed that cognition intimately depends on the functioning of the cerebral cortex. Understanding the neural basis of cognition therefore will likely require knowledge of cortical operations at all organizational levels, which may usefully be grouped as microscopic, mesoscopic, and macroscopic. The cellular mechanisms of cortical neurons operate at the microscopic scale and are measured by a host of techniques targeted at that level. Individual cortical neurons contribute to cognitive function, however, by joining in the cooperative actions of neural networks, which operate at the mesoscopic and macroscopic scales. At the microscopic scale, the cooperative fraction of any single neuron's total activity may be exceedingly small, but the cooperative activity of the network exerts effects that are relevant for cognition. The mesoscopic level concerns the cooperative activity of neurons locally in ensembles and area networks, and the macroscopic level concerns the cooperative activity of neurons globally in large-scale networks and entire systems. Thus, many important cortical functions reside in the operations of neural networks and are measured by specialized techniques targeted at the mesoscopic and macroscopic levels.

The event-related potential (ERP) is a neural signal that reflects coordinated neural network activity. The cortical ERP provides a window onto the dynamics of network activity in relation to a va-

riety of different cognitive processes at both mesoscopic and macroscopic levels on a time scale comparable to that of single-neuron activity. Cortical ERPs arise from synchronous interactions among large numbers of participating neurons. These include dense local interactions involving excitatory pyramidal neurons and inhibitory interneurons, as well as long-range interactions mediated by axonal pathways in the white matter. (See **NEUROANATOMY IN A COMPUTATIONAL PERSPECTIVE**.) Multiple feedback loops involving both excitatory and inhibitory interactions typically cause ERPs to be oscillatory, meaning that they fluctuate within bounds around a central value. Depending on the types of interaction that occur in a specific behavioral condition, cortical networks may display different states of synchrony, causing their ERPs to oscillate in different frequency bands, designated delta (0–4 Hz), theta (5–8 Hz), alpha (9–12 Hz), beta (13–30 Hz), and gamma (31–100 Hz).

The physiological basis of the cortical ERP lies in fields of potential generated by interacting neurons (Lopes da Silva, 1991). Field potentials are largely dendritic in origin, resulting from the summed extracellular currents generated by electromotive forces (EMFs) in the dendrites of synchronously active cortical neurons, primarily pyramidal cells. The EMFs, arising from synaptic activation of postsynaptic ion channels, circulate current in closed loops across the cell membrane and through the intracellular and extracellular spaces. Summed closed-loop currents generated by an ensemble of neighboring neurons flow across the external resis-

tance to form the local ensemble mean field potential (Freeman, 2000).

Depending on the location and size of the recording and reference electrodes, recorded cortical field potentials integrate neural activity over a range of spatial scales: from the intracortical local field potential (LFP) to the intracranial electrocorticogram (ECoG) to the extracranial electroencephalogram (EEG). The LFP (Figure 1) is the most spatially localized signal, integrating the field potential on a submillimeter scale; the ECoG integrates on a submillimeter to millimeter scale; and the EEG integrates over centimeters. The term “field potential” will be used here in reference to the general class of signal subsuming the LFP, ECoG, and EEG. (The intracellular components of the same closed-loop currents that give rise to field potentials are responsible for the closely related magnetic fields, recorded extracranially as the magnetoencephalogram, or MEG.)

A general problem in the investigation of ERPs is that field potential recordings most often contain a combination of potentials, in unknown proportions, from multiple sources. Thus, in addition to the ERP, which is derived from specific networks associated with a behavioral event, the field potential typically also contains potentials derived from the more general field activity of large neural populations. Owing to their fortuitous geometric arrangements and synchronous behavior, these later potentials are mixed with the ERP waveform. Thus, a primary task of all ERP studies is to extract the event-related portion of the recorded field potential. The next section deals with some basic methodology by which this is accomplished for different kinds of ERP.

ERP Varieties and Their Analysis

Whether reflecting mesoscopic or macroscopic activity, the cortical ERP is an electrical signal generated by neuronal networks in relation to a behaviorally significant event. (The corresponding event-related magnetic field has many of the same dynamic and functional properties as the ERP.) Two general classes of ERP are distinguished by whether the relevant event is discrete or continuous. In the case of discrete events, the associated transient ERP is analyzed in short epochs that are time-locked to the event. In the case of continuous events, which usually are periodically modulated sensory stimuli such as a visual flicker, the concurrent steady-state ERP is analyzed in a relatively long time segment.

The traditional approach to the analysis of transient ERPs is to consider the ERP as a characteristic waveform that occurs in relation to the behaviorally significant discrete event. As a simplifying assumption, the ERP waveform is usually treated as if it possesses the same amplitude and phase each time that the event is repeated on multiple trials, although recent analysis shows that

this assumption may not always be valid (Truccolo et al., 2002). Nonetheless, as was discussed above, the recorded single-trial field potential contains contributions from network activity that are both associated (ERP signal) and not associated (noise) with the event. Therefore, averaging of the single-trial field potential time series, time-locked to the event, is commonly employed to extract the ERP from the non-event-related noise. When the relevant event is a sensory stimulus, such phase-locked ERPs are called “evoked.” Averaged evoked potentials (Figure 2) are most commonly described in terms of the succession of waveform components that follow stimulus presentation. These components are typically identified according to their polarity (positive or negative) and their time latency following stimulus onset. (Note that the time latency is equivalent to phase in this context.)

Transient ERP waveform components having variable phase may also reliably occur in relation to the repeated event. In this case, time series averaging does not reveal the ERP but instead is destructive, since components of opposite polarity on successive trials tend to be canceled. Non-phase-locked ERPs are referred to as “induced” when they occur following a stimulus and “spontaneous” in the period prior to a stimulus or motor response. This type of ERP may be effectively analyzed by averaging the frequency content of single-trial time series rather than the time series themselves.

Non-phase-locked transient event-related phenomena are detected as frequency-specific changes in the ERP time series. These phenomena may consist of either an event-related increase or decrease of power in one or more of the aforementioned frequency bands. Since the level of ERP power is typically considered to reflect the degree of synchrony within local neuronal populations, a power increase is called event-related synchronization, and a power decrease is called event-related desynchronization (Pfurtscheller and Lopez da Silva, 1999). Frequency analysis has the further advantage of allowing measurement of event-related phase synchronization of ERPs from different cortical sites (Varela et al., 2001). ERP phase synchronization in different frequency ranges has been identified as a fundamental neural correlate of basic sensory and motor processes, as well as higher cognitive processes such as perception and recall of semantic entities. (See SYNCHRONIZATION, BINDING AND EXPECTANCY.)

The study of steady-state ERPs also depends on a variant of frequency analysis. Field potentials recorded during periodically modulated sensory stimulation are narrow-bandpass filtered around the frequency of the driving periodicity to derive the steady-state (periodic) ERPs. Variations in the amplitude and phase of the steady-state ERP are interpreted in terms of driving frequency, spatial location, and behavioral state.

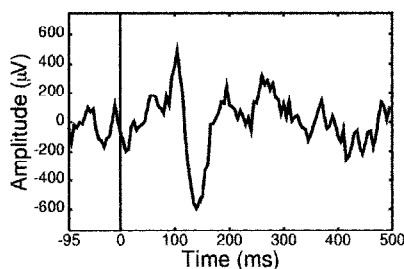


Figure 1. A local field potential (LFP) recorded from the posterior parietal cortex of a macaque monkey in relation to a visual stimulus presented on a display screen for 100 ms, starting at time 0. The LFP was recorded from a chronically implanted bipolar transcortical electrode consisting of 51- μ m-diameter Teflon-coated platinum wires with 2.5-mm tip separation.

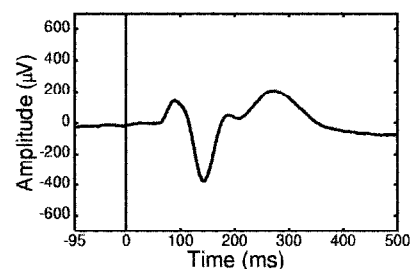


Figure 2. The averaged event-related potential from the same posterior parietal cortex site as in Figure 1, computed from an ensemble of 888 trials. Note the flat prestimulus baseline as compared to the single trial in Figure 1. This illustrates the fact that rhythmic prestimulus activity that is not phase-locked to the stimulus is canceled out by the averaging process.

The Theory of Large-Scale Cortical Networks

Evidence from a variety of sources indicates that neural networks in the cerebral cortex are organized both locally in anatomically segregated areas and on a large scale encompassing multiple distributed areas (Bressler, 2002). Although research on cortical network properties is still in its infancy, a rough depiction of some basic operational features is now possible. Local-area networks process and store information related to specialized sensory, motor, and executive functions, and local synaptic interactions lead to the manifestation of coherent spatial ERP patterns in these specialized informational domains. These interactions also modify the local synaptic matrix with learning. The modified synaptic matrix exerts an essential control on pattern formation in the local-area network by attracting its dynamics to learned (attractor) patterns. In this regard, artificial neural networks that operate according to attractor dynamics bear a resemblance to cortical networks at the local level. (See COMPUTING WITH ATTRACTORS.)

An essential element of overall cortical network function, however, is missing from most artificial network models. Following training on pattern recognition problems, traditional artificial neural networks converge to fixed solutions for a given class of input patterns. Although this behavior has well-known advantages for pattern recognition, it represents an excessive processing rigidity, since these networks lack the ability to adapt to changing external constraints such as are found in real-world situations. Adaptability, in this sense, is a distinguishing feature of normal cortical function.

Theoretical considerations suggest that processing adaptability in the cerebral cortex derives from an essential property of large-scale network dynamics called *metastability*. Cortical metastability refers to a state of dynamic balance among multiply interacting local networks in which the tendency for independent local expression is offset by the tendency for large-scale entrainment (Bressler and Kelso, 2001). The property of metastability permits local networks that are interconnected within the large-scale network architecture of the cortex to coordinate their activities without becoming locked in a fixed pattern of coordination from which they cannot escape.

The ability of local-area networks to form transient coordination relations may represent a basic cortical mechanism for the rapid and flexible association of information from different informational domains. (See ASSOCIATIVE NETWORKS.) It is to be expected that the concurrent coordination of multiple local-area networks imposes conjoint constraints on the spatiotemporal patterning of activity in each local network. The imposition of such constraints may have the important effect of creating associations between activity patterns in different informational domains during the learning process, through the modification of synapses of axons that project from one local network to another. These learned associations would then act during recall on the attractor dynamics of multiple interacting local area networks, causing them to reach a conjunction of consensual patterns that represents an integration of their information. (See COMPUTING WITH ATTRACTORS.)

ERP Evidence for Large-Scale Cortical Network Organization

The theoretical considerations presented in the previous section lead to predictions about the large-scale cortical network organization underlying cognition. One straightforward prediction is that cognitive states should be characterized by unique configurations of interdependent cortical areas in large-scale networks. A confirmation of this prediction is found in the spatial patterning of coactivated cortical areas seen with functional brain imaging techniques such as PET and fMRI. (See COVARIANCE STRUCTURAL EQUATION

MODELING.) Like these neuroimaging procedures, ERPs can provide information about the spatial distribution of large-scale network activity underlying a cognitive function. Moreover, because ERPs reflect neurodynamics on a fast time scale (that is inaccessible to current brain imaging technologies), ERPs can also reveal elementary neural subprocesses that subserve that cognitive function. This section uses working memory to illustrate how ERP results can relate large-scale network activity to different subprocesses of a cognitive function.

Working memory consists of several subprocesses for which prominent averaged ERP waveform components have revealed distinct underlying large-scale networks (McEvoy et al., 1998). The mismatch negativity is an early poststimulus ERP component that reflects the maintenance of sensory working memory in the auditory modality. It is elicited by auditory stimuli having physical acoustic properties that deviate from prior (standard) stimuli registered in auditory memory. Occurring between 80 and 200 ms after presentation of deviant auditory stimuli, thus overlapping the N1 and P2 components, the mismatch negativity is isolated by computing the difference wave between averaged ERPs evoked by deviant and standard stimuli. The mismatch negativity is subserved by a large-scale network that includes, in addition to auditory cortical areas, dorsolateral prefrontal cortex, which may serve to control the maintenance of sensory memory in the auditory cortex following one stimulus for comparison with subsequent stimuli (Alain et al., 1998).

A second ERP component, the P3b, occurring roughly 300 ms poststimulus, also results from the comparison of target stimuli with the content of working memory. However, rather than being tuned to the physical characteristics of stimuli, the widely distributed cortical network underlying the P3b is involved in the categorization of stimuli as significant events. Network strength has been found to reflect the degree of consonance resulting from comparison of stimulus attributes with a maintained "expectation" (Kok, 2001).

A third ERP component, related to semantic memory, is the negative-going N400. It occurs between 200 and 500 ms after presentation of a potentially meaningful information-bearing stimulus and varies systematically according to the preexisting context that is established by semantic and long-term memory influences. Specifically, N400 amplitude is reduced as a function of associative, semantic, and repetition priming within or across sensory modalities (Kutas and Federmeier, 2000). Variation of its scalp-recorded topographic distribution with task and stimulus type suggests that the N400 reflects the construction of meaning by cross-modal interactions in a widely distributed neural network. This view is supported by intracranial evidence that the N400 arises from similar waves of activity in multiple brain areas, particularly in the temporal and prefrontal cortices, during the retrieval of information from semantic memory.

Deeper insight into the dynamic organization of large-scale networks underlying working memory comes from studies of the phase synchronization between ERPs from distributed cortical areas. For example, long-range ERP phase synchronization has been reported in the theta frequency range between prefrontal and posterior association areas when subjects retain verbal and spatial items for short periods of time (Sarnthein et al., 1998) and in the beta frequency range between extrastriate areas when they retain visual object representations (Tallon-Baudry et al., 2001). These studies suggest that large-scale cortical network function is based not just on the co-activation of distributed neuronal ensembles, but also on the active coordination of ensemble activity, observable as ERP phase synchronization.

Finally, other ERP types have been used to examine the neural correlates of working memory load. In one investigation, the

steady-state visual ERP elicited by a diffuse 13-Hz visual flicker was used to study memory load during the retention period of an object working memory task (Silberstein et al., 2001). The steady-state visual ERP exhibited a load-dependent increase in amplitude at frontal and occipitoparietal sites. By comparison, in a study of event-related synchronization and desynchronization, significant effects of memory load were found in the frontal lobe during a visual sequential letter task (Krause et al., 2000). Event-related synchronization was found at theta frequencies during the initial stages of stimulus processing, whereas event-related desynchronization was observed at alpha frequencies.

Discussion

The cortical ERP reflects the coordinated behavior of large numbers of neurons in relation to a meaningful externally or internally generated event. Single neurons are actively coordinated in the operations of ensembles, local-area networks, and large-scale networks. ERP studies provide a unique avenue of approach to the dynamics of coordination in the cortex at the mesoscopic and macroscopic levels of organization. ERP analysis is an indispensable complement to single-cell neurophysiology and whole-head neuroimaging techniques and can supply a rich source of criteria for neural network modeling efforts.

ERP studies have shown that local cortical area networks are able to synchronize and desynchronize their activity rapidly with changes in cognitive state. These synchronization changes occur between neurons located both within individual local networks and in different local networks. The ability of local area networks to repeatedly reconfigure their activity patterns under constraint of large-scale coordinating influences may allow them to increase the degree of consensus of those local patterns in a short period of time, thereby causing the cortical system as a whole to evolve toward the solution of computational problems. Since it normally operates in a metastable dynamic regime, the cortex is able to balance the coordinated and independent behavior of local networks to maintain the flexibility of this process. When incorporated into artificial neural network designs, a similar computational process could prove useful in avoiding the processing rigidity of many current network models. A metastable large-scale neural network design that recruits and excludes subnetworks according to their ability to reach consensual local patterns has the potential to implement behavioral schema and adapt to changing environmental conditions. Such a system would represent an important advance in machine cognition.

Road Map: Cognitive Neuroscience

Related Reading: Covariance Structural Equation Modeling; EEG and MEG Analysis; Hippocampal Rhythm Generation; Schema Theory; Synchronization, Binding and Expectancy

References

- Alain, C., Woods, D. L., and Knight, R. T., 1998, A distributed cortical network for auditory sensory memory in humans, *Brain Res.*, 812:23–37.
- Bressler, S. L., 2002, Understanding cognition through large-scale cortical networks, *Curr. Dir. Psychol. Sci.*, 11:58–61.
- Bressler, S. L., and Kelso, J. A., 2001, Cortical coordination dynamics and cognition, *Trends Cogn. Sci.*, 5:26–36.
- Freeman, W. J., 2000, Mesoscopic neurodynamics: From neuron to brain, *J. Physiol. Paris*, 94:303–322. ♦
- Kok, A., 2001, On the utility of P3 amplitude as a measure of processing capacity, *Psychophys.*, 38:557–577.
- Krause, C. M., Sillanmaki, L., Koivisto, M., Saarela, C., Haggqvist, A., Laine, M., and Hamalainen, H., 2000, The effects of memory load on event-related EEG desynchronization and synchronization, *Clin. Neurophysiol.*, 111:2071–2078.
- Kutas, M., and Federmeier, K. D., 2000, Electrophysiology reveals semantic memory use in language comprehension, *Trends Cogn. Sci.*, 4:463–470.
- Lopes da Silva, F., 1991, Neural mechanisms underlying brain waves: From neural membranes to networks, *Electroenceph. Clin. Neurophysiol.*, 79:81–93. ♦
- McEvoy, L. K., Smith, M. E., and Gevins, A., 1998, Dynamic cortical networks of verbal and spatial working memory: Effects of memory load and task practice, *Cereb. Cortex*, 8:563–574.
- Pfurtscheller, G., and Lopez da Silva, F. H., 1999, Event-related EEG/MEG synchronization and desynchronization: Basic principles, *Clin. Neurophysiol.*, 110:1842–1857.
- Sarnthein, J., Petsche, H., Rappelsberger, P., Shaw, G. L., and von Stein, A., 1998, Synchronization between prefrontal and posterior association cortex during human working memory, *Proc. Natl. Acad. Sci. USA*, 95:7092–7096.
- Silberstein, R. B., Nunez, P., Pipingas, A., Harris, P., and Danieli, F., 2001, Steady state visually evoked potential (SSVEP) topography in a graded working memory task, *Int. J. Psychophysiol.*, 42:219–232.
- Tallon-Baudry, C., Bertrand, O., and Fischer, C., 2001, Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance, *J. Neurosci.*, 21:RC177.
- Truccolo, W. A., Ding, M., Knuth, K. H., Nakamura, R., and Bressler, A., 2002, Trial-to-trial variability of cortical evoked responses: Implications for the analysis of functional connectivity, *Clin. Neurophysiol.*, 113:206–226.
- Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J., 2001, The brainweb: Phase synchronization and large-scale integration, *Nat. Rev. Neurosci.*, 2:229–239. ♦

Evolution and Learning in Neural Networks

Stefano Nolfi

Introduction

Evolution and learning are two forms of adaptation that operate on different time scales. Evolution is capable of capturing relatively slow environmental changes that might encompass several generations. Learning allows an individual to adapt to environmental changes that are unpredictable at the generational level. Moreover, evolution operates on the genotype, but learning affects the phenotype and phenotypic changes cannot directly modify the genotype. Recently, the study of artificial neural networks subjected

both to an evolutionary (see EVOLUTION OF ARTIFICIAL NEURAL NETWORKS) and a lifetime learning process has received increasing attention. These studies (see also Nolfi and Floreano, 1999) have been conducted with two different purposes: (1) looking at the advantages, in terms of performance, of combining two different adaptation techniques; (2) understanding the role of the interaction between learning and evolution in natural organisms. The general picture emerging from this body of research suggests that, within an evolutionary perspective, learning has several different adaptive functions:

- It might help and guide evolution by channeling the evolutionary search toward promising directions. For example, learning might significantly speed up the evolutionary search.
- It might supplement evolution by allowing individuals to adapt to environmental changes that cannot be tracked by evolution because they occur during the lifetime of the individual or within few generations.
- It might allow evolution to find more effective solutions and facilitate the ability to scale up to problems that involve large search space.

Learning also has costs. In particular, it might increase the unreliability of evolved individuals (Mayley, 1997). Because learned abilities are also determined by learning experiences, learning individuals might fail to acquire necessary abilities in unfavorable conditions.

How Learning Might Help and Guide Evolution

Hinton and Nowlan (1987) provided a clear and simple demonstration of how learning might influence evolution even if the learned characteristics are not communicated to the genotype. The authors considered a simple case in which (1) the genotype of the evolving individuals consists of 20 genes that encode the architecture of the corresponding neural networks and (2) just a single architecture (i.e., a single combination of gene values) confers added reproductive fitness. Individuals have a genotype with 20 genes that can assume two alternative values (0 or 1). The only combination of genes that provides a fitness value above 0 consists of all 1s. In this extreme case, the probability of finding the good combination of genes is very small, given that the fitness surface looks like a flat area with a spike corresponding to the good combination. Indeed, on such a surface, artificial evolution does not perform better than random search—finding the right combination is akin to discovering a needle in a haystack. The fitness surface metaphor is often used to visualize the search space on an evolutionary algorithm. Any point on the search space corresponds to one of the possible combinations of genetic traits, and the height of each point on the fitness surface corresponds to the fitness of the individual with the corresponding genetic traits.

The addition of learning simplifies the evolutionary search significantly. One simple way to introduce learning is to assume that, in the learning individual, genes can have three alternative values [0, 1, and ?], where question marks indicate modifiable genes whose value is randomly selected within [0, 1], each time step of the individual's lifetime. By comparing learning and nonlearning individuals, one can see that performance increases throughout generations much faster in the former. The addition of learning, in fact, enlarges and smoothes the fitness surface area around the good combination, which can be discovered much more easily in this case by the genetic algorithm. This is because not only the right combination of alleles but also combinations having in part the right alleles and in part unspecified (learnable) alleles might report an average fitness greater than 0. (Fitness increases monotonically with the number of fixed right values because the time needed to find the right combination is inversely proportional, on average, to the number of learnable alleles.) According to Hinton and Nowlan (1987, p. 496), "It is like searching for a needle in a haystack when someone tells you when you are getting close." (On this point, see also BASAL GANGLIA; REINFORCEMENT LEARNING.) A variation of this model has been used to study the interaction between evolution, learning, and culture (Hutchins and Hazlehurst, 1991).

The Hinton-Nowlan model is an extremely simplified case that can be analyzed easily, but it makes several unrealistic assumptions: (1) There is no distinction between genotype and phenotype; (2) learning is modeled as a random process that does not have any

directionality; (3) there is no distinction between the learning task (i.e., the learning functions that individuals try to maximize during their lifetimes) and the evolutionary task (i.e., the selection criteria that determine which individuals are allowed to reproduce). Nolfi, Elman, and Parisi (1994) conducted further research that showed how learning and evolution display other forms of mutually beneficial interactions when these limitations are released.

Nolfi et al. (1994) studied the case of artificial neural networks that "live" in a grid world containing food elements. Networks evolve (to become fitter at one task) at the population level and learn (a different task) at the individual level. In particular, individuals are selected on the basis of the number of food elements they are able to collect (evolutionary task) and their capacity to predict the sensory consequences of their motor actions during their lifetime (learning task).

The genotype of the evolving individuals encoded the initial weights of a feedforward neural network which, at each time step, receives sensory information from the environment (the angle and the distance of the nearest food element and the last planned motor action), determines a given motor action selected within four options (move forward, turn left, turn right, or stay still), and predicts the next sensory state (the state of the sensors after the planned action is executed). Sensory information is used both as input and as teaching input for the output units encoding the predicted state of the sensors—the new sensory state is compared with the predicted state, and the difference (error) is used to modify the connection weights through backpropagation. As in the case of the Hinton-Nowlan model, modifications due to learning are not transferred back into the genotype.

The experimental results showed that (1) after a few generations, by learning to predict, individuals increased their performance not only with respect to their ability to predict but also with respect to their ability to find food (i.e., learning produced a positive effect on evolution even if the learning and the evolutionary tasks were different), and (2) the ability to find food increased faster and achieved better results in the case of learning populations than in the case of control experiments in which individuals were not allowed to learn during their lifetime. Further analysis demonstrated that the first observation can be explained by considering that evolution tends to select individuals that are located in regions of the search space where the learning and evolutionary task are dynamically correlated (i.e., where learning-induced changes that produce an increase in performance with respect to the learning task also produce positive effects, on average, with respect to the evolutionary task). And the second observation can be explained by considering that, once learning channels evolution toward solutions in which the learning task and the evolutionary task are dynamically correlated, it allows individuals to recover from deleterious mutations (Nolfi, 1999).

Consider, for example, two individuals, *a* and *b*, that are located in two distant locations in weight space but have the same fitness at birth; i.e., the two locations correspond to the same height on the fitness surface (Figure 1). However, individual *a* is located in a region where the fitness surface and the learning surface are dynamically correlated—a region in which movements that result in an increase in height with respect to the learning surface cause an increase with respect to the fitness surface on average. Individual *b*, on the other hand, is located in a region where the two surfaces are not dynamically correlated. If individual *b* moves in weight space, it will go up in the learning surface but not necessarily in the fitness surface. Because of learning, the two individuals will move during their lifetime in a direction that improves their learning performance, i.e., a direction in which their height on the learning surface tends to increase. This implies that individual *a*, which is located in a dynamically correlated region, will end up with a higher fitness than individual *b* and will therefore have a better chance to be selected. The final result is that evolution will have a

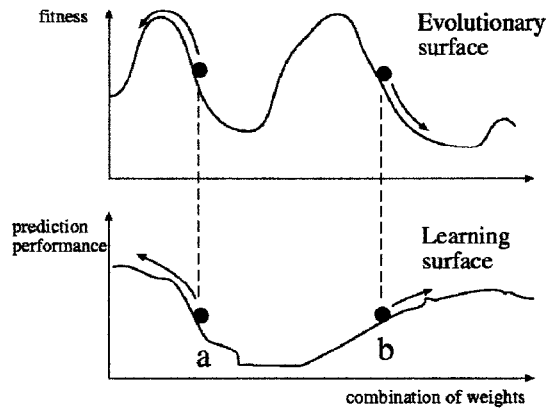


Figure 1. Fitness surface for the evolutionary task and performance surface for the learning task (sensory prediction) for all possible weight matrices. Movements due to learning are represented as arrows.

tendency to progressively select individuals that are located in dynamically correlated regions. In other words, learning forces evolution to select individuals that improve their performance with respect to both the learning and the evolutionary task.

Adapting to Changing Conditions on the Fly

As claimed above, learning might complement evolution by providing a means to master changes that occur too fast to be tracked by the evolutionary process. However, the combination of learning and evolution deeply alters both processes; thus, in individuals that evolve and learn, adaptive characteristics emerge as the result of the interaction between evolutionary and lifetime adaptation and cannot be traced to just one of the two processes.

Nolfi and Parisi (1997) evolved neural controllers for a small mobile robot that was asked to explore an arena (60×20 cm) surrounded by walls. The robot was provided with eight infrared sensors that could detect walls up to a distance of about 4 cm and two motors that controlled the two corresponding wheels. The colors of the walls switched from black to white and vice versa each generation because the activity of the infrared sensors is highly affected by the color of the reflecting surface (white walls reflect much more than black walls). Thus, to maximize their exploration behavior, evolved robots should modify their behavior on the fly. In the dark-walled environment, in fact, robots need to move very carefully whenever their sensors are activated; this is because dark walls are detected only when they are very close. In the white-walled environment, robots must begin to avoid walls only when their sensors are strongly activated; this facilitates exploration of the area close to the walls.

Individuals learn during their lifetime by means of self-generated teaching signals. The genotype of the evolving individuals encodes the connection strengths of two neural modules: (1) a teaching module that, each time step, receives the state of the sensors as input and produces a teaching signal as output; (2) an action module that receives the state of the sensors as input and produces motor actions as output. The self-generated teaching signal is used to modify the connection strengths of the action module (for a similar architecture, see Ackley and Littman, 1991). This implies that both the initial behavior produced by the evolving individuals and what the individuals learn are results of the evolutionary process, and neither is determined by the experimenter.

Evolved robots displayed an ability to discriminate between the two types of environments and to modify their behavior accordingly, thereby maximizing their exploration capability. The anal-

ysis of the results revealed that this ability arose from a complex interaction between the evolutionary and learning process. For example, evolved individuals displayed an inherited ability to behave so as to enhance the perceived differences between the two environments. This in turn allows the learning process to progressively modify the behavior of the robots so as to adapt to the different environmental conditions.

More generally, this research and that of others has shown that evolution, in the case of individuals able to change during their lifetime as a result of learning, does not tend to directly develop an ability to solve a problem; rather, it tends to develop a predisposition to acquire such ability through learning.

Other experiments conducted by coevolving two competing populations of predator and prey robots (Nolfi and Floreano, 1998) emphasized how lifetime learning might allow evolving individuals to achieve generality, i.e., the ability to produce effective behavior in a variety of circumstances. In these experiments, predators consisted of small mobile robots provided with infrared sensors and a linear camera with a view angle of 36° , which allowed them to detect prey. Prey consisted of mobile robots of the same size; these robots had only infrared sensors, but their maximum available speed was set to twice that of the predators. Each individual was tested against different competitors for ten trials. Predators scored one point for each trial in which they caught prey while prey scored one point for each trial in which they escaped predators.

In this experimental situation, both populations change through the generations as predators and prey face ever-changing, progressively more complex challenges. Interestingly, the authors observed that, in this situation, evolution alone displayed severe limitations; progressively more effective solutions could be developed only by allowing evolving individuals to adapt on the fly through a form of lifetime learning. Indeed, any fixed strategy could master only a limited number of different types of competitors; therefore, only by combining evolution and learning were the authors able to synthesize individuals capable of dealing with competitors adopting qualitatively different strategies. Indeed, by evolving learning individuals, the authors observed the emergence of predators able to detect the current strategy adopted by the prey and to modify their behavior accordingly.

Evolving the Learning Rules

Floreano and Urzelai (2000) conducted a set of experiments in which the genotype of the evolving individuals encoded the learning properties of the neurons of the corresponding neural network. These properties included one of four possible Hebbian learning rules, the learning rate, and the sign of all the incoming synapses of the corresponding neuron. When the genotype is decoded into a neural controller, the connection strengths were set to small random values. After some generations, the genetically specified configuration of learning rules tended to produce changes in the synaptic strengths that allow individuals to acquire the required competencies through lifetime learning. By comparing the results obtained with this method with a control experiment in which the strength of the synapses was directly encoded into the genotype, the authors observed that evolved controllers able to adapt during lifetime can solve certain tasks faster and better than standard nonadaptive controllers. Moreover, they demonstrated that their method scales up well to large neural architectures.

The authors applied this method to evolve neural controllers for mobile robots. Interestingly, the analysis of the synaptic activity of the evolved controllers showed that several synapses did not reach a stable state but kept changing all the time. In particular, synapses continued to change even when the behavior of the robot became rather stable.

Similar advantages have been reported by Husband et al. (1999), who evolved a type of neural network in which neurons, which

were distributed over a 2D surface, emitted “gases” that diffused through the network and modulated the transfer function of the neurons in a concentration-dependent fashion, thus providing a form of plasticity.

Finally, the experiments performed by Di Paolo (2000) showed how learning could play the role of a homeostatic process whereby evolved neural networks adapt in order to remain stable in the presence of external perturbations.

Discussion

The interaction between learning and evolution deeply alters both the evolutionary and the learning processes. Evolution in interaction with learning displays dynamics very different from those observed in evolution alone. While in nonlearning individuals the characters that are selected through evolution directly incorporate an ability to produce successful behaviors, in learning individuals they incorporate a predisposition to learn, i.e., a predisposition to acquire necessary abilities through learning. This predisposition to learn may consist of the following:

1. The presence of starting conditions that canalize learning in the right direction. Evolution may select initial weight matrices or network architectures that cause a better and/or a faster learning (Belew, McInerney, and Schraudolph, 1992). This happens either when the learning task and the evolutionary task are the same or when they differ. In the latter case, evolution selects not only individuals that have a predisposition to learn better, but also individuals that, by learning a given task, improve their performance with respect to the evolutionary task.
2. An inherited tendency to behave in such a way that the individual is exposed to the appropriate learning experiences. Evolution tends to select characters that produce initial behaviors that enhance the possibility of learning and/or increase the probability of acquiring adaptive characters through learning. In other words, evolution tends to select individuals whose initial behavior is suitable for learning and not necessarily for solving the evolutionary task.

Similarly, learning within an evolutionary perspective has quite different characteristics from learning studied in isolation, as in “traditional” connectionist research. In individuals that learn but are not subjected to an evolutionary process (e.g., neural networks trained with supervised methods), learning is usually accomplished by ignoring the characters of the individual prior to learning (which are typically generated at random); but in evolving plastic individuals, learning exploits such starting conditions. Moreover, when the learning process itself (i.e., what it is learned during lifetime) is subjected to evolution and not determined in advance, learning does not necessarily tend to incorporate the right solution to the problem; rather, it tends to pull the learning individual in a direction that, given the initial state of the individual, maximizes the chances of adapting to the current environment.

The study of learning within an evolutionary perspective is still in its infancy. But in forthcoming years, it might have an enormous impact on our understanding of how learning and evolution operate in nature. In particular, this type of research might shed light on the ability to learn from others, and more generally on the co-evolution between brain structure and cultural processes such as natural language (for initial explorations of such issue, see Cangelosi, 2001).

Road Maps: Learning in Artificial Networks; Neuroethology and Evolution

Related Reading: Evolution of Artificial Neural Networks; Evolution of Genetic Networks; Language Evolution and Change; Locomotion, Vertebrate

References

- Ackley, D. H., and Littman, M. L., 1991, Interaction between learning and evolution, in *Proceedings of the Second Conference on Artificial Life* (C. G. Langton et al., eds.), Reading, MA: Addison-Wesley, pp. 487–509.
- Belew, R. K., McInerney, J., and Schraudolph, N. N., 1992, Evolving networks: Using the genetic algorithm with connectionistic learning, in *Proceedings of the Second Conference on Artificial Life* (C. G. Langton et al., Eds.), Reading, MA: Addison-Wesley, pp. 511–548.
- Cangelosi, A., 2001, Evolution of communication and language using signals, symbols and words, *IEEE Trans. Evolutionary Computation*, 5(2):93–101.
- Di Paolo, E. A., 2000, Homeostatic adaptation to inversion in the visual field and other sensorimotor disruptions, in *From Animals to Animats: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* (J.-A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, and S. W. Wilson, Eds.), Cambridge, MA: MIT Press, pp. 440–449.
- Floreano, D., and Urzelai, J., 2000, Evolutionary robotics with on-line self-organization and behavioral fitness, *Neural Networks*, 13:431–443.
- Hinton, G. E., and Nowlan, S. J., 1987, How learning guides evolution, *Complex Systems*, 1:495–502. ♦
- Husband, P., Smith, T., Jakobi, N., and O'Shea, M., 1999, Better living through chemistry: Evolving GasNets for robot control, *Connection Science*, 3–4:185–210.
- Hutchins, E., and Hazlehurst, B., 1991, Learning in the cultural process, in *Artificial Life II* (C. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, Eds.), Reading, MA: Addison-Wesley, pp. 689–706.
- Mayley, G., 1997, Landscapes, learning costs, and genetic assimilation, *Evolutionary Computation*, 4:213–234.
- Nolfi, S., 1999, How learning and evolution interact: The case of a learning task which differs from the evolutionary task, *Adaptive Behavior*, 2:231–236.
- Nolfi, S., and Floreano, D., 1998, Co-evolving predator and prey robots: Do “arm races” arise in artificial evolution? *Artificial Life*, 4:311–335.
- Nolfi, S., and Floreano, D., 1999, Learning and evolution, *Autonomous Robots*, 1:89–113. ♦
- Nolfi, S., and Parisi, D., 1997, Learning to adapt to changing environments in evolving neural networks, *Adaptive Behavior*, 1:75–98.
- Nolfi, S., Elman, J. L., and Parisi, D., 1994, Learning and evolution in neural networks, *Adaptive Behavior*, 1:5–28.

Evolution of Artificial Neural Networks

Stefano Nolfi and Domenico Parisi

Introduction

Artificial neural networks may be either computational models of biological nervous systems or computational systems inspired, per-

haps loosely, by neurobiology. Natural organisms, however, possess not only nervous systems but also genetic information stored in the nucleus of their cells (genotype). The nervous system is part of the phenotype derived from this genotype through a process

called development. The information specified in the genotype determines aspects of the nervous system that are expressed as innate behavioral tendencies and predispositions to learn. When neural networks are viewed in the broader biological context of artificial life (i.e., the attempt to synthesize life-like phenomena within computer and other artificial media), they tend to be accompanied by genotypes and to become members of evolving populations of networks in which genotypes are inherited from parents to offspring (Parisi, 1997).

Artificial neural networks can be evolved by using evolutionary algorithms (Holland, 1975; Koza, 1992). An initial population of different artificial genotypes, each encoding the free parameters (e.g., the connection strengths, the architecture of the network, the learning rules, or some combination thereof) of a corresponding neural network, is created randomly. The population of networks is evaluated in order to determine the performance (fitness) of each individual network. The fittest networks are allowed to reproduce by generating copies of their genotypes with the addition of changes introduced by genetic operators such as mutations (random changes of a few genes that are selected randomly) or crossover (the combination of parts of the genotype derived from two reproducing networks). This process is repeated for a number of generations until a network that satisfies the performance criterion set by the experimenter is obtained (for a review of methodological issues, see Yao, 1993).

The genotype might encode all the free parameters of the corresponding neural network or only the initial values of the parameters and/or other parameters that affect development and learning. In the former case, networks are entirely specified in the genotype and change only phylogenetically as a result of the modifications introduced by genetic operators during reproduction. In the latter case, networks also change ontogenetically (i.e., during the period in which they are evaluated) as a result of both genetic and environmental factors. In this article we review examples of networks that undergo developmental processes such as neural growth. For an analysis of networks that are able to adapt to the environment as a result of a form of lifetime learning see *EVOLUTION AND LEARNING IN NEURAL NETWORKS*.

Evolution and Development

A cornerstone of biology is the distinction between the inherited genetic code (genotype) and the corresponding organism (phenotype). What is inherited from the parents is the genotype. The phenotype is the complete individual that is formed according to the instructions specified in the genotype.

Evolution is critically dependent on the distinction between genotype and phenotype, and on their relation, i.e., the genotype-to-phenotype mapping. The fitness of an individual, which affects selective reproduction, is based on the phenotype; but what is inherited is the genotype, not the phenotype. Furthermore, while the genotype of an individual is a single entity, the organism is a continuum of different phenotypes taking form during the genotype-to-phenotype mapping process, each derived from the previous one under genetic and environmental influences.

When the genotype-to-phenotype mapping process takes place during an individual's lifetime, we speak of development. In this case, each successive phenotype corresponding to a given stage of development has a distinct fitness. The fitness of a developing individual is a complex function of these developmental phases. Evolution must ensure that all these successive forms are viable and, at the same time, that they comprise a well-formed sequence in which each form leads to the next until a mostly stable (adult) form is reached. This puts various constraints on evolution, but it also offers new means for exploring novelty. Small changes in the developmental rates of different components of the phenotype, for example, can have huge effects on the resulting phenotype. Indeed,

it has been hypothesized that in natural evolution changes affecting regulatory genes that control the rates of development played a more important role than other forms of change such as point mutations (Gould, 1977; see also *EVOLUTION OF GENETIC NETWORKS*).

Although the role of genotype-to-phenotype mapping and development has been ignored in most experiments involving artificial evolution, awareness of its importance is now increasing. According to Wagner and Altenberg (1996, p. 967), "In evolutionary computer science it was found that the Darwinian process of mutation, recombination and selection is not universally effective in improving complex systems like computer programs or chip designs. For adaptation to occur, these systems must possess *evolvability*, i.e., the ability of random variations to sometimes produce improvement. It was found that evolvability critically depends on the way genetic variation maps onto phenotypic variation, an issue known as the representation problem."

Genetic Encoding

To evolve neural networks one should decide how to encode the network in the genotype in a manner suitable for the application of genetic operators. In most cases, phenotypic characteristics such as synaptic weights are coded in a uniform manner so that the description of an individual at the level of the genotype assumes the form of a string of identical elements (such as binary or floating point numbers). The transformation of the genotype into the phenotypic network is called genotype-to-phenotype mapping.

In direct encoding schemes there is a one-to-one correspondence between genes and the phenotypic characters subjected to the evolutionary process (Yao, 1993). Aside from its biological implausibility (see *EVOLUTION OF THE ANCESTRAL VERTEBRATE BRAIN*), simple one-to-one mapping has several drawbacks. One problem, for example, is scalability. Since the length of the genotype is proportional to the complexity of the corresponding phenotype, the space to be searched by the evolutionary process increases quadratically with the size of the network (Kitano, 1990).

Another problem of direct encoding schemes is the impossibility of encoding repeated structures (such as network composed of several subnetworks with similar local connectivity) in a compact way. In one-to-one mappings, in fact, elements that are repeated at the level of the phenotype must be repeated at the level of the genotype as well. This affects not only the length of the genotype and the corresponding search space, but also the evolvability of individuals. A full genetic specification of a phenotype with repeated structures, in fact, implies that adaptive changes affecting repeated structures should be independently rediscovered through changes introduced by the genetic operators.

Growing Methods

The genotype-to-phenotype process in nature is not just an abstract mapping of information from genotype to phenotype; it is also a process of physical growth (both in size and in physical structure). Thus, taking inspiration from biology, one may decide to encode in the genotype growing instructions. The phenotype is progressively built by executing the inherited growing instructions.

Nolfi, Miglino, and Parisi (1994) used a growing encoding scheme to evolve the architecture and the connection strengths of neural networks that controlled a small mobile robot. These controllers consisted of a collection of artificial neurons distributed over a 2D space with growing and branching axons (Figure 1, *Top*). Inherited genetic material specified instructions that controlled the axonal growth and the branching process of neurons. During the growth process, when a growing axonal branch of a particular neuron reaches another neuron, a connection is established between the two neurons. The bottom of Figure 1 shows the network re-

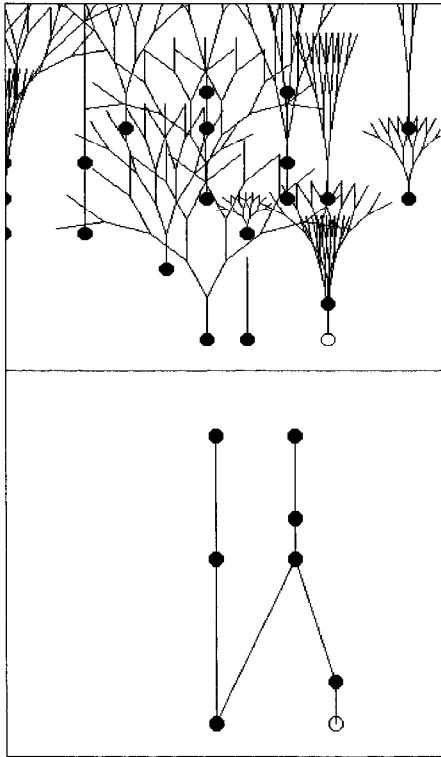


Figure 1. Development of an evolved neural network. *Top*, The growing and branching process of the axons. *Bottom*, The resulting neural network after removal of nonconnecting branches and the elimination of isolated neurons and groups of interconnected neurons.

sulting from this growth process after the elimination of isolated and nonfunctional neurons. Axons grew and branched only if the activation variability of the corresponding neurons was larger than a genetically specified threshold. This simple mechanism is based on the idea that sensory information coming from the environment has a critical role in the maturation of the connectivity of the biological nervous system and, more specifically, that the maturation process is sensitive to the activity of single neurons (Purves, 1994). Therefore, both genetic and environmental factors influenced the developmental process; i.e., the actual sequence of sensory states experienced by the network influenced the process of neural growth.

This method allows the evolutionary process to select neural network topologies that are suited to the task chosen (see *LEARNING NETWORK TOPOLOGY*). Indeed, analysis of the evolutionary process shows how improvements are due to changes affecting both the neural architecture and the connection weights. Moreover, if some aspects of the task are allowed to vary during the evolutionary process, evolved genotypes display an ability to develop into different final phenotypic structures that are adapted to current conditions.

Cellular Encodings

In natural organisms the development of the nervous system begins with an in-folding of the ectodermic tissue to form the neural crest. This structure gives rise to the mature nervous system through three phases: the genesis and proliferation of different classes of neurons by cellular duplication and differentiation, the migration of the neurons toward their final destination, and the growth of neurites (axons, dendrites). The growing process described in the previous sec-

tion therefore characterizes very roughly only the last of these phases. A number of attempts inspired by the seminal work of Lindenmayer (1971) have been made to include other aspects of this process in artificial evolutionary experiments.

Cangelosi, Nolfi, and Parisi (1994), for example, extended the model described in the previous section by adding a cell division and migration stage to the already existing stage of axonal growth. The genotype is a collection of rules governing the process of cell division (a single cell is replaced by two daughter cells) and migration (the new cells can move in the 2D space). The genotype-to-phenotype process therefore starts with a single cell which, by undergoing a number of duplication and migration processes, produces a collection of neurons arranged in a 2D space. These neurons grow their axons and establish connections until a neural controller is formed.

Gruau (1994) proposed a genetic encoding scheme for neural networks based on a cellular duplication and differentiation process. The genotype-to-phenotype mapping starts with a single cell that undergoes a number of duplication and transformation processes, ending up in a complete neural network. In this scheme the genotype is a collection of rules governing the process of cell divisions (a single cell is replaced by two daughter cells) and transformations (new connections can be added and the strengths of the connections departing from a cell can be modified). In this model, therefore, connection links are established during the cellular duplication process.

The instructions contained in the genotype are represented as a binary-tree structure as in genetic programming (Koza, 1992). During the genotype-to-phenotype mapping process, the genotype tree is scanned starting from the top node of the tree, then following each ramification. The top node represents the initial cell that undergoes a set of duplication processes to produce the final neural network. Each node of the genotype tree encodes the operations that should be applied to the corresponding cell, and the two subtrees of a node specify the operations that should be applied to the two daughter cells. The neural network is progressively built by following the tree and applying the corresponding duplication instructions. Terminal nodes of the tree (i.e., nodes that have no subtrees) represent terminal cells that will not undergo further duplications. Gruau also considered the case of genotypes formed by many trees where the terminal nodes of a tree may point to other trees. This mechanism allows the genotype-to-phenotype process to produce repeated phenotypic structures (e.g., repeated neural subnetworks) by re-using the same genetic information. Trees that are pointed to more than once, in fact, will be executed more times. This encoding method has two advantages: (1) Compact genotypes can produce complex phenotypic networks; (2) evolution may exploit phenotypes in which repeated substructures are encoded in a single part of the genotype. By evolving neural controllers for a simulated hexapod robot able to walk, the author showed that the problem could be solved only by using a genetic encoding that allows for the possibility of re-using the same genetic information to encode repeated substructures (i.e., similar subnetworks controlling the legs) or by using incremental evolution (i.e., by first evolving oscillator networks to control a single limb, then evolving the coordinating circuitry to yield walking [see also Lewis et al., 1992]).

Discussion

Artificial evolution can be seen as a learning algorithm for training artificial neural networks. From this point of view, one distinctive feature is the limited amount of feedback required. Supervised learning algorithms require immediate and detailed desired answers as feedback. Reinforcement learning algorithms require less—only a judgment of right or wrong, which need not be immediate. Viewed as a learning algorithm, artificial evolution requires still

less—only an overall evaluation of the performance of the network over the entire evaluation period. A second distinctive feature is that any parameter of the neural network (e.g., the connection strengths, the network topology, the learning rules, the transfer function of the neurons) can be subjected to the evolutionary process.

Although systematic comparison between artificial evolution and other algorithms has not yet been done, it is reasonable to claim that artificial evolution tends to produce better results when detailed feedback is not available. This is the case, for example, for neural networks that should control mobile robots (Nolfi and Floreano, 2000). In this case, in fact, although the experimenter can provide a general evaluation of how much the behavior of a robot approximates the desired behavior, he or she usually cannot indicate what the robot should do at each time step to produce a desired behavior. Moreover, artificial evolution might prove more effective when certain features of the network (such as the network topology or the transfer functions) that cannot be properly set by hand are crucial. Artificial evolution, in fact, provides a way to co-adapt different types of parameters. Artificial evolution also has drawbacks, such as the time needed to conduct the evolutionary process and the lack of formal criteria for designing effective fitness functions.

The analogy with natural evolution, however, can also be considered more strictly. In this case, the evolutionary process is not seen as an abstract training algorithm but as a process that mimics some of the key aspects of the evolutionary process in nature. From this perspective, neural networks tend to be viewed as a part of a population of artificial organisms that adapt autonomously by interacting with the external environment.

This body of research might contribute to the understanding of natural systems by identifying the key characteristics of natural evolution that make it so successful in producing the extraordinary variety of highly adapted life forms present on the planet. Examples include a better understanding of the role of incremental or staged evolution (i.e., how evolution of animals adapted to one biological niche then provides the basis for further evolution into other niches [Harvey, 1993; Lewis et al., 1992]); the role of competitive co-evolution (i.e., how the evolution of two competing populations with coupled fitness may reciprocally drive each other to increasing levels of complexity [Nolfi and Floreano, 2000]); the importance of pre-adaptation (i.e., the possibility of evolving a predisposition to acquire an ability to solve a given problem during lifetime rather than, directly, an ability to solve such a problem [see EVOLUTION AND LEARNING IN NEURAL NETWORKS]).

Road Maps: Learning in Artificial Networks; Neuroethology and Evolution

Related Reading: Evolution and Learning in Neural Networks; Evolution of the Ancestral Vertebrate Brain; Locomotion, Vertebrate

References

- Cangelosi, A., Nolfi, S., and Parisi, D., 1994, Cell division and migration in a "genotype" for neural networks, *Network—Computation in Neural Systems*, 5:497–515.
- Gould, S. J., 1977, *Ontogeny and Phylogeny*, Cambridge, MA: Harvard University Press.
- Gruau, F., 1994, Automatic definition of modular neural networks, *Adaptive Behavior*, 3:151–183.
- Harvey, I., 1993, Evolutionary robotics and SAGA: The case for hill crawling and tournament selection, in *Artificial Life 3: Proceedings of the Santa Fe Conference* (C. Langton, Ed.), Reading, MA: Addison-Wesley, pp. 299–326.
- Holland, J. J., 1975, *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press.
- Kitano, H., 1990, Designing neural networks using genetic algorithms with graph generation system, *Complex Systems*, 4:461–476.
- Koza, J. R., 1992, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, MA: MIT Press.
- Lewis, M. A., Fagg, A. H., and Solidum, A., 1992, Genetic programming approach to the construction of a neural network for control of a walking robot, in *Proceedings of the IEEE International Conference on Robotics and Automation*, New York: IEEE Press, pp. 2618–2623.
- Lindenmayer, A., 1971, Developmental systems without cellular interactions, their language and grammars, *J. Theor. Biol.*, 30:455–484.
- Nolfi, S., and Floreano, D., 2000, *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*, Cambridge, MA: MIT Press/Bradford Books. ♦
- Nolfi, S., Miglino, O., and Parisi, D., 1994, Phenotypic plasticity in evolving neural networks, in *Proceedings of the International Conference: From Perception to Action* (D. P. Gaussier and J.-D. Nicoud, Eds.), Los Alamitos, CA: IEEE Press, pp. 146–157.
- Parisi, D., 1997, Artificial life and higher level cognition, *Brain Cogn.*, 34:160–184. ♦
- Purves, D., 1994, *Neural Activity and the Growth of the Brain*, Cambridge, UK: Cambridge University Press.
- Wagner, G. P., and Altenberg, L., 1996, Complex adaptations and the evolution of evolvability, *Evolution*, 50:967–976.
- Yao, X., 1993, A review of evolutionary artificial neural networks, *Int. J. Intelligent Systems*, 4:203–222. ♦

Evolution of Genetic Networks

Kirk W. Beisel and Bernd Fritzsche

Introduction

With the completion of an initial draft of the more than 30,000 genes of the human genome we can now establish all the information to make a human from the coded sequences, which are based on a four-nucleotide alphabet. Man has achieved a significant catalog useful for understanding the basis of all evolution through the comparisons of whole genomes ranging from yeast and worms to mice and humans. At present, we can read those four letters and are, to some extent, able to understand words (genes) or phrases (known gene networks). However, most of our current higher level of understanding of the meaning of the DNA sequences and their subsequent organization into gene networks is comparable to an

understanding of a poem by Goethe read by a six-year-old. Like the poem, the meta-language of information coding in the human genome needs to be understood above and beyond our current limited insight (Venter et al., 2001). Herein, we offer an outline that specifies some of the computational problems in modeling genetic networks, which can direct the establishment of a diversity of neuronal networks in the brain. Since these neuronal networks are composed of a wide variety of cell types, the final fate or end stage of each cell type represents the outcome of a dynamic amalgamation of gene networks. Genetic networks not only determine the cell fate acquisition from the original stem cell, but also govern the contact formation between the cell populations of a given neuronal network.

In many respects there are numerous intriguing parallels between the establishment and functioning of genetic networks with those of neuronal networks, which can range from simple (on and off switch) to extremely complex (computer logic gate). To appreciate the full complexity of organismic development we outline below how intracellular and cell-cell interactions modify the complexity of gene interactions involved in genetic networks. Such interactions will achieve an altered status of cell function and, ultimately, the connection alterations in the formation of neuronal networks.

Evolving from Operons to Promoters with Enhancers and Suppressors

When Jacob and Monod published their now famous lac operon model of the regulation of the bacterial gene transcription, they initiated what has now become the main problem of developmental biology: unraveling the proximate causes for gene activation and silencing (repression) at the appropriate level. As such, a full understanding of this regulatory process is essential for any higher-level understanding of the information decoding process that unfolds the genome information to form a human being. Simply said, if a muscle cell precursor activates the wrong genetic network, it will form bone around it, like bone cells. The effect of such a developmental “error” would be catastrophic for the organism.

The lac operon (Figure 1) is a gene that consists of two regions. One contains the DNA elements responsible for regulation of gene transcription, and the other encodes for the transcribed protein(s). Regulation of gene transcription to form messenger RNA (mRNA), which is in turn the basis for translating the information coded in the DNA into a protein, is the way the cell regulates the availability of information coded in the genome. The regulatory region is composed of DNA elements, which can be categorized as operator (enhancer/silencer) and promoter sites. Promoters are regions in which RNA transcription is initiated. The operator sites, containing enhancer and silencer elements, are DNA sequences that require proper binding of regulatory proteins to either stimulate or inhibit the promoter region. Initiating transcription requires utilization of regulatory elements within a gene’s genomic structure. In the case of the lac operon system, this is achieved by having an operator strand of DNA that initiates gene transcription, controlled by a repressor protein (lacI), which in turn is controlled by an inducer (lactose). The lacI and lactose have specific affinities for each other, such that in the presence of the inducer, the repressor changes its configuration and cannot bind to the operator DNA. This allows transcription of the gene. The product of the gene is an enzyme that metabolizes the inducer (lactose), thus freeing the repressor

(lacI) to block transcription. Once all lactose is metabolized, the repressor binds to the operator, preventing the latter from initiating transcription of an enzyme that is no longer needed. In general terms, both negative (repression) and positive (activation) controlling elements exist for gene transcription.

This basic system of bacterial gene expression regulation has evolved into a much more complicated regulatory process that involves genetic networks rather than single genes. This increase in complexity is further complicated by the evolution of interrupted DNA sequences in multicellular organisms that form exons (parts of the DNA that, when transcribed, are exported from the nucleus and subsequently translated into a protein) and introns (parts of DNA between the exons, which are not translated). This situation does not allow a simple specification of units of information as continuous strands of DNA that specify unique proteins. In general, a single gene specifies each protein, even if this gene is transcribed into a variety of mRNA species through differential splicing of various exons that gives rise to various proteins. It must be recognized that not all genes encode for proteins.

We next have to look at the level of transcription initiation in multicellular organisms, which can have greater complexity compared with that demonstrated by the lac operon. The promoter region requires a basic set of DNA binding proteins to initiate transcription, which is common for all promoter regions. However, for this complex to function, it requires association with one or more specific transcription factors. This transcription factor complex also targets to specific promoter sites. The presence of such transcription factors is therefore essential for the activity mediated by the enhancer and silencer sites on the promoter region of a gene. Such transcription factors regulate the temporal and tissue specific expression of any differentially regulated gene. However, a given enhancer may activate a number of genes, whereas several enhancer sites may be linked to a given promoter region. In addition, the same enhancer may activate transcription of some genes while simultaneously suppressing that of other genes. Most transcription factors can bind to specific DNA sequences, have an activation domain that acts on gene transcription, and also interact with the basic transcription complex as well as other transcription factors. In essence, the promoter region of a gene computes the overall expression of all enhancer/silencer sites that can modify its activity and regulates how much and for how long transcription of a gene occurs. In conclusion, most gene activation/suppression is not on a single gene level, but rather affects multiple genes simultaneously. We propose to use the term *genetic network* for those interactively regulated genes. Essentially, these genetic networks and their complex, spatiotemporally restricted interactions with the DNA provide the basis of the nonlinear relationship of genotype to phenotype.

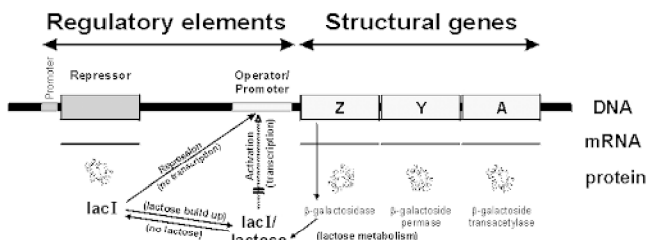


Figure 1. The lac operon model, which shows the regulation of transcription of the Z (β -galactosidase), Y (β -galactoside permease), and A (β -galactoside transacetylase) genes. The promoter regions for the repressor protein and the three structural genes are shown. If no lactose is present, the repressor protein (lacI) will bind with the operator/promoter sites and prevent transcription. If lactose is present, transcript occurs, since lactose binds with lacI and prevents lacI binding with the operator/promoter site. β -Galactosidase will metabolize lactose and thus allows lacI to bind to the promoter to prevent further transcription (negative feedback).

Modeling Genetic Networks and Their Evolution

In single-celled organisms many of the biochemical pathways represent a hierarchal pathway and as such may represent an intracellular genetic network in its most simplistic form. As organisms evolved, more complex genetic networks formed to control and modify cell-cell interactions and functions. Such networks allowed cells to evolve beyond simple substrate interactions in a continuously active cell and permitted the formation of different cell types, each specialized for specific tasks. Instead of evolving distinct sets of genes for each specific cell type, networks consisting of variable mixes of genetic modules evolved to allow rapid evolution of multiple cell types by rearranging those modules. The establishment of genetic networks was possible only through the evolution of multiple regulatory elements to control gene expression. Once established, a genetic network can be duplicated, then modified to form repeated modules or cassettes.

Beyond modularity of genetic networks, developing organisms and the brain also consist of repeated modules. For example, the cerebellum consists of such repetitive modules and has long been viewed as a paradigm for neuronal information processing. Likewise, reiterative development in terms of segmentation and segment transformation has appealed to developmental biologists. This mechanism has allowed us to understand how evolution works by gradually transforming developmental sequences in different segments to generate individuality from commonality. Because of this general interest, we know a great deal of the molecular development of segments and how the basic machinery is altered in various segmented animals as well as how modifiers of development can alter segmental fate (Robert, 2001).

Several models have been developed to theoretically examine genetic networks and how they can impact cellular patterning. Two basic approaches are the standard Boolean network model and the continuous models that approximate neural-like connectionist architectures or biochemical networks of interacting molecules. Salazar-Ciudad and colleagues (Salazar-Ciudad, Newman, and Sole, 2001a; Salazar-Ciudad, Sole, and Newman, 2001b) recently provided a mathematical model for some of those interactions. Building on a previous model to describe the formation of stripes in the fly embryo, Salazar-Ciudad et al. (2001a, b) analyzed properties of various gene networks to form patterns. The basic idea is a reaction-diffusion system that factors in the concentration of a gene product, the interactions of genes (both positive and negative), thresholds of gene responses, and diffusion between cells. Both qualitative and quantitative effects can be mathematically represented so that proximal and distal cell-cell interactions and the resulting effects on the ensuing cellular patterning can be predicted. Salazar-Ciudad et al. (2001a, b) showed that many properties of these larger networks of up to 30 interacting gene products can be broken down into subsets of a genetic network that can form basic patterns such as stripes. Moreover, many of those subsets, called modules, can be combined in various ways to form more complex patterns. Interestingly, this model is close to some models for information processing in neural networks (Salazar-Ciudad et al., 2001a, b).

Building on these properties, they generate two model networks based on the interactions of two diffusible factors that have been identified in setting up segments in insects (Salazar-Ciudad et al., 2001a, 2001b). In one case they assume that one of these diffusible factors increases expression of both factors whereas the other factor inhibits expression of both factors (emergent network). In the second (hierarchic) network, there are no direct or indirect reciprocal relationships between the gene products (Figure 2). Based on these two network properties, they show how either a hierarchic network or an emergent network can represent a realistic model of segmentation. The strength of the emergent network lies in its resilience against change and can result in complex repetitive patterns (three or more stripes). In contrast, the hierarchic network provides a closer relationship between genotype and phenotype and this allows for a more rapid implementation of variation. In addition, hierarchic networks form less complicated repetitive patterns (three or fewer stripes). Moreover, switching between those two networks during developmental steps can provide flexibility in cell fate acquisition and pattern formation. Other patterning processes that lead to the formation of checkerboard patterns can most easily be simulated by employing the lateral inhibition of cell fate assignment provided by the ubiquitous Delta-Notch system of lateral inhibition, again largely simulating network properties well known in computational neurobiology. The genetic networks discussed thus far are essentially suited for simple strings of cells or two-dimensional (2D) sheets of cells and can at best roughly approximate the complexity of three-dimensional organs such as the brain.

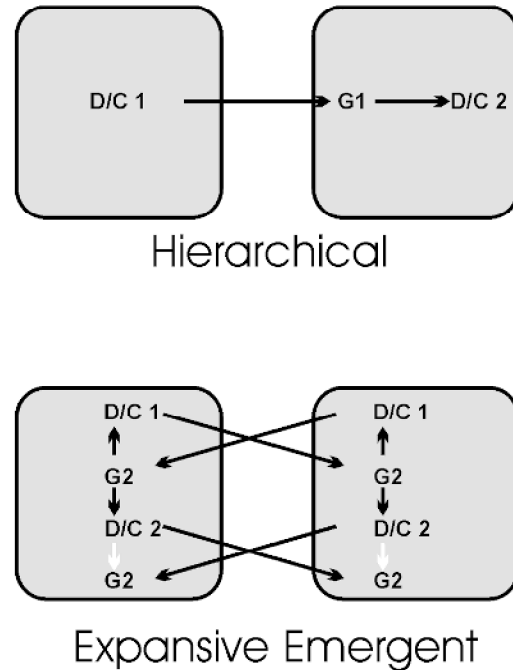


Figure 2. Two interactive modules, which show the interactions between two cells or cell clusters. Diffusible or surface contacting genes (D/C1; D/C2) and nondiffusible gene products (G1; G2) have both activating (black arrows) and inhibiting (white arrows) interactions in the emergent network.

Transcription Factors Belong to a Number of Families

Transcription factors are characterized by a DNA binding domain and by a domain that permits protein-protein interactions to interact with other transcription factors and the transcription activation complex. Many transcription factors can be grouped into larger families depending on certain conserved DNA binding motifs. Perhaps the best-known transcription factors are the homeodomain (Helix-Turn-Helix) proteins, which contain a highly conserved sequence of 60 amino acids. This gene family is present in most eukaryotic organisms and plays a prominent role in controlling the genetic determination of development and implementation of the genetic body plan. Some of the other more famous DNA binding domains are the basic Helix-Loop-Helix (bHLH) domain, the basic Leucine zipper and the Zinc finger domain. Advances in comparative analysis of human, fly, worm, and yeast genomes suggest that, as expected, many transcription factor families have been greatly enlarged in the course of human evolution. For example, while the overall human genome is only about twice as large as the fly genome, and five times as large as the yeast genome, a specific Zinc finger transcription factor family has expanded almost 20 times that of yeast (Venter et al., 2001). Others, like the Forkhead domain, the bHLH domain, and the homeobox domain, have expanded from yeast to humans 10 to 25 times.

In addition to these transcription factors, an expansion of secreted factors that regulate transcription via their specific receptors has played a major role in the evolution of multicellular organisms. Outside of hormones, those secreted factors can be grouped into four families: the Fibroblast Growth Factor (*FGF*) family, the Hedgehog (*Hh*) family, the Wingless (*Wnt*) family, and the Transforming Growth Factor- β (*TGF- β*) superfamily. As is to be expected, these families have been disproportionately enlarged in humans compared to flies (24:1 *FGFs*; 3:1 *Hh*; 18:6 *Wnt*; 29:6

TGF- β). This signifies that expansion of the human genome is not so much a simple duplication of all genes, but rather, an expansion of modulating capacity of gene expression to generate unique contexts in which genes are activated.

A number of researchers consider that some of these contexts involve what has been dubbed “master control genes.” Such genes produce transcription factors that, if expressed in areas where they are normally not expressed, can turn those areas into an organ comparable to that in which the normal expression pattern of that gene is apparently involved. A classic example is the *Pax6* gene (reviewed in Pichaud and Desplan, 2002). *Pax6*, together with other genes, is essential for eye formation and forms an interactive gene network. If the mammalian *Pax6* gene is expressed instead of its fly homolog in flies, it can direct fly eye development. Moreover, if it is overexpressed, it can turn areas, like skin at joints, into eye-like structures. It must be noted that eyes form in flies from ectoderm. Thus, only the fly’s ectoderm is competent to respond to the enhanced presence of *Pax6*. In analogy to the “grandmother” neuron concept in neurobiology, some people have adopted the view that a single gene can switch on and govern development of an entire organ—in this case, a fly’s eye. Most recent research has already modified those initial claims; it has shown that a number of genes [e.g., *eyeless1* (*Eya1*) and *sine oculis* homeobox (*Drosophila*) homolog 3 (*Six3*)], if disrupted by targeted deletion of essential parts of their protein coding regions (knockout), can also cause loss of eyes. Although *Pax6*, *Eya1*, and *Six3* are expressed in a number of organs, they are able to form eyes only if co-expressed in distinct areas. Thus, the genetic context in which a gene is expressed is undeniably important. By changing the context in which a transcription factor is expressed, it can be utilized in a variety of genetic networks and result in the development of a wide range of cell types. This cellular context can also be extended to the spatial and temporal patterning of precursor cells. This is exemplified by the generation of distinct *Drosophila* neuroblast cell lineages, which depend on the timing of the sequential expression of the transcription factors from a common neural precursor (Isshiki et al., 2001).

We have now established that context-dependent gene activation and developmental regulation of dynamically interacting networks of transcription factors are the likely causes for development to take a given pathway. Clearly, if *Pax6* were overexpressed in the endodermal lining of the gut, it would be unable to establish eye formation. The endoderm of the gut is not competent to develop an eye because it lacks the context in which *Pax6* gene activation can govern eye formation. In the following section we explore an example of those networks and their transcriptional regulation.

Cochlea of the Inner Ear: A Paradigm for 2D Developmental Networks Utilizing Boundary Formation

A realistic application of such models would be for the development of an organ with two dimensions conveying crucial structural differences that are functionally meaningful. One such model would be the developmental network that generates the limbs and hands. Here we focus on another, clearly two-dimensionally organized organ—the mammalian cochlea—which converts sound energy into electric signals that convey frequency- and intensity-specific sound information. The mammalian cochlea is a spiral organ in which a functional longitudinal, tonotopic gradient is the basis for frequency-specific sound analysis. The cytoarchitecture of the cochlea is significantly altered across the longitudinal axis, where changes in specific cell types are quite dramatic. The cochlea can also be subdivided radially into the spiral ganglion, organ of Corti (with one row of inner and three rows of outer hair cells), and the lateral wall, which includes the stria vascularis. In essence,

the cochlea represents a 2D organ with concentric and longitudinal differences in cell types. The adult cochlear structure requires at least two concurrent and overlapping developmental networks: one for generating the elongation of the cochlea and another for engendering the radial changes in cell types. In addition, given the existence of functionally distinct rows of single cells (one row of inner hair cells, three rows of outer hair cells), some of the reaction-diffusion models computed by Salazar-Ciudad et al. (2001b), which involve the presence of sharp boundaries, could well apply for cochlear development (Figure 3).

On the molecular side, a number of factors that are crucial for distinct steps in cochlear development have been identified. Several bHLH genes are known to control cell fate determination in the cochlea. The first of these factors to appear in development, *ngn1*, is essential for all sensory neuron formation. *Math1* is essential for all hair cell formation. *Hes1* and *Hes5* are specific for supporting cell formation out of pluripotent supporting cell/hair cell precursors (Zine et al., 2001). Expression of *Math1* drives, through the *Delta/Notch* system for lateral inhibition, the upregulation of *Hes1/5* in supporting cells. Expression of *Hes1/5* is essential to suppress *Math1*; in *Hes1/5* null mutants, some supporting cells assume a hair cell fate, as evidenced by additional hair cell rows (Zine et al., 2001). The exact mechanism(s) for the formation of the spiral ganglion, hair cells, and the supporting cells in their specific topology are still relatively unknown. However, there is evidence for molecular and possibly cellular interactions in assigning distinct cell lines to neuronal and hair cell phenotypes. The mechanisms for upregulation of a specific bHLH gene are tied into those that suppress upregulation of others. Thus, once a cell has upregulated, say, *Math1*, it cannot simultaneously upregulate *Hes1* or *ngn1*. In other words, once upregulated, one of the bHLH factors will activate the genetic network that ultimately leads to the specific cell differentiation.

A gradient exists with respect to the last mitosis of hair cells (first in the apex, last in the base), whereas the process of proliferation and differentiation of spiral ganglion cells starts in the base and finishes in the apex. Interestingly, this apical to basal hair cell proliferation gradient differs from almost all other developmental gradients, which always run from base to apex (e.g., Fariñas et al., 2001). These two complementary developmental genetic networks could be the foundation for the establishment of countergradients of gene expression. The resulting countergradient could lead to a

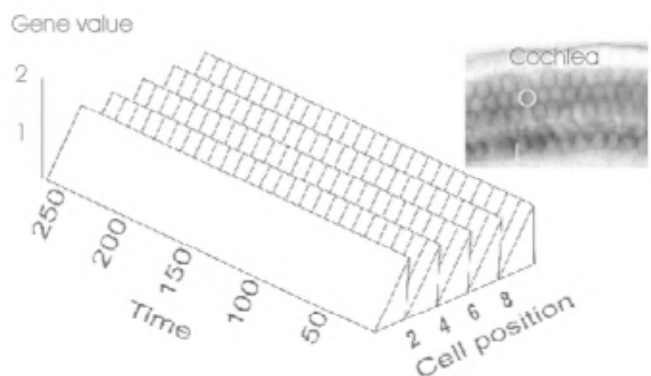


Figure 3. The outcome of a simulation using the nonexpansive emergent module is compared to the arrangement of four sensory hair cells in the cochlea (I, one row of inner hair cells; O, three rows of outer hair cells) as revealed by BDNF expression. It is possible that as few as four different genes and their interactions determine the pattern of the hair cell distribution in the cochlea. (Modified from Salazar-Ciudad et al., 2001a, and Fariñas et al., 2001.)

greater disparity of gene expression between the base and apex of the cochlea. A number of regulatory proteins and/or their receptors are expressed in the inner ear during embryonic and postnatal development and could influence the basal to apical longitudinal developmental expression patterns. These are the diffusible factors and their receptors, *BDNF/trkB* and *NT3/trkC*, *Jag2/Notch*, and *FGF8/FGFR-2(IIIb)* and the transcription factor *GATA3*. For example, in the course of hair cell differentiation the acetylcholine receptor channel $\alpha 9$ gene is upregulated in a basal to apical sweep starting at approximately embryonic day 15.5. In general, ion channel expression patterns in the inner ear follow the general developmental sweep from base to apex first in IHC, then in OHCs. Furthermore, this apparently expansive network may have quantitative effects on downstream gene expression, which could be displayed by differential continuous and discontinuous longitudinal gradients (Beisel et al., 2000).

In many respects, the longitudinal gradient in the cochlea parallels the dorsal-ventral network of bHLH expression in spinal cord and brain. In the brain, a *Mash1* expression domain is ventral to an *ngn* expression domain, which is capped by a *Math1* expression domain. These domains are mutually inhibitory (Gowan et al., 2001) and are directly involved in linking pathways of neurogenesis and regional specification to the formation of distinct functional longitudinal zones (Bermingham et al., 2001). One likely possibility for pattern formation in the cochlea is that the inner ear builds on these developmental regulatory networks. Thus, *ngn1* is expressed in spiral sensory neurons prior to *Math1* expression and inhibits its cellular upregulation, which would specify hair cell formation. Instead of forming longitudinal stripes, as in the brain, the interaction between those genes and the cells carrying them is now forming radially distinct, concentric columns of sensory neurons (more medial, *ngn1* dependent) and hair cells (more lateral, *Math1* dependent). Temporal extension of this interaction by repetitive utilization of this patterning network will result in the longitudinal extension of the cochlea as exemplified in its spiraling growth. Superimposing on this module other developmental modules for finer organization of the existing macropattern will organize the large disparities between the cochlear cell types and the cytoarchitecture.

Beyond these basically 2D models of cellular developmental organization, we next need to investigate 3D pattern formation, which is the real hallmark of brain development.

Realistic Developmental Networks and Their Evolution Involved in the Central Nervous System

The brain can be broken down into longitudinal and transverse compartments. We have already introduced some of the molecular mechanisms for longitudinal functional column formation. Transverse boundaries are in the form of neuromeres, which coincide with specific gene expression domains such as Hox genes. One such transverse boundary has been identified as a major organizing center of molecular and cellular interactions, the midbrain/hindbrain boundary (MHB; Liu and Joyner, 2001). This area is crucial for a number of specific neuronal aspects of the vertebrate brain, notably the cerebellum, the area of the brain that contains the most neurons (Wang and Zoghbi, 2001). This MHB is induced by a negative feedback loop involving two transcription factors, *Otx2* and *Gbx2*, which are expressed in adjacent domains (Figure 4). Absence of *Otx2* reduces the entire forebrain/midbrain formation to the level of the otocyst; absence of *Gbx2* eliminates the MHB and expands the midbrain toward the level of the otic vesicle. Within the expression domain of *Gbx2* an upregulation of *FGF8* occurs, which is inhibited by *Otx2*. But while *FGF8* inhibits *Otx2*, it promotes *Wnt1*, *Engrailed*, and *Pax2/5* expression. Despite all this information, critical steps in this reasonably complicated in-

teractive network are not yet understood. One obstacle hindering a precise model for the data is that the expression of other genes is eventually lost in a specific knockout (null) mutant (e.g., the *Gbx2* null). The associated genes appear to be upregulated initially in the proper spatiotemporal pattern, suggesting that the underlying patterns specifying the upregulation of these genes are not yet known. However, it is clear that the above outlined genes are essential for the continued expression of the other genes, as indicated in Figure 4.

Previously, we introduced the context dependence of gene expression and its regulation. It is noteworthy that none of the MHB genes, either alone or in combination, can induce formation of a posterior midbrain or a cerebellum in areas such as the caudal hindbrain or the forebrain. This suggests that another set of as yet undiscovered transcription factors must be expressed in the cells of the MHB area to render them competent to respond to the transcription factors outlined above. Those very genes may also be the ones that upregulate the initial expression of *Otx2* and *Gbx2*.

Evolution of the MHB seems to have occurred in steps. An *Otx2* homolog is expressed in the anterior part of the brain vesicle in most chordate species that have been analyzed (Hullond et al., 2000). However, neither spatial nor temporal expressions of all the molecular members necessary to form the mammalian MHB seem to be in place. Specifically, although a homolog of *Wnt1* exists in all chordates, it is not expressed in the typical vertebrate pattern in each chordate (Holland, Holland, and Schubert, 2000). Moreover, the evolution of crucial members of this boundary, such as *FGF8*, is unlikely to be conserved across phyla, as flies have only one *FGF* compared to the 24 *FGFs* known for man (Venter et al., 2001). In addition, formation of the cerebellum and many associated structures further requires the presence of *Math1* (Bermingham et al., 2001). Thus, although certain chordates do have some of the genes that regulate the MHB formation and these genes are expressed in a topologically comparable pattern, even the presence of all those players requires additional implementations of other genes to generate the cellular basis to build a cerebellum. In other words, if fully analyzed, the MHB example may elucidate how an existing expression genetic network implements more genes and expands its regulatory basis to transform midbrain and hindbrain neuronal tissue into a cerebellum.

Discussion

Patterning is an evolving process that can generate new patterns by modifying existing genetic networks (Davidson et al., 2002). Thus, cell fate within those genetic networks can be further modified from existing patterning processes by merging other patterning modules

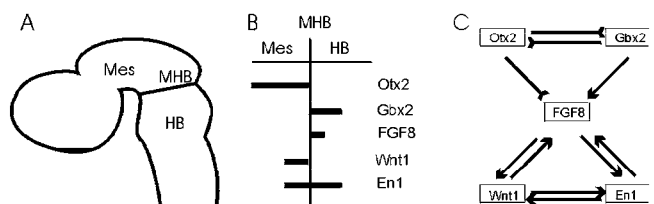


Figure 4. The midbrain/hindbrain boundary (A) is one of the best-understood boundaries in brain development. At least five genes (B) are interacting to form this boundary in both positive (arrows) and negative interactions (arrow tail; C). Each gene is expressed in specific stripes at or across the boundary (B). En1, engrailed 1; FGF8, fibroblast growth factor 8; Gbx2, gastrulation brain homeobox 2; Otx2, orthodenticle homolog 2; Wnt1, wingless related MMTV integration site 1. (Adapted from Wang and Zoghbi, 2001, and Liu and Joyner, 2001.)

into an existing genetic network as well as by implementing new downstream players. Currently, the increasing use of knockout (null mutant) mice has greatly facilitated our understanding of genetic networks in development of the central and peripheral neural systems. In many cases these studies demonstrated that knocking out a single gene can affect the expression of hundreds of genes, confirming that genes do not work in isolation. In other null mutants, however, no overt effects are observed, suggesting that biological redundancy may be playing a role. The use of transgenic mice, involving gene dosage or in other cases dominant-negative mutations, may provide additional approaches to elucidate the role of genetic networks in formation and function of neuronal networks. However, these models must be interpreted with caution. Genes, which affect development to yield a variety of CNS deformations, provide limited insights into “how the brain works” because many of these mutations are early lethal and hence cannot be studied. The anticipated extensive use of conditional mutations, in which a gene is deleted in a limited area of the brain only, will likely change this picture dramatically in the near future.

Road Map: Neuroethology and Evolution

Related Reading: Axonal Path Finding; Evolution of Artificial Neural Networks; Evolution of the Ancestral Vertebrate Brain

References

- Beisel, K. W., Nelson, N. C., Delimont, D. C., and Fritzsch, B., 2000, Longitudinal gradients of KCNQ4 expression in spiral ganglion and cochlear hair cells correlate with progressive hearing loss in DFNA2, *Brain Res. Mol. Brain Res.*, 82:137–149.
- Bermingham, N. A., Hassan, B. A., Wang, V. Y., Fernandez, M., Banfi, S., Bellen, H. J., Fritzsch, B., and Zoghbi, H. Y., 2001, Proprioceptor pathway development is dependent on Math1, *Neuron*, 30:411–422.
- Fariñas, I., Jones, K. R., Tessarollo, L., Vigers, A. J., Huang, E., Kirstein, M., de Caprona, D. C., Coppola, V., Backus, C., Reichardt, L. F., and Fritzsch, B., 2001, Spatial shaping of cochlear innervation by temporally regulated neurotrophin expression, *J. Neurosci.*, 21:6170–6180.
- Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Rust, A. G., Pan, Z., Schilstra, M. J., Clarke, P. J., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., Bolouri, H., 2002, A genomic regulatory network for development, *Science* 295:1669–1678.
- Gowan, K., Helms, A. W., Hunsaker, T. L., Collisson, T., Ebert, P. J., Odom, R., and Johnson, J. E., 2001, Crossinhibitory activities of *ngn1* and *math1* allow specification of distinct dorsal interneurons, *Neuron*, 31:219–232. ♦
- Holland, L. Z., Holland, N. N., and Schubert, M., 2000, Developmental expression of *AmphiWnt1*, an amphioxus gene in the *Wnt1*/wingless subfamily, *Dev. Genes Evol.*, 210:522–524.
- Isshiki, T., Pearson, B., Holbrook, S., and Doe, C. Q., 2001, *Drosophila* neuroblasts sequentially express transcription factors, which specify the temporal identity of their neuronal progeny, *Cell*, 106:511–521.
- Liu, A., and Joyner, A. L., 2001, Early anterior/posterior patterning of the midbrain and cerebellum, *Annu. Rev. Neurosci.*, 24:869–896. ♦
- Pichaud, F., Desplan, C., 2002, Pax genes and eye organogenesis, *Curr. Opin. Genet. Dev.* 12:430–434
- Robert, J. S., 2001, Interpreting the homeobox: Metaphors of gene action and activation in development and evolution, *Evol. Dev.*, 3:287–295.
- Salazar-Ciudad, I., Newman, S. A., and Sole, R. V., 2001a, Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships, *Evol. Dev.*, 3:84–94.
- Salazar-Ciudad, I., Sole, R. V., and Newman, S. A., 2001b, Phenotypic and dynamical transitions in model genetic networks. II. Application to the evolution of segmentation mechanisms, *Evol. Dev.*, 3:95–103. ♦
- Venter, J. C., Adams, M. D., Myers, E. W., et al., 2001, The sequence of the human genome, *Science*, 291:1304–1351. ♦
- Wang, V. Y., and Zoghbi, H. Y., 2001, Genetic regulation of cerebellar development, *Nat. Rev. Neurosci.*, 2:484–491.
- Zine, A., Aubert, A., Qiu, J., Therianos, S., Guillemot, F., Kageyama, R., and de Ribaupierre, F., 2001, *Hes1* and *Hes5* activities are required for the normal development of the hair cells in the mammalian inner ear, *J. Neurosci.*, 21:4712–4720.

Evolution of the Ancestral Vertebrate Brain

Bernd Fritzsch

Introduction

Earlier work on the evolution of the vertebrate brain centered on the description and functional analysis of adult brains, with limited attempts to project the origin of various nuclei back to the ventricular surface, thereby generating flat, two-dimensional maps (Nieuwenhuys, ten Donkelaar, and Nicholson, 1997). The remarkable variations in size and shape of parts of the brain were often viewed as evidence of a progressive increase in complexity. The usual implication drawn was that information processing between the sensory input and the motor output became more sophisticated as the number of neurons increased. However, mere size is a tricky issue, as humans and dolphins share both absolute and relative size ratios, but we have no objective way to show which brain is more complex.

In this context it is important to understand that fairly little is known with respect to variations in connections between neurons from topologically comparable areas in different animals. This is unfortunate since altered connections between neurons are crucial functionally relevant aspects of brain variation.

In part this is so because there are always several ways to implement a certain function within existing connections using molecular and cellular mechanisms. These mechanisms are only partly

understood (Koch and Laurent, 1999). It would therefore be naive to work out all connections of every neuron assuming that this alone will help to understand the function of the system in question. Ideally, one would need to have this information and record simultaneously from all neurons involved in a given behavior to unravel the computational properties of even small neuronal networks. Such networks might be those formed by the 302 neurons in the worm *Caenorhabditis elegans*, or the 30 cells of the stomatogastric ganglion. The next best thing is recording from individual neurons under situations of stable behavior, and using these data to simulate a network that can perform the same task using the measured parameters. An excellent example is the fictive swimming in lampreys. Another example is the use of cortical activity to govern robotic movements. For obvious reasons, achieving this in the ancestral vertebrate brain is impossible, and we have to find other ways to gain insights into the origin of the nervous system that evolved into the vertebrate brain.

In this overview I discuss the possible origin of the vertebrate brain, then present a possible ontogenetic way through which the problem of homology/homoplasy may be minimized. In addition, possible mechanisms to diversify the ancestral vertebrate brain are suggested, and data concerning structural changes whose functional implications are as yet unclear are presented

When and How Did the Vertebrate Brain Form?

The origin of the vertebrate brain is tightly coupled to the origin of the vertebrate head, which likely happened around 600 million years ago (Knoll and Carroll, 1999). Next to nothing is known about the central nervous system of these ancestral, bilaterally symmetric animals. We can only assume that it consisted of a simple, dorsally located tube of nerve cells with a barely recognizable specialization of what would become the vertebrate brain at its anterior end.

However, the emerging synthesis of comparative morphology, development, and paleontology offers a new approach to understanding vertebrate brain evolution. Adult diversity is now viewed as the outcome of divergent genetic developmental mechanisms. Thus, resolving brain evolution through comparison of adult structures can be aided by comparative neuroembryology, that is, by comparing adult structures with their specific development and the genes that guide such development. Such comparative embryological data in conjunction with gene expression patterns have aided in clarifying major issues of brain development (Holland and Holland, 1999).

The vertebrate brain forms through invagination of ectoderm (the embryonic "skin") to form a neural tube. During further development the central nervous tissue becomes polarized and then subdivides into compartments, each characterized by a specific pattern of gene expression (Figure 1). It is this latter aspect that may eventually allow us to obtain a map of expression of homologous genes (based on nucleotide sequence identity) to identify topologically comparable and thereby homologous parts of the developing vertebrate brain, irrespective to their structural similarities. Thus far, only rudiments of this map are known, and the relationship of gene expression patterns to the structural evolution of topologically comparable areas is largely unknown (Bermingham et al., 2001). In a few years we will have the molecular delineation of identical subdivisions of the vertebrate brain based on the nested expression of homologous genes. This is a prerequisite to sorting out how identical areas of the brain (defined by their gene expression patterns) have evolved structural differences that serve specific ad-

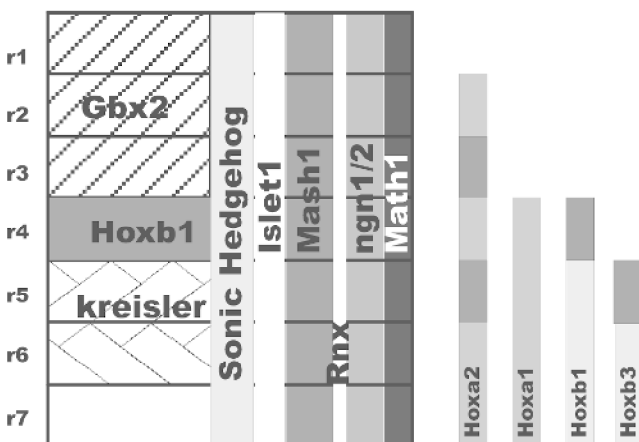


Figure 1. The expression of some longitudinal (Mash1, ngn1/2, Math1, Rnx) and transverse transcription factors (Gbx2, Hoxb1, kreisler) is shown. These expression domains produce an orthogonal grid in which brainstem neuronal phenotypes are uniquely specified. On the left is an idealized flat mounted embryonic day 11 mouse hindbrain. r1–r7 indicate rhombomeres. Note that homeobox genes (Hox) end at different levels with varying expression intensity. Sonic hedgehog defines the floor plate between the two halves of the brain. (Genes and their expression are adapted from Cordes, 2001, and Qian et al., 2001.)

aptations unique to a specific animal. However, it also appears that molecular identification of major subdivisions of the vertebrate brain, such as forebrain, midbrain, and hindbrain, is possible in all chordates, based on the nested expression of certain homologous genes (Holland and Holland, 1999). Other genes, while conserved in their sequence, appear to have altered their developmental expression patterns in various groups of animals. For example, the left-right asymmetry determination in mice and chickens uses the same molecules but in a different pattern. Moreover, there is ample evidence that developmental genes can be co-opted into novel developmental pathways, either alone or in addition to their original task, (see EVOLUTION OF GENETIC NETWORKS). Moreover, a specific gene can be of conserved importance for forebrain development and evolution but also essential for specific aspects of ear development (Fritsch, Signore, and Simeone, 2001).

The essence of these findings is that the developmental patterning of the vertebrate brain evolved over 600 million years ago and has been rather stable in many aspects of developmental gene expression. Nevertheless, compartments of the forming brain, once specified in their relative position through conserved gene expression, have diverged to form the unique anatomy of a given species.

The vertebrate head (and brain) evolved in part owing to novel embryonic material, neural crest and placodes, which contribute to all sensory systems of the head. Neural crest is embryonic neuronal material that emerges from the forming neural tube and undergoes extensive migration to form branchial arches, bones, teeth, and peripheral ganglia. Placodes are cake-like epidermal thickenings that contribute sensory neurons to cranial ganglia. Most recently, specific molecules have been identified that are associated with neural crest development. Interestingly, these molecules have also been identified in vertebrate relatives that lack neural crest. This provides molecular evidence that neural crest precursors may exist in the brain and spinal cord of these animals (Corbo et al., 1997). What is important here is that molecular markers may delineate identical populations of neurons that can evolve a novel morphology and function.

The view proposed here is that the evolution of the brain is a sequence of developmental variations leading to modified adult structures. These modified structures serve somewhat different functions that, in turn, are chosen in the process of natural selection (Raff, 1996). Thus the role of mutation in the selection of novel structures and functions is rather indirect (Fritsch, 1998a). Arguably, the motor system, the common output of the brain, is the best understood part of the brain, as we can clearly define the output in terms of measuring the generated movement and can actually simulate the movement. In addition, we know more about its evolution (Fritsch, 1998a) and the molecular governance of its compartmentalization (Cordes, 2001) than in any other part of the brain. In the following section I will focus on the evolution of some cranial motor neurons.

Evolution of the Brainstem Oculomotor System

Recent research has shown that developmental selector genes (transcription regulation factors: see EVOLUTION OF GENETIC NETWORKS) play an important role in the differentiating vertebrate brain. These networks appear to form a space map that possibly orchestrates the activation of topologically appropriate structural genes in longitudinal columns and their rostrocaudal subdivisions. Some of these transcription regulators, known as basic helix-loop-helix genes (bHLH genes), may play a role in the dorsoventral patterning of the differentiating brain and spinal cord (Bermingham et al., 2001), while others, known as homeotic genes, play a role in rostrocaudal subdivisions (Cordes, 2001). These genes were identified first in the fruit fly, where they are related, for example, to homeotic transformation (changes in the developmental fate of compartments). Subsequently, homeotic and other selector genes,

such as bHLH genes were identified in vertebrates, nonvertebrates, and plants. In vertebrates, these genes apparently govern (1) the formation of longitudinal columns of, for example, taste-related nuclei (Qian et al., 2001), and (2) the regionalization of these columns into domains destined to serve a specific cranial nerve (Corde, 2001). Thus, a developing neuron will undergo activation of a different set from the 30,000+ genes available to differentiate according to its position within this spacemap. This differentiation leads to formation of classes of neurons that will have a characteristic morphology and set of afferents and efferents. However, the detailed connectivity of individual neurons varies as a result of both the regularities and happenstance of development.

The following discussion highlights what is known about the variable organization of a simple system composed of three neuron populations that constitute the vestibulo-ocular pathway, the oculomotor system. The oculomotor system is a unique model because both the input (gravistatic and angular acceleration from the ear) and the output (movement of the eyes) can be quantified and put in the appropriate behavioral context. In its basic pattern, the system has six eye muscles to move the eye and three sets of motor neurons to drive these muscles. A series of bilaterally projecting interneurons (named vestibular nucleus neurons) mediates the input from the ear to the motor neurons. Sensory input from the ear travels via the semicircular canal (for angular acceleration) and two or more gravistatic sensors (for linear acceleration, including earth gravity). Four of the six eye muscles in jawed vertebrates are innervated by the oculomotor motor neurons, but only three out of six eye muscles in lampreys are so innervated (Fritsch, 1998b). The abducens innervates only one of the six eye muscles (but sometimes an additional muscle, see later discussion) in jawed vertebrates, but it innervates two eye muscles in lampreys (Figure 2). The remaining muscle is innervated by the trochlear motor neurons in all vertebrates. In the ear, jawless vertebrates, such as lampreys, have no horizontal canal, which is one of the major inputs into the oculomotor system of jawed vertebrates that drives compensatory horizontal eye movement.

Thus, the oculomotor system shows two major evolutionary changes in vertebrates: one is related to changes in the eye muscles and their innervation and the other is related to the ear and its additional formation of a semicircular canal. Interestingly, the connections within the brainstem from the inner ear projection to the oculomotor neurons seem to be fairly constant among all vertebrates (Fritsch, 1998b; Baker, 1998). This common feature apparently provides enough built-in plasticity to accommodate changes on the input and output side without major reorganization of the interconnections.

It is unknown how computation of horizontal acceleration is performed and then transformed into the different coordinates of ocular muscles in lampreys. This vertebrate lacks a horizontal canal but obviously has a functional oculomotor system (Figure 2). If we are to understand the selective advantage (if any) of this reorganization in jawed vertebrates, we have to unravel the structural/functional relationship in this likely primitive pattern of the three-neuron vestibulo-ocular arch in lampreys, then compare this system with the two functionally equivalent vestibulo-ocular systems of jawed vertebrates (Figure 2). Such a comparison could provide exemplary insight into the functional constraints that may accompany the structural transformation of homologous systems. Interestingly, this system has retained its basic function, vestibular governance of eye movement, while at the same time modifying some of its properties.

The major driving force behind this eye muscle reorganization may not have been the need of the system to achieve a different function. Rather, “accidental” developmental changes may have led to a subdivision of eye muscles and their nerves (Fritsch, 1998b). Alternatively, variation in the inner ear (Fritsch et al., 2001) may have been the force driving eye muscle reorganization. Once such

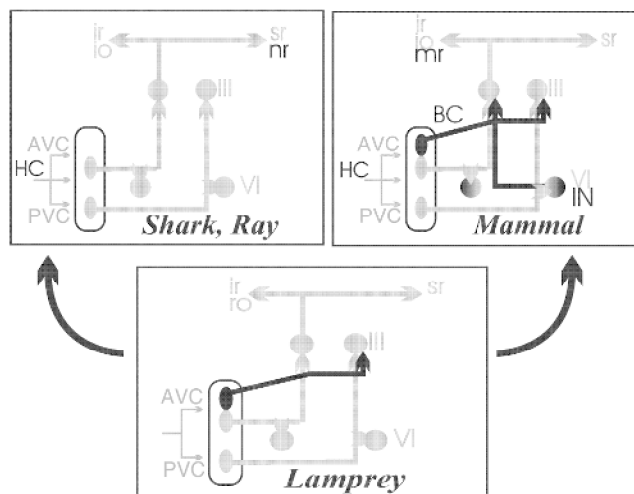


Figure 2. Reorganization of the oculomotor system among vertebrates. Lampreys have three eye muscles (inferior rectus [ir], rostral oblique [ro], and superior rectus [sr]), jawed vertebrates have four eye muscles innervated by cranial nerve III. However, in sharks and rays the nasal rectus (nr) receives a crossed innervation, whereas the functionally equivalent medial rectus (mr) of mammals receives an uncrossed innervation. All vertebrates have a rostral population in the vestibular nuclei that projects uncrossed and a caudal population that projects crossed to oculomotor neurons. Lampreys and mammals have an additional rostral population that projects through the brachium conjunctivum (BC) to the cranial nerve III nucleus. Note that lampreys have no horizontal canal (HC), and that only mammals have internuclear neurons (IN) from cranial nerve VI to III that provide the basis for conjugated horizontal eye movements. This task must be performed differently in lampreys and sharks. AVC, anterior vertical canal; PVC, posterior vertical canal; io, inferior oblique muscle. Dashed line indicates the midline. (Modified from Fritsch, 1998b.)

changes were in place, the entire oculomotor system implemented the computational alterations generated by this “new” sensory structure into the preexisting connection pattern. This change was nevertheless enough to achieve implementation of horizontal gaze control. Understanding the adaptive implications of these changes would require a detailed comparison of neuroanatomical, physiological, and behavioral data to gain insights into the ecophysiological context of adaptation of this system.

How Does the Brain Change Its Pattern?

Thus far I have largely dealt with factors that keep neuronal development constant. However, despite the overall similarity in the regional subdivisions, it is clear that specific neuronal connections differ between species and perform different and sometimes novel functions. Logically, there are only three possible ways in which neurons can be made available to perform new functions:

1. Through increased proliferation of an existing population that forms redundant and therefore potentially uncommitted neurons.
2. Through loss of an old input and/or target, which frees neurons from their previous functional constraints and allows them to adopt a new commitment, thus adapting their function to changes in the sensory or motor system to which they are linked.
3. Through de novo formation of a novel set of neurons.

The last possibility requires that neurons somehow escape the pattern of gene activation normally mediated through the spatially restricted expression of developmental selector genes (i.e., they

were set aside) (Knoll and Carroll, 1999). This could happen in three ways: (1) by alteration of selector gene expression through, for example, upstream changes in regulatory factor gradients that activate these genes, (2) by mutation of selector genes, resulting in changes in the expression pattern and thereby changes in the pattern of activation of downstream genes (see EVOLUTION OF GENETIC NETWORKS), or (3) by mutation of downstream genes to respond differently to selector genes. There is growing evidence that among chordates many selector genes may have a rather stable pattern of expression (Holland and Holland, 1999). Thus, differences in the activation of downstream genes or mutations in downstream genes may be a major mechanism by which nervous system organization is varied.

Increased proliferation and formation of more neurons clearly happened in the evolution of vertebrates, as exemplified by the relatively (brain to body weight) and absolutely larger-sized brains of humans compared to bony fishes. Moreover, specific areas may in some vertebrates become the largest part of the entire brain, while other species may entirely lack these areas. For example, the valvulae cerebelli are unique to bony fishes and can become the largest part of the bony fish brain, much as the forebrain has become the largest part of the human brain (Nieuwenhuys et al., 1997). It is thought that this growth of the valvulae cerebelli is related to the unique electroreceptive sense of these bony fishes, a sense all land vertebrates have lost. In contrast, the growth of the cerebellum in four-legged vertebrates is in part related to the growth of the forebrain (which provides a major input through the pontine nuclei).

This increased proliferation could come about through a simple mutation in genes regulating the proliferation of precursor cells. For example, if four divisions occur in a precursor cell population of ancestors, adding one round of mitosis could double the total number of neurons. Subsequently, some neurons of this initially identical population could develop a different identity. How this new identity could be achieved is still unclear, but one possibility might be that the enlarged population of postmitotic neurons could see a slightly different gradient of selector gene products and could therefore achieve their different phenotypes (Raff, 1996). Thus, neurons were developmentally set aside (Knoll and Carroll, 1999). The differences in function would be obtained by (1) reaching a different target and/or (2) segregation of their perikarya and dendrites (and thus input) through differential migration.

A second possible scenario for the formation of uncommitted neurons may be the loss of either the target or the input. Such a loss would eliminate the constraints normally acting on these neurons and thus would allow them to evolve a new function. A well-known example of this scenario is the evolution of the middle ear ossicles of land vertebrates. There is good comparative and developmental evidence suggesting that these ossicles are derived from former jaw-supporting ossicles. Once their original functional constraint, supporting the jaws, was lost, they underwent a radical change in function.

Experimental reorganization of input changes the function of existing neuronal networks and entire sensory systems. The mechanosensory lateral-line and the electroreceptive system of ampullary organs of fishes and many amphibians is lost in most land vertebrates. Although there is evidence for evolutionary loss of both inputs and targets, it is not yet proven that the neurons freed by such a process from their previous functional constraints are in fact modified to perform a novel, different function. However, it is clear that experimental changes of inputs change the function of existing neuronal networks (Pallas, 2001). Thus, while experimental evidence tends to support the notion of functional changes in existing neuronal networks, there is little evolutionary evidence to support it, and the effect of such a process for brain evolution has not yet been worked out.

From New Neurons to New Functions

Irrespective of how new neurons evolved, they have to achieve new input/output relations to mediate any new function. Clearly, migration of neurons into a new position is a widespread phenomenon in the developing brain. It is widely agreed that neuronal migration correlates with the formation of novel input to this differently positioned subset of neurons. Migration can bring neurons from a dorsal part of the brain to the ventral part, or from one side to the other (Fritzsche, 1998a). One likely scenario is that homologous neurons with comparable function can differ in their position, as occurs in the cases of the laminar nucleus of birds and the medial superior olive in mammals. Both are relay neurons for the auditory pathway, but in birds (and reptiles) they are dorsal, next to the primary auditory nuclei, whereas in mammals they are ventral, near the base of the hindbrain. These differences in position can be reconciled with the assumed homology of these populations by showing that indeed, both nuclei arise dorsally but migrate ventrad in mammals, and that their initial formation is under the control of homologous genes. This change in position of neurons will affect the computation only if the input or output is changed.

In contrast to the well-accepted role of migration in achieving a novel input, how a novel target is reached by the axon is much more controversial. Some ideas have emanated from the undeniable fact of widespread, exuberant projection to different targets during development in birds and mammals. Out of these many targets a single or few targets will be selected by neuronal cell death and axonal pruning. In their ultimate version, such ideas require that no new connections ever form in the brain, arguing that differences are achieved exclusively through the differential loss of connections present in ancestral forms (Ebbesson, 1984). Although appealing, this idea fails to integrate the available data on the development of the peripheral and central nervous system, which show a rather precise selection of pathway choices by a navigating axon with limited developmental "error." Certain molecules have been identified that guide commissural systems in vertebrates and invertebrates alike. Other molecules specify the topology of the retinal projection onto the midbrain. Thus, beyond specifying the identity of topologically restricted populations of neurons, molecules also specify topologically restricted connections within the brain. However, maps generated by different sensors of the world around us form in the brain based on different principles. In the visual system, mapping of the eye onto the midbrain is achieved molecularly by using matching gradients of molecules that provide positional identity of neurons in the retina and target them toward specific areas in the tectum of the midbrain. In contrast, the space map of the hearing system is generated by computing time and intensity differences between both ears into an auditory space map. Thus, there likely will not be a uniform molecular principle for brain map formation in the various sensory modalities. However, the various maps, once established, can be brought into register with each other in certain areas of the brain, such as the roof of the midbrain.

One of the most striking examples of novel pathway selection is the growth of axons from the mammalian cortex to the spinal cord to form the corticospinal tract. The pyramidal neurons are the only cortical neurons in vertebrates that do project to the lumbar spinal cord and thus are considered by most researchers to be a novel tract. Any other interpretation would require the assumption that this tract is ancestral to vertebrates and was lost in all lineages except mammals. Even more compelling evidence for the *de novo* formation of pathways comes from fiber outgrowth into the periphery. Although these mechanisms were rejected for almost a century, it is now clear that motor neuron axons do not pass through the "ventral roots" of certain chordates. Instead of nerve fibers leaving the spinal cord, muscle fibers project with noncontractile processes to the "ventral roots" of the spinal cord of these small marine

animals. One of the major steps in the evolution of the vertebrate brain and spinal cord was then to have motor neurons, which themselves project through the ventral root to reach their target muscles (Fritzsche, 1998a). Although it is still unclear when and how this novel invasion of axons into the periphery happened, those motor neurons had to find a pathway where they had never been before. Moreover, the addition of novel tissue, such as the ear, may have caused redirection of growing fibers.

In the hindbrain of chickens, motor neurons initially form two paramedian strips (Cordes, 2001). Motor neurons within a restricted region project their axons to the facial root to exit the brain (Figure 1). Within the facial nerve some axons reroute to reach the developing ear and function as stato-acoustic efferents. Instead of innervating striated muscle fibers as the facial motor neurons do, they innervate hair cells of the ear (Karis et al., 2001). Moreover, this pathway of efferent fibers to the ear differs among vertebrates. This rerouting within the nerve occurs in chickens and frogs, but not in mammals, where it happens within the brain. In addition, the initially overlapping populations of efferent neurons segregate into different positions within the hindbrain through differential migration.

It is unclear how much the reorganization of the facial motor neurons to the ear to form the stato-acoustic efferents depended on the formation of the ear itself and a concomitant suppression of differentiation of parts of branchial arch-derived muscle fibers. Such losses of target muscles could conceivably have forced a subset of facial motor neurons either to innervate a new target and become the efferents to the ear or to disappear entirely. This highlights but one possibility of how connections within the brain can be changed and the resulting function of the network altered. More detailed connectional data are needed to show how often such changes have happened in the brain.

In conclusion, there is both evidence for ongoing invasion of certain fibers into novel territories and evidence for rather precise pathway selection. This precludes the idea that evolution picked from a completely randomized network. Certain areas, like the forebrain, may in fact benefit from a less constrained development that enables them to form a wider array of initial connections from which only certain connections will be selected in a later developmental step.

Variation in Functionally Relevant Details

Nervous tissue not only undergoes modifications in its long-range connections and formation of new cells, it also shows numerous cellular reorganizations. Such reorganization could manifest as changes in the degree of branching of dendrites, neurotransmitter(s), postsynaptic receptors, and stratification into distinct laminae, or as the absence of these structures in topologically comparable areas of the brain (Nieuwenhuys et al., 1997). For example, the gustatory nuclei in closely related species of bony fishes may be laminated or not. Or the neurons of the cerebellar nuclei, which are all assembled in the white matter in mammals, may become stratified with Purkinje cells into a single cortical layer in bony fishes. What (if any) functional implications these differences in cellular assembly may offer is still unclear, and the variations may represent nothing more than alternative designs to compute the same information. More physiology is needed to distinguish what each network is good for. This is particularly obvious in the stratification of the midbrain tectum of many vertebrates, which apparently can be secondarily reduced without appreciable functional deficits (Nieuwenhuys et al., 1997). Thus, salamanders may have a much less stratified midbrain tectum and yet are fully able to govern the protrusion of their tongue, as accurately as a reptile or a frog, which have a more stratified midbrain. In other words, we may be looking at structural organizations that do little for the underlying function but keep us baffled at how they vary.

Another example is provided by the differences in organization but similarities in function and long-range connections in the forebrain of birds and mammals. Obviously, comparable functions can be accomplished with either a laminar, cortical organization or with a set of interconnected groups of neurons. It has been suggested that “laminar organization of populations is an alternative means of organizing populations of neurons” (Nieuwenhuys et al., 1997). It appears, then, that some of the highly appreciated, laminar networks are but one way in which the brain can implement a specific functional circuit. It is entirely possible that the morphological differences in the forebrain organization of birds and mammals are not driven by their physiology and adaptivity to behavioral tasks but rather reflect alternative developmental strategies that were implemented simply because the outcome is a viable organism.

Discussion

This brief overview has stressed some of the emerging developmental principles presumed to govern the remarkably stable overall pattern of the vertebrate brain throughout its evolution while providing enough room for plasticity in both local interactions as well as long-range connections. The developmental analysis has not yet reached a level that can causally explain how local neuronal assemblies, such as cortical columns, come about during development and how the connections between those neurons are organized to adapt the organism to various tasks.

However, it is proposed that complex interactive genetic networks not only guide overall development but offer room for self-organization of these neuronal assemblies within a set of limitations imposed by the developmental selector genes. Thus, whereas some behavior, such as a reflex, is directly translated from the genes through development into a specific set of connections forming the reflex arc, other connections have enough redundancy to be modified in their response by experience and are able to learn. Evolution not only selects modifications in specific connections that allow for a more adapted response, it also selects neuronal networks that allow for learning and thus adaptation during the life of an organism within the genetic limitations in which the brain and its connections developed.

Looking at the evolution of the brain as the evolution of a system that has both genetically and behaviorally hardwired components and other components that can change their response properties allows us to understand the “adaptiveness” of the brain from a less restricted, more dynamic perspective. Adaptation may be viewed as a compromise between change and development in a system that struggles to stay on top of its changing adaptive landscape (Raff, 1996) by balancing both components in various ways. Clearly, hardwiring a response provides superior response speed but restricts adaptation to change during an organism’s lifetime. In contrast, learned responses, while initially slower, allow adaptation over the lifetime of an organism. Modeling such systems must take into account that optimization is constrained by at least these two facts, and thus may never be achieved in any living system to the extent that it can be achieved in artificial systems. Although there have been interesting models of the EVOLUTION OF ARTIFICIAL NEURAL NETWORKS (q.v.), their biological relevance has been somewhat limited by the use of selection on features, such as individual synaptic weights, that are not necessarily regulated by separate genes in biological nervous systems. Thus, an interesting challenge will be to develop a new generation of models in which evolution occurs within genetically plausible parameters, such as numbers of neurons, variation in long-range output connections, and variation in inputs into the dendrites.

Road Map: Neuroethology and Evolution

Background: Vestibulo-Ocular Reflex

Related Reading: Evolution of Genetic Networks

References

- Baker, R., 1998, From genes to behavior in the vestibular system, *Otolaryngol. Head Neck Surg.*, 119:263–275. ♦
- Bermingham, N. A., Hassan, B. A., Fernandez, M., Banfi, S., Bellen, H. J., Fritzsche, B., and Zoghbi, H. Y., 2001, Development of the proprioceptor pathway is MATH1-dependent, *Neuron*, 30:411–422.
- Corbo, J. C., Erives, A., Di Gregorio, A., Chang, A., and Levine, M., 1997, Dorsoroventral patterning of the vertebrate neural tube is conserved in a protochordate, *Development*, 124:2335–2344.
- Cordes, S. P., 2001, Molecular genetics of cranial nerve development in mouse, *Nat. Rev. Neurosci.*, 2:611–623. ♦
- Ebbesson, S. O. E., 1984, Evolution and ontogeny of neural circuits, *Brain Behav. Sci.*, 7:321–366.
- Fritzsche, B., 1998a, Of mice and genes: Evolution of vertebrate brain development, *Brain Behav. Evol.*, 52:207–217. ♦
- Fritzsche, B., 1998b, Evolution of the vestibulo-ocular system, *Otolaryngol. Head Neck Surg.*, 119:182–196.
- Fritzsche, B., Signore, M., and Simeone, A., 2001, *Otxl* null mutants show partial segregation of sensory epithelia comparable to lamprey ears, *Dev. Genes Evol.*, 211:388–396.
- Holland, L. Z., and Holland, N. D., 1999, Chordate origins of the vertebrate central nervous system, *Curr. Opin. Neurobiol.*, 9:596–602.
- Karis, A., Pata, I., van Doorninck, J. H., Grosveld, F., de Zeeuw, C. I., de Caprona, D., and Fritzsche, B., 2001, Transcription factor GATA-3 alters pathway selection of olivocochlear neurons and affects morphogenesis of the ear, *J. Comp. Neurol.*, 429:615–630.
- Knoll, A. H., and Carroll, S. B., 1999, Early animal evolution: Emerging views from comparative biology and geology, *Science*, 284:2129–2137. ♦
- Koch, C., and Laurent, G., 1999, Complexity and the nervous system, *Science*, 284:96–98.
- Nieuwenhuys, R., ten Donkelaar, H. J., and Nicholson, C., 1997, *The central Nervous System of Vertebrates*, Berlin: Springer-Verlag, p. 2200.
- Pallas, S. L., 2001, Intrinsic and extrinsic factors that shape neocortical specification, *Trends Neurosci.*, 24:417–423.
- Raff, R. A., 1996, *The shape of Life*, Chicago: University of Chicago Press, p. 520.
- Qian, Y., Fritzsche, B., Shirasawa, S., Chen, C.-L., and Ma, Q., 2001, Formation of brainstem catecholaminergic neurons and first order relay visceral sensory neurons is dependent on RNK, *Genes Dev.*, 15:2533–2545.

Eye-Hand Coordination in Reaching Movements

Valérie Gaveau, Phillipe Vindras, Claude Prablanc,
Denis Pélisson, and Michel Desmurget

Introduction

Despite a century of research, the neural mechanisms involved in eye-hand coordination during reaching movements are still largely unknown. This article addresses this question and describes the mechanisms whereby a visual input is transformed into a motor command. To this end, we consider the different problems that the nervous system has to solve to generate a movement, namely, localizing the target, creating a motor plan, and correcting the ongoing movement if necessary.

Representation of Target Position

Reaching toward a visual target requires transformation of visual information about target position with respect to the line of sight into a frame of reference suitable for the planning of hand movement. This problem is classically decomposed in analytic steps that provide target position information in an eye, head, and ultimately bodily frame of reference. For the sake of clarity, we follow this progression to describe the mechanisms encoding retinal information and extraretinal signals of eye-in-orbit and head-on-trunk positions.

Visual information initially signals the angle separating the target and the line of sight. The reliability of this retinal signal is constrained by the spatial anisotropy of the retina and visual system. Because of the gradient of visual acuity, the encoding of a target location with respect to the line of sight degrades when the stimulus falls in the peripheral visual field. This relative inaccuracy of signals from the peripheral retina can be illustrated by hand-pointing errors observed when the movement is performed while the foveating saccade is prevented. Despite this limitation, it is the peripheral part of the retina that is most often involved in the initial localization of a visual target.

In addition to the retinal signal, the position of the eye in the orbit is necessary to encode the location of the target in a body-centered frame of reference. Paradoxically, without a retinal signal, orbital eye position appears to be only coarsely encoded by extraocular signals. Indeed, when subjects are required to point in dark-

ness in the direction of their eyes, the final hand position correlates with eye position, but the scatter is much higher than when the target is a luminous spot (Bock, 1986). Thus, it appears that retinal and extraretinal signals do not simply add but also interact with each other, and that accurate encoding of target location requires concomitant foveal and extraretinal signals. Compatible with this hypothesis are recent studies suggesting that gaze position could influence the encoding of target location for limb motions (Soechting, Engel, and Flanders, 2001).

At the neurophysiological level, the search for interactions between retinal and extraretinal information has stimulated many studies on the neural code of target internal representations. Two different conceptions have emerged: single-unit coding and distributed coding. The single-unit coding concept of integration hypothesizes the existence of individual neurons encoding information about target position, irrespective of eye position. In support of this hypothesis, individual neurons representing symbolic parameters, such as target location in a body-centered reference system, have been described in several studies. For instance, the neuronal activities described by Duhamel et al. (1997) have been shown to encode the position of a visual target respectively in a head-centered frame of reference. This latter coding might hypothetically result from an ultimate stage of coordinate transformation necessary to direct the hand toward a target.

In contrast to the single-unit concept, the distributed coding hypothesis assumes a statistical combination of elementary information about retinal eccentricity and eye position within large neuronal populations. A growing body of evidence of population-based interactions between retinal and extraretinal information supports this concept. Thus, electrophysiological recordings in the parietal cortex of awake monkeys have shown that the activity of the reach-related cells was influenced by eye position information (Batista et al., 1999). With the aid of neural network modeling, Andersen and colleagues (1997) showed that these characteristics of individual neuron discharges are compatible with the existence in the parietal cortex of a distributed code for egocentric target localization.

It may be worth mentioning at this point that the single-unit and distributed concepts are not mutually exclusive. Indeed, symbolic information generated by distributed neuronal populations may ultimately converge at the output level to provide a single-unit representation. For example, the distributed model of Andersen and colleagues yields an output signal of target position relative to the head that is represented at a single-unit level. In addition, a recent electrophysiological study of ventral intraparietal neurons in the monkey showed that the visual response of single units reveals a continuum between head-centered coding and retinotopic coding, leading to the hypothesis that “space may be represented in the cortex both at the population level and at the single cell level” (Duhamel et al., 1997).

How head position signals are integrated with retinal and eye position signals has stimulated less neurophysiological investigation. In 1995, Brochie et al. reported that the visual response of parietal cortex neurons is modulated by the direction of gaze (integrating both eye and head components). This result suggests that the distributed coding hypothesis of target relative to the head can be generalized to visual target encoding in trunk-centered coordinates. Thus, target-related information in a body reference system seems to be distributed in large neuronal populations.

Planning Movement Trajectory

It is generally admitted that spatiotemporal invariances can give insight into how visually directed movements are planned and controlled by the nervous system. As an example, consider the task of pointing, from a given initial position, toward visual targets distributed within the workspace. In such a situation, two types of regularities can be expected: (1) *extrinsic regularities*, such that the movement displays invariant features in the Cartesian space (e.g., a straight-line path irrespective of the movement direction or amplitude), and (2) *intrinsic regularities*, such that the movement displays invariant features in one of the intrinsic spaces (e.g., a linear relation between joint angle variations). Interestingly, because the relationships between the extrinsic and intrinsic variables are complex and nonlinear, these two potential types of invariances generally cannot occur at the same time. That is, when the hand trajectory is invariant in the extrinsic space, it displays a consistent variability in the intrinsic space, and vice versa.

The task of pointing from a given starting position toward visual targets distributed within the workspace was initially studied by Morasso (1981). Morasso found that joint covariation patterns varied systematically as a function of movement direction, whereas Cartesian hand paths were always roughly straight. Based on this extrinsic stability, he concluded that (1) the hand trajectory in Cartesian space was the primary variable computed during movement planning, and (2) the joint covariation pattern constituted a dependent variable defined secondarily in order to allow the hand to move along the planned trajectory. Further evidence supporting this view was found in the demonstration that motor planning was a parametric process involving an independent specification of the Cartesian amplitude and Cartesian direction of the upcoming movement (Vindras and Viviani, 1998), as would be expected if these two components were planned separately.

The hypothesis that the hand always follows a straight line path in external space during visually directed reaching has recently been challenged in several studies showing that hand movements can be significantly curved and that the amount of curvature can vary with the movement direction. For instance, Osu et al. (1997) investigated visually directed movements under two conditions: no path instruction (NI), and instruction to move the hand along a straight line (SI). They found that subjects generated much straighter movements in SI than in NI. As indicated by electromyographic activity (EMG), this difference could not be related to an increase in arm stiffness, which suggested that path curvature

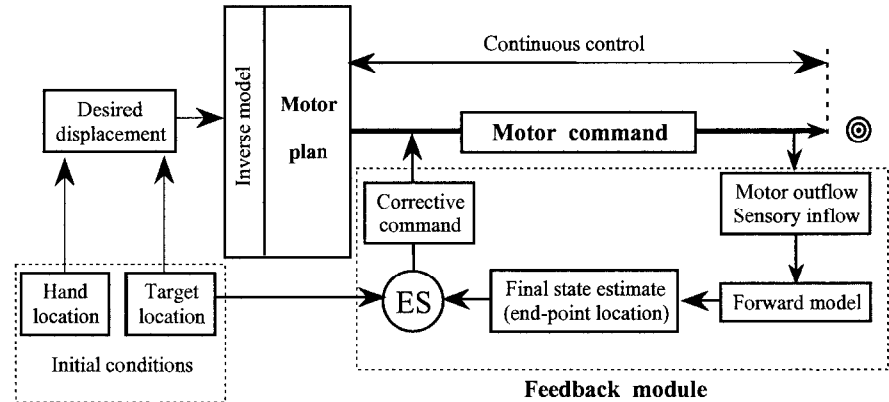
was really reflective of the movement planning process. A similar conclusion was reached by Desmurget et al. (1999), who showed, however, that the results reported by Osu et al. were valid only for unconstrained three-dimensional movements. No difference was observed between NI and SI for planar movements. This result might suggest that unconstrained movements, unlike planar movements, are not programmed to follow a straight-line path in the extrinsic space. With respect to this conclusion, several hypotheses have been proposed to explain how unconstrained movements are planned by the central nervous system (CNS). The most popular of these hypotheses suggests that hand trajectory is specified as a vector in the joint space (Flanders, Helms-Tillery, and Soechting, 1992). According to this view, the spatial characteristics of the target are initially converted into a set of arm and forearm orientations. The movement from the starting posture to the target posture is then implemented on the basis of an “angular error vector” whose components represent the difference between the starting and target angles for each joint. During the movement, joint angle variations are not controlled independently, but in a synergic way (temporal coupling). The movement curved path observed in the task space results directly from this temporal coupling.

Eye-Hand Coordination and the Need for On-Line Trajectory Control

For an external observer, the relative coordination of eye, head, and hand during goal-directed reaching appears sequential. When a subject points to a visual target in peripheral space, the eyes move first, followed by the head and ultimately the hand. Because eye movement duration is brief, the gaze generally arrives at the target before or around the time of hand movement onset. Although this sequential organization was initially thought to have a functional foundation, it was subsequently shown to result primarily from inertial factors (Desmurget and Grafton, 2000). Indeed, the EMG discharge is generally synchronized for the eye, head, and arm during fast reaching movements, indicating that the motor command is sent to these different effectors in parallel (the arm moves last simply because it has the greatest inertia). It follows that the motor command initially sent to the arm is based on an extrafoveal visual signal that has been shown to be inaccurate (see first section). At the end of the ocular saccade, which roughly corresponds to the onset of hand movement, the target location can be recomputed on the basis of foveal information. As shown by Prablanc and Martin (1992), this updated visual signal is used by the nervous system to adjust the ongoing trajectory. To demonstrate this point, the authors used a double-step pointing paradigm in which the target location was slightly modified during the course of the ocular saccade when there was saccadic suppression (i.e., the target jump was not perceived consciously by the subject). Results showed that the hand path, which was initially directed to the first target, diverged smoothly toward the second target. Interestingly, corrections were detectable about 110 ms after hand movement onset, showing that hand trajectory was amended very early. As shown by Prablanc and Martin, these corrections were similar whether or not the moving limb was visible to the subject. This suggests that nonvisual feedback loops represent the main process through which extrinsic errors are corrected.

Because of the existence of consistent delays in sensorimotor loops, the rapid path corrections observed during reaching movements cannot be attributed to sensory information only. They can only rely on a “forward model” of the arm dynamics. The idea behind this concept is that the motor system can progressively learn to estimate its own behavior in response to a given command. By integrating information related to the initial movement conditions, the motor outflow, and the sensory inflow, the forward model can determine, and even predict in advance, the probable position and

Figure 1. Forward model of arm dynamics for controlling hand movements. To reach a target, the nervous system has to elaborate a motor plan based on initial conditions (locations of the hand and target). During the execution of the motor command, a forward model of the dynamics of the arm is generated by integration of a copy of motor outflow and sensory inflow. This model then generates an estimation of the movement end-point location. Discrepancies between estimation and target location cause an error signal, which triggers a modulation of the motor command. (From Desmurget, M., and Grafton, S., 2000, Forward modeling allows feedback control for fast reaching movements, *Trends Cognit. Sci.*, 4:423–431. © Elsevier Publishing Co.; reproduced with permission.)



velocity of the effector, thus making feedback strategies possible for fast reaching movements.

This idea was recently operationalized by Desmurget and Grafton (2000), who proposed a simple model through which fast feedback control might be achieved (Figure 1). According to these authors, the forward model of the arm's dynamics generated during a movement is used to predict the movement's end point. By comparing this prediction with the actual target location, the motor system can directly estimate the movement's final accuracy. When a discrepancy is detected between the movement's predicted final location and the target location, an error signal is generated and a corrective command is issued.

The past two decades have been dominated by the hypothesis that reaching movements are primarily under preprogrammed control and that sensory feedback loops exert only a limited influence at the very end of the trajectory. As a consequence, functional investigations have focused primarily on the cerebral structures participating in motor preparation and execution, yielding few insights into the functional anatomy of on-line movement guidance. This latter issue was recently investigated by our group using positron emission tomography (PET) (Desmurget et al., 2001). Seven subjects were required to look at (Eye) or look and point to (EyeArm) visual targets whose location either remained stationary or changed undetectably during the ocular saccade. The latter condition allowed us to increase the amount of correction to be generated during the movement. The functional anatomy of nonvisual feedback loops was identified by comparing the reaching condition involving large corrections (Jump) with the reaching condition involving small corrections (Stationary), after subtracting the activations associated with saccadic movements and hand movement planning [(EyeArm-Jumping minus Eye-Jumping) minus (EyeArm-Stationary minus Eye-Stationary)]. In agreement with earlier observations (Prablanc and Martin, 1992), behavioral recordings indicated that the subjects were both accurate at reaching toward the stationary targets and able to update their movement smoothly and early in response to the target jump. PET difference images showed that these corrections were mediated by a restricted network involving the posterior parietal cortex (PPC), the cerebellum, and the primary motor cortex (M1). As shown in Figure 2, the parietal activation was located in the left intraparietal sulcus, in a region that is generally considered the rostral part of the PPC. The cerebellar activation occurred in the right anterior parasagittal cortex, in a region associated with the production of arm movements. The frontal activation was located in the arm-related area of M1. The contribution of the PPC to movement guidance has recently been confirmed by several studies showing that on-line movement corrections are suppressed when this structure is lesioned or prevented from exerting its function through the application of a transcranial magnetic pulse at the onset of hand movement (Desmurget and Grafton, 2000).

Although the role of the PPC, the cerebellum, and the motor cortex in movement guidance has not been totally elucidated, a general model can be proposed. At a first level, one may suggest that the role of the PPC is to compute a motor error by comparing the target location and the estimated movement end point. This hypothesis is based on two main observations about the PPC, namely, (1) that the PPC has access to a representation of both the target and current hand location through afferent information coming from different sensory modalities and the main motor struc-

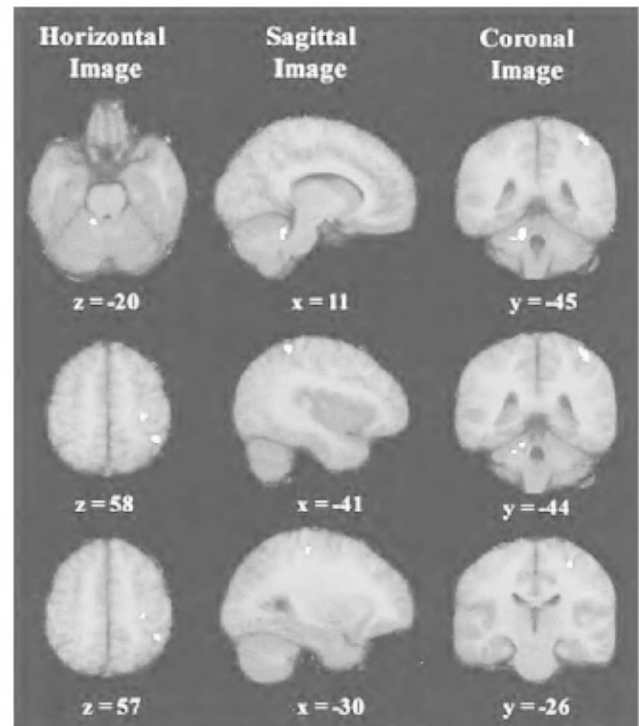


Figure 2. Areas of brain activation superimposed on a mean MRI in Talairach coordinates. On the sagittal images, positive values of x designate the hemisphere ipsilateral to the reach arm (right one), and negative values of x designate the contralateral hemisphere (left one). Top row is centered on the cerebellar activation site; middle row is centered on the PPC activation site; bottom row is centered on the precentral activation site. (From Desmurget, M., et al., 2001, Functional anatomy of nonvisual feedback loops during reaching: A position emission tomography study, *J. Neurosci.*, 21:2919–2928. © Elsevier Publishing Co.; reproduced with permission.)

tures, and (2) that the PPC is critical for merging the arm- and target-related signals into a common frame of reference (Andersen et al., 1997). Once computed, the motor error needs to be converted into an effective motor command (inverse computation). Converging evidence indicates that this function may be carried out by the cerebellum. In particular, it has been shown that patients with cerebellar lesions are impaired in defining the pattern of muscle activation required to direct the hand along a specific path, and that inverse models are represented within the cerebellum (Wolpert, Miall, and Kawato, 1998). At the extremity of the system, the cerebellar corrective signal might influence the ongoing motor command by modulating the neural signal issued by the primary motor cortex. In agreement with this view, it is known that the primary motor cortex receives substantial input from the cerebellum via the ventrolateral thalamus. Also, it is thought that the motor system is organized in a relative hierarchy such that the primary motor cortex is mainly involved in low-level aspects of motor control. Of course, the previous observations do not imply that other areas are not involved in movement guidance. A role for the basal ganglia, for instance, was proposed in a recent study on patients with Huntington's disease (Smith, Brandt, and Shadmehr, 2000). This role was not, however, confirmed by our group in subsequent work showing that Parkinson's disease patients were able to smoothly and quickly correct ongoing movement in an unconscious double-step study identical to the one used by Prablanc and Martin (1992).

Conclusion

Our knowledge of the neural processes underlying eye-hand coordination has greatly improved over the past two decades. However, our understanding is still far from exhaustive, and many issues remain to be addressed. Current approaches will benefit from the development of independent motor theories within specific experimental and theoretical contexts. Our understanding of the processes whereby a visual input is transformed into a motor command would be greatly improved by comparative studies involving heterogeneous approaches as well as various areas of brain theory and neural networks. The complementary contributions gathered in the *Handbook* are a first step in this promising direction.

Road Map: Mammalian Motor Control

Background: Motor Control, Biological and Theoretical

Related Reading: Collicular Visuomotor Transformations for Gaze Control; Grasping Movements: Visuomotor Transformations; Reaching Movements: Implications for Computational Models

References

- Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J., 1997, Multimodal representation of space in the posterior parietal cortex and its use in planning movements, *Annu. Rev. Neurosci.*, 20:303–330. ♦
- Batista, A. P., Buneo, C. A., Snyder, L. H., and Andersen, R. A., 1999, Reach plans in eye-centered coordinates, *Science*, 285:257–260.
- Bock, O., 1986, Contribution of retinal versus extraretinal signals towards visual localization in goal-directed movements, *Exp. Brain Res.*, 64:476–482.
- Brotchie, P. R., Andersen, R. A., Snyder, L. H., and Goodman, S. J., 1995, Head position signals used by parietal neurons to encode locations of visual stimuli, *Nature*, 375:232–235.
- Desmurget, M., and Grafton, S., 2000, Forward modeling allows feedback control for fast reaching movements, *Trends Cognit. Sci.*, 4:423–431. ♦
- Desmurget, M., Grea, H., Grethe, J. S., Prablanc, C., Alexander, G. E., and Grafton, S. T., 2001, Functional anatomy of nonvisual feedback loops during reaching: A positron emission tomography study, *J. Neurosci.*, 21:2919–2928.
- Desmurget, M., Prablanc, C., Jordan, M. I., and Jeannerod, M., 1999, Are reaching movements planned to be straight and invariant in the extrinsic space? *Q. J. Exp. Psychol.*, 52A:981–1020.
- Duhamel, J. R., Bremmer, F., BenHamed, S., and Graf, W., 1997, Spatial invariance of visual receptive fields in parietal cortex neurons, *Nature*, 389:845–848.
- Flanders, M., Helms-Tillery, S. L., and Soechting, J. F., 1992, Early stages in sensori-motor transformations, *Behav. Brain Sci.*, 15:309–362. ♦
- Morasso, P., 1981, Spatial control of arm movements. *Exp. Brain Res.*, 42:223–227.
- Osu, R., Uno, Y., Koike, Y., and Kawato, M., 1997, Possible explanations for trajectory curvature in multijoint arm movements, *J. Exp. Psychol. Hum. Percept. Perform.*, 23:890–913.
- Prablanc, C., and Martin, O., 1992, Automatic control during hand reaching at undetected two-dimensional target displacements, *J. Neurophysiol.*, 67:455–469. ♦
- Smith, M. A., Brandt, J., and Shadmehr, R., 2000, Motor disorder in Huntington's disease begins as a dysfunction in error feedback control, *Nature*, 403:544–549.
- Soechting, J. F., Engel, K. C., and Flanders, M., 2001, The Duncker illusion and eye-hand coordination, *J. Neurophysiol.*, 85:843–854.
- Vindras, P., and Viviani, P., 1998, Frames of reference and control parameters in visuo-manual pointing, *J. Exp. Psychol. Hum. Percept. Perform.*, 24:569–591.
- Wolpert, D. M., Miall, R. C., and Kawato, M., 1998, Internal models in the cerebellum, *Trends Cognit. Sci.*, 2:338–347. ♦

Face Recognition: Neurophysiology and Neural Technology

Rolf P. Würtl

Introduction

The ability to recognize other individuals is a major prerequisite for human social interaction and hence a rather important brain function. The most prominent cue for that recognition is the face. The ability to recognize persons from their faces is part of a spectrum of related skills that includes face segmentation (i.e., finding faces in a scene or image), estimation of the pose, estimating the direction of gaze, and evaluating the person's emotional state. This

article focuses on recognition of identity. A more detailed treatment of the other aspects can be found in FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM.

Neurophysiology

From neuropsychological studies of patients with brain injuries, it is known that there are subsystems in the brain that are specialized for face processing. Brain injury can lead to loss of the ability to

recognize faces, a deficit called *prosopagnosia*, while leaving recognition of general objects intact. The opposite dissociation is reported in Moscovitch, Winocur, and Behrmann (1997), in a patient with intact face recognition together with highly impaired general object recognition. Various stunning perceptual demonstrations show that faces are perceived differently when viewed upside down or as photographic negatives. Those image manipulations make little difference for the perception of general objects but can modify the perception of identity and expression considerably. These findings lead to the assumption that different brain circuits are used for processing general objects and for processing faces, but there is also considerable evidence that not only faces receive special treatment, but all object classes for which there is high expertise (Gauthier, Behrmann, and Tarr, 1999).

Other studies show that patients with prosopagnosia who exhibit no conscious recognition of facial identity still exhibit an unconscious reaction to familiar faces, which is revealed by changes in skin conductance. This mechanism seems to play a major role in the emotional reaction to facial stimuli.

Single-unit recordings of activity in the inferotemporal cortex of macaque monkeys have revealed neurons with a high responsiveness to the presence of a face, an individual, or the expression on a face (see Desimone, 1991, for a review). Although the notion of the optimal stimulus for a cell is very hard to probe experimentally, some of these cells are as close to grandmother cells (see ASSOCIATIVE NETWORKS) as the experimental evidence gets.

In humans, cells that become active when a familiar face is seen have been identified in the inferotemporal gyrus and the fusiform gyrus in both hemispheres. Their clusters do not form anatomically well-defined subregions but are neighbored by modules of different specificity, and their location and extent vary considerably among individuals.

A good account of the current knowledge about face recognition in the human brain is given by Haxby, Hoffman, and Gobbini (2000), whose model refines a cognitive model by Bruce and Young (see Young, 1998, chap. 3) and attaches anatomical locations to its modules. Haxby et al. propose a *core system* for face processing that consists of three interconnected modules. The first, located in the inferotemporal occipital gyrus, is responsible for the early extraction of features relevant for faces. The second, in the superior temporal sulcus, codes for the changeable properties of faces, such as the direction of gaze, lip movement, expression, and the like. Identity as an invariant face property is processed in the lateral fusiform gyrus. This core system communicates with other parts with a need for facial information, such as attention modules, auditory cortex, and emotional centers. The essence of face recognition—to link the visual information to a name and biographical knowledge about particular persons—is carried out in the anterior temporal lobe. These other parts make up the *extended system*.

Computational Theory

As for all cases of object recognition, the main problem to be solved by a face recognition procedure is *invariance*. The same face can produce very different images with changes in position, pose, illumination, expression, partial occlusion, background, and so forth. The task of the recognition system is to generalize over all these variations and capture only the identity.

This sort of invariant recognition is a quotidian property of natural brains but does not come very naturally in current artificial neural network models. Even the simplest case, invariance under translations in the input plane, is difficult to obtain. One major approach starts with the observation that complex cells generalize about small translations of the signal. This can be iterated, and leads to hierarchical networks such as the NEOCOGNITRON: A MODEL FOR VISUAL PATTERN RECOGNITION (q.v.). A huge advantage of such purely feedforward networks is their speed of processing.

Very little is known about how invariant recognition can be learned from examples and generalized to other instances. In an abstract sense, the important long-term goal is to teach a network precisely the invariances required for a given problem domain. This is directly relevant for face recognition, because invariance under expression and slight deformations are very difficult to capture analytically.

If the only invariance required is translation, then template matching (see OBJECT RECOGNITION) can solve the problem rather efficiently. A stored pattern (which we will call “model”) is compared to an image by shifting the model across the image and taking the scalar product with appropriate normalization at all possible image locations. The maximum of the resulting matrix can serve as a similarity measure between both images.

In order to extend this method to the more complicated invariances involved in face recognition, the notion of a *correspondence map* is helpful (Figure 1). Correspondence, central to many problems in computer vision, can be defined as follows: *Point pairs from two given images of the same face correspond if they originate from the same point on the physical face.*

Once these correspondences have been established for sufficiently many points, an invariant similarity measure between model and object can be defined as the sum or average over the similarities of local features of all corresponding point pairs. Because the points on the real face are not accessible to either the brain or a computer, these correspondences can only be estimated on the basis of image information. Strictly speaking, correspondences are defined only between images of the same person, but all faces are sufficiently similar in structure that the notion can be extended to correspondence maps between different faces. These maps have many applications beside recognition (see FACE RECOGNITION: PSYCHOLOGY AND CONNECTIONISM).



Figure 1. Correspondence maps provide a basis for a similarity measure between two facial images, which is used for person identification. They also deliver information about pose, size, and expression, and are crucial for animation. Their computation is difficult and rarely perfect. The figure shows selected correspondences obtained with the algorithm from Würtz (1997).

A system to recognize a person out of a collection of known ones can proceed as follows. Correspondence maps are estimated between the given image and all stored models, similarities are calculated on the basis of the correspondence maps, and the model with the highest similarity is picked as the recognized person. A measure for the *reliability* of the recognition can be derived by a simple statistical analysis of the series of all similarity values.

Because correspondence finding is a slow process, the database of known individuals must be organized in such a way that the need for correspondence finding is minimized. Furthermore, it should not be applied to arbitrary images, but some filtering must select image portions that are likely to contain a face for processing and recognition.

Summarizing the computational theory reveals the following building blocks for a successful face recognition system:

1. A representation of the facial images
2. A method of solving the correspondence problem
3. A similarity measure derived from a pair of images and a correspondence map
4. Organization of the database of known individuals
5. Filtering of the visual data (face finding)

For general reviews of face recognition systems, see Grudin (2000) and Chellappa, Wilson, and Sirohey (1995).

Image Representation

Many models for face recognition work directly on image gray values or retinal images. In this case, the correspondence problem becomes particularly difficult, as many points from very different locations share the same pixel value without actually corresponding to each other. A possible remedy consists in combining local patches of pixels. The larger the patch, the more this ambiguity is reduced. On the other hand, the features become more sensitive to distortions and changes in background and thus are of less value for the other required invariances. Patch building may also include linear combinations of pixel values. In this context, *Gabor functions* (see GABOR WAVELETS AND STATISTICAL PATTERN RECOGNITION) as a model of simple and complex cells in V1 have turned out to be a good compromise between locality and robustness and are well-suited for correspondence finding.

The possibility of processing the amplitudes and phases of the Gabor wavelet responses separately is very useful for face processing. Amplitudes (which model the activity of complex cells) vary rather smoothly across the image, and so do the similarities of all image features to a single one. Consequently, they provide smooth similarity landscapes well-suited for matching templates or single feature vectors. The phases, on the other hand, vary as rapidly as dictated by their center frequency and proceed roughly linearly on image paths in the respective direction. Therefore, they can be used to estimate correspondences with subgrid accuracy (Wiskott et al., 1997; Würtz, 1997).

An important alternative for image representation is to use local features that are derived directly from the statistics of facial images. A prominent example is the neural network-based *local feature analysis* (Penev and Atick, 1996), which allows learning local descriptors by minimizing their correlation. This results in a sparse code adapted for the class represented by the training examples.

Correspondence Finding

The representation of a face in terms of local features serves two purposes. First, correspondences must be estimated on the basis of feature similarity, and second, the feature similarities constitute the image similarity. In principle, different features can be used for both purposes.

Because of the ambiguities discussed above, simplifying assumptions must be made about the correspondence maps. A good candidate for such an assumption is *neighborhood preservation*. Consequently, algorithms for correspondence finding usually optimize a combined objective function that favors similarity between local features and smoothness of the correspondence map.

One implementation of this procedure is *elastic graph matching* (EGM) (Lades et al., 1993), in which stored models are represented as graphs vertex-labeled with vectors of local Gabor responses and edge-labeled with a distance constraint. The correspondence problem can be solved by optimizing the similarity between model graph and a (topologically identical) graph in the image in terms of similarity of both edge and vertex labels. This is a high-dimensional optimization problem that is usually simplified by applying a hierarchy of possible graph transformations. It starts with pure translation, later adds scale changes, and finally adds local displacements. In the first steps, Gabor amplitudes are used exclusively, which leads to smooth similarity landscapes and allows separating the different steps.

An alternative method, one that makes use of the pyramidal form of Gabor wavelet transform, is *Gabor pyramid matching* (Würtz, 1997). It starts with standard template matching of the Gabor amplitudes on a sparse grid and low spatial frequency and refines the results using higher spatial frequencies. Thus, neighborhood preservation is not explicitly coded into an objective function but is inherited from the undistorted matching on low frequencies. Very precise correspondences can be obtained by subsequent subgrid estimation using the Gabor phases. This method allows much better background suppression, because the need to know local features for each feature point on all scales is eliminated.

Memory Organization

The importance of memory organization is due to the computational expense of the inevitable correspondence estimation, which should not be carried out separately on all stored models. Consequently, it is necessary to evaluate correspondences *between* the stored models. Adding this idea to EGM results in the so-called *bunch graph* (Wiskott et al., 1997). In that data structure, each vertex is labeled with one local feature vector from each person in the database, and care has to be taken during creation of the bunch graph that these feature vector are indeed taken from corresponding points. In addition to different matching schemes, bunch graphs can be used in two major modes. In one mode, it is assumed that the person to be recognized is indeed in the bunch graph, and is selected according to similarity. Alternatively, the feature most similar to the given image can be selected for each vertex separately, leading to a composition of the face image in terms of the local features of all persons in the bunch graph. Moreover, the vertices can carry additional information, such as sex, beardedness, or a genetic disease of the person they belong to. By majority voting, a decision about that feature for completely unknown persons can be made.

Eigenfaces (Turk and Pentland, 1991) are another technically successful approach to face recognition. Gray-value images of faces are prealigned by an optical flow method and then subjected to PRINCIPAL COMPONENT ANALYSIS (q.v.), which can be interpreted as a neuronal method. It turns out that a few components are sufficient to recognize identity. Recognition proceeds by projecting the image to be classified onto these components and applying a classifier to the resulting low-dimensional vector. Calculating the PC representation from a database of persons is rather time consuming, but projection and classification are very fast. This shows that the major strength of the eigenface method lies in very efficient memory organization.

Neuronal Models

On the technical side, a large variety of neural network models have been applied to the problem of face recognition. They usually start from well-aligned faces with little variation. See Gong et al. (2000) for a good discussion of the application of neural classifiers and an excellent treatment of technical approaches to face recognition.

It is currently not known if there is neuronal machinery in the brain to explicitly estimate correspondences. However, DYNAMIC LINK ARCHITECTURE (q.v.) can be used to solve the correspondence problem, as follows (Lades et al., 1993). Two layers of neurons that represent the image space in model and image, respectively, are fully interconnected by dynamic links. They have an internal wiring that supports moving localized blobs of activity. The development of links is supported by feature similarity and synchronous activation of the connected neurons. The link dynamics then converge to a correspondence mapping. It has been extended by a competition between a multitude of model layers to a full-blown neural face recognition system. This system is sped up by a coarse-to-fine strategy working on the Gabor pyramid. The speedup is due to the possible parallelism between all refinement steps. That system also shows good background invariance, because model and image representation are the same as for pyramid matching.

Face Finding

Having found a correct correspondence map from a stored model into an image in principle implies that segmentation has also been solved. However, applying correspondence-based techniques like bunch graph matching to arbitrary images yields plenty of misclassifications: depending on the parameters, either many faces go undetected or many nonfaces pass as faces.

It seems very difficult to encode the notion of a general face into a program or data structure. Therefore, for *finding* faces in images or video sequences, neural net classifiers are widely used. Typically, a whole set of segmented and roughly normalized face images is used to train a network. Then the network is applied to all points of an image to provide a face/nonface decision. A good review of the facefinding literature can be found in Hjelmås and Low (2000).

Discussion

A correspondence map is a very general model of the variation of appearance of a face. Its estimation is computationally intensive, and the currently known neuronal models that can implement it cannot account for the rapidity of human recognition, even if run on highly parallel hardware like real neurons. The advantage of correspondence maps lies in the fact that much more information, such as the actual position, pose, and expression, can be determined from them.

The quality of technical face recognition systems is difficult to judge. On small data sets (less than 100 individuals), even naive template matching may yield respectable recognition rates. The use of standard databases, which is inevitable for achieving fair com-

parisons, raises the danger of overadapting the classifiers to the data. To prevent this, the Army Research Laboratory has set up a standard comparison procedure called the FERET test (Philips et al., 2000), in which where the major part of 14,126 images of 1,199 individuals is withheld for independent testing.

With this database, eight competitors underwent a test with the additional information about the (hand-labeled) eye position. Only two competitors, a bunch graph-based system and an eigenface-based system, took the realistic test on the images without any extra information. Both systems performed equally well on the data set with given eye coordinates, and the bunch graph-based system clearly won on the more difficult examples without additional information (Philips et al., 2000). These results underscore the need for very good correspondence estimation for successful face recognition.

Road Map: Vision

Background: Gabor Wavelets and Statistical Pattern Recognition

Related Reading: Dynamic Link Architecture; Face Recognition: Psychology and Connectionism; Object Recognition

References

- Chellappa, R., Wilson, C. L., and Sirohey, S., 1995, Human and machine recognition of faces: A survey, *Proc. IEEE*, 83:705–740. ♦
- Desimone, R., 1991, Face-selective cells in the temporal cortex of monkeys, *J. Cognit. Neurosci.*, 3:1–8.
- Gauthier, I., Behrmann, M., and Tarr, M., 1999, Can face recognition really be dissociated from object recognition? *J. Cognit. Neurosci.*, 11:349–370.
- Gong, S., McKenna, S. J., and Psarrou, A., 2000, *Dynamic Vision*, London: Imperial College Press. ♦
- Grudin, M. A., 2000, On internal representations in face recognition systems, *Pattern Recogn.*, 33:1161–1177.
- Haxby, J. V., Hoffman, E. A., and Gobbini, M. I., 2000, Distributed human neural system for face perception, *Trends Cognit. Sci.*, 4:223–233. ♦
- Hjelmås, E., and Low, B. K., 2000, Face detection: A survey, *Comput. Vision Image Understanding*, 83:236–274.
- Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W., 1993, Distortion invariant object recognition in the dynamic link architecture, *IEEE Trans. Comput.*, 42:300–311. ♦
- Moscovitch, M., Winocur, G., and Behrmann, M., 1997, What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition, *J. Cognit. Neurosci.*, 9:555–604.
- Penev, P. S., and Atick, J. J., 1996, Local feature analysis: A general statistical theory for object representation, *Network*, 7:477–500.
- Philips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J., 2000, The FERET evaluation methodology for face-recognition algorithms, *IEEE Trans. Pattern Anal. Machine Intell.*, 22:1090–1104.
- Turk, M., and Pentland, A., 1991, Eigenfaces for recognition, *J. Cognit. Neurosci.*, 3:71–86.
- Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C., 1997, Face recognition by elastic bunch graph matching, *IEEE Trans. Pattern Anal. Machine Intell.*, 19:775–779.
- Würtz, R. P., 1997, Object recognition robust under translations, deformations and changes in background, *IEEE Trans. Pattern Anal. Machine Intell.*, 19:769–775. ♦
- Young, A., 1998, *Face and Mind*, Oxford, Engl.: Oxford University Press.

Face Recognition: Psychology and Connectionism

Alice J. O'Toole

Introduction

Faces provide humans with rich information about their owners. From a brief glance across a dimly lit room, we can recognize the face of a friend. Faces also provide us with information important for social interaction, including the sex, race, approximate age, and current mood of a person. Humans can extract this information nearly instantaneously from the complex three-dimensional (3D) surface of the head. The skills humans have in perceiving and recognizing faces are even more impressive when we consider the number and diversity of people we must remember as individuals, and the fact that all faces share the same basic set of features (eyes, nose, and mouth), arranged in roughly the same configuration. Quantifying or even specifying the information that makes a face unique is a challenging problem.

Automatic or computer-based approaches to modeling face recognition have a long and relatively successful history among connectionist models of visual recognition. Commercial versions of facial recognition systems are now widely available and have already been employed in a variety of security applications. Although still not at the level of human performance on all tasks, with minimal viewpoint and illumination variation, the performance of these models compares favorably with the performance of humans (see FACE RECOGNITION: NEUROPHYSIOLOGY AND NEURAL TECHNOLOGY).

This article begins with a brief history of connectionist approaches to face recognition. Next, the broad range of tasks to which these models have been applied will be presented. Relating the models to psychological theories is accomplished by breaking the problem into the subtasks of representing faces and retrieving them from memory. Human and model performance can be compared along these dimensions. Finally, the article concludes with an overview of the challenges facing computational models of face processing at the beginning of the twenty-first century.

History and Theoretical Background

Connectionist face recognition models are some of the first successful computational models of a visual recognition task. Perhaps because of the homogeneity of faces as a class of objects, and the utility of recognition within the class of faces, the problem of modeling face recognition has been more tractable than modeling object recognition (see OBJECT RECOGNITION). Early face recognition models date from the beginnings of associative models of memory. In the 1970s, Kohonen (1977) used faces to illustrate the potential of a linear autoassociative network to act as a parallel distributed memory for images. He used simple pixel-based encodings of faces and showed that the model could selectively retrieve memorized faces using partial or occluded memory keys. The work of Kohonen using autoassociative memories with faces provided the cornerstone for most current computational models of face recognition. The autoassociative memory brings together important aspects of the psychology of face recognition and the varied, yet related, computational approaches to the problem.

The key to this connection is that autoassociative memories can be shown to implement PRINCIPAL COMPONENT ANALYSIS (PCA; q.v.), the most common of current connectionist or statistical models of face recognition (cf. Phillips et al., 2000). PCA is a technique for representing a number of correlated variables using a lesser number of uncorrelated or orthogonal variables. Applied to a set of face representations (e.g., images, 3D surfaces), PCA produces

a set of orthogonal "feature" axes, known variously as principal components (PCs), eigenvectors, eigenfeatures, or eigenfaces (Turk and Pentland, 1991). These axes are ordered according to the proportion of variance they explain in the set of faces analyzed. A critical property of PCA is that any face in the set of analyzed or "learned" faces can be represented or reconstructed as a weighted combination of PCs. This property is shared by more recent techniques, such as nonnegative matrix factorization, that produce a sparser representation of faces than PCA (Lee and Seung, 1999; see also SPARSE CODING IN THE PRIMATE CORTEX).

From a psychological point of view, it is reasonable to think of PCs as features, and of the weights required for reconstructing a face as its *feature values*. PCA and connectionist-style models have properties reminiscent of human memory for faces. First, the memory is distributed rather than localized. This means that individual face representations interfere with each other in a way that allows for natural confusions between similar faces. Second, at the level of groups of faces, the proportion of variance associated with individual PCs is indicative of the importance of individual features for describing the set of faces. This enables the statistical properties of a particular face-learning history to affect the performance of the model. For example, connectionist models that vary in the racial composition of the training set can simulate the "other-race effect" for human memory (O'Toole et al., 1994). This is the well-known finding that we recognize faces of our own race more accurately than faces of other races.

An important theoretical focus of the last decade has been the integration of connectionist models of face recognition into the *face space theory* of human face processing (Valentine, 1991). Valentine's face space theory posits that human memory for faces can be thought of metaphorically as a multidimensional face space. At the center of the space is the average or "prototype" face. Individual faces are represented as points in the space, with the axes of the space defining the features with which faces are encoded. The distance between any two faces is, therefore, a measure of their similarity.

The face space model accounts for some common psychological findings for face recognition. For example, it is well known that the faces rated by subjects as typical are recognized less accurately than faces rated as distinctive. The face space theory explains this difference based on the probable distribution of faces in the space. Typical faces are close to the average face, where the space is "crowded." Distinctive faces are located in the sparser parts of the space, away from the average. Typical faces, therefore, are more easily confused with other faces than are distinctive faces.

It has become increasingly clear that the majority of connectionist models of face recognition, including those based on PCA, implement a physical version of Valentine's (1991) face space. These physical face spaces provide powerful tools for testing psychological theories about the way humans represent and retrieve faces from memory.

Face-Processing Tasks

Before proceeding, it is worth defining the range of face-processing tasks to which connectionist models have been applied. *Face recognition* involves a judgment about whether or not a face has been learned previously (i.e., is in the model's memory). This is distinct from *face identification*, which requires the retrieval of semantic information about a person whose face has been recognized (i.e., a name, or context or previous encounter). For *face verification*, a

face is presented along with identifying information such as a name badge, and the model or machine must affirm or reject the face identification. This task has become increasingly important in recent years with the development of automatic security systems.

In addition to recognizing and identifying the faces of people we know, we must also be able to visually categorize faces along a number of general dimensions, including sex, race, age, and facial expression. Connectionist models have been applied to nearly all of these tasks, with varying levels of success (see O'Toole, Wenger, and Townsend, 2001, for a detailed review of face classification models). Of particular note, recent applications of connectionist models to the task of processing facial expression have become important in the context of human-computer interaction applications. They are also valuable tools for addressing the theoretical issues involved in human perception of emotion. Recent work by two groups have made inroads into the problem using PCA (Calder et al., 2001) and backpropagation models (Dailey, Cottrell, and Adolphs, 2001). Although there is still active debate on the psychological mechanisms involved in the processing of expressions, both models have achieved good performance and offer insight into the human solution to the problem.

Representing Faces

Much of the scientific dialogue associated with connectionist models of face processing over the past 5 to 10 years has focused not on the algorithms, but on the representations to which they are applied. Although the basic models have remained relatively simple, the representations to which they have been applied have evolved quickly and have provided fascinating insights into the complexities of representing the information in faces. As noted, face space models are powerful tools for testing the validity of different underlying representations of faces. The general idea is that the models operate by deriving a face space that is sensitive to the statistical structure of the faces analyzed. However, the way in which these inputs are encoded (e.g., by images, 3D measures, geometrical measures) has a substantial impact on the layout of the face space, and thereby alters the distances between individual faces. This is easy to understand intuitively, as follows. Imagine two faces with highly similar shapes but very different coloring (pigmentation). Representing the faces using images will result in a rather different similarity estimate than representing the faces with 3D surface measures.

The face representations used most commonly in connectionist models of face processing can be divided into three types: (1) raw image codes, (2) partially aligned "pre-morph" codes, and (3) fully corresponded "pre-morph" codes (cf. O'Toole et al., 2001, for additional details).

The first connectionist models represented faces as raw images (Sirovich and Kirby, 1987). Although the faces were spatially scaled and aligned in the image so that the eyes or nose of all faces coincided, this code has the obvious disadvantage of being only minimally tolerant of changes in viewpoint and illumination. Notwithstanding, raw image codes, within these limits, have proved highly effective and accurate at the task of face recognition and have largely remained the standard in state-of-the-art computational models of face recognition (Phillips et al., 2000). Connectionist models based on these codes also provide a reasonable approximation to human judgments about the distinctiveness of a face and human accuracy in recognizing faces (see O'Toole et al., 1994).

More effective ways of aligning faces have been developed based on partially aligned "pre-morph" codes. These codes define a partial correspondence between the features of all faces and help bridge the gap between computational models of face recognition and computer graphics-based face synthesis. Pre-morph codes define a representation that can support a morphable transition be-

tween any two faces. Craw and Cameron (1991) first introduced the idea of a pre-morph code for face representation by dividing a face representation into *shape* and *shape-free* encoding. Though not yet exploited to its full potential, this method is valuable to psychologists because it allows independent manipulation of the shape and shape-free components of the face in psychological experiments. Face shape is defined by the spatial positions of a set of facial landmarks (corners of the eyes, mouth, etc.) in an image. Next, a complementary shape-free code of a face is created by morphing (or warping) the face into the average face shape, computed over a large number of faces. The encoding of an individual face, therefore, includes both its shape and shape-free parts. Hancock, Burton, and Bruce (1996) used this separated face code with PCA for modeling human face recognition and found a better fit to human performance than the purely image-based codes used previously.

Dividing the representation of faces in this way enables a separate analysis of the structure of a face and its image-based information, but is costly in terms of the preprocessing required. Successful use of these codes involves locating and marking (often by hand) a relatively large number of facial landmark points that match on all faces. A related representation, inspired by early visual processing, overcomes this problem in an interesting way. This approach characterizes the work of von der Malsburg and his colleagues over the years (see, e.g., Okada et al., 1998, and FACE RECOGNITION: NEUROPHYSIOLOGY AND NEURAL TECHNOLOGY). Okada et al. used banks of Gabor jets—filters with varying orientations and resolutions—to sample face images at multiple locations. The Gabor jets simulate the spatial frequency filtering that occurs in early visual processing in the cortex. Using a dynamic link architecture, face recognition occurs by allowing the relative locations of the Gabor jets to migrate to fit individual faces. In some ways, this implements the partial alignment discussed previously in a biologically plausible way. This model has been shown to support highly accurate face verification and is the basis of commercial identification security systems (see Phillips et al., 2000).

To simulate face synthesis and to implement a face space that is completely continuous (i.e., locations in the space are "possible" faces), however, true correspondences between meaningful facial landmarks are necessary. In recent work using elaborated optic flow algorithms, Blanz and Vetter (1999) have been able to completely automate the feature-matching process to produce "fully corresponded" face representations, using laser scans of human heads. *Fully corresponded* means that the matching process is carried out not only on the landmark points, but on *all* of the sample points. Specifically, Blanz and Vetter aligned the x , y , z and r , g , b sample points for a large number of individuals. From this corresponded representation of the laser scans, they defined a 3D morphable model of faces by analyzing the data with PCA. This model readily fits into the multidimensional face space theory outlined previously, but employs a sophisticated and nearly complete representation of the perceptual information in faces. Psychologists can use this model to precisely manipulate the information in faces. In particular, 3D shape versus 2D reflectance information in faces can be varied selectively. These manipulations have been used to test the contribution of 2D versus 3D information for human face recognition (see O'Toole et al., 2001, for a review of findings).

The morphable model was used to predict high-level face adaptation effects in human perception. Leopold et al. (2001) showed that trajectories can be defined through this space by connecting a face with the average face, and continuing the trajectory to the "other side of the mean" (Figure 1). This defines an "anti-face" or "opposite" to the original face, because it inverts all of the coordinates (feature values) on the PCs. For example, dark faces have light-colored anti-faces, and round, chubby faces have long, skinny anti-faces. Leopold et al. showed that the perception of a face can be facilitated by simply pre-viewing its anti-face for a few seconds.

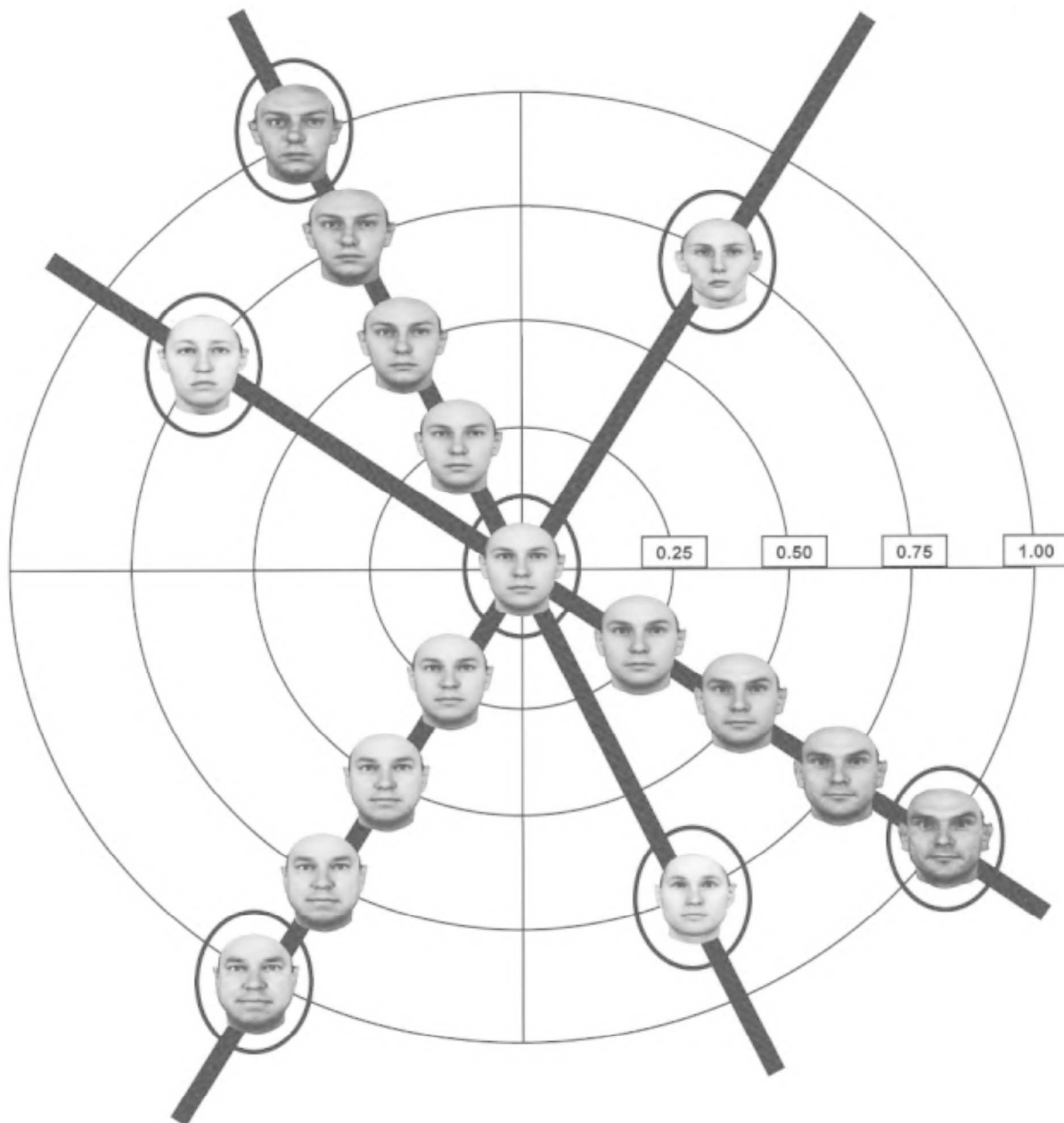


Figure 1. A face space created by the morphable model of Blanz and Vetter (1999) shows the continuous trajectory between a face and the mean or prototype of the face space. Anti-caricatures lie between the face and the prototype. On the other side of the mean is the anti-face, which is an op-

posite (i.e., all feature values are reversed) of the original face. (From Leopold, D. A., O'Toole, A. J., Vetter, T., and Blanz, V., 2001, Prototype-referenced shape encoding revealed by high-level aftereffects, *Nature Neurosci.*, 4:89–94. Reprinted with permission.)

This effect has been compared to simpler visual aftereffects, such as the perception of green after viewing red. Trajectories through this complicated face space, therefore, can be used to predict human performance in a rather precise way. Leopold et al. relate these findings to the activity of face-selective cells in inferotemporal cortex, which may encode high-level shape properties of objects.

In summary, face representations have evolved in sophistication and have provided connectionist and statistically based models with better and more precise information about faces. As these representations have improved, so too has the power of the models for simulating aspects of human performance with faces.

Retrieving and Categorizing Faces

In the context of a computational model, face recognition involves a decision about whether or not a face has been learned previously. For the face space framework, this works as follows. The learned faces are points in a multidimensional space. Recognition is implemented by projecting a test face into this space to locate a match. Projection involves the representation of the test face using the features or eigenvectors derived from learned faces. If a good match is found—i.e., if the location of the test face in the space is close to a stored face—then the face should be “recognized.” If there is

no neighboring face sufficiently close to the test face, the face should be declared “novel.” This way of testing face recognition algorithms produces “hit” and “false alarm” data very similar to those acquired in a standard human face recognition experiment.

It is clear that anything that alters the appearance of the face between learning and test time and that is included in the representation (e.g., illumination for a pixel-based code) will affect the accuracy of recognition. This interaction between the model representation and the retrieval process can be used to make predictions about the factors that will determine human accuracy for faces. These predictions can be used to test hypotheses about the kinds of representations employed in the visual system. At present, a variety of algorithms and representations are available. The work of assessing the accord between human and model performance in the context of psychological theories of these processes is under way.

Discussion

Connectionist models of face recognition have a long and distinguished history. Their development has been spurred by the many and varied psychological issues inherent to the perception of human faces, and by a quickly emerging set of industrial applications. These include applications for security systems and biometric analysis of person identification. Connectionist algorithms are available for face recognition, verification, and classification. In the future, we can expect to see systems that track moving people and faces being merged into the better-developed face-processing algorithms discussed here.

Road Maps: Psychology; Vision

Related Reading: Principal Component Analysis; Face Recognition: Neurophysiology and Neural Technology; Object Recognition, Neurophysiology

References

- Blanz, V., and Vetter, T., 1999, A morphable model for the synthesis of 3d faces, in *SIGGRAPH'99 Proceedings*, ACM, Computer Society Press, pp. 187–194.
- Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S., 2001, A principal component analysis of facial expressions, *Vision Res.*, 41:1179–1208.
- Craw, I., and Cameron, P., 1991, Parameterising images for recognition and reconstruction, in *Proceedings of the 22nd Annual Cognitive Science Society Conference*, Mahwah, NJ: Erlbaum.
- Hancock, P. J. B., Burton, A. M., and Bruce, V., 1996, Face processing: Human perception and principal components analysis, *Memory Cognit.*, 24:26–40.
- Kohonen, T., 1977, *Associative Memory*, New York: Springer-Verlag.
- Leopold, D. A., O'Toole, A. J., Vetter, T., and Blanz, V., 2001, Prototype-referenced shape encoding revealed by high-level aftereffects, *Nature Neurosci.*, 4:89–94.
- Lee, D. D., and Seung, H. S., 1999, Learning the parts of objects by non-negative matrix factorizations, *Nature*, 401:788–791.
- Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., and von der Malsburg, C., 1998, The Bochum/USC face recognition system and how it fared in the FERET Phase III test, in *Face Recognition: From Theory to Applications* (H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, Eds.), Berlin: Springer-Verlag.
- O'Toole, A. J., Deffenbacher, K. A., Valentin, D., and Abdi, H., 1994, Structural aspects of face recognition and the other-race effect, *Memory Cognit.*, 22:208–224.
- O'Toole, A. J., Wenger, M. J., and Townsend, J. T., 2001, Quantitative models of perceiving and remembering faces: Precedents and possibilities, in *Computational, Geometric, and Process Perspectives on Facial Cognition* (M. J. Wenger and J. T. Townsend, Eds.), Mahwah, NJ: Erlbaum, pp. 1–38. ♦
- Phillips, P. J., Moon, H., Rizvi, S., and Rauss, P., 2000, The FERET evaluation method for face recognition algorithms, *IEEE Trans. Pattern Recog. Machine Intell.*, 22:1090–1104. ♦
- Sirovich, L., and Kirby, M., 1987, Low dimensional procedure for characterization of human faces, *J. Opt. Soc. Am.*, A4:518–519.
- Turk, M., and Pentland, A., 1991, Eigenfaces for recognition, *J. Cognit. Neurosci.*, 3:71–86.
- Valentine, T., 1991, A unified account of the effects of distinctiveness, inversion, and race in face recognition, *Q. J. Exp. Psychol.*, 43A:161–204.

Fast Visual Processing

Simon J. Thorpe and Michèle Fabre-Thorpe

Introduction

How long does the visual system take to process an image? Yarbus's pioneering studies of scanning eye movements in the 1960s showed that we typically make about three saccades a second. This finding implies that a few hundred milliseconds is enough for visual analysis and programming the next eye movement. By the 1970s, work by Irv Biederman and Molly Potter had shown that much information can be extracted from briefly glimpsed scenes, even at presentation rates of around 10 frames/s, a technique known as rapid sequential visual presentation (RSVP). In the early 1980s, one of the prime motivations for the development of connectionist and PDP models was Jerry Feldman's “100-step limit.” Feldman argued that since many complex cognitive tasks can be performed in about half a second, and since interspike intervals for neurons are seldom shorter than 5 ms, the underlying algorithms should involve no more than about 100 sequential, though massively parallel, steps. These numbers were only ballpark figures and unrelated

to any particular task, but Feldman's strategy has recently been taken a step further by measuring processing time on specific high-level visual tasks. In combination with knowledge about underlying anatomy and physiology, such information can provide insights into the processing strategies used by the brain.

There is an important distinction in neural computation between feedforward processing models and those with recurrent connections that allow feedback and iterative processing. Pure feedforward models (e.g., multilayer perceptrons, or MLPs) can operate very quickly in parallel hardware. But many authors, including, for example, Ullman (1996), Rao and Ballard (1999), and Grossberg (2001), have argued that sophisticated visual processing requires the interplay of bottom-up and top-down mechanisms. Such views are supported by the anatomy of the visual system, since neurons at virtually every level are influenced not only by feedforward projections but also by extensive feedback connections from later stages and horizontal connections within each layer. Should one conclude that, because the visual system has recurrent connections,

pure feedforward mechanisms have no role to play? Perhaps not, because even in systems that use extensively recurrent connections, the very fastest behavioral responses might essentially depend on a single feedforward processing wave. This article considers how detailed measurements of processing speed can be combined with anatomical and physiological constraints to constrain models of how the brain performs particular computations.

Measuring Processing Speed

The ultimate test for processing speed lies in behavior. If animals can make reliable behavioral responses to specific categories of stimuli with a particular reaction time, there can be no argument about whether the processing has been done. When a fly reacts to displacements of the visual world by changing wing torque 30 ms later, 30 ms is clearly enough for both visual processing and motor execution. Fast behavioral reactions are not limited to insects. For example, tracking eye movements are initiated within 70–80 ms in humans and within around 50 ms in monkeys (Kawano, 1999), and vergence eye movements, required to keep objects within the fixation plane, have latencies of around 85 ms in humans and less than 60 ms in monkeys (Miles, 1997). Such low values probably reflect the relatively simple visual processing needed to detect stimulus movement and the short path lengths in the oculomotor system. How fast could behavioral responses be on tasks that require more sophisticated visual processing?

In 1996, we reported fast behavioral responses on a challenging task for the visual system (Thorpe, Fize, and Marlot, 1996). Presented with color photographs flashed for only 20 ms, subjects had to release a button as quickly as possible if the image contained an animal, and not respond otherwise. Target and nontarget images were extremely varied, with targets including mammals, birds, fish, and insects in their natural environments. Furthermore, no image was shown more than once, which forced subjects to process each image from scratch with minimal contextual help. Despite all these constraints, accuracy was high (around 94%), with mean reaction times (RTs) typically around 400 ms.

While mean RT might be the obvious candidate for measuring processing speed, another useful value is the minimal time needed to complete the task. Using RT distributions, this value can be defined as the first time bin at which correct responses start to significantly outnumber erroneous responses to nontargets. Faster responses occurring with no bias toward targets are presumably anticipations triggered before stimulus categorization was completed. Remarkably, in the animal categorization task, these minimal response times can be less than 250 ms.

It might be thought that images associated with fast responses constitute a subpopulation particularly easy to analyze. However, we found no obvious features that characterize rapidly categorized images (Fabre-Thorpe et al., 2001), implying that, even with highly varied and unpredictable images, the entire processing sequence from photoreceptor to hand movement can be completed in under 250 ms. Remarkably, rhesus monkeys can also perform this task, but their minimal RTs are even faster, around 170–180 ms (Fabre-Thorpe, Richard, and Thorpe, 1998). As in the tracking and vergence eye movement studies mentioned earlier, humans take nearly 50% longer than their monkey cousins to perform a given task.

Such data clearly impose upper limits on the time needed for visual processing. However, they do not directly reveal how long visual processing takes because reaction times also include response execution. How much time should we allow for the motor part of the task? To get at this question, event-related potentials (ERPs) and magnetoencephalography (MEG) recordings can be used to track information processing between stimulus and response. For example, during performance of the animal categorization task, simultaneously recorded ERPs showed that the average response to correct target trials diverged sharply from the average

response to correct nontarget trials at about 150 ms post stimulus. This remarkably robust differential ERP response appears specifically related to target detection and occurs well in advance of even the fastest behavioral responses. This value of 150 ms for visual categorization leaves no more than 100 ms for motor execution when behavioral responses occur at around 250 ms.

Processing speed can also be assessed at the level of single neurons by determining the point at which they start to show selectivity for particular visual inputs. By examining the information contained in neuronal responses at different times and in different visual structures, one can follow how processing develops over time. Surprisingly, using response latency to track the time course of visual processing is a relatively recent technique in experimental neuroscience. Nevertheless, by 1989 it was clear that the onset latencies of selective visual responses were a major constraint on models (Thorpe and Imbert, 1989). Face-selective neurons had been described in monkey inferotemporal cortex (IT) with typical onset latencies of around 100 ms. Beyond the visual system as such, neurons in the lateral hypothalamus were known to respond selectively to food after only 150 ms. Although these earlier studies suggested that visual processing could be very fast, they did not specifically determine at which point the neuronal response was fully selective. In 1992 it was demonstrated that even the first 5 ms of the responses in IT neurons could be highly selective to faces (Oram and Perrett, 1992). Since IT neurons can start responding before 100 ms, such data can be used to assign firm limits to the processing time required to reach a certain level of analysis.

Implications for Computational Models

Determining the minimal time required to perform a particular task or computation is not enough to constrain models of the underlying mechanisms without taking into account underlying anatomy and physiology. Even if monkeys can perform an abstract categorization task in as little as 180 ms, this feat would tell us little about the underlying computations if the brain was an unstructured assembly of neurons. However, the underlying anatomical pathways involved are actually quite well known (see VISUAL SCENE PERCEPTION; OBJECT RECOGNITION, NEUROPHYSIOLOGY). Figure 1 illustrates a plausible route involving the various stages in the so-called ventral processing stream that leads to the inferotemporal cortex (V1-V2-V4-PIT-AIT), where neurons selective for complex visual forms are found. But since IT has no direct outputs to the motor system, information has to travel via such areas as prefrontal (PFC), premotor (PMC), and motor cortices (MC) before reaching the spinal cord.

On the basis of this diagram, we can start using neurophysiological data to estimate the processing time available at each step. For example, neurons in V1 can start firing 40 ms after stimulus onset, although 60 ms would be more typical. In IT, the earliest responses start at around 80 ms, with 100 ms being more typical. Given the number of processing steps involved, it would appear that the earliest responses in each area must be produced on the basis of only about 10 ms of activity in the previous stage.

How can such data be used to constrain models of processing? Bear in mind that this 10 ms value includes several components: synaptic transmission, integration of information by postsynaptic cells, spike initiation, and propagation to the next stage. Estimates of conduction velocity for intracortical axons suggest that conduction delays might be considerable. At 1–2 m/s, it would take 15–30 ms simply for information to propagate from V1 to IT (Nowak and Bullier, 1997), a substantial proportion of the 40 ms latency shift between the two areas. Moreover, few if any neurons that receive inputs from the preceding stage project directly to the next stage. In other words, processing in each cortical area almost certainly involves more than one layer of synapses, reducing further the amount of time available at each step. Together, such data imply

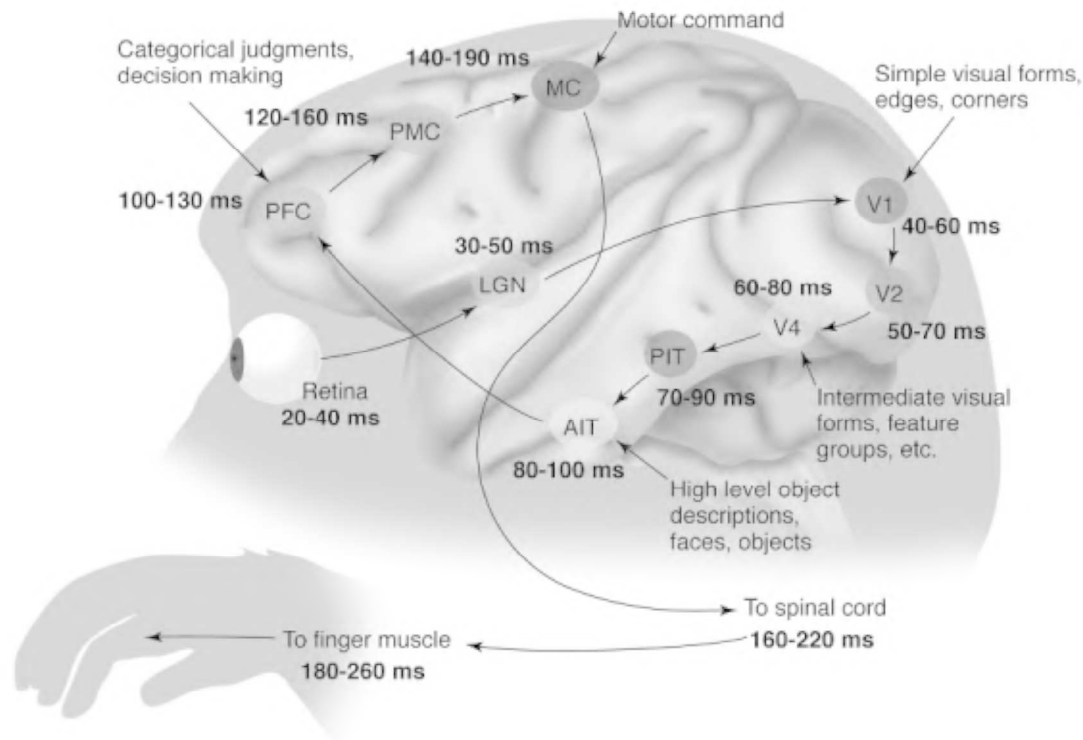


Figure 1. A possible input-output pathway for performing go/no-go visual categorization tasks in monkeys. Information passes from retina to lateral geniculate nucleus (LGN) before arriving in cortical area V1. Further processing occurs in areas V2, V4, and in the posterior and anterior inferotemporal cortex (PIT and AIT) before being relayed to the prefrontal (PFC),

premotor (PMC), and motor cortices (MC). Finally, motor neuron activation in the spinal cord triggers hand movement. For each area, the two numbers provide approximate latency values for the earliest responses and for a typical average response. (From Thorpe and Fabre-Thorpe, *Science*, 2001. Reprinted with permission.)

that the earliest responses in any particular area are presumably based on just a few milliseconds of activity in neurons at earlier levels.

This conclusion has some profound implications for cortical computation. First, it implies that, at least in the case of the earliest responses in any particular structure, there will be little time for iterative processing involving recurrent loops at previous stages. Since neurons in areas such as IT have responses that can be selective from the very start, it must presumably be possible to generate such responses on the basis of a pure feedforward pass through the system. This view is also supported by a recent study that analyzed IT responses to sequences of images presented at high rates. As frame rate was progressively increased, response strength dropped, but some neurons were still able to respond selectively at a 72-Hz frame rate, i.e., with images lasting only 14 ms (Keyser et al., 2001). With an onset latency of around 100 ms, it would appear that as many as seven separate images were being processed at the same time in a sequential pipeline, each stage of the visual system effectively processing a different image.

A second major consequence of these temporal constraints concerns the nature of the neural code. It is generally assumed that neurons encode information in their firing rates. Indeed, in the vast majority of artificial neural networks, the spike trains of real neurons are replaced by a single continuous value, often fixed in the range 0 to 1, supposed to represent firing frequency. But how accurately can one determine a firing rate with only a few milliseconds to listen to the output of each neuron? Given that neurons only rarely fire above 100 Hz, very few will generate more than one spike in the time available. This rules out counting several

spikes from the same neuron, or determining the interval between two spikes. Of course, one could measure firing rates across a population of neurons, but this would be an expensive strategy. Yet some alternative coding strategies can operate very efficiently, even with only one spike per neuron. For example, one could use the order in which neurons fire and the fact that the most strongly activated neurons tend to fire first (Thorpe, Delorme, and Van Rullen, 2001). Whatever the solution, it is clear that rapid processing poses a major challenge for computational neuroscience.

Discussion

By combining the detailed time course of visual processing with information about anatomical organization and the characteristics of individual neurons, including their conduction velocities, we can go well beyond the global statements made by Feldman in the early 1980s. Specifically, we have seen that the earliest, highly selective responses of neurons in high-order areas appear to depend almost entirely on a feedforward wave of processing that is so fast that each neuron may well get to fire only one spike. Although this may seem far-fetched, there are simulations showing that simple feedforward networks of neurons that generate only one spike can perform sophisticated tasks that include detecting, localizing, and identifying faces in natural images (Thorpe et al., 2001).

On the other hand, vision cannot be reduced to a feedforward pass. Visual perception is much more complex than simply pressing a button when a particular category of object is present, and there is increasing evidence that this very rapid processing may well be largely unconscious. The formation of a fully segmented,

conscious percept may well require far more complex processing than can be achieved in 100–150 ms, involving extensive use of feedback connections. Here again, temporal constraints can help determine the computations that can be performed with only feedforward processing and those that require recurrent mechanisms. For example, neuronal response properties that are not present at the onset of the response but take several tens of milliseconds to develop have recently been reported (Sugase et al., 1999; Lamme and Roelfsema, 2000). This progressive shaping of particular response properties is a hallmark of processing that requires recurrent mechanisms and feedback connections. In such cases, the initial very fast feedforward pass can act as a seeding process that allows subsequent processing to be performed in an intelligent top-down way.

This line of research, in which the details of the time course of visual processing are used to distinguish between different theoretical models of how the brain computes, constitutes a particularly clear example of the way in which experimental and theoretical work in brain theory can complement each other.

Road Map: Vision

Related Reading: Object Recognition, Neurophysiology; Visual Attention; Visual Scene Perception

References

- Fabre-Thorpe, M., Delorme, A., Marlot, C., and Thorpe, S., 2001, A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes, *J. Cognit. Neurosci.*, 13:171–180.
- Fabre-Thorpe, M., Richard, G., and Thorpe, S. J., 1998, Rapid categorization of natural images by rhesus monkeys, *NeuroReport*, 9:303–308.
- Grossberg, S., 2001, Linking the laminar circuits of visual cortex to visual perception: Development, grouping, and attention, *Neurosci. Biobehav. Rev.*, 25:513–526.
- Kawano, K., 1999, Ocular tracking: Behavior and neurophysiology, *Curr. Opin. Neurobiol.*, 9:467–473.
- Keyser, C., Xiao, D. K., Foldiak, P., and Perrett, D. I., 2001, The speed of sight, *J. Cognit. Neurosci.*, 13:90–101. ♦
- Lamme, V. A. F., and Roelfsema, P. R., 2000, The distinct modes of vision offered by feedforward and recurrent processing, *Trends Neurosci.*, 23:571–579.
- Miles, F. A., 1997, Visual stabilization of the eyes in primates, *Curr. Opin. Neurobiol.*, 7:867–871.
- Nowak, L. G., and Bullier, J., 1997, The timing of information transfer in the visual system, in *Extrastriate Cortex in Primates* (J. Kaas, K. Rockland, and A. Peters, Eds.), New York: Plenum Press, pp. 205–241. ♦
- Oram, M. W., and Perrett, D. I., 1992, Time course of neural responses discriminating different views of the face and head, *J. Neurophysiol.*, 68:70–84.
- Rao, R. P., and Ballard, D. H., 1999, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects, *Nature Neurosci.*, 2:79–87.
- Sugase, Y., Yamane, S., Ueno, S., and Kawano, K., 1999, Global and fine information coded by single neurons in the temporal visual cortex, *Nature*, 400:869–873.
- Thorpe, S., Delorme, A., and Van Rullen, R., 2001, Spike-based strategies for rapid processing, *Neural Netw.*, 14:715–725. ♦
- Thorpe, S., Fize, D., and Marlot, C., 1996, Speed of processing in the human visual system, *Nature*, 381:520–522.
- Thorpe, S. J., and Imbert, M., 1989, Biological constraints on connectionist models, in *Connectionism in Perspective* (R. Pfeiffer, Z. Schreier, F. Fogelman-Soulié, and L. Steels, Eds.), Amsterdam–North Holland, pp. 63–92.
- Ullman, S., 1996, *High-Level Vision: Object Recognition and Visual Cognition*, Cambridge, MA: MIT Press. ♦

Feature Analysis

Michael J. Morgan

Introduction

A popular idea is that natural images can be decomposed into constituent objects, which are in turn composed of features. The space of all possible images is vast, but natural images occupy only a small corner of this space, and images of significant objects such as animals or plants occupy a still smaller region. The visual brain has evolved to analyze only the interesting regions of image space.

A feature description of an image reduces the number of dimensions required to describe the image. An image is a two-dimensional (N by N) array of pointwise (or pixelwise) intensity values. If the number of possible pixel values is p , then the number of possible images is a set \mathfrak{X} , of size pN^2 . To distinguish all possible images having N by N pixels, we need a space of N^2 dimensions, which is too large in practice to search for a particular image.

The core idea behind feature analysis is that in real images, objects can be recognized in a space \mathfrak{Y} with a much smaller number of dimensions (a smaller dimensionality) than \mathfrak{X} . The space \mathfrak{Y} is a *feature space*, and its dimensions are the features. A simple example of a feature space is color space, in which all possible colors can be specified in a three-dimensional space, with axes $L-M$, $L+M-S$ and $L+M+S$, and L , M , and S are the photon catches of the long-, medium-, and short-wavelength receptors, respectively. The reason why a three-dimensional space suffices to distinguish the very much higher-dimensional space of surface reflectance spectra is that there is huge redundancy in natural spectra.

The reflectance at a given wavelength is highly correlated with reflectance at nearby wavelengths. We seek similar redundancies in space that will allow dimensional reduction of images.

Features Are Not Necessarily Localized

Note that in this very general framework, there is no implication that features are spatially localized. Features could, for example, be Fourier components. The global Fourier transform has the same dimensionality as the original image and is thus not, according to the present definition, a feature space. But if we throw away Fourier components that are unimportant in distinguishing objects or if we quantize phase, dimensional reduction has been achieved, and we have a feature space. The familiar example of JPEG compression involves a feature space.

To distinguish between spatially localized and nonlocalized features, we shall follow physicists in calling the former *particles* and the latter *waves*. Wavelets (see Olshausen and Field, 1996) are hybrids that are waves within a region of the image but otherwise are particles. Another important distinction is between particles that have place tokens and those that do not. Although all particles have places in the image, it does not follow that these places will be represented by tokens in feature space. It is entirely feasible to describe some images as a set of particles of unknown position. Something like this happens in many descriptions of texture. A very active source of debate in visual psychophysics has been the

extent to which the visual system uses place tokens (sometimes called local signs). For example, if the distance between two points *A* and *B* is seen as greater than the distance between two other points *C* and *D*, does this imply that there are place tokens for *A*, *B*, *C*, and *D*, or is some other mechanism involved (Morgan and Watt, 1997)?

The feature concept has proved useful in an impressive variety of different contexts.

Features in Ethology

Lorenz and Tinbergen's concept of the innate releasing mechanism (IRM) with its releasing stimulus foreshadowed much later work in behavior and physiology. The red spot at the base of the herring gull beak, which the young attack to get food from the parent, and the silhouette of the hawk/goose, which elicits fear when moved only in the hawk direction, are classic features that have entered folk psychology.

Features in Physiology

The classic paper "What the Frog's Eye Tells the Frog's Brain" (Lettvin et al., 1959) popularized the idea that special low-level sensory analyzers might exist for the purpose of responding to simple input features, the canonical example being the response of bug-detecting retinal ganglion cells to a small, moving spot. Hubel and Wiesel (1977) described bar and edge detectors in the visual cortex of cat and introduced an influential feature analysis scheme in which hierarchies of mechanisms would combine elementary features into ever increasingly complex objects. The hierarchical scheme, although not without its critics, was supported by the discovery of neurons in inferotemporal cortex (IT) responding selectively to images of complex objects such as a face or hand. Further studies of IT with simpler shapes found a columnar organization of IT with cells having similar response properties organized in repeating columns (Tanaka, 1996) (see Figure 1). The proposal that a ventral pathway leading to IT is responsible for object recognition gains support from lesioning and functional brain-imaging studies, although the idea of a single area for feature analysis is almost certainly too simple (Logothetis and Sheinberg, 1996).

Concept Learning

Most dogs bark and have hair, tails, and ears. But not all dogs bark, and Wittgenstein famously pointed out that some concepts such as

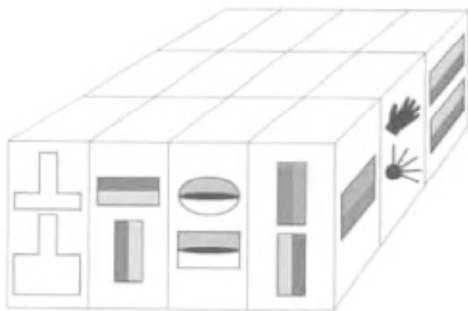


Figure 1. Columnar organization of inferotemporal cortex, based on work of Fujita et al. (see review by Tanaka, 1996). Columns of cells selective for similar complex shapes are interspersed with columns of cells unresponsive to these stimuli but selective to different complex shapes. (Source: Reproduced with permission from Stryker, M. P. (1992), *Nature (Lond.)*, 360:301. Note: In the original, different gray levels are in color, to show color preferences of the cells.)

"a game" have no necessary features. Philosophers and animal psychologists met his challenge to the feature concept with the polymorphous concept (Watanabe, Lea, and Ditttrich, 1993), an *n*-dimensional feature space in which instances of a concept occupy a subspace without sharp boundaries. Considerable research effort has been devoted to investigating the abilities of animals to learn both natural and artificially polymorphous concepts, a key issue being whether a linear model of feature combination will serve.

Cognitive Psychology

The idea that certain features can be analyzed at a preconscious level has proved fertile. Using the technique of visual search, Treisman (1988) suggested that only certain elementary features, and not their combinations, could serve as preconscious markers (Figure 2). This idea proved especially popular when linked, on rather slender evidence, with the discovery of specialized prestriate cortical areas in monkey devoted to the analysis of color and motion. However, the simple dichotomy between a fast, parallel search for features and a slow, serial search for combinations of features has come to be questioned (see the caption for Figure 2).

Image Compression

The earliest pictures to be sent across the transatlantic telegraph took more than a week to transmit. Engineers soon reduced this time to three hours by encoding the image more economically. Image compression techniques are divided into those that preserve all the information in the original (error-free) and lossy techniques, which try to transmit only the important features (Gonzalez and Woods, 1993). An early pioneer of speeding up telegraph transmission by a lossy feature decomposition was Francis Galton, who proposed a set of features for transmitting face profiles over the telegraph using only four telegraphic "words." Appropriately for the man who invented the fingerprint, he saw this primarily as a forensic aid in sending profiles of wanted criminals around the world. Galton's feature space lays stress on five *cardinal points*. These are the notch between the brow and the nose, the tip of the nose, the notch between the nose and the upper lip, the parting of the lips and the tip of the chin (Figure 3).

The remainder of this review illustrates the concept of feature spaces and describes key issues, some resolved and others not.



Figure 2. Searching for a single "odd man out" is easy (A) when the target has a very different orientation from the background elements or (B) when it has a different spatial frequency. Search times do not increase with the number of background elements (parallel search), provided that the orientation difference is sufficiently large. With smaller orientation differences (<10 degrees), search times do increase with the number of background elements, indicating a serial search. It might be thought that orientation and spatial frequency are easy search features because they are represented in primary visual cortex. However, (C) the conjunction of a particular spatial frequency and orientation is much harder to find, despite the fact that many V1 neurons are jointly tuned to orientation and frequency. If contrast is randomized (D), the search becomes harder still. The most powerful generalization about search is that it becomes harder when the number of different background elements increases. A standard ("back pocket") texture segmentation mechanism that responds to local contrasts in orientation, frequency, contrast, and color can explain most of these findings. (Source: Figures by J. A. Solomon.)

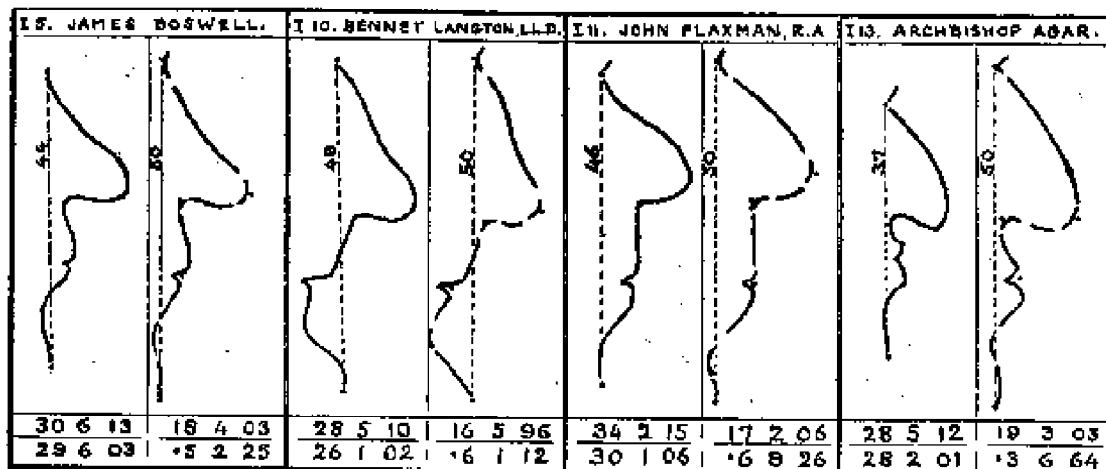


Figure 3. Face profiles described (left member of each pair) and reconstructed (right member of each pair) from 20 telegraphic numbers arranged in five groups of four (bottom). The code uses the relationships between five cardinal points on the profile. Galton was perhaps the first to use the

term *cardinal points* to describe a multidimensional feature space for object recognition. (Source: From Galton, F., 1910, *Nature (Lond.)*, March 31, 127–130.)

Principal Component Analysis of Faces

Galton was the first to treat faces as mathematical objects and to add them photographically. The *average face* was held to be especially beautiful, possibly because smallpox scars and other blemishes were removed. Galton also presented the average faces of groups such as rapists, clergymen, and athletes. Any particular face could then be correlated with each average in turn and described by a vector \mathbf{V} . This vector will tell us the extent of resemblance of that face to the mean rapist, the mean clergyman, the mean athlete, and so on. The average images are a *basis set* for describing all faces. Dimensional reduction has been achieved, as long as the dimension of \mathbf{V} is less than that of the original image space, N^2 . Whether the clergyman-athlete-rapist space is a good one is another matter. It is probably not, because the dimensions are correlated. The aim of most feature analysis, including principal component analysis (PCA), is to ensure that the dimensions of the feature space are uncorrelated, or, in other words, that the axes in the space are *orthogonal*.

The idea behind PCA for faces (or Karhunen-Loeve expansion) is to find a set of features called *eigenvectors* that span the subspace of images in which faces lie (Turk and Pentland, 1991). Each eigenvector has dimensions N^2 and is a linear combination of a set of training faces, each of dimension N by N pixels. Equivalently, each face in the training set is a linear combination of the N^2 eigenvectors. To achieve dimensional reduction, only the most important eigenvectors are chosen as a basis set. These are the vectors that correlate most highly with the members of the training set. The most important, in this sense, is the average image, as defined by Galton.

Since the eigenvectors resemble faces when they are represented as 2D images, they can be called *eigenfaces*. Examples are shown in Figure 4. Once the eigenfaces have been created, each face in the training set can be described by a set of numbers representing its correlation to each of the eigenfaces in turn. If seven eigenfaces are chosen to span the face space, each face will be described by a vector of seven numbers instead of by its pixel values. A huge dimensional reduction has been achieved. The problem of recognizing a face is now a simple one of pattern recognition: finding the vector in memory that it most closely resembles.

PCA has been used for face detection, face recognition, and sex classification. Effective caricatures can be derived by exaggerating

the differences of a face from the average. An intriguing experiment by Leopold, O'Toole, and Blanz (2001) suggests that the brain may use a feature space to identify faces. Observers were trained to discriminate "Adam," who was described by a vector in face space, from "Anti-Adam," who had the directly opposite vector. After adapting to Adam for some minutes, observers were more likely to classify the average face (midway between Adam and Anti-Adam) as Anti-Adam than as Adam. The inference is that there exist feature detectors that are tuned in face space and that identification depends on a population code comprising these detectors.

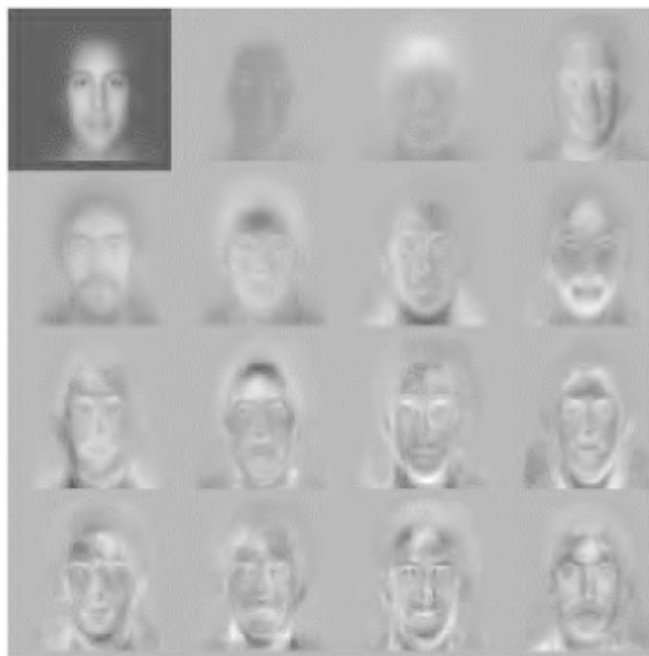


Figure 4. Eigenfaces for face recognition. (Source: Reproduced with permission from the interactive MIT Media Lab Web site: <http://www.white.media.mit.edu/vismod/demos/facerec/basic.html>)

Fourier Freaks and Feature Creatures

In broad terms, we now see that the code for early vision will be cracked when we find the basis set used by the brain for describing the image. One such basis set is the Fourier transform: sinusoids of differing frequency, orientation, and phase are the eigenfunctions of linear systems (Turk and Pentland, 1991). Following the application of certain key ideas in linear systems theory to vision (Robson, 1980), there was much debate about whether feature analysis or Fourier analysis was the preferred vehicle for understanding biological vision. This entertaining but ultimately fruitless debate is now essentially dead and buried. It made sense only when the "Fourier freaks" maintained that objects were recognized exclusively from their global amplitude spectrum. Since no one will now admit to having thought such a thing, it is pointless to provide historical detail.

Contrary to the view that the Fourier amplitude spectrum carries the important information in natural images, it is now recognized that the amplitude spectrum of most images is very similar (Field, 1987). The interesting features of images such as the boundaries of objects are represented in the relative *phases* of Fourier components. An edge or a bar in the image is a place where Fourier components undergo constructive interference. If the global amplitude spectra of two different images such as faces are interchanged, the hybrid images look like the images from which their phase spectra are derived (Figure 5). However, this is true only for the global Fourier transform. If the image is decomposed into a number of overlapping patches and each patch is transformed, it is the amplitude rather than the phase information in the patches that determines the appearance of the image if the patch size is sufficiently small (Figure 5). This is self-evidently true when the patch size is a single pixel, because the transform contains only the DC level and no phase information. But the limit is reached before a single pixel. Further work is needed to determine how the limiting patch size varies in different images and whether it is determined by cycles/image or cycles/degree of visual angle.

The consensus view now is that Fourier analysis in the visual system is limited to a local or *patchwise* Fourier analysis and that this is performed by neurons in V1 with localized, oriented, and spatial-frequency-tuned receptive fields (Robson, 1980). The idea

of patchwise spatial frequency analysis fits in well with the architectural division of V1 into *hypercolumns*, each containing a full range of orientations and spatial frequencies, with a scatter of their receptive fields within a region of the image (Hubel and Wiesel, 1977). According to this model, the receptive fields of simple cells in V1 provide a *basis set* for describing local properties of the image, comparable to the wavelet transform in image processing (Olshausen and Field, 1996). Putting this simply, the idea is that an object can be recognized locally in an image by a series of numbers representing its effect on the activity of a population of detectors tuned individually in orientation and spatial frequency. Dimensional reduction is achieved because pointwise intensity values have been discarded and replaced by a more economical code. Just how many types of receptive field are needed to provide a satisfactory basis set for describing natural images is a question of equal interest to psychophysics and image processing.

PCA is far from being the only way to find a suitable basis set for natural images. Receptive fields like those in V1 emerge naturally as a basis set from a learning algorithm that seeks a sparse code for natural images (Olshausen and Field, 1996). The aim of sparse coding is to have each image activate the smallest possible number of members of the basis set. In physiological terms, the aim would be to have as many neurons as possible not activated at all by the image. Using this approach, Olshausen and Field derived a basis set having an impressive similarity to the receptive fields of V1 neurons (Figure 6).

The Primal Sketch

The ability of line drawings to convey shape is very strong evidence that the feature space for object recognition may be of drastically reduced dimensionality, compared to the space of all possible images. Just a few lines drawn on a flat surface can suggest the face of a well-known person or the idea "no skateboarding allowed." Impressed by the effectiveness of cartoons, David Marr (1982) proposed that the earliest stages of vision transform the continuous gray-level image into a neural cartoon, which he called "the primal sketch." Since cartoons emphasize primarily the outlines of shapes, the main objective of the primal sketch is to find *edges* in the image, edges being the loci of points on the 3D object where an object



Figure 5. Images of two political theorists (rightmost panels), one of whose ideas have been recently discredited. Each thinker contributes his or her phase from the Fourier transform to each of the images on the left; the amplitude information comes from the other face. The Fourier transform is not global but is rather derived from overlapping patches, the size of which

decreases (64, 32, 16, 8, and 4 pixels) from left to right. When the patch size is large, the appearance of the image is dominated by phase information; when it is small, the appearance is dominated by the amplitude of the Fourier components. At intermediate patch sizes, the appearance is composite. (Source: Reproduced with permission from Morgan et al., 1991.)

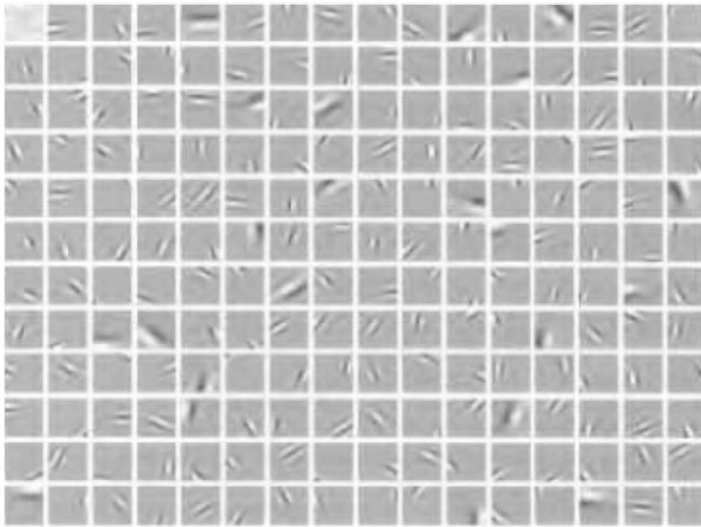


Figure 6. A basis set derived from ten 512 by 512 natural images of the American Northwest by a training algorithm aimed at maximizing the sparseness of the representation. The basis set bears a striking similarity to the receptive fields of V1 neurons, suggesting that they too form an efficient basis set for describing natural images. (Source: Reproduced with permission from Olshausen and Field, 1996.)

occludes itself or objects that are farther away (edge of a cube), or where there is a rapid change in the direction of the tangent plane to the object surface (outline of the nose on a face), or where there is some abrupt change in surface reflectance (boundary of the iris). It is a remarkable fact that much of the edge structure of an object in the image depends on its 3D structure—remarkable because we recognize objects from a variety of viewpoints or from sketches made from a variety of viewpoints.

If we consider an edge such as the outline of the nostril on a face, we shall find in its image a sudden change in luminance, which is not predictable from the gradual changes around it. A discontinuity is conveniently found by a local maximum or minimum in the first spatial derivative of the luminance profile or, equivalently, a *zero-crossing* in the second derivative. Ernst Mach (followed by William McDougall in his quaint “drainage” theory) was the first to conjecture that the visual system uses second derivatives to find features in the luminance profile, his evidence being the appearance of “Mach bands” on the inflection points in luminance ramps. Marr and Hildreth (see Marr, 1982) proposed that the receptive fields of retinal ganglion cells and simple cells in V1 make them nearly ideal second-derivative operators (Laplacians of Gaussians) and that their function is to produce the primal sketch. Different sizes of receptive field produce cartoons at different *spatial scales*, corresponding to the different frequencies in the Fourier transform but agreeing on the position of zero-crossings at the most significant points in the image. This recalls the fact that an edge or a bar in the image is a place where Fourier components undergo constructive interference. The primal sketch neatly uses wavelets to locate edges and turns them into particles.

A wide variety of phenomena have been used to investigate the nature of the “spatial primitives” or features in human vision. These include Mach bands, the Chevreul illusion, the apparent location of bars and edges in gratings and plaids with different spatial frequency components, and edge blur discrimination. Various primitives have been considered, such as zero-crossings, zero-bounded regions, and local energy maxima. Although it is now possible to predict the apparent location of edges and bars in images with a fair degree of accuracy, there is no consensus as yet about the nature or existence of primitives or about the way in which they are combined over spatial scale (Morgan and Watt, 1997).

Summary

Features are useful for describing natural images because the latter have massive informational redundancy. Image space itself is too

vast to search directly. Feature analysis depends on the proposition that the search for particular objects can be concentrated in a subspace of image space: *the feature space*. Biologists expect that there will be special sensory mechanisms for searching just the right subspace for a particular task. Ethologists and animal learning theorists concur. Useful hints about likely feature spaces may be obtained from engineers working on image compression. Although in the past feature analysis was contrasted with Fourier analysis, the modern synthesis is that a patchwise Fourier analysis by localized receptive fields in primary visual cortex (V1) provides the primitive basis set for the feature space of vision. These form the basis set for the elaboration of neurons responding selectively to geometrical features in area TE of the inferotemporal cortex, and these in turn form the basis for object recognition in different but overlapping areas of IT.

Road Map: Vision

Related Reading: Gabor Wavelets and Statistical Pattern Recognition; Global Visual Pattern Extraction; Object Recognition, Neurophysiology; Orientation Selectivity

References

- Field, D. J., 1987, Relations between the statistics of natural images and the response properties of cortical cells, *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 4(12):2379–2394.
- Gonzalez, R., and Woods, R., 1993, *Digital Image Processing*, Reading, MA: Addison-Wesley. ♦
- Hubel, D. H., and Wiesel, T. N., 1977, Functional architecture of the macaque monkey visual cortex: Ferrier Lecture, *Proc. R. Soc. Lond. B Biol. Sci.*, 198:1–59. ♦
- Leopold, D., O’Toole, A. T., and Blanz, V., 2001, Prototype-references shape encoding revealed by high-level aftereffects, *Nature Neurosci.*, 4:89–94.
- Letttvin, J. Y., Maturana, R. R., McCulloch, W. S., and Pitts, W. H., 1959, What the frog’s eye tells the frog’s brain, *Proc. Inst. Rad. Eng.*, 47:1940–1951.
- Logothetis, N. K., and Sheinberg, D. L., 1996, Visual object recognition, *Ann. Rev. Neurosci.*, 19:577–621. ♦
- Marr, D., 1982, *Vision*, San Francisco: W.H. Freeman. ♦
- Morgan, M. J., Ross, J., and Hayes, A., 1991, The relative importance of local phase and local amplitude in patchwise image reconstruction, *Biol. Cybern.*, 65:113–119.
- Morgan, M. J., and Watt, R. J., 1997, The combination of filters in early

- spatial vision: A retrospective analysis of the MIRAGE model, *Perception*, 26:1073–1088. ♦
- Olshausen, B., and Field, D., 1996, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, 381:607–609.
- Robson, J. G., 1980, Neural images: The physiological basis of spatial vision, in *Visual Coding and Adaptability* (C. S. Harris, Ed.), Hillsdale, NJ: Lawrence Erlbaum, pp. 177–214. ♦
- Tanaka, K., 1996, Inferotemporal cortex and object vision, *Ann. Rev. Neurosci.*, 19:109–139. ♦

- Treisman, A. M., 1988, Features and objects: The 14th Bartlett Memorial Lecture, *Q. J. Exp. Psychol. A*, 40:201–237. ♦
- Turk, M., and Pentland, A., 1991, Face recognition using eigenfaces, in *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA: IEEE Computer Society Press, pp. 586–591.
- Watanabe, S., Lea, S., and Ditttrich, W., 1993, What can we learn from experiments in pigeon concept formation?, in *Vision, Brain and Behaviour in Birds* (H. Zeigler and H.-J. Bischof, Eds.), Cambridge, MA: MIT Press. ♦

Filtering, Adaptive

John J. Shynk

Introduction

Adaptive filtering is an active area of research that has found widespread use in numerous signal processing and communications applications (Haykin, 2002). These applications include a wide range of important problems, such as noise canceling and noise reduction, channel equalization, co-channel signal separation, system identification, pattern recognition, fetal heart monitoring, and array processing (Ljung and Söderström, 1983; Qureshi, 1985; Giannakis, 1999). Adaptive filters are particularly useful for applications in which the underlying statistics are unknown or nonstationary, which is usually the case in practice (Widrow and Stearns, 1985). The parameters of an adaptive filter are adjusted to “learn” or track signal and system variations according to a performance criterion that is determined by the needs of the specific application.

The field of adaptive filtering was derived partly from work on neural networks and adaptive pattern recognition (Lippmann, 1987; Haykin, 1999). In pattern recognition applications, such as speech recognition and image classification, a neural network can be trained on a wide range of representative input patterns (called the training set). If the training set is sufficiently large, the neural net is capable of successfully classifying new patterns; for example, distorted patterns corrupted by noise will be mapped to the most similar pattern in the training set, as defined by some error criterion. Classification is performed by a threshold or decision device that quantizes the neural net output to one of many different levels representing the various classes.

An adaptive filter can be viewed as a signal combiner consisting of a set of adjustable weights (or coefficients represented by a polynomial) and an algorithm (learning rule) that updates these weights using the filter input and output, as well as other available signals. The filter may include internal signal feedback, whereby delayed versions of the output are used to generate the current output, and it may contain some nonlinear components. The single-layer perceptron is a well-known type of adaptive filter that has a binary output nonlinearity; it is also referred to as *adaline* (for *adaptive linear neuron*) (Widrow and Lehr, 1990) (see PERCEPTONS, ADALINES, AND BACKPROPAGATION). A multilayer perceptron contains many single-layer perceptrons interconnected to form a network that can implement a complex nonlinear system or represent a multidimensional signal pattern (Rumelhart and McClelland, 1986).

It is beyond the scope of this article to cover in depth the different types of adaptive filter configurations and the large number of adaptive algorithms. Instead, we will focus on the most widely used adaptive filter architecture and describe in some detail two representative adaptive algorithms. The least-mean-square (LMS) algorithm (Widrow and Stearns, 1985) computes the coefficients of

a tapped delay line (TDL), which is basically a shift register with adjustable coefficients. The constant modulus algorithm (CMA) (Godard, 1980; Treichler and Agee, 1983) also computes the coefficients of a TDL, but, unlike the LMS algorithm, it does not require a training sequence. As we shall see, both algorithms are stochastic-gradient methods that iteratively search for the minimum of well-defined performance (cost) functions. Example CMA performance functions are plotted to reveal their nonquadratic shapes, and computer simulations of the CMA weight trajectories are shown to illustrate the algorithm’s learning behavior for an equalization application.

Adaptive Filter Components

An adaptive filter consists of two components: (1) an *adjustable set of weights* that filter or process the input signal and (2) an *adaptive algorithm* that modifies the weights to minimize a performance measure. To be more precise, define the weight vector

$$W(n) \triangleq [w_1(n), w_2(n), \dots, w_N(n)]^T \quad (1)$$

and the input signal vector

$$X(n) \triangleq [x_1(n), x_2(n), \dots, x_N(n)]^T \quad (2)$$

where the argument n denotes discrete time and the superscript T is matrix/vector transpose. This particular form of the signal vector contains N distinct inputs and would be appropriate for multidimensional applications such as pattern recognition and array processing. However, there are many signal processing applications involving only one input signal, such as channel equalization, echo cancellation, and system identification. For these cases, the input signal vector would instead be

$$X(n) \triangleq [x(n), x(n-1), \dots, x(n-N+1)]^T \quad (3)$$

where delayed versions of the input signal $x(n)$ are stored in a shift register or TDL.

The (scalar) filter output is given by the inner product

$$y(n) = W^T(n)X(n) = X^T(n)W(n) \quad (4)$$

It is typically used by the adaptive algorithm in a feedback mechanism that determines how $W(n)$ should be adjusted with each new input vector $X(n)$. In some applications, $y(n)$ is processed by a nonlinear device that yields a decision statistic for an underlying signal contained in the input $X(n)$. For example, in a digital communications application with binary (± 1) transmitted symbols in an additive white Gaussian noise channel (Proakis, 2001), the nonlinearity could be the signum function:

$$\text{sgn}(y(n)) = \begin{cases} +1, & y(n) \geq 0 \\ -1, & y(n) < 0 \end{cases} \quad (5)$$

where we have arbitrarily assigned $\text{sgn}(0) = +1$. This detector uses the sign of the received signal $y(n)$ to determine the most likely transmitted symbol: $+1$ or -1 . The signum function is also used for Rosenblatt's training algorithm (Lippmann, 1987), while a soft nonlinearity such as the hyperbolic tangent is used in the back-propagation algorithm for the multilayer perceptron (Widrow and Lehr, 1990).

The error signal in Figure 1 depends on the application and the performance function; we will provide details of the error in a subsequent section when the adaptive algorithms are discussed. However, at this point we should mention that the error can be generated in (at least) two basic ways. If a training signal $d(n)$ is available (such as in the previously mentioned digital communications application), then the error is a measure of how much $y(n)$ differs from $d(n)$; in this scenario, the training signal is sometimes called the *desired response* in the adaptive filtering literature. For some problems, however, a training signal is not available (also in the digital communications application); in this case, the error is computed in a *blind* manner, which means that it depends only on the adaptive filter input $X(n)$ and output $y(n)$, without any explicit desired-response information.

Applications

Two representative applications of adaptive filtering are described in this section. Many signal processing applications fall into one of these general models. For example, adaptive channel equalization belongs to the field of deconvolution, and some co-channel signal separation techniques utilize an array processing formulation.

Channel Equalization

The configuration for adaptive channel equalization is shown in Figure 2 (Qureshi, 1985). Observe that the adaptive filter is placed in *cascade* with the unknown channel. The goal of the adaptive filter is to mitigate the channel distortion (e.g., multipath propagation) and, in effect, estimate the channel *inverse*. The input of the adaptive filter is obtained from the channel output, which may be corrupted by an additive (Gaussian) noise process $v(n)$. The training signal (desired response) is the transmitted signal $x(n)$. However, since the channel and the adaptive filter each have a finite non-zero delay, it is necessary that the transmitted signal be delayed by an appropriate amount $\Delta > 0$ such that $d(n) = x(n - \Delta)$.

Clearly, since the purpose of any communication system is to transmit information (unknown to the receiver) across the channel,

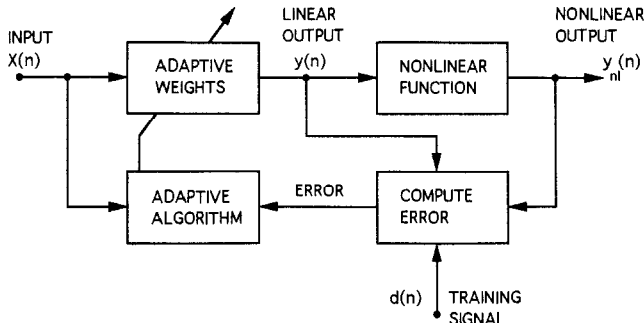


Figure 1. Adaptive filter components.

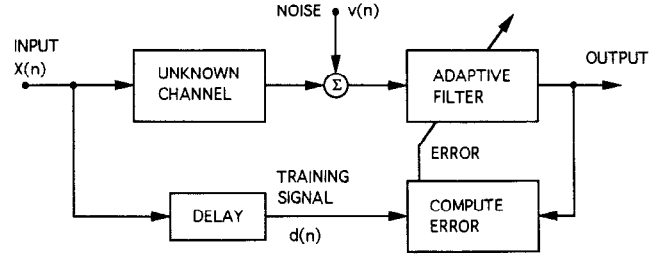


Figure 2. Adaptive channel equalization.

the configuration in Figure 2 for training the adaptive filter is not always feasible. Typically, there is a short training period at the beginning of transmission (start-up) whereby a signal known at the receive end is transmitted and used by the adaptive filter to adjust its weights. When the error rate is sufficiently reduced at the end of training, information can be successfully transmitted across the channel. During this time, the adaptive algorithm stops updating; or, alternatively, a decision-directed or blind adaptive algorithm can be employed that requires knowledge only of certain statistical properties of the transmitted signal (Haykin, 2002).

Co-channel Signal Separation

In cellular communication systems, co-channel interference is becoming an increasingly important issue as the number of subscribers grows (Giannakis, 1999). Co-channel interference is due to frequency reuse, whereby multiple cells operate on the same carrier frequency. To mitigate co-channel interference, it would be desirable to incorporate adaptive antennas that have a directional (beam-forming) capability to separate several co-channel signals, thus allowing for greater frequency reuse. In recent years there has been much interest in blind co-channel signal separation algorithms for antenna arrays that can adapt their parameters without using a training signal.

A block diagram of an adaptive array known as the constant modulus (CM) array (Gooch and Lundell, 1986) is shown in Figure 3, where the antenna elements are uniformly spaced and omnidirectional. It consists of two components: (1) a conventional adaptive antenna (beamformer) with weights updated by CMA, and (2) an adaptive signal canceler that removes from the input the co-channel signal captured by the CM array. The weights of the signal canceler are updated by the LMS algorithm using the received array signals as the desired response. A multistage implementation of this architecture, consisting of multiple beamformer/canceler systems in cascade, can be used to separate and recover several co-channel signals (one co-channel signal per stage).

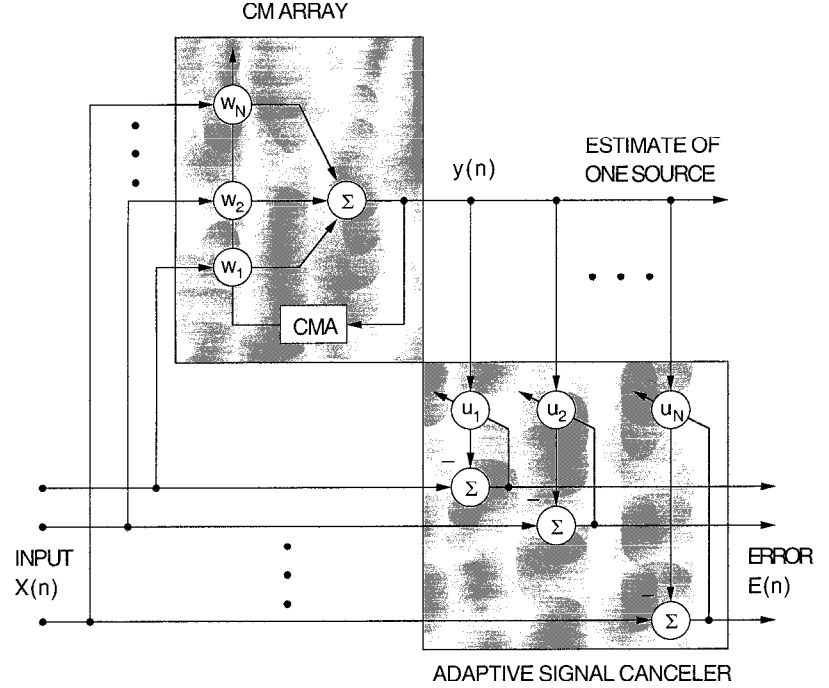
Adaptive Algorithms

In this section, we consider two performance functions: the squared error, which employs a training sequence, and a constant modulus formulation, which yields a blind adaptive algorithm. Both algorithms have the following general form:

$$W(n+1) = W(n) + \mu[-\nabla(n)] \quad (6)$$

where $\nabla(n)$ is the *gradient* with respect to $W(n)$ of the corresponding performance function. The positive step size μ controls the transient and steady-state convergence properties of the adaptive weights. The goal of an adaptive algorithm is to adjust $W(n)$ such that it converges to a minimum (stationary point) of the performance function.

Figure 3. Co-channel signal separation.



Least-Mean-Square Algorithm

Consider the following performance function:

$$C_{LMS}(n) = e^2(n) \quad (7)$$

where the error is

$$e(n) = d(n) - y(n) \quad (8)$$

and $d(n)$ is the training sequence. Note that $C_{LMS}(n)$ is a stochastic variable that fluctuates around the mean square error (MSE), given by

$$\xi_{LMS} = E[e^2(n)] \quad (9)$$

where $E[\cdot]$ denotes statistical expectation. For this reason, the weight update is referred to as a *stochastic-gradient* algorithm. $C_{LMS}(n)$ can be viewed as a very simple estimate of the ensemble average ξ_{LMS} . (In this article, we interchangeably refer to ξ and its stochastic estimate $C(n)$ as performance functions.) Using Equations 4 and 8, and noting that $\partial(W^T(n)X(n))/\partial W(n) = X(n)$, the gradient of Equation 7 is

$$\nabla_{LMS}(n) = -2X(n)e(n) \quad (10)$$

resulting in the least-mean-square (LMS) algorithm (Widrow and Stearns, 1985):

$$W(n+1) = W(n) + 2\mu X(n)e(n) \quad (11)$$

Thus, with each new set of data $\{X(n), d(n)\}$, the error $e(n)$ is computed and the filter weights are updated in an attempt to learn the underlying signal statistics. The algorithm converges on average (i.e., in the mean) when $E[X(n)e(n)] = 0$ (the zero vector) which, after substituting $e(n)$, yields the following unique stationary point:

$$W_{LMS} = R^{-1}P \quad (12)$$

where $R \triangleq E[X(n)X^T(n)]$ and $P \triangleq E[X(n)d(n)]$ are signal correlations (an $N \times N$ matrix and an $N \times 1$ vector, respectively). Observe that the algorithm utilizes the output $y(n)$ before it is pro-

cessed by any subsequent nonlinear device (as illustrated in Figure 1).

Constant Modulus Algorithm

The constant modulus algorithm (CMA) is a blind equalization technique that attempts to restore the constant modulus property of certain communication signals (the binary digital communication signal with values ± 1 is such an example). The performance function is a measure of the deviation of the modulus of the equalizer output $y(n)$ from a predetermined constant $r > 0$ according to Treichler and Agee (1983):

$$C_{CMA}(n) = \|y(n)\|^p - r^{p/q} \quad (13)$$

where p and q are positive integers, equal to 1 or 2, resulting in four versions of CMA. The scalar r is usually chosen such that the gradient of ξ_{CMA} (the expectation of Equation 13) with respect to the coefficients $W(n)$ is zero when the channel is perfectly equalized. (Note that although $C_{CMA}(n)$ depends on p and q , in order to simplify the notation, we do not indicate this explicitly.)

The gradient of Equation 13 is

$$\nabla_{CMA}(n) = pqX(n)y(n)|y(n)|^{p-2} \times (|y(n)|^p - r^{p/q})^{q-1} \text{sgn}[|y(n)|^p - r^{p/q}] \quad (14)$$

where $\text{sgn}[\cdot]$ is the signum function previously defined. For convenience, consider the case of $p = q = 2$. Substituting Equation 14 into Equation 6 yields the 2-2 version of CMA:

$$W(n+1) = W(n) - 4\mu X(n)y(n)(|y(n)|^2 - r^2) \quad (15)$$

This recursion does not depend on a training sequence; the update is a function only of $X(n)$ and $y(n) = W^T(n)X(n)$. Note that Equation 15 is similar to Equation 11, except that the scalar error $e(n)$ has been replaced by $-2y(n)(|y(n)|^2 - r^2)$.

Unfortunately, since $\xi_{CMA} = E[C_{CMA}(n)]$ is not a quadratic function of $W(n)$ (as is ξ_{LMS}), it is not possible to derive a simple general expression for the stationary point W_{CMA} . In fact, the stationary

point is not necessarily unique, and there may be local as well as global minima (the gradient is zero at a local minimum, but a local minimum does not correspond to the overall minimum value of the performance function). However, for the sake of completeness, if we assume that the input $X(n)$ is a Gaussian random vector with zero mean, then it can be shown that the stationary points are given by the following implicit expression (Shynk and Chan, 1993):

$$W_{CMA}^T R W_{CMA} = r^2/3 \quad (16)$$

Thus, there is an infinity of solutions, all of which achieve the global minimum. Note, however, that this result does not apply in general, because $X(n)$ might be generated by an underlying constant modulus signal (which CMA is designed to handle), in which case $X(n)$ would not be exactly Gaussian.

Convergence Results

In this section, we present some examples of the performance surfaces for CMA, as well as representative trajectories of the CMA weights during adaptation. A *performance surface* is a three-dimensional plot of the performance function versus two of the weight vector components.

CMA Performance Surfaces

Suppose that CMA attempts to equalize a channel with the following transfer function (written using z -transform notation): $C(z) = 1 - 0.6z^{-1} + 0.36z^{-2}$. Assume that the transmitted signal is a sequence of binary symbols ± 1 that are independent and equally likely (i.e., Bernoulli with $\Pr(+1) = \Pr(-1) = 1/2$). The elements of the corresponding correlation matrix R are $R(1, 1) = R(2, 2) = 1.4896$ and $R(1, 2) = R(2, 1) = -0.8160$. Figure 4A shows the performance surface of a two-weight ($N = 2$) equalizer for the CMA 2-2 performance function. Observe that there are two global minima and two local minima (there is also a local maximum at the origin $W = 0$). Thus, depending on how the algorithm is initialized, the adaptive weights may converge to any one of these four minima, two of which are not optimal. Similar results are shown in Figure 4B for the CMA 1-1 performance function. In contrast, the performance surface for the LMS algorithm (which uses training) is a paraboloid with a unique global minimum and no local minima (provided R is positive definite) (Widrow and Stearns, 1985).

CMA Weight Trajectories

In adaptive algorithms that employ a gradient update mechanism, such as LMS and CMA, a trade-off exists between the convergence rate and the steady-state value of the MSE (even though the MSE is not the performance function for CMA, it is typically used as a benchmark measure for the performance of a blind algorithm). Both of these properties are controlled by the step size μ , as well as the statistics of the input $X(n)$. When μ is increased, the rate of convergence also increases, but so does the steady-state MSE. Thus, when comparing the convergence behavior of different stochastic-gradient algorithms, it is necessary that the steady-state and transient properties be considered together. For example, the convergence rates of two algorithms can be compared once the step sizes are chosen so that the algorithms achieve the same steady-state MSE.

Figure 5A shows this trade-off for CMA 2-2 for a channel equalization application. These “learning curves” were obtained by averaging the squared error between the equalizer output $y(n)$ and the desired (transmitted) signal $d(n)$ (with an appropriate delay) over ten independent computer runs; the resulting trajectories were then smoothed by a moving average filter. Observe that the steady-state

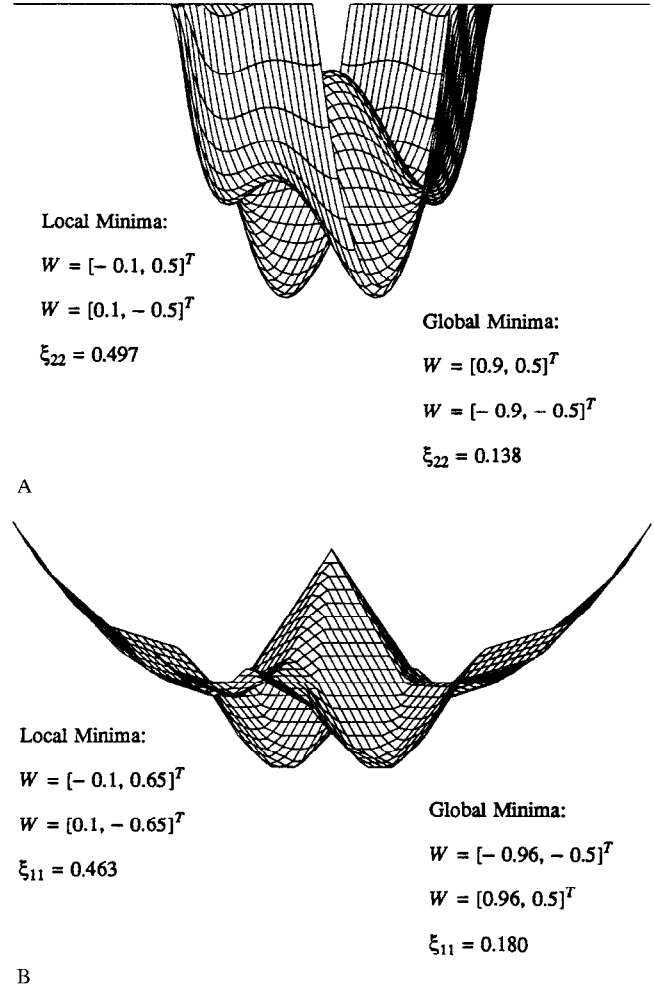
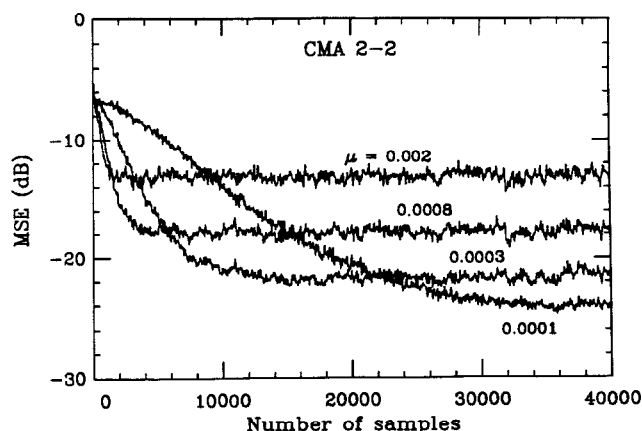


Figure 4. Performance surfaces ξ_{CMA} . A, CMA 2-2 ($p = q = 2$). B, CMA 1-1 ($p = q = 1$). (From Shynk, J. J., and Chan, C. K., 1993, Performance surfaces of the constant modulus algorithm based on a conditional Gaussian model, *IEEE Trans. Signal Proc.*, 41:1965–1969. © 1993, IEEE. Reprinted with permission.)

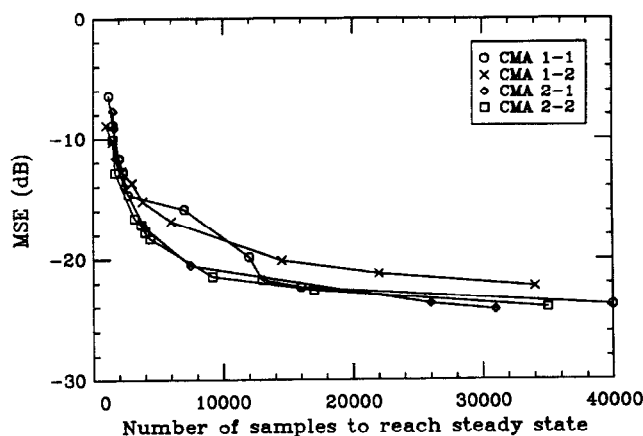
MSE and the convergence rate increase with increasing μ . In order to compare algorithms without having to explicitly consider the step size, several simulations can be performed for the algorithms, each simulation with a different value of μ . From this series of simulations, the number of samples required to reach steady state can be determined, along with the steady-state value of the MSE. Using this information, the steady-state MSE can be plotted versus the convergence time for each algorithm. Examples of these “performance curves” for the four versions of CMA are shown in Figure 5B (Shynk et al., 1991). Notice that the performance curve for CMA 1-2 is generally higher than the other curves, indicating that it requires (for this example) more iterations to achieve the same steady-state MSE for a wide range of step size values.

Discussion

Adaptive filters are widely used for a variety of signal processing applications. Although several configurations are possible, the linear combiner shown in Figure 1 is the most common, and the LMS algorithm, which is known to be robust (Haykin, 2002), is the most popular means of adjusting the adaptive weights. There are several



A



B

Figure 5. A, CMA learning curves. B, CMA performance curves. (From Shynk, J. J., Gooch, R. P., Giridhar, K., and Chan, C. K., 1991, A comparative performance study of several blind equalization algorithms, in *Proceedings of the SPIE Conference on Adaptive Signal Processing*. © 1991, SPIE. Reprinted with permission.)

variations of the LMS algorithm, including those that have less complexity or improved convergence properties. For example, CMA is a blind stochastic-gradient algorithm that can be used instead of the LMS algorithm when an explicit training sequence is not available. The recursive-least-squares (RLS) algorithm is an adaptive algorithm based on the method of least squares that offers faster convergence rates (compared with the LMS algorithm), but at the expense of an increased computational complexity (Haykin, 2002).

The adaptive filter configuration described in this article is the basic component of a multilayer perceptron. These additional layers provide greater nonlinear modeling capabilities, which is usually necessary for complex applications such as speech and image processing. Stochastic-gradient algorithms are typically used to adjust the weights of a multilayer perceptron. They are similar to the adaptive algorithms described in this article, but they have an additional degree of complexity owing to the cascade of layers. One such algorithm, known as the backpropagation algorithm (Rumelhart and McClelland, 1986), has been successfully applied to a number of signal processing problems (Widrow and Lehr, 1990).

Road Map: Applications

Background: Perceptrons, Adalines, and Backpropagation

Related Reading: Forecasting; Kalman Filtering; Neural Implications; Recurrent Networks; Learning Algorithms

References

- Giannakis, G. B., Ed., 1999, Highlights of signal processing for communications, *IEEE Signal Proc. Mag.*, 16:14-50. ♦
- Godard, D. N., 1980, Self-recovering equalization and carrier tracking in two-dimensional data communication systems, *IEEE Trans. Commun.*, COM-28:1867-1875.
- Gooch, R. P., and Lundell, J. D., 1986, The CM array: An adaptive beamformer for constant modulus signals, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York: IEEE, pp. 2523-2526.
- Haykin, S., 1999, *Neural Networks: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice-Hall. ♦
- Haykin, S., 2002, *Adaptive Filter Theory*, 4th ed., Upper Saddle River, NJ: Prentice-Hall. ♦
- Lippmann, R. P., 1987, An introduction to computing with neural nets, *IEEE ASSP Mag.*, 4:4-22. ♦
- Ljung, L., and Söderström, T., 1983, *Theory and Practice of Recursive Identification*, Cambridge, MA: MIT Press.
- Proakis, J. G., 2001, *Digital Communications*, 4th ed., New York: McGraw-Hill. ♦
- Qureshi, S. U. H., 1985, Adaptive equalization, *Proc. IEEE*, 73:1349-1387. ♦
- Rumelhart, D. E., and McClelland, J. L., Eds., 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press. ♦
- Shynk, J. J., and Chan, C. K., 1993, Performance surfaces of the constant modulus algorithm based on a conditional Gaussian model, *IEEE Trans. Signal Proc.*, 41:1965-1969.
- Shynk, J. J., Gooch, R. P., Giridhar, K., and Chan, C. K., 1991, A comparative performance study of several blind equalization algorithms, in *Proceedings of the SPIE Conference on Adaptive Signal Processing*, Bellingham, WA: SPIE, pp. 1565:102-117. ♦
- Treichler, J. R., and Agee, B. G., 1983, A new approach to multipath correction of constant modulus signals, *IEEE Trans. Acoust. Speech Signal Proc.*, ASSP-31:459-472.
- Widrow, B., and Lehr, M. A., 1990, 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation, *Proc. IEEE*, 78:1415-1441. ♦
- Widrow, B., and Stearns, S. D., 1985, *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall. ♦

Forecasting

Lyle H. Ungar

Introduction

Forecasting the future values of sequences of observations is, in many ways, ideally suited for neural networks. Large amounts of

data may be available, and the underlying relationships are often nonlinear and unknown. Neural nets, mostly of the standard backpropagation type (see BACKPROPAGATION: GENERAL PRINCIPLES), have been used with great success in many forecasting applications,

including forecasting electricity load, freeway traffic volume, solar cycles, milk yields, tourism demand, grain drying times, ambient air quality, exchange rates, inflation, unemployment, disease epidemics, fish stock levels, sea surface temperatures, sales volumes, flood occurrence in Moravia, and rainfall in Bangladesh. However, in not all of such cases do neural networks outperform conventional ARMA models. This article looks at the use of neural nets for forecasting, with particular attention to understanding when they perform better or worse than other technologies.

The success of neural networks in forecasting depends significantly on the characteristics of the process being forecast. One may want to predict minute-by-minute progress of a chemical reaction, hour-by-hour power usage (load) for an electric power utility, daily weather, monthly prices of products and inventory levels, and quarterly or yearly sales and profits. These problems differ in the quantity and type of information available for forecasting, and hence call for different forecasting techniques. One also needs to choose an appropriate network architecture.

Forecasting problems can be characterized on a number of dimensions: (1) Is a single series of measurements used, as is often done in conventional forecasting, or are multiple related measurements available? (2) Are the data seasonal or not? Monthly or quarterly data such as sales volume or energy use often show strong seasonal variation, while annual data or data measured each second or minute do not. (3) The number of observations and (4) the degree of randomness (signal/noise ratio) of the process also strongly limit the complexity of the model that can be fit. If data are only available annually for the past 10 or 20 years, and if no measurement is available for most of the disturbances, one should not expect to be able to fit a complex model such as a neural network. This is unfortunately the case for many forecasting problems such as those represented in the Makridakis collection (described below). (5) Finally, for some forecasting problems, one only requires prediction a single time step in the future, while for others, multiple time step forecasts are required. This has implications for the method used to train the neural network.

Before looking at neural networks, we will briefly review conventional forecasting methods. Forecasting has mostly been done using one of two different classes of methods, depending on whether the data are seasonal or not. For monthly data, such as sales or unemployment levels, the seasonal variation is often removed by dividing the series by an index representing the historical seasonal variation. For example, dividing the unemployment rate for each month (perhaps averaged over several years) by the average annual unemployment rate gives an index that indicates monthly variations. This index will have an average value of one. Dividing the actual unemployment rate in a given month by the index for that month gives the seasonally adjusted unemployment rate, which shows overall trends after typical monthly variations are accounted for. A linear or exponential regression (i.e., fitting the data as a linear or exponential function of time), or some form of smoothing such as a moving average, can then be used to make predictions of the deseasonalized unemployment. Actual levels are then forecast by multiplying these base predictions by the index for the month being forecast (Makridakis, Wheelwright, and McGee, 1983).

In contrast, for many complex processes such as chemical plant production, robots, or stock prices, the best prediction of the near future is obtained by using an appropriately weighted combination of recent measurements of the variable being predicted and other correlated variables. The most widely used approach is the Autoregressive Moving Average (ARMA) model. For example, to predict the value of a variable y (such as a temperature or a pressure or a stock price) at time $t + 1$ using past values of y and of a second variable z , one would use a linear regression to fit a model of the form

$$y_{t+1} = c_0 + c_1 y_t + c_2 y_{t-1} + c_3 y_{t-2} + \dots + c_n z_t + c_{n+1} z_{t+1} + \dots \quad (1)$$

Note that ARMA models differ from the linear regression models mentioned above in that they are functions of previous variables rather than of time.

Neural networks can be used to learn a nonlinear generalization of ARMA models of the form

$$y_{t+1} = f(y_t, y_{t-1}, y_{t-2}, \dots, z_t, z_{t+1}, \dots) \quad (2)$$

When the process is nonlinear and sufficient data are available, the neural networks will provide a more accurate model than the linear ARMA model. See Box and Jenkins (1970) for extensive descriptions of conventional ARMA models and the Box-Jenkins modeling approach, which involves picking a model of the form of Equation 1 with some subset of the coefficients set to zero. Later in this article we summarize the results of a number of studies that compare ARMA and neural network models.

Two other modeling methods are also often used by engineers, Kalman filtering and Wiener-Volterra series. Kalman filters (see KALMAN FILTERING: NEURAL IMPLICATIONS) assume a known model structure in which the parameters and their covariance, which is modeled explicitly, may be changing over time. Kalman filters are good for modeling relatively simple but noisy processes, but, unlike neural networks, they do not form nonparametric models that can accurately forecast the behavior of nonlinear systems. Wiener-Volterra series are polynomial expansions fitted to past data. As such, they, like neural nets, can approximate arbitrary functions. However, for models with multiple inputs they require more data than neural networks to obtain an equal level of accuracy.

Using Neural Nets for Forecasting

Neural networks are most often used to fit ARMA-style models of raw time series data from one or more measurements, but they can also be used as a piece of larger forecasting systems, such as in combination with deseasonalizing (i.e., forecasting a time series from which the seasonal component has been removed, as described above). Even for the simpler ARMA-style models, attention to the method is required if one is making forecasts multiple time steps in the future rather than a single time step.

Direct Versus Recurrent Prediction

A simple form of multistep forecasting is direct prediction (Figure 1A), in which a network takes past values as inputs and has separate outputs for predictions one, two, and more time steps in the future. Alternatively, one can train a network to predict one time step in the future and then use the network recursively to make multistep predictions (Figure 1B). Such networks are sometimes called *externally recurrent networks*, in contrast to networks that have internal memory. Direct forecasting networks are easier to build than externally recurrent nets because they do not require unfolding in time (described below), but the predictions are generally less accurate, since they have more parameters that must be fit from the same limited data.

The obvious way to train a network such as is used in Figure 1B is to minimize the error on the one-time-step predictions. Unfortunately, this does not give optimal networks for multistep predictions. To better understand this somewhat confusing point, consider the case of a simple linear ARMA model:

$$y_{t+1} = c_0 + c_1 y_t + c_2 y_{t-1} \quad (3)$$

A two-step-ahead prediction would then take the form

$$y_{t+2} = c_0 + c_1(c_0 + c_1 y_t + c_2 y_{t-1}) + c_2 y_t \quad (4)$$

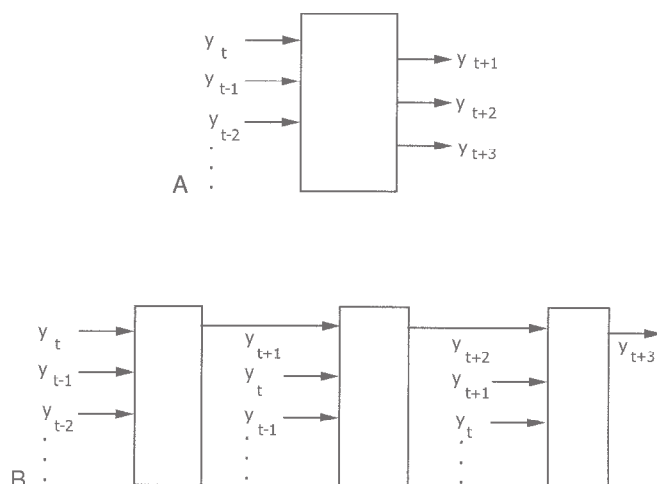


Figure 1. A, Direct prediction using a neural network. B, Recurrent one-step-ahead prediction using a neural network repeatedly.

Selecting coefficients c_0 , c_1 , and c_2 to minimize the prediction error for the one-step-ahead error yields a different equation than selecting the same coefficients to minimize the error in the two-step prediction. (Note that the former is a linear regression problem, whereas the latter requires nonlinear regression because the coefficients multiply each other.) More accurate long-range predictions are obtained by training to minimize the multistep prediction error. The solution using backpropagation uses the same unfolding in time or other solution methods as for internally recurrent networks (see RECURRENT NETWORKS: LEARNING ALGORITHMS). This and related issues are covered in detail in books on conventional system identification methods (e.g., Ljung and Torsten, 1983). Much good work has been done using recurrent nets to model time series (e.g., Mozer, 1994).

Combining Neural Networks with Other Methods

There are a number of ways in which neural networks can be combined with data preprocessing techniques, first principles (mechanistic) with partial models of the process being forecast, and with other forecasting techniques. Most commonly, if there is a strong seasonal component to the data, the data may be deseasonalized and the neural net used to forecast the basic trend. It may appear pointless to use a seasonal index when it is well known that neural networks can approximate arbitrary functions, which should include any seasonal variation. Experience indicates that if sufficient data are available, this is true, but that for shorter time series, deseasonalizing gives more accurate forecasts.

Similarly, when modeling complex physical systems, much better forecasts can be obtained with much less data when prior knowledge (e.g., in the form of mass, energy, or kinematic constraints on the variables, or in terms of monotonic relations between measured and forecast variables) is built into the network (Psychogios and Ungar, 1992). In a typical example, the equations governing a fermentation reactor are known except for the growth kinetics of the cells (e.g., yeast) in the reactor. If a neural network is used just to approximate the growth kinetics rather than to model the whole system, models are learned that are more accurate and that extrapolate better to operating regimens where no data are available. Such hybrid or "gray box" methods are popular in science and engineering.

Neural networks can also be used in conjunction with conventional forecasting methods. For example, one can often produce

more accurate forecasts by providing several conventional forecasts as input to the neural network. In this case, the network serves partly as a combining method in which the network produces a weighted average of the different forecasts (Foster, Collopy, and Ungar, 1992). Such combining of forecasts is widely practiced in the forecasting community, mostly with relatively arbitrary combining weights.

Assessing Neural Nets for Forecasting

There are several difficulties in assessing forecasting methods. The most serious is that the results of a single forecast tell little about whether the method will be superior for other forecasts. In testing any method, it is important to have a large set of representative time series on which the methods will be tested. An example of such a collection of time series that has been widely used to compare forecasting methods is the Makridakis competition, or M-competition, model (Makridakis et al., 1982). This competition included 1,001 series and evaluated 24 forecasting methods. The series were taken from a variety of organizations in a number of countries and included macroeconomic, microeconomic, industrial, and demographic data such as production levels, net sales, unemployment, spending, GNP, vital statistics, and infectious disease incidence. The series included yearly, quarterly, and monthly series, but no series arising from securities or commodities trading. These time series all involve only a single variable and do not provide correlated variables, which might enhance the predictions.

One must also decide which error criteria to use. The most obvious criterion, and the one that is optimized by standard neural networks, is minimization of the mean squared prediction error. This criterion has the property that a small number of unusual series may have a large effect on the error. In looking at combined errors for different time series, one must, of course, also normalize for the different magnitudes of the series. Thus, forecasters often measure performance by using measures that are more robust to outliers or atypical time series.

Three error measures that have proved particularly robust are the percentage of time a method had a lower absolute error than the "no-change" forecast (or "percent better"), the relative absolute error (or RAE), and the median absolute percent error (or mdAPE). The RAE is calculated as the geometric mean across all series i of

$$RAE_i = \frac{\sum_{t=1}^T |\tilde{x}(t) - x(t)|_i}{\sum_{t=1}^T |x(0) - x(t)|_i} \quad (5)$$

where $\tilde{x}(t)$ is the forecast and $x(t)$ represents the true value of the series at time t . The RAE represents a comparison over the forecast horizon T for series i of the absolute error of the forecast method, compared to the no-change or random walk forecast. One then calculates a geometric mean over all the series:

$$RAE = \left[\prod_{i=1}^n RAE_i \right]^{1/n} \quad (6)$$

The median average percent error is defined as the median across all series i of

$$APE_i = \frac{1}{T} \sum_{t=1}^T 100 \frac{|\tilde{x}(t) - x(t)|_i}{|x(t)|_i} \quad (7)$$

Good forecast performance is reflected in higher "percent betters" and lower RAEs and mdAPEs.

In assessing neural networks for forecasting, one must compare the accuracy of the neural networks with that of other statistical

tools such as exponential smoothing (for a single time series) or linear ARMA models (for several correlated time series). Surprisingly, many studies fail to compare neural network forecasts with well-made conventional forecasts.

Table 1 lists some applications in which neural networks have been used for forecasting. Almost all of the studies used standard backpropagation networks with less than a dozen inputs and less than a dozen hidden nodes, with the exact architecture being selected by trial and error. Also, most of the studies used data from a single source, and most of the authors evaluated their results on the basis of the mean squared error on out-of-sample forecasts (i.e., error when forecasting data other than that used for building the model). Table 1 does not include any studies using chaotic time series such as from the Mackey-Glass equation, which give little insight into neural network forecasts of realistic data. See Vemuri and Rogers (1994) for a good collection of reprints of a wide variety for neural network forecasting studies, including all studies cited in Table 1 that are not listed in the references. There is also an extensive literature on neural network forecasting for process control (see *PROCESS CONTROL* in the First Edition). Process control and robotics applications have seen some of the most successful use of neural networks for forecasting, as the processes involved are often sufficiently multivariable and nonlinear to warrant the use of neural networks but sufficiently well characterized and free of noise to allow accurate models to be built.

Dangers in Using Forecasts

Forecasts rely on a number of assumptions. They assume that the system that is modeled remains constant, i.e., that the model that held when the model was built still applies when the forecast is made. If the system structure is evolving over time, techniques from adaptive control may be more appropriate. It is also implicitly assumed when forecasting using neural networks with multiple inputs that the covariance structure of the inputs will remain constant. This presents a major difficulty when modeling systems that have

feedback in them, if the feedback structure is variable. For example, consider a house controlled by a thermostat. One will typically find that the heater will be on more often when the house is cold (this is, after all, what the heating system is designed to do). Forecasts of future house temperature can be accurately made using historical temperature measurements. If, however, these forecasts are used as part of the control scheme (the thermostat), then instability often results, since the forecasts fail to account for the new thermostat behavior. Similar situations often occur in economics and marketing, where forecasts can result in new laws being passed or in new prices being charged (and resulting actions by competitors), thus invalidating the original forecast. Unfortunately, there is generally little that one can do other than monitoring forecasts and distrusting them or collecting more data, if the process being forecast changes. (This is true in linear regression as well, where it is impossible to tell which of two highly correlated inputs is responsible for changes in an output, but at least one can easily detect the problem in linear problems by examining the uncertainty on the regression coefficients, whereas it is usually concealed in neural nets.)

Discussion

Neural networks have many demonstrated successes as forecasting tools and a smaller number of documented failures. All the usual warnings about model building apply. In particular, to build a good model, one needs good data. When the data are noisy and occur in short series, neural networks often fail to do better than simple forecasting techniques. For example, the 181 yearly series of the M-competition, which have a mean length of 19 data points on which to base a prediction, do not provide a good basis for complex nonlinear models. Neural networks generally give significant improvements over conventional forecasting methods when applied to monthly data in the M-competition set but not when applied to yearly data (Hill, O'Connor, and Remus, 1996). This is probably due to the high ratio of noise to data in the yearly data.

It may also be the case that the data are truly random or that the key independent variables are not being measured. Research suggests that this is true of the stock market (White, 1988). If this is true, then neural networks will not produce useful market forecasts, although they may help sell forecasting products. Several fund managers claim that they are getting superior predictions using neural networks, but for obvious competitive reasons, they generally do not provide enough information to test the claims. Moody (1998) provides a good discussion of the issues in forecasting the economy.

Neural networks have proved successful in a number of applications such as forecasting prices (Chakraborty et al., 1992), product demand (Chitra, 1993), electric utility loads (Yu, Moghaddamjoo, and Chen, 1992), and inventory levels (see Table 1). Such problems are characterized by ample measurements with a relatively high signal-to-noise ratio. In most cases, substantially better performance is obtained by using several related inputs to the network. For example, in forecasting wheat prices in three cities, superior performance was found by using recent wheat prices and measures of the local earning power. Similarly, in forecasting demand for polypropylene production, several macroeconomic variables were fed into the network. On longer, more deterministic time series, such as measuring the progress of a chemical reaction, neural networks have been shown to be a relatively accurate means of forecasting even chaotic series (Hudson et al., 1990; Lapedes and Farber in Vemuri and Rogers, 1994).

All of the applications cited above use standard backpropagation networks, occasionally with some degree of structure built into the network. For example, in the currency exchanges, excess weights were eliminated, while for forecasting wheat prices, past values of prices in three different cities were used to predict the logarithm

Table 1. Forecasting Using Neural Nets: Sample Results

Application	Authors	Results	Compared with
Car sales, airline passengers	Tang et al.	NNet better for longer-term forecast; Box-Jenkins better for shorter	Box-Jenkins
Currency exchange rates	Weigend et al.	NNets better	Random guessing
Electric load forecasting	Park et al.	NNet better	Currently used technology (unclear what)
Electrochemical reaction	Hudson et al.	Prediction looks good	—
Flour prices	Chakraborty et al.	NNets better than ARMA	ARMA
Polypropylene sales	Chitra	NNets slightly better than ARMA	ARMA
Stock prices	White	NNets provide no benefit	Random walk
Widely varied (Makridakis collection)	Foster et al.	NNets better on quarterly data, worse on annual data	Many exponential smoothing and deseasonalizing methods

of flour prices. All have demonstrated better performance than conventional forecasting methods, except when only short time series were available (10 to 30 data points) or when it was unclear if there was an underlying model other than a biased random walk (e.g., stock prices). However, the gains in accuracy over conventional forecasting methods are often relatively small, and overfitting is a common problem. Many companies are now using neural networks for problems such as demand forecasting. When sufficient data are available and care is taken to avoid overfitting, neural networks work well.

Road Map: Applications

Related Reading: Kalman Filtering: Neural Implications; Recurrent Networks: Learning Algorithms

References

- Box, G., and Jenkins, G., 1970, *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day. ♦
- Chakraborty, K., Mehrotra, K., Mohan, C. K., and Ranka, S., 1992, Forecasting the behavior of multivariate time series using neural networks, *Neural Netw.*, 5:961–970. ♦
- Chitra, S. P., 1993, Use neural networks for problem solving, *Chem. Eng. Prog.*, April, pp. 44–52. ♦
- Foster, B., Collopy, F., and Ungar, L. H., 1992, Neural network forecasting of short noisy time series, *Comput. Chem. Eng.*, 16:293–298. ♦
- Hill, T., O'Connor, M., and Remus, W., 1996, Neural network models for time series forecasts, *Manage. Sci.*, 42:1082–1092.
- Hudson, J. L., et al., 1990, Nonlinear signal processing and system identification: Applications to time series from electrochemical reactions, *Chem. Eng. Sci.*, 45:2075–2981.
- Ljung, L., and Torsten, S., 1983, *Theory and Practice of Recursive Identification*, Cambridge, MA: MIT Press.
- Makridakis, S., et al., 1982, The accuracy of extrapolation (time series) methods: Results of a forecasting competition, *J. Forecast.*, 1:111–153.
- Makridakis, S., Wheelwright, S., and McGee, V., 1983, *Forecasting: Methods and Applications*, New York: Wiley. ♦
- Moody, J., 1998, Forecasting the economy with neural nets: A survey of challenges and solutions, *Lecture Notes Comput. Sci.*, 1524:347–371.
- Mozer, M. C., 1994, Neural net architectures for temporal sequence processing, in *Time Series Prediction* (A. S. Weigend and N. A. Gershenfeld, Eds.), Menlo Park, CA: Addison-Wesley, pp. 243–264.
- Psychogios, D. C., and Ungar, L. H., 1992, A hybrid neural network: First principles approach to process modeling, *Am. Inst. Chem. Eng. J.*, 38:1499–1512.
- Vemuri, V. R., and Rogers, R. D., 1994, *Artificial Neural Networks: Forecasting Time Series*, Los Alamitos, CA: IEEE Computer Society Press. ♦
- White, H., 1988, *Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns*, in *Proceedings of the IEEE International Conference on Neural Networks*, San Diego, p. II-451.
- Yu, D. C., Moghaddamjo, A. R., and Chen, S.-T., 1992, Weather sensitive short-term load forecasting using a nonfully connected artificial neural network., *IEEE Trans. Power Syst.*, 7:1098–1105.

Gabor Wavelets and Statistical Pattern Recognition

John Daugman

Introduction

Starting around 1960, for about three decades investigation into the functioning of the mammalian primary visual cortex was dominated by recordings from single neurons. Using relatively simple stimuli such as oriented bars of light (e.g., Hubel and Wiesel, 1962, 1974), the apparent coding dimensions underlying spatial vision were mapped out by measuring tuning curves of individual neural responses as functions of stimulus parameters. Although methods later moved on, with innovations such as population recordings, noninvasive imaging with photovoltaic dyes, and novel anatomical techniques, the single-unit recording paradigm left a rich legacy of data that lent itself to modeling in engineering terms such as filtering, feature extraction, transform coding, and dimensionality reduction.

In this framework, the key functional concept is that of a neuron's *receptive field*, which specifies that region of two-dimensional (2D) visual space in which image events or structure can influence the neuron's activity. More exactly, the neuron's *receptive field profile* indicates the relative degree to which the cell is excited or inhibited by the distribution of light as a function of its spatial position within the receptive field. Through careful measurements with precisely defined stimuli, the receptive field profile of a *linear* neuron (one obeying proportionality and superposition in its responses to stimuli) reveals how it will respond to *any* pattern and allows the neuron to be analyzed in signal processing terms as a filter. The powerful mathematical tools of linear systems analysis (including Fourier analysis) are the basis of such extrapolations, subject always to the assumption of linearity. More recent findings of adaptive, nonlinear, remote interactions between visual neurons "beyond the classical receptive field" undermine the linear filter perspective and may even call into question the whole notion that

a neuron has a stable receptive field profile. Nevertheless, impressive practical results have been achieved in engineering applications of one such model inspired by the classical receptive field data. This article reviews the model that has come to dominate the classical description of cortical simple cells and their inputs to complex cells, and it reviews some successful applications of that scheme within computer vision and statistical pattern recognition.

Receptive Fields and 2D Gabor Wavelets

Typical two-dimensional receptive field profiles of simple cells in the feline visual cortex (Jones and Palmer, 1987) are shown in the top row of Figure 1. There are arguably five major degrees of freedom (i.e., independent forms of variation) spanned by the spatial receptive field structure of such neural populations. These can be regarded as defining the dimensions of the spatial visual code at this cortical level. The first two degrees of freedom are the *location* of a neuron's receptive field, defined by retinotopic coordinates (x , y). The third is the *size* of its receptive field (which can be described using a single scalar diameter, provided we view variation in the field width/length aspect ratio as a secondary population structure). The fourth is the *orientation* of the boundaries separating excitatory and inhibitory regions, as seen in Figures 1 and 3, normally also corresponding to the direction of receptive field elongation. The fifth is the *symmetry*, which may be even or odd, or some linear combination of these two canonical functions. (Any function can be decomposed into the sum of an even function plus an odd function, and their relative amplitudes define a continuum that allows this fifth dimension to be regarded as *phase*.)

These degrees of freedom in the spatial visual code also correspond to certain dimensions of the "cortical architecture" (rules of topographic and modular organization), although such structure is

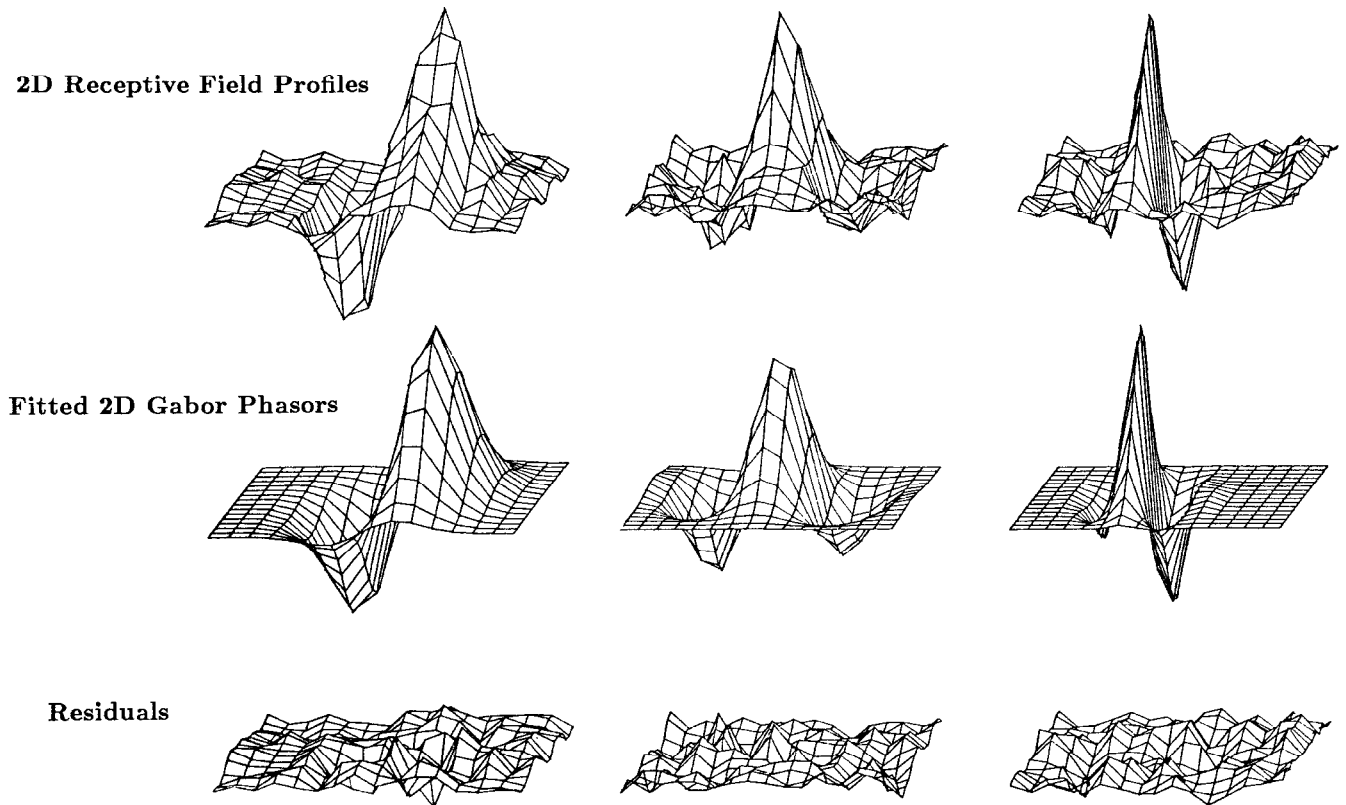


Figure 1. Typical 2D receptive field profiles of simple cells found in cat visual cortex, from measurements by Jones and Palmer (1987). The raw receptive field profiles (top row) are well-described by the 2D Gabor wave-

let model (middle row) in 97% of the cells studied, yielding residuals (bottom row) that are indistinguishable from random error in chi-squared tests.

less pronounced for some variables than for others. The (x, y) position coordinates of receptive fields in visual space form systematic (although nonconformal) topographic maps of these two dimensions across the cortical surface. Third, subpopulations of neurons that share the same orientation preference are grouped together into columns, and successive columns rotate systematically in preferred angle ("sequence regularity"). A similar structure exists for the grouping of cells by the dominant eye from which they receive input, the ocular dominance columns. These rather crystalline organizational principles were originally documented in seminal papers by Hubel and Wiesel (1962, 1974). Of the remaining two degrees of freedom, there is some evidence for pairwise grouping by quadrature (90°) phase relationship (Pollen and Ronner, 1981), and also some evidence for anatomical grouping by field size, either in different cortical layers or in adjacent columns analogous to those for orientation.

One benefit of identifying the primary degrees of freedom in a spatial image code is that it allows us to characterize the coding strategy in information-theoretic terms. The information-carrying capacity associated with each degree of freedom is a function of the number of individually resolvable states for that dimension. These define a kind of "information budget" that can be allocated in alternative ways among the different available degrees of freedom. Certain inescapable conflicts arise, however, that limit the extent to which some combinations of information can be simultaneously resolved. These conflicts take the form of an "uncertainty principle," whose mathematical form (Daugman, 1985) is just a 2D generalization of the one familiar from quantum physics in the famous work of Weyl and Heisenberg. One such conflict, or trade-off, will be intuitively clear from considering the oval receptive

fields in Figure 3. Orientation resolution (the "sharpness" of orientation tuning) would be enhanced by making the ovals longer, but this would reduce their resolution for spatial location in that direction. Similarly, increasing the field width by adding more cycles of undulation would sharpen the tuning for spatial frequency, but at the cost of lost resolution for spatial location in this direction. The optimal solution for these trade-offs, achieving maximal *conjoint* resolution of image information in both 2D spatial and 2D spectral terms, is the family of complex-valued 2D Gabor wavelets. These were first introduced into vision modeling by Daugman (1980, 1985) as a generalization of the 1D elementary functions, "logons," originally proposed for signal expansions by Gabor (1946).

This family of complex-valued 2D wavelets defining filters with minimal conjoint uncertainty have the following parameterized form in the (x, y) space domain:

$$G(x, y) = e^{-[(x-x_0)^2/\alpha^2 + (y-y_0)^2/\beta^2]} e^{-2\pi i[u_0(x-x_0) + v_0(y-y_0)]} \quad (1)$$

where (x_0, y_0) specify position in the image, (α, β) specify the filter's effective width and length, and (u_0, v_0) specify the filter's modulation wave vector, which can be interpreted in polar coordinates as spatial frequency $\omega_0 = \sqrt{u_0^2 + v_0^2}$ and orientation (or direction) $\theta_0 = \arctan(v_0/u_0)$. The real and imaginary parts of this complex filter function describe associated pairs of simple cells in "quadrature phase" (90° phase relation), as were discovered by Pollen and Ronner (1981). The middle row of Figure 1 shows three examples of the real or imaginary parts of the complex filter of Equation 1, with parameters chosen to fit the experimentally measured receptive field profiles shown in the top row. Neural record-

ings by Jones and Palmer (1987) confirmed that this family of functions provided good fits to the receptive field profiles of about 97% of the simple cells whose 2D profiles they measured in cat visual cortex. (It should be noted, however, that other investigators have preferred other functions, such as differences of several offset Gaussians, which, having additional fitting parameters, offered better fits to their data.) The top row of Figure 1 illustrates three of the 131 simple-cell 2D receptive field profiles measured by Jones and Palmer. The bottom row shows the residuals obtained by subtracting the best-fitting 2D Gabor wavelet component (middle row) from each measured profile. For nearly all of the cells studied, these residuals were indistinguishable from random error in chi-squared tests. Although alternative analytic forms could be chosen to fit the available 2D receptive field data, there can be no doubt that the 2D Gabor wavelet model specifies an efficient set of coding primitives capturing 2D spatial location, orientation, size (or frequency), and phase (or symmetry) in a natural way.

The 2D Fourier transform $F(u, v)$ of a 2D Gabor wavelet, which reveals its spectral response selectivity in the Fourier plane, has exactly the same functional form as the space-domain function (i.e., it is “self-Fourier”), but with the parameters just interchanged or inverted:

$$F(u, v) = e^{-[(u-u_0)^2\alpha^2 + (v-v_0)^2\beta^2]} e^{2\pi i[\lambda_0(u-u_0) + y_0(v-v_0)]} \quad (2)$$

Thus the 2D Fourier power spectrum $F(u, v)F^*(u, v)$ of a 2D Gabor wavelet is simply a bivariate Gaussian centered on (u_0, v_0) . Hence its peak response occurs for an orientation θ_0 and spatial frequency ω_0 as defined earlier, corresponding to the excitatory/inhibitory structure of the receptive field, as one would expect. Some authors have questioned the relevance of the Gabor wavelet property of optimal conjoint resolution in these two domains, or the specialness of the variance metric on which the measure of uncertainty is based. Perhaps the best reply is Aristotle’s dictum that “vision is knowing what is where.” The extraction of local image structure in terms of oriented undulatory primitives provides information in 2D spectral terms about “what,” and the resolution of positional information indicates “where.” If we wish to extract visual information simultaneously in terms of both what and where, as Aristotle said, then under the Heisenberg uncertainty principle we cannot do better than to construct our spatial visual code from 2D Gabor wavelets. It would appear that the evolution of the mammalian visual cortex may have been shaped by this criterion and thus converged on the coding primitives that optimize it.

Compact Image Coding and 2D Gabor Transforms

Besides their optimality in terms of the uncertainty relation, 2D Gabor wavelets have many practical properties. They can be used to form a complete and compact image code, as a self-similar 2D wavelet expansion basis, despite their nonorthogonality (Daugman, 1988). Although Gabor wavelets do not technically satisfy the original admissibility conditions for wavelets such as orthogonality and strictly compact support, their practical advantages as coding primitives are not much diminished. For example, they can achieve significant image compression, with appropriate parameterization for wavelet dilations, rotations, and translations. If we take $\Psi(x, y)$ to be a chosen generic 2D Gabor wavelet as specified above in Equation 1, which may be called a “mother wavelet,” then we can generate from this one function a complete self-similar family of “daughter wavelets” through the generating operation

$$\Psi_{mpq\theta}(x, y) = 2^{-2m}\Psi(x', y') \quad (3)$$

where the substituted variables (x', y') incorporate dilations of the wavelet in size by octave factors 2^{-m} , translations in position (p, q) , and rotations through angle θ :

$$x' = 2^{-m}[x \cos(\theta) + y \sin(\theta)] - p \quad (4)$$

$$y' = 2^{-m}[-x \sin(\theta) + y \cos(\theta)] - q \quad (5)$$

It is noteworthy that as consequences of the similarity, shift, and modulation theorems of 2D Fourier analysis, together with the rotation isomorphism of the 2D Fourier transform, all of these effects of the generating function (Equation 3) applied to a 2D Gabor mother wavelet $\Psi(x, y) = G(x, y)$ in generating the 2D Gabor daughter wavelets $\Psi_{mpq\theta}(x, y)$ will have just corresponding or reciprocal effects on the wavelet’s 2D Fourier transform $F(u, v)$ without any other change in functional form (Daugman, 1985). This family of 2D wavelets, and their 2D Fourier transforms, is each closed under the transformation groups of dilations, translations, rotations, and convolutions.

Any image can be represented completely in terms of such a basis of elementary expansion functions. An example of this in a progressive sequence is provided in Figure 2, showing the benchmark “Lena” image reconstructed from increasing numbers of 2D Gabor wavelets. It is interesting that even when only 100 or 500 wavelets are present and distributed across the entire image, already the primary facial features such as the eyes are discernible. Since facial features are essentially just localized undulations, parameterized for scale, position, orientation, and symmetry—that is, the same as the parameterizations of the Gabor wavelets themselves—it is perhaps not surprising that very efficient face codes can be constructed from such wavelets.

An error that occurs frequently in the literature is a confusion between Gabor *projection coefficients* (obtained merely by taking the convolution or inner product of each image region onto a local Gabor wavelet) and Gabor *expansion coefficients* (those needed to reconstruct the image as a linear combination of Gabor wavelets). Because these wavelets are not an orthogonal set (i.e., their mutual inner products are not zero), the expansion coefficients are not the same as the projection coefficients, nor can they easily be obtained from them. For this reason, it is incorrect to refer to the result of image convolutions with Gabor wavelets as a Gabor transform, since the resulting representation is not invertible. One approach for obtaining the Gabor expansion coefficients needed for an invertible image representation (thus defining a true *Gabor transform*) is a relaxation network method introduced in Daugman (1988). The progressive stages of image reconstruction shown in Figure 2 are based on expansion coefficients obtained by that relaxation network.

Interesting issues arise concerning how the “information budget” in a visual code should be allocated. For example, because all Gabor wavelets are indexed by their 2D location, the parameters that specify each wavelet’s orientation and spatial frequency can be sequenced much more sparsely than Fourier components in a Fourier transform. Moreover, the necessary density of sampling in orientation and frequency is in a trade-off with the needed density of sampling in position (i.e., how much the wavelets overlap each other). The exact rules for the sampling densities necessary in these various parameters in order to obtain a *complete* image code are dictated by *frame theory*. Whereas, for example, the 2D Fourier transform must sample the frequency plane along a uniform Cartesian grid, a 2D Gabor transform can sample the frequency plane on just a log-polar grid. This great reduction in sampling density for the higher frequencies is purchased by the wavelet position parameters. Illustrations of self-similar 2D Gabor representations of images, obtained with varying numbers of wavelet orientations (six, four, three, and two orientations in the sampling set), may be



Figure 2. Illustration of the completeness of 2D Gabor wavelets as image coding primitives. The benchmark “Lena” picture is reconstructed by progressive numbers of wavelets in linear combination, of 25, 100, 500, and 10,000. The primary facial features are effectively represented by just a handful of such wavelets. However, because of their nonorthogonality, the wavelets require coefficients that differ from the simple inner product projection of the image onto them.

found in Daugman (1988). These sorts of considerations may be able to answer such longstanding neurobiological questions as “Why are there orientation columns in the cortex, and why is orientation sampled with those bandwidths and intervals? Why do receptive fields overlap this much?” Further neurobiological issues are raised by the fact that the Gabor wavelets are nonorthogonal. The consequences of this include paradoxes in the classical interpretation of what it is that a neuron’s receptive field actually enables it to encode about an image. In particular, the classical view that a linear neuron’s response (which is determined by the inner product of its receptive field profile with the local image) signifies the “relative presence” of its own structure in the local image region is paradoxical: it implies an incorrect image representation by the ensemble of neurons, given their mutual nonorthogonality.

Facial Analysis and Recognition

Whereas simple cells are regarded as linear filters whose phase sensitivity is clearly determined by their alternating pattern of a few excitatory and inhibitory regions, the so-called “complex” cells that receive input from them have no such phase sensitivity, yet their orientation and spatial frequency tuning are similar to that of simple cells. A natural model therefore supposes that the inputs to complex cells come from quadrature pairs of simple cells, taking the sum of their squared responses, as shown at the top of Figure 3. This nonlinear combination not only achieves a weak kind of

translation invariance (in that the complex cell responds to the stimulus but is indifferent to its phase, or position within the receptive field) but, more important, this arrangement can play a useful role in feature extraction for pattern recognition. This idea is illustrated for the case of a face image in Figure 4 (see the figure caption for a detailed explanation). Since major facial features are essentially localized undulations of a certain scale (frequency) and orientation, it is perhaps not surprising that facial features are easily represented and detected by operations using 2D Gabor wavelets.

This idea has been elaborated further in a number of full face recognition systems based on Gabor wavelets (e.g., Lades et al., 1993). Besides encoding a face by the coefficients of projection of particular regions of the face (centered on fiducial points) onto clusters of multiscale, oriented Gabor wavelets (renamed “jets”), the Lades scheme organizes the data into an elastic graph that can accommodate some distortion. A graph-matching technique searches for matches of the Gabor wavelet projection coefficients while allowing graph distortions corresponding to limited changes in facial expression, perspective angle, and pose. But the approach remains “appearance-based” (i.e., it is a 2D image representation for faces, not a 3D object representation), and as such it is susceptible to changes in perspective geometry and illumination geometry. Such variations in image capture conditions affect the wavelet projection coefficients in a manner for which the graph matching cannot compensate and is not invariant. For similar reasons, all current face recognition algorithms can work only under con-

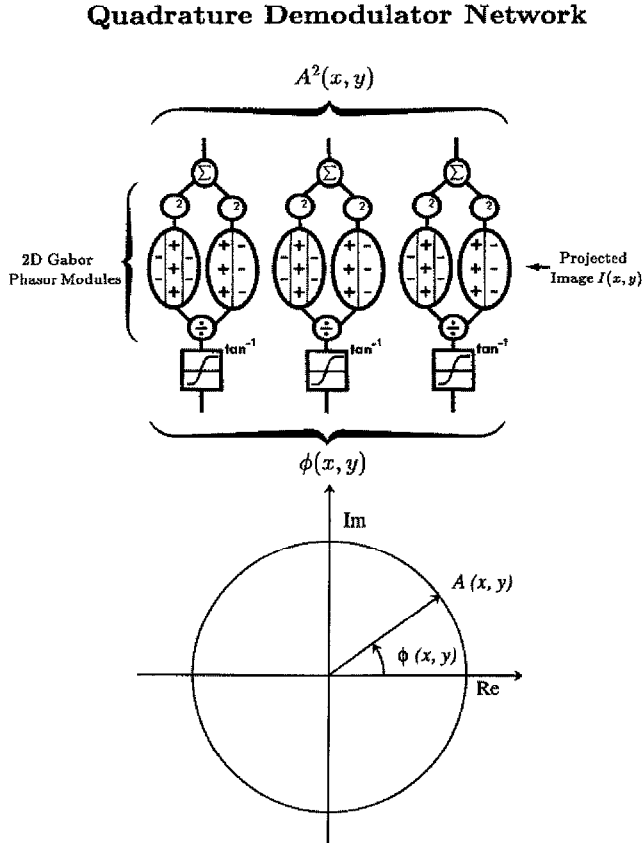


Figure 3. The 2D Gabor wavelet quadrature demodulation network. Even- and odd-symmetric receptive fields, of the kind associated with cortical simple cells, subserve a phasor resolution of information in the complex plane. The sum of the squares of quadrature simple-cell responses extracts an amplitude function $A(x, y)$ or modulus (top of network), while the ratio of their responses resolves the local phase function $\phi(x, y)$ (bottom of network). Such amplitude-and-phase descriptions of image structure can be very useful in computer vision.

strained imaging conditions, with fixed illumination and perspective geometry and relatively fixed expressions. Moreover, in realistic tests, the best systems have error rates approaching 50% when comparing images taken just 1 year apart.

Phase and Amplitude Coding of Texture and Complex Patterns

An important goal of vision is *dimensionality reduction*: creating a succinct and useful representation of image structure having much lower dimension than the raw image itself. In a sense, standard edge detection strategies are examples of this idea, since edge maps can signify object structure, and they clearly have much lower dimension than the raw pixel count. However, many naturally occurring objects, such as faces and bodies, lack the planar or geometrical forms of manufactured objects that generate simple edge maps; instead, they are defined by continuous-tone structure, textures, and undulations, which are not well captured by detecting edges. We saw earlier that facial features are efficiently represented and detected by 2D Gabor wavelets; a similar subspace projection approach using these wavelets was successfully applied by Shustorovich (1994) to the problem of classifying and recognizing handwritten characters. A more general representation for image infor-

mation (Daugman and Downing, 1995) reduces its dimensionality by *decorrelating* it not only in amplitude but also in phase, making use of the intriguing quadrature phase relationship found in the neurobiological recordings from cortical simple cells. As portrayed in Figure 3, the quadrature simple-cell structure not only supports an energy, or modulus, computation emerging from the top of the network as $A^2(x, y)$, but the same paired 2D Gabor receptive fields also support a computation of local phase $\phi(x, y)$, shown emerging from the bottom of the network. The arctangent-like “squashing function” that operates on the ratio of the simple-cell responses is a common feature of many neural network models, but here it serves trigonometrically to resolve a phase angle in the complex plane, as indicated in the phasor diagram at the bottom.

The representation of images in terms of local phase $\phi(x, y)$ and local amplitude $A(x, y)$ is a form of *predictive coding* that takes as its prediction the locally prevalent scale and orientation of image structure, and encodes the full detailed pattern as modulations of that prediction (Daugman and Downing, 1995). This lends itself not only to compact image coding but also to texture segmentation (i.e., the division of an image into regions defined by some local homogeneity of texture). The analysis of texture and its use for image segmentation are important topics in statistical pattern recognition. 2D Gabor wavelets have played dominant roles here, as reviewed in Bovik, Clark, and Geisler (1990) and Navarro, Taberner, and Cristobal (1996). An alternative approach that explicitly computes Gabor phase rather than just energy for effective texture segmentation is given in du Buf (1990).

Iris Recognition

In this section, we illustrate the principles discussed in this chapter with a practical application that is now coming into wide international use: the automatic visual recognition of persons by their iris patterns. Details about the algorithms that locate an iris and segment it from other tissues, mapping it into a doubly dimensionless coordinate system with invariance for size, translation, and pupil dilation, are given in Daugman (2001). The iris pattern is then demodulated by 2D Gabor wavelets (see Figure 3) in order to extract its *phase sequence* $\phi(x, y)$, with these phase values quantized very coarsely into only the nearest quadrant of the complex plane. This sets two bits of phase information for each wavelet applied in a particular location. An “IrisCode” comprising 2,048 such bits of pattern phase information is then compared against an enrolled database of other IrisCodes in search of a match. These comparisons are performed at the speed of 100,000 IrisCodes per second by the decision network shown in Figure 5. This network transforms the pattern recognition problem into a simple test of statistical independence on the 2D Gabor wavelet phase sequences derived from the patterns.

Results from 9.1 million comparisons between different iris patterns are given in Figure 6, based on images acquired at kiosks in Britain, the United States, Japan, and Korea in public trials of these algorithms over a 3-year period. In these trials, as well as in tests conducted by independent government laboratories (the largest involving 2.73 million iris comparisons), there has never been a single reported false match. The reason is because the 2D Gabor IrisCode extracts about 250 degrees of freedom, whose combinatorics generate binomial distributions with extremely rapidly attenuating tails. Since comparisons of phasor bits are Bernoulli trials whose values of p and q depend on whether a pair of IrisCodes comes from the same or from different eyes, the confidence levels associated with recognition decisions are determined by cumulatives of the binomial probability density of observing a fraction $x = m/N$ “true” exclusive-OR outcomes in N comparisons:

$$f(x) = \frac{N!}{m!(N-m)!} p^m q^{(N-m)} \quad (6)$$

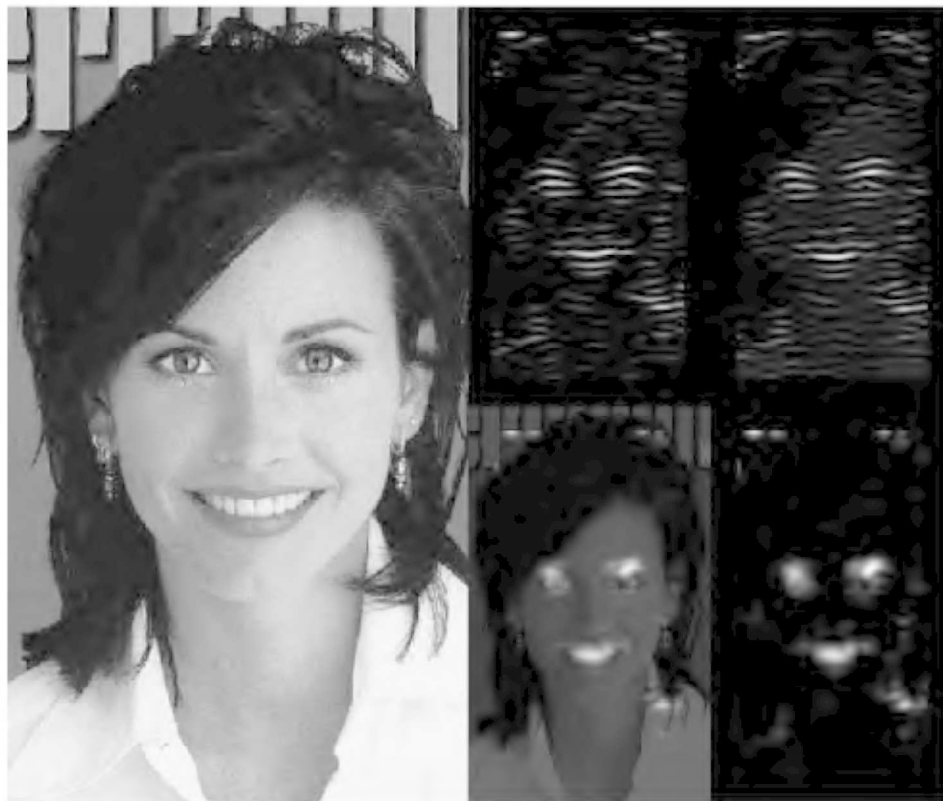


Figure 4. Illustration of facial feature detection by the quadrature demodulator network shown in Figure 3. *Left*, Input image. *Right* (clockwise from top left), the real part of the result of convolution with a 2D Gabor wavelet; the imaginary part from the same convolution (both of these representing the phase-sensitive simple-cell responses); the squared modulus $A^2(x, y)$, representing complex cell response; and this result superimposed on the original (faint) image, illustrating feature detection and localization.

The solid curve in Figure 6 superimposed on the raw data distribution is a plot of the Equation 6 binomial, and it provides a remarkably exact fit. It shows that there is vanishingly small prob-

ability that two different iris patterns could agree just by chance in more than about two-thirds of their bits, i.e., produce a fractional Hamming distance smaller than about 0.33. But images acquired

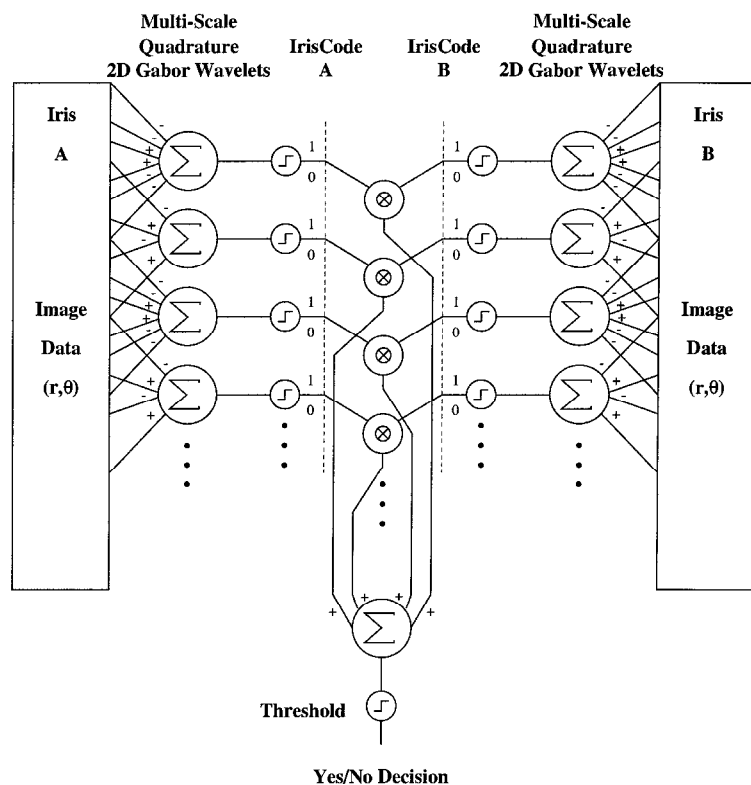


Figure 5. The comparison and recognition network used to make decisions about the identity of iris patterns. In effect, this network transforms the problem of pattern recognition into a test of statistical independence on the iris pattern phase sequences extracted by 2D Gabor wavelet demodulation (Figure 3).

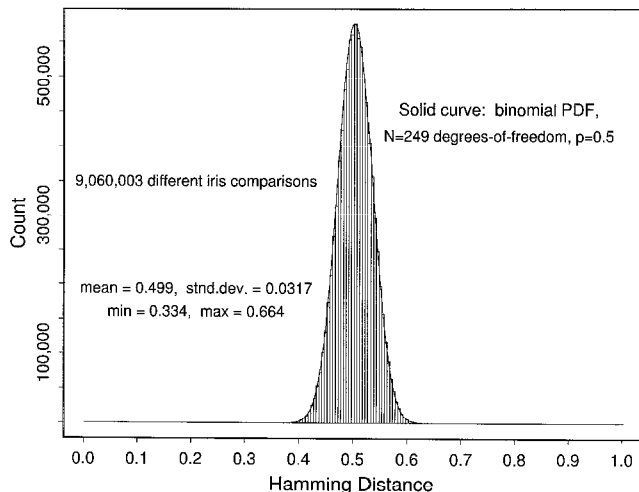


Figure 6. Results of 9.1 million comparisons between human iris patterns. Their Hamming distances are binomially distributed (solid curve, Equation 6). The rapidly decaying tails of such factorial distributions make it almost impossible for two different IrisCodes to disagree by chance in less than about a third of their bits (Hamming distance < 0.33). Thus, the failure of a simple test of statistical independence in this application of 2D Gabor wavelets allows reliable human identification with great tolerance for poor imaging.

from a given iris at different times and under different conditions score Hamming distances well below this, typically in the 0.10–0.15 range. Thus, this complex yet stable textural signature can provide a very accurate basis for automatically recognizing personal identity (in lieu of using PINs, cards, keys, passwords, or documents) for purposes such as border control, building entry, cash machines, computer login, authentication, and security measures in general. All current iris recognition systems installed worldwide use the 2D Gabor wavelet encoding, demodulation, and decision networks described here (Figures 3 and 5). Recent installations of this system include Heathrow Airport, Amsterdam-Schiphol, Washington-Dulles, and Charlotte Airports, for both passenger screening and control of access to restricted areas.

Discussion

The role of 2D Gabor wavelets in the visual sciences began as a model proposed in 1980 for cortical simple-cell 2D receptive field profiles. Today these wavelets are used pervasively in computer vision, image processing, and pattern recognition (for an in-depth review, see Navarro et al., 1996), even though more recent investigations in neuroscience perhaps call into question the very idea that visual neurons even possess stable receptive field profiles. The benefits of performing image coding and analysis using these elementary detectors include the opportunity to describe image struc-

ture in terms of local phase and energy, allowing demodulation, which lends itself well to solving pattern recognition problems. Some practical applications now in widespread use include facial and texture analysis, and personal identification by automatic, real-time recognition of iris patterns. These examples illustrate the fruitful interaction that can occur between ideas originating in brain theory and ideas about artificial neural networks.

Road Map: Vision

Related Reading: Face Recognition: Neurophysiology and Neural Technology; Feature Analysis; Orientation Selectivity

References

- Bovic, A. C., Clark, M., and Geisler, W. S., 1990, Multi-channel texture analysis using localized spatial filters, *IEEE Trans. Pattern Anal. Machine Intell.*, 12:55–73.
- Daugman, J. G., 1980, Two-dimensional spectral analysis of cortical receptive field profiles, *Vision Res.*, 20:847–856.
- Daugman, J. G., 1985, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *J. Opt. Soc. Am. A*, 2:1160–1169. See also: Daugman, J. G., 1993, Quadrature-phase simple-cell pairs are appropriately described in complex analytic form, *J. Opt. Soc. Am. A*, 10:375–377.
- Daugman, J. G., 1988, Complete discrete 2D Gabor transforms by neural networks for image analysis and compression, *IEEE Trans. Acoust. Speech Sign. Process.*, 36:1169–1179.
- Daugman, J. G., 2001, Statistical richness of visual phase information: Update on recognizing persons by iris patterns, *Int. J. Comput. Vision*, 45:25–38.
- Daugman, J. G., and Downing, C. J., 1995, Demodulation, predictive coding, and spatial vision, *J. Opt. Soc. Am. A*, 12:641–660.
- du Buf, J. M. H., 1990, Gabor phase in texture discrimination, *Sign. Process.*, 21:221–240.
- Gabor, D., 1946, Theory of communication, *J. Inst. Electr. Eng.*, 93:429–457.
- Hubel, D. G., and Wiesel, T. N., 1962, Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex, *J. Physiol. (Lond.)*, 160:106–154.
- Hubel, D. G., and Wiesel, T. N., 1974, Sequence regularity and geometry of orientation columns in the monkey striate cortex, *J. Comp. Neurol.*, 158:267–293.
- Jones, J. P., and Palmer, L. A., 1987, An evaluation of the 2D Gabor filter model of simple receptive fields in cat striate cortex, *J. Neurophysiol.*, 58:1233–1258.
- Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W., 1993, Distortion invariant object recognition in the dynamic link architecture, *IEEE Trans. Comput.*, 42:300–311.
- Navarro, R., Tabernero, A., and Cristobal, G., 1996, Image representation with Gabor wavelets and its applications, *Adv. Imaging Electron Phys.*, 97:1–84. ♦
- Pollen, D. A., and Ronner, S. F., 1981, Phase relationships between adjacent simple cells in the visual cortex, *Science*, 212:1409–1411.
- Shustorovich, A., 1994, A subspace projection approach to feature extraction: The 2D Gabor transform for character recognition, *Neural Netw.*, 7:1295–1301.

Gait Transitions

James J. Collins

Introduction

Legged animals typically employ multiple gaits for terrestrial locomotion. Bipedal, for example, walk, run, and hop, whereas quad-

rupeds commonly walk, trot, and bound. Animals make transitions between different gaits depending on their speed and the terrain. Experimental studies have demonstrated that animal locomotion is controlled, in part, by a central pattern generator (CPG), which is

a network of neurons in the central nervous system (CNS) capable of producing rhythmic output (Shik and Orlovsky, 1976; Grillner, 1981; Pearson, 1993). (The control of locomotion, however, is not purely central; e.g., the output of a locomotor CPG is modulated by feedback from the periphery.) Shik and colleagues, for instance, showed that mesencephalic cats could exhibit a walking gait on a treadmill when the midbrain was electrically stimulated. Moreover, they found that such preparations could switch between different gaits if either the stimulation strength or the treadmill speed was varied.

Although the aforementioned studies established the existence of rhythm-generating networks in the CNS, a vertebrate CPG for legged locomotion remains to be identified or isolated. As a result, little is known about the specific characteristics of the neurons and interconnections making up such systems. Consequently, researchers have resorted to using modeling techniques to gain insight into the possible functional organization of these networks. The most popular approach has involved the analysis of systems of coupled oscillators. Coupled-oscillator models have been used to control the gaits of bipeds (Bay and Hemami, 1987; Taga, Yamaguchi, and Shimizu, 1991), quadrupeds (Stafford and Barnwell, 1985; Schöner, Jiang, and Kelso, 1990; Collins and Stewart, 1993a; Collins and Richmond, 1994), and hexapods (Beer, 1990; Collins and Stewart, 1993b).

The neural mechanisms underlying gait changes are not well understood. A key question in this regard is whether gait transitions involve (1) switching between different CPGs, or (2) bifurcations of activity in a single CPG. In this article, we discuss a number of modeling approaches that have been developed to explore the feasibility of using either one or the other of these mechanisms to generate gait transitions in coupled-oscillator networks.

A Neuromodulatory Approach

As a model for legged-locomotion control, Grillner (1981) proposed that each limb of an animal is governed by a separate CPG, and that interlimb coordination is achieved through the actions of interneurons that couple together these CPGs. Within this scheme, gait transitions are produced by switching between different sets of coordinating interneurons; that is, a locomotor CPG is reconfigured to produce different gaits.

Grillner's proposed strategy has been adopted, in spirit, by several CPG modeling studies. Stafford and Barnwell (1985), for example, used a similar approach in a study of quadrupedal locomotion. They considered a CPG model that was composed of four coupled networks of oscillators. Each network controlled the muscle activities of a limb of a model quadruped. Stafford and Barnwell showed that this model could produce the walk, trot, and bound. In addition, they demonstrated that the walk-to-trot and walk-to-bound transitions could be generated by changing the relative strength of certain interoscillator connections or by eliminating others altogether. (Transitions in the reverse direction, e.g., bound-to-walk, were not reported.) Along similar lines, Bay and Hemami (1987) used a CPG network of four coupled van der Pol oscillators to control the movements of a segmented biped. Each limb of the biped was composed of two links, and each oscillator controlled the movement of a single link. Bipedal walking and hopping were simulated by using the oscillators' output to determine the angular positions of the respective links. Transitions between out-of-phase and in-phase gaits were generated by changing the nature of the interoscillator coupling; for example, the polarities of the network interconnections were reversed to produce the walk-to-hop transition.

This approach is, in principle, physiologically reasonable. For instance, the notion that supraspinal centers may call on functionally distinct sets of coordinating interneurons to generate different

gaits is plausible but not yet experimentally established. In addition, from a different but relevant perspective, it has been shown that rhythm-generating neuronal networks can be modulated—reconfigured—through the actions of neuroamines and peptides, and that they are thereby enabled to produce several different motor patterns (see Pearson, 1993, and CRUSTACEAN STOMATOGASTRIC SYSTEM), at least in invertebrate preparations.

A Synergetic Approach

Synergetics deals with COOPERATIVE PHENOMENA (q.v.) in non-equilibrium systems (Haken, Kelso, and Bunz, 1985). In synergetics, the macroscopic behavior of a complex system is characterized by a small number of collective variables, which in turn govern the qualitative behavior of the system's components.

Schöner et al. (1990) used a synergetic approach in a study of quadrupedal locomotion. They analyzed a network model that was made up of four coupled oscillators. Each oscillator represented a limb of a model quadruped. Three relative phases—the phase differences between the right-front and the left-front, left-hind, and right-hind oscillators, respectively—were used as collective variables to characterize the system's interlimb-coordination patterns. Gait transitions were modeled as nonequilibrium phase transitions, which, in this case, could also be interpreted as bifurcations in a dynamical system (see the next section). Schöner et al. demonstrated that various four-component networks could produce and switch (abruptly or gradually) between different gaits, such as the gallop, trot, and pace, if the coupling terms that operated on the relative phases were varied. Importantly, this work predicted that gait transitions should be accompanied by loss of stability; that is, signs of instability, such as spontaneous gait transitions, should arise near a switching point. Phenomena of this sort have been observed experimentally; the decerebrate cats in the Shik and Orlovsky study, for example, could, near the trot-gallop transition point, switch back and forth spontaneously between the trot and gallop.

This approach is significant in that it relates system parameter changes and stability issues to gait transitions. Its primary weakness, however, is that the physiological relevance of the aforementioned relative-phase coupling terms is unclear. This remains an open issue.

A Group-Theoretic Approach

The traditional approach for modeling a locomotor CPG has been to set up and analyze, either analytically or numerically, the parameter-dependent dynamics of a hypothesized neural circuit. Collins and Stewart (1993a, 1993b), however, approached this problem from the perspective of group theory. Specifically, they considered various networks of symmetrically coupled nonlinear oscillators and examined how the symmetry of the respective systems leads to a general class of phase-locked oscillation patterns. Within this approach, the onset of a given pattern is modeled as a symmetric Hopf bifurcation, and transitions between different patterns are modeled as symmetry-breaking bifurcations of various kinds. In standard Hopf bifurcation, the dynamics of a nonlinear system change as some parameter is varied and a stable steady state becomes unstable, "throwing off" a limit cycle (or periodic solution). At a symmetric analog of a Hopf bifurcation, which is appropriate for symmetric dynamical systems, one or more periodic solutions, usually several, bifurcate. There may also be secondary branches of solutions and other more complicated bifurcations. Successive bifurcations tend to break more and more symmetry; i.e., they lead to states with less and less symmetry. Importantly, the pattern of bifurcations that can occur and the nature of the periodic states that arise through such bifurcations are controlled primarily by the symmetries of the system.

The theory of symmetric Hopf bifurcation thus predicts that symmetric oscillator networks with invariant structure can sustain multiple patterns of rhythmic activity. From the standpoint of CPGs, this prediction challenges the notion that a network's coupling architecture needs to be altered to produce different oscillation patterns. Importantly, the symmetry-breaking analysis is independent of the details of the oscillators' intrinsic dynamics and the inter-oscillator coupling. (The production of periodic states through symmetric Hopf bifurcation, however, does depend on the variation of some suitable system parameter.) This approach thus provides a framework for distinguishing model-independent features (attributable to symmetry alone) from model-dependent features.

Collins and Stewart used this approach to study the dynamics of symmetric networks of two, four, and six coupled oscillators. These networks were considered as models for bipedal, quadrupedal, and hexapodal locomotor CPGs, respectively. They demonstrated that many of the generic phase-locked oscillation patterns for these models correspond to animal gaits. They also showed that transitions between these gaits could be modeled as symmetry-breaking bifurcations occurring in such systems. These studies led to natural hierarchies of gaits, ordered by symmetry, and to natural sequences of gait bifurcations (Figure 1). This work thus related observed gaits and gait transitions to the organizational structures of the underlying CPGs.

This approach is significant in that it provides a novel mechanism for generating gait transitions in locomotor CPGs. Its primary disadvantage, however, is that its model-independent features cannot provide information about the internal dynamics of individual oscillators. In particular, the stability of the predicted gait patterns and the conditions under which one is selected over another depend on the specific parameters of the model under investigation.

A Hardwired Network Approach

Motivated by the predictions of the above group-theoretic approach, Collins and Richmond (1994) conducted a series of computer experiments with a symmetric, hardwired locomotor CPG model that consisted of four coupled oscillators. They demonstrated that it was possible for such a network to produce multiple phase-locked oscillation patterns that correspond to three quadrupedal gaits: the walk, the trot, and the bound. Transitions between the different gaits were generated by varying the driving signal or by altering internal oscillator parameters. As observed in real animals (Alexander, 1989), transitions between the walk and trot, which were generated by varying the intrinsic frequency of the CPG oscillators and the amplitude of the driving signal, could be either gradual or abrupt, depending on the nature of the parameter variation. Similar parameter changes could also shift the CPG

model from either the walk or the trot into the bound. However, once the CPG model was in bound, it maintained that gait even if the system parameters were returned to their original values for either walk or trot, i.e., there was "total" hysteresis in the network's dynamics. To produce transitions from bound, it was necessary to subject two of the CPG oscillators to an increased driving stimulus before the system parameters were changed to those of the desired gait. (Experimental data that indirectly support such a strategy for generating transitions from bound were provided by Afelt, Blaszczyk, and Dobrzecka, 1983. Specifically, they found that the initiation of the gallop-to-trot transition in dogs was characterized by kinematic changes in a *single pair* of diagonal limbs.) Importantly, the above *in numero* results were obtained without changing the relative strengths or polarities of the system's interconnections; i.e., the network maintained an invariant coupling architecture. Collins and Richmond (1994) also showed that the ability of the hardwired CPG network to produce and switch between multiple gaits was, in essence, a model-independent phenomenon: three different oscillator models—the Stein neuronal model, the van der Pol oscillator, and the FitzHugh-Nagumo model—and two different coupling schemes were incorporated into the network without impeding its ability to produce the three gaits and the aforementioned gait transitions. This general finding was likely attributable to the symmetry of the network, which was maintained in all the numerical experiments.

Earlier, Beer (1990) had designed a hardwired CPG network for controlling hexapodal locomotion (see LOCOMOTION, INVERTEBRATE). In Beer's model, each leg of a model cockroach was controlled by a circuit made up of one pacemaker neuron, two sensory neurons, and three motor neurons. The pacemaker neurons of adjacent leg-controller circuits inhibited one another. If the pacemaker neurons of the network were identical, then the model could generate the tripod gait. To produce metachronal-wave gaits (in which waves of leg movements sweep from the back of the animal to the front), Beer varied the intrinsic frequencies of the pacemaker neurons such that the natural frequency of the back-leg pacemakers was lower than that of the middle-leg pacemakers, which was lower than that of the front-leg pacemakers. With this arrangement, the progression speed of the model cockroach could be changed, and transitions between different gaits could be produced by varying the tonic level of activity of a single command neuron, which was connected to every leg-controller circuit. This model's ability to generate and switch between different gaits was a direct consequence of the interactions between its coupled pacemaker neurons and their respective central and afferent inputs.

A similar model, made up of six coupled unit oscillators, was developed by Taga et al. (1991) to control bipedal locomotion. In this case, each unit oscillator controlled a single joint, i.e., an ankle, knee, or hip, of a multi-link biped. As with Beer's model, the CPG network was driven by a tonic activation signal, and each unit oscillator received feedback about the state of the system's limbs. With this arrangement, the biped's speed could be changed, and abrupt transitions between walking and running could be generated by varying the amplitude of the network's activation signal. Interestingly, these gait transitions exhibited hysteresis; i.e., the walk-to-run transition occurred at a faster progression speed than did the reverse transition. Similar hysteretic behavior has been observed in humans (Alexander, 1989). Taga et al., unfortunately, did not report on the model's ability to switch between out-of-phase gaits (i.e., walking and running) and in-phase gaits (i.e., hopping).

In these studies, gait transitions were produced by varying the CPG's driving signal. From a physiological standpoint, this pattern-switching mechanism is reasonable; e.g., experimental studies have shown that the output of a locomotor CPG can be modified by changes to its descending inputs. Nonetheless, it is important to note that the exact form of the driving signal or signals

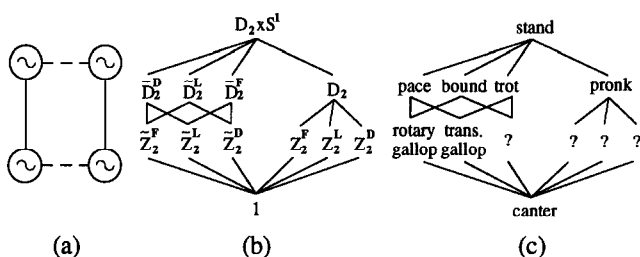


Figure 1. A, A rectangularly symmetric network of four coupled oscillators. The solid and dashed lines represent two forms of coupling. B, Patterns of symmetry breaking for the network in A. The respective group-theoretic symbols are described in Collins and Stewart (1993a). C, Quadrupedal gaits corresponding to the patterns in B.

acting on a locomotor CPG is unknown. Similarly, it is unclear how externally applied stimulation signals are transmitted to locomotor CPGs. For instance, although the stimulation signal in the Shik and Orlovsky study was amplitude modulated (to produce gait transitions), this does not necessarily mean that the resulting descending signals were also amplitude modulated. In addition, although the results of the Shik study were largely independent of the stimulation frequency, there is evidence that frequency-modulated stimulation signals can affect the output of locomotor CPGs. Lennard and Stein (1977), for example, electrically stimulated the dorsolateral funiculus in spinal and intact turtles and found that an increase in the stimulus frequency resulted in an increased repetition rate of hindlimb swimming movements. Finally, it should be reiterated that it is most likely erroneous to assume (as it has been in several CPG modeling studies) that the net driving signal of a locomotor CPG consists only of descending influences from supraspinal centers. The results from several experimental studies indicate that a CPG “driving” signal may also consist of afferent inputs from peripheral sensory organs (Pearson, 1993).

Discussion

The discussed modeling studies fall into two camps: (1) gait transitions are produced by changing the relative strength or polarity of the interoscillator coupling in a CPG; i.e., “different” CPGs are used to produce different gaits; or (2) gait transitions are generated by changing the CPG’s driving signal; i.e., bifurcations in a single CPG are used to generate different gaits. Both of these pattern-switching mechanisms are physiologically plausible, and they each lead to realistic locomotor patterns; e.g., in most of these studies, the stepping frequency or progression speed of the model animal increased when the CPG network switched to “faster” gaits. However, for a consistent theory of gait transitions to emerge, additional experimental data about the functional organization and operation of locomotor CPGs will have to be obtained. In particular, work is needed: (1) to determine whether a locomotor CPG uses functionally distinct sets of coordinating interneurons to produce different motor patterns, (2) to establish the extent to which neuromodulatory mechanisms are employed in vertebrate motor systems, and (3) to clarify the nature of the peripheral and descending inputs that influence the output of a locomotor CPG. Further experimentation is also needed to examine the possible role of bifurcation in gait transitions. In this regard, future investigations should explore the extent of hysteresis in gait transitions and the occurrence of increased instabilities near switching points, as well as consider more

extensively the effects of system-parameter variation on gait-transition dynamics.

[Reprinted from the First Edition]

Road Map: Motor Pattern Generators

Background: I.3. Dynamics and Adaptation in Neural Networks

Related Reading: Dynamics and Bifurcation in Neural Nets; Locomotion, Vertebrate; Spinal Cord of Lamprey: Generation of Locomotor Patterns

References

- Afelt, Z., Blaszczyk, J., and Dobrzecka, C., 1983, Speed control in animal locomotion: Transitions between symmetrical and nonsymmetrical gaits in the dog, *Acta Neurobiol. Exp.*, 43:235–250.
- Alexander, R. McN., 1989, Optimization and gaits in the locomotion of vertebrates, *Phys. Rev.*, 69:1199–1227. ♦
- Bay, J. S., and Hemami, H., 1987, Modeling of a neural pattern generator with coupled nonlinear oscillators, *IEEE Trans. Biomed. Eng.*, 34:297–306.
- Beer, R. D., 1990, *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology*. San Diego: Academic Press.
- Collins, J. J., and Richmond, S. A., 1994, Hard-wired central pattern generators for quadrupedal locomotion, *Biol. Cybern.*, 71:375–385.
- Collins, J. J., and Stewart, I. N., 1993a, Coupled nonlinear oscillators and the symmetries of animal gaits, *J. Nonlin. Sci.*, 3:349–392. ♦
- Collins, J. J., and Stewart, I., 1993b, Hexapodal gaits and coupled nonlinear oscillator models, *Biol. Cybern.*, 68:287–298.
- Grillner, S., 1981, Control of locomotion in bipeds, tetrapods and fish, in *The Handbook of Physiology*, section 1: *The Nervous System*, vol. II, *Motor Control* (V. B. Brooks, Ed.), Bethesda, MD: American Physiological Society, pp. 1179–1236.
- Haken, H., Kelso, J. A. S., and Bunz, H., 1985, A theoretical model of phase transitions in human hand movements, *Biol. Cybern.*, 51:347–356.
- Lennard, P. R., and Stein, P. S. G., 1977, Swimming movements elicited by electrical stimulation of turtle spinal cord: I. Low-spinal and intact preparations, *J. Neurophysiol.*, 40:768–778.
- Pearson, K. G., 1993, Common principles of motor control in vertebrates and invertebrates, *Annu. Rev. Neurosci.*, 16:265–297. ♦
- Schöner, G., Jiang, W. Y., and Kelso, J. A. S., 1990, A synergetic theory of quadrupedal gaits and gait transitions, *J. Theoret. Biol.*, 142:359–391.
- Shik, M. L., and Orlovsky, G. N., 1976, Neurophysiology of locomotor automatism, *Phys. Rev.*, 56:465–501.
- Stafford, F. S., and Barnwell, G. M., 1985, Mathematical models of central pattern generators in locomotion: III. Interlimb model for the cat, *J. Motor Behav.*, 17:60–76.
- Taga, G., Yamaguchi, Y., and Shimizu, H., 1991, Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment, *Biol. Cybern.*, 65:147–159.

Gaussian Processes

Chris K. I. Williams

Introduction

Much of the work in the field of artificial neural networks concerns the problem of supervised learning. Here we may be interested in regression problems (by which we mean the prediction of some real-valued variable(s)), or classification problems (predicting a class label) given the values of some input variables. Due to factors such as measurement noise, it is necessary to take a statistical view of the learning problem.

Given (possibly noisy) observations of a function at n points, it is necessary to impose extra assumptions about the function if there is to be hope of predicting its value elsewhere. Here we take a

Bayesian approach, placing a prior probability distribution over possible functions and then letting the observed data “sculpt” this prior into a posterior using the available data. The Bayesian approach can provide solutions to several problems, such as local optima in weight space, the setting of regularization parameters, overfitting, and model selection (see MacKay, 1992; Neal, 1996; and BAYESIAN METHODS AND NEURAL NETWORKS).

One can place a prior distribution $P(\mathbf{w})$ on the weights \mathbf{w} of a neural network to induce a prior over functions $P(y(\mathbf{x}; \mathbf{w}))$ but the computations required to make predictions are not easy, owing to the nonlinearities in the system, and one needs to resort to analytic approximations or Monte Carlo methods. Gaussian processes are

a way of specifying a prior directly over function space; it is often simpler to do this than to work with priors over parameters. Gaussian processes (GPs) are probably the simplest kind of function space prior that one can consider, being a generalization of finite-dimensional Gaussian distributions over vectors.

A finite-dimensional Gaussian distribution is defined by a mean vector and a covariance matrix. A GP is defined by a *mean function* (which we shall usually take to be identically zero), and a *covariance function* $C(\mathbf{x}, \mathbf{x}')$, which indicates how correlated the value of the function y is at \mathbf{x} and \mathbf{x}' . This function encodes our assumptions about the problem (for example, that the function is smooth and continuous) and will influence the quality of the predictions.

Gaussian process prediction is illustrated in Figure 1. The upper panel shows a sample of five functions drawn from the prior. The lower panel shows five samples from the posterior after two observations have been made; notice that the posterior is tightly constrained near the observations, but varies more widely further away. Essentially, what has happened is that prior samples not consistent with the observations have been eliminated. The crucial computational point is that it is not necessary to draw samples to make predictions; for regression problems, only linear algebra is required. Below we give more detail on this computation, discuss how to use GPs for classification problems, and describe how data can be used to adapt the covariance function to the given prediction problem.

Further discussion of Gaussian processes is available in Schölkopf and Smola (2001), MacKay (1998), and Williams (1998).

Gaussian Processes

Formal Definition

A stochastic process is a collection of random variables $\{Y(\mathbf{x}) | \mathbf{x} \in X\}$ indexed by a set X . In our case X will often be \mathbb{R}^d , where d is the number of inputs. The stochastic process is specified by giving

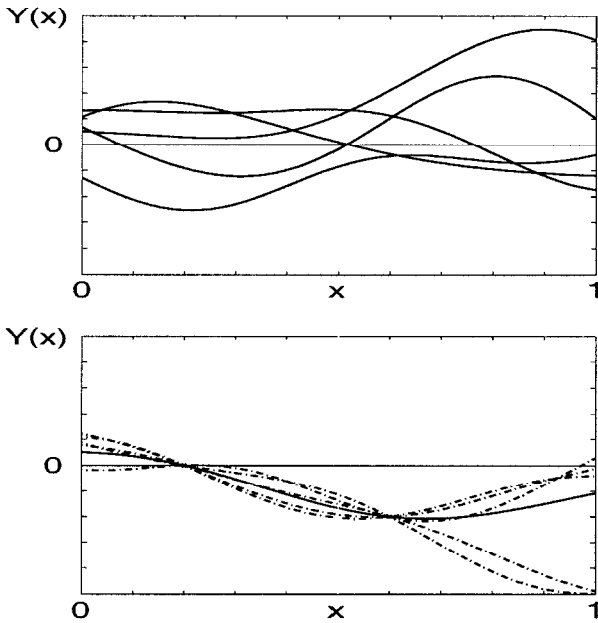


Figure 1. *Top*, Five samples from a Gaussian process prior. *Bottom*, Five samples from the Gaussian process posterior (shown as dot-dash lines) and the posterior mean (solid line), after observing the data points $(0.2, 0)$ and $(0.6, -1)$.

the joint probability distribution for every finite subset of variables $Y(\mathbf{x}_1), \dots, Y(\mathbf{x}_k)$ in a consistent manner. A Gaussian process (GP) is a stochastic process for which any finite set of Y -variables has a joint multivariate Gaussian distribution. A GP is fully specified by its mean function $\mu(\mathbf{x}) = E[Y(\mathbf{x})]$ and its covariance function $C(\mathbf{x}, \mathbf{x}') = E[(Y(\mathbf{x}) - \mu(\mathbf{x}))(Y(\mathbf{x}') - \mu(\mathbf{x}'))]$. For a multidimensional input space, a Gaussian process may also be called a Gaussian random field.

Below we consider Gaussian processes that have $\mu(\mathbf{x}) \equiv 0$. A non-zero $\mu(\mathbf{x})$ can be incorporated into the framework at the expense of a little extra complexity.

Example: Bayesian linear regression. Consider the model $y(\mathbf{x}) = \sum_{i=1}^m w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$, where $\{\phi_i\}$ is a set of fixed basis functions and \mathbf{w} is a vector of “weights.” Let \mathbf{w} have a Gaussian distribution with mean $\mathbf{0}$ and covariance Σ . Then $\mu(\mathbf{x}) = E[y(\mathbf{x})] = E[\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})] = 0$ as $E[\mathbf{w}] = \mathbf{0}$. As the mean is zero we have that $C(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}^T(\mathbf{x}) E[\mathbf{w} \mathbf{w}^T] \boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}^T(\mathbf{x}) \Sigma \boldsymbol{\phi}(\mathbf{x}')$. For example, using basis functions 1 and the components of \mathbf{x} along with $\Sigma = I$ gives $C(\mathbf{x}, \mathbf{x}') = 1 + \mathbf{x} \mathbf{x}'$.

In the case of a finite dimensional model we can make predictions using calculations in the parameter space (of dimension m), or a GP prediction (which is n -dimensional, where n is the number of data points). For $m < n$ the parameter space method is preferable, but for many useful covariance functions (see, e.g., Equation 1) m is infinite and the GP method is necessary.

Covariance Functions

The only constraint on the covariance function is that it should generate a non-negative definite covariance matrix for any set of points in X . This gives wide scope, and different choices of $C(\mathbf{x}, \mathbf{x}')$ can give rise to such differing priors as straight lines of the form $y = w_0 + w_1 x$ (as discussed above) to the very rough and jagged sample paths associated with a Wiener process (a model for Brownian motion) or an Ornstein-Uhlenbeck process.

One very common form of covariance function is the *stationary* covariance function, where $C(\mathbf{x}, \mathbf{x}')$ is a function of $\mathbf{x} - \mathbf{x}'$. The use of stationary covariance functions is appealing if one would like the predictions to be invariant under shifts of the origin in input space. For example, in one dimension letting $h = x - x'$, the covariance of the Ornstein-Uhlenbeck process is $C_{OU}(h) = v_0 e^{-h/\lambda}$, where v_0 sets the overall variance of the process and λ sets a length scale in the input space. Another example of a stationary covariance function is the “squared exponential” covariance function $C_{SE}(h) = v_0 \exp(-h^2/\lambda^2)$ (sometimes called the “Gaussian” covariance function).

One commonly-used covariance function for inputs in \mathbb{R}^d is

$$C(\mathbf{x}, \mathbf{x}') = v_0 \exp \left\{ - \sum_{i=1}^d \frac{(x_i - x'_i)^2}{\lambda_i^2} \right\} \quad (1)$$

This is simply the product of d squared-exponential covariance functions, but with different length scales on each dimension. The general form of the covariance function expresses the idea that cases with nearby inputs will have highly correlated outputs, and the λ parameters allow a different distance measure for each input dimension. For irrelevant inputs, the corresponding λ_i will become large, and the model will effectively ignore that input. This is closely related to the automatic relevance determination (ARD) idea of MacKay and Neal (Neal, 1996).

The term *kernel function* used in the support vector machines literature is broadly equivalent to the covariance function. Further information on kernel/covariance functions can be found in Schölkopf and Smola (2001, chaps. 4 and 13), MacKay (1998), Williams (1998), and references therein.

Gaussian Processes for Regression Problems

In the previous section we discussed the properties of Gaussian processes. We now assume that we have input points $\mathbf{x}^n = \mathbf{x}_1, \dots, \mathbf{x}_n$ and target values $\mathbf{t} = t_1, \dots, t_n$ and wish to predict the function value y_* corresponding to an input \mathbf{x}_* . We assume that the target values t_i are obtained from the corresponding function value y_i by means of additive Gaussian noise, i.e., $t_i = y_i + \varepsilon_i$ for $i = 1, \dots, n$, where ε_i is an independent zero-mean Gaussian random variable of variance σ_v^2 . (The generalization to different variances at each location is straightforward, but notationally a bit more complex.) As the prior is a Gaussian process, the prior distribution over the y_i 's is given by $\mathbf{Y} \sim N(\mathbf{0}, K)$, where K is the $n \times n$ covariance matrix with entries $K_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$. It is then easy to show that the prior distribution over the targets is $N(\mathbf{0}, K + \sigma_v^2 I_n)$ where I_n is the $n \times n$ identity matrix.

To make a prediction for y_* we now need to consider the $n + 1$ -dimensional vector, which consists of the n variables in \mathbf{t} with the variable y_* appended, and condition on \mathbf{t} to obtain $P(y_* | \mathbf{t})$. As conditional distributions of jointly Gaussian variables are also Gaussian, it is clear that this distribution will be Gaussian, and our task is to compute the mean $\hat{y}(\mathbf{x}_*)$ and variance $\hat{\sigma}^2(\mathbf{x}_*)$. It turns out that

$$\hat{y}(\mathbf{x}_*) = \mathbf{k}^T(\mathbf{x}_*)(K + \sigma_v^2 I_n)^{-1} \mathbf{t} = \sum_{i=1}^n \alpha_i C(\mathbf{x}_i, \mathbf{x}_*) \quad (2)$$

$$\hat{\sigma}^2(\mathbf{x}_*) = C(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^T(\mathbf{x}_*)(K + \sigma_v^2 I_n)^{-1} \mathbf{k}(\mathbf{x}_*) \quad (3)$$

where $\mathbf{k}(\mathbf{x}_*)$ is the $n \times 1$ vector of covariances $(C(\mathbf{x}_1, \mathbf{x}_*), \dots, C(\mathbf{x}_n, \mathbf{x}_*))^T$, and $\boldsymbol{\alpha} = (K + \sigma_v^2 I_n)^{-1} \mathbf{t}$. Unpacking Equation 2, we see that the prediction function $\hat{y}(\mathbf{x}_*)$ is a linear combination of the kernel functions $C(\mathbf{x}_i, \mathbf{x}_*)$, with coefficients given by the appropriate entries of the vector $\boldsymbol{\alpha}$.

Equations 2 and 3 require the inversion of an $n \times n$ matrix, which is in general an $O(n^3)$ operation. When n is of the order of a few hundred, then this is quite feasible with modern computers. However, once $n \sim O(1000)$, these computations can be quite time-consuming, and much recent research effort has gone into developing approximation methods; see Tresp (2001) for a review. Note that in special cases (notably when the input space is \mathbb{R} and for certain Markovian kernels), the necessary calculations can be carried out in linear time (see Wahba, 1990, for further details).

The use of Gaussian processes for regression problems has been studied extensively by Carl Rasmussen (in Rasmussen, 1996) and in his Ph.D. thesis (available at <http://www.cs.utoronto.ca/~carl/>). He carried out a careful comparison of the Bayesian treatment of Gaussian process regression with several other state-of-the-art methods on a number of problems and found that its performance is comparable to that of Bayesian neural networks as developed by Neal (1996), and consistently better than the other methods tested.

Adapting the Covariance Function

Given a covariance function, it is straightforward to make predictions for new test points. However, in practical situations we are unlikely to know which covariance function to use. One option is to choose a parametric family of covariance functions (with a parameter vector $\boldsymbol{\theta}$) and then to search for parameters that give good predictions.

Adaptation of $\boldsymbol{\theta}$ is facilitated by the fact that the log likelihood $l = \log P(\mathbf{t} | \boldsymbol{\theta})$ can be calculated analytically as

$$l = -\frac{1}{2} \log \det(K + \sigma_v^2 I_n) - \frac{1}{2} \mathbf{t}^T (K + \sigma_v^2 I_n)^{-1} \mathbf{t} - \frac{n}{2} \log 2\pi \quad (4)$$

This is just the log likelihood of the vector \mathbf{t} under a Gaussian with mean $\mathbf{0}$ and covariance $K + \sigma_v^2 I_n$. The evaluation of the likelihood and its partial derivatives with respect to the parameters takes time $O(n^3)$, unless special structure in the problem can be exploited. Given l and its derivatives with respect to $\boldsymbol{\theta}$, it is straightforward to feed this information to an optimization package in order to obtain a local maximum of the likelihood.

One can also combine $P(\mathbf{t} | \boldsymbol{\theta})$ with a prior $P(\boldsymbol{\theta})$ to yield a Bayesian approach to the problem. Another approach to adapting $\boldsymbol{\theta}$ is to use the cross-validation (CV) or generalized cross-validation (GCV) methods, as discussed in Wahba (1990).

Relationship to Other Methods

Prediction with Gaussian processes is certainly not a very recent topic; the basic theory goes back at least as far as the work of Wiener and Kolmogorov in the 1940s on time series. Gaussian process prediction is also well known in the geostatistics field (see Cressie, 1993), where it is known as “kriging,” although this literature naturally has focused mostly on two- and three-dimensional input spaces.

As mentioned above, there is a close relationship between Bayesian approaches and regularization theory. This connection was described in Kimeldorf and Wahba (1970), and further details can be found in Wahba (1990), Poggio and Girosi (1990), and GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS.

When the covariance function $C(\mathbf{x}, \mathbf{x}')$ depends only on $h = \|\mathbf{x} - \mathbf{x}'\|$, the predictor derived in Equation 2 has the form $\sum_i c_i C(\|\mathbf{x} - \mathbf{x}_i\|)$ and may be called a *radial basis function* (or RBF) network. This is one derivation of RBFs, which are described in more detail in RADIAL BASIS FUNCTION NETWORKS.

The Gaussian process approach adds a stochastic process view to the regularization viewpoint, giving us “error bars” on the prediction (Equation 3), an expression for $P(\mathbf{t} | \boldsymbol{\theta})$ and its derivatives, and allows us to use the Bayesian machinery for hierarchical models.

Gaussian Processes for Classification Problems

Given training data and an input \mathbf{x} , the aim of a classifier is to predict the corresponding class label. This may be done by simply predicting a class label (“hard” classification), or by outputting an estimate of the posterior probabilities for each class $P(k | \mathbf{x})$ (“soft” classification), where $k = 1, \dots, C$ indexes the C classes. Naturally, we require that $0 \leq P(k | \mathbf{x}) \leq 1$ for all k and that $\sum_k P(k | \mathbf{x}) = 1$. A naive application of the regression method for Gaussian processes using, say, targets of 1 when an example of class k is observed and 0 otherwise will not obey these constraints. Soft classification has the advantage that the posterior probability estimates can be used in a principled fashion with loss matrices, rejection thresholds, and so on.

For the two-class classification problem it is only necessary to represent $P(1 | \mathbf{x})$, since $P(2 | \mathbf{x}) = 1 - P(1 | \mathbf{x})$. An easy way to ensure that the estimate $\pi(\mathbf{x})$ of $P(1 | \mathbf{x})$ lies in $[0, 1]$ is to obtain it by passing an unbounded value $y(\mathbf{x})$ through an appropriate function that has range $[0, 1]$. A common choice is the logistic function $\sigma(z) = 1 / (1 + e^{-z})$ so that $\pi(\mathbf{x}) = \sigma(y(\mathbf{x}))$. The input $y(\mathbf{x})$ to the logistic function will be called the *activation*. In the simplest method of this kind, logistic regression, the activation is simply computed as a linear combination of the inputs, plus a bias, i.e., $y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$. Using a Gaussian process or other flexible methods allows $y(\mathbf{x})$ to be a nonlinear function of the inputs.

For the classification problem with more than two classes, a simple extension of this idea using the “softmax” function gives the predicted probability for class k as

$$\pi(k|\mathbf{x}) = \frac{\exp y_k(\mathbf{x})}{\sum_m \exp y_m(\mathbf{x})} \quad (5)$$

For the rest of this section we shall concentrate on the two-class problem; extension of the methods to the multiclass case is relatively straightforward.

Defining a Gaussian process prior over the activation $y(\mathbf{x})$ automatically induces a prior over $\pi(\mathbf{x})$. To make predictions for a test input \mathbf{x}_* when using fixed parameters in the GP we would like to compute $\hat{\pi}_* = \int \pi_* P(\pi_* | \mathbf{t}, \boldsymbol{\theta}) d\pi_*$, which requires us to find $P(\pi_* | \mathbf{t}) = P(\pi(\mathbf{x}_*) | \mathbf{t})$ for a new input \mathbf{x}_* . This can be done by finding the distribution $P(y_* | \mathbf{t})$ (y_* is the activation of π_*) as given by

$$P(y_* | \mathbf{t}) = \int P(y_* | \mathbf{y}) P(\mathbf{y} | \mathbf{t}) d\mathbf{y} = \frac{1}{P(\mathbf{t})} \int P(y_* | \mathbf{y}) P(\mathbf{y}) P(\mathbf{t} | \mathbf{y}) d\mathbf{y} \quad (6)$$

where $\mathbf{y} = (y_1, \dots, y_n)$ denotes the activations corresponding to the data points. $P(\pi_* | \mathbf{t})$ can then be found from $P(y_* | \mathbf{t})$ using the appropriate Jacobian to transform the distribution. When $P(\mathbf{t} | \mathbf{y})$ is Gaussian, then the integral in Equation 6 can be computed exactly to yield Equations 2 and 3. However, the usual expression for $P(\mathbf{t} | \mathbf{y}) = \prod_i P(t_i | y_i)$ and $P(t_i | y_i) = \pi_i$ if $t_i = 1$ and $P(t_i | y_i) = (1 - \pi_i)$ for $t_i = -1$ for classification data (where the t 's take on values of 1 or -1) means that the marginalization to obtain $P(y_* | \mathbf{t})$ is no longer analytically tractable. Faced with this problem, we can either use an analytic approximation to the integral in Equation 6 or use Monte Carlo methods to approximate it. These two approaches will be considered in turn.

First, we note that $P(y_* | \mathbf{t})$ is mediated through $P(\mathbf{y} | \mathbf{t})$ and that $P(y_* | \mathbf{y})$ is Gaussian, so that obtaining information about $P(\mathbf{y} | \mathbf{t})$ is the essential step. It is easy to find the maximum of this distribution by optimizing $\log P(\mathbf{y}) + \log P(\mathbf{t} | \mathbf{y})$ with respect to \mathbf{y} , e.g., with a Newton-Raphson iteration. It can be shown that the optimization problem is convex. This yields the *maximum a posteriori* estimator \mathbf{y}^{MAP} . We could build a classifier based on \mathbf{y}^{MAP} by calculating $y_*^{MAP}(\mathbf{x}_*)$ as the mean of $P(y_* | \mathbf{y}^{MAP})$. This can then be fed through the logistic function to obtain an approximation to $\hat{\pi}_*$. This MAP solution is the one used in spline-smoothing approaches to classification (Wahba, 1990).

One can also make a Gaussian approximation to $P(\mathbf{y} | \mathbf{t})$ with mean \mathbf{y}^{MAP} and inverse covariance matrix $-\nabla \nabla \log P(\mathbf{y} | \mathbf{t})$. This yields a Laplace approximation to the integral in Equation 6.

Neal (1998) has developed an MCMC method for the Gaussian process classification model. This works by generating samples from $P(\mathbf{y} | \mathbf{t})$ by updating each of the n individual y_i 's sequentially using Gibbs sampling. This sampling process can be also be interleaved with sampling for the parameters $\boldsymbol{\theta}$. As with the regression problem, there has been much work on approximation schemes for large data sets; see Tresp (2001) for further details.

Classifiers using splines have been used extensively on a wide variety of problems, see Wahba (1990) and references in GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS. Gaussian process classifiers using MCMC sampling over $\boldsymbol{\theta}$ have been described in Williams and Barber (1998).

Relationship to Support Vector Machines

We have seen that the *maximum a posteriori* solution \mathbf{y}^{MAP} is obtained by minimizing $\Psi(\mathbf{y}) = -\log P(\mathbf{y}) - \log P(\mathbf{t} | \mathbf{y})$. This expression can be refined using $-\log P(\mathbf{t} | \mathbf{y}) = -\sum_i \log P(t_i | y_i)$ and $-\log P(t_i | y_i) = \log(1 + e^{-t_i y_i})$ to give

$$\Psi(\mathbf{y}) = \frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} + \sum_i \log(1 + e^{-t_i y_i}) + c \quad (7)$$

where c is a constant independent of \mathbf{y} . The criterion optimized by the support vector machine (SVM) learning algorithm (Vapnik, 1995) is very similar, but with $g_{GP}(z) \stackrel{\text{def}}{=} \log(1 + e^{-z})$ replaced by $g_{SVM}(z) \stackrel{\text{def}}{=} [1 - z]_+$, where $[x]_+ = \max(x, 0)$. These are both monotonically decreasing functions of z , which are linear for $z \rightarrow -\infty$. They both decay to zero as $z \rightarrow \infty$, but the main difference is that the g_{SVM} takes on the value 0 for $z > 1$, while g_{GP} asymptotes to 0 as $z \rightarrow \infty$. The SVM optimization problem is convex, but inequality constraints mean that it is quadratic programming problem.

By replacing g_{GP} with g_{SVM} we obtain $y^{SVM}(\mathbf{x}_*)$ instead of $y^{MAP}(\mathbf{x}_*)$. To make a “hard” ($+1/-1$) prediction, we simply take the predicted class label as $\text{sgn}(y(\mathbf{x}_*))$. This is the SVM classifier. The effect of the flat region of g_{SVM} is to introduce *sparsity* into the prediction of the corresponding $y^{SVM}(\mathbf{x}_*)$, where only those data points with $t_i y_i \leq 1$ contributing; these are known as the support patterns. Note that g_{SVM} is not interpretable as a negative log likelihood as it does not normalize properly. For further discussion, see Wahba (1999) and SUPPORT VECTOR MACHINES.

Discussion

In this article we have seen how Gaussian process priors over functions (which are in general infinite-dimensional objects) can be used in a computationally efficient manner to make predictions.

Methods such as Gaussian processes and support vector machines have come to be known by the umbrella term of *kernel machines* (see Schölkopf and Smola, 2001). The web site <http://www.kernel-machines.org/> has extensive links to research publications and software in this area.

One key issue concerning obtaining good performance with kernel methods is the choice of kernel. The squared-exponential kernel is widely used in practice, but it encodes only a general notion of smoothness. For particular problems, incorporation of prior/domain knowledge requires “kernel engineering.” A second key issue for kernel methods is developing good approximation algorithms for large data sets.

Road Map: Learning in Artificial Networks

Background: Bayesian Methods and Neural Networks

Related Reading: Generalization and Regularization in Nonlinear Learning Systems; Radial Basis Function Networks; Support Vector Machines

References

- Cressie, N. A. C., 1993, *Statistics for Spatial Data*, New York: Wiley.
- Kimeldorf, G., and Wahba, G., 1970, A correspondence between Bayesian estimation of stochastic processes and smoothing by splines, *Ann. Math. Statist.*, 41:495–502.
- MacKay, D. J. C., 1992, A practical Bayesian framework for backpropagation networks, *Neural Comput.*, 4:448–472.
- MacKay, D. J. C., 1998, Introduction to Gaussian processes, in *Neural Networks and Machine Learning* (C. M. Bishop, Ed.), New York: Springer-Verlag. ♦
- Neal, R. M., 1996, *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics 118, New York: Springer-Verlag.
- Neal, R. M., 1998, Regression and classification using Gaussian process priors (with discussion), in *Bayesian Statistics 6* (J. M. Bernardo et al., Eds.), Oxford, Engl.: Oxford University Press, pp. 475–501.
- Poggio, T., and Girosi, F., 1990, Networks for approximation and learning, *Proc. IEEE*, 78:1481–1497.
- Rasmussen, C. E., 1996, A practical Monte Carlo implementation of Bayesian learning, in *Advances in Neural Information Processing Systems 8* (D. S. Touretzky, M. Mozer, and M. E. Hasselmo, Eds.), Cambridge, MA: MIT Press, pp. 598–604.

- Schölkopf, B., and Smola, A., 2001, *Learning with Kernels*, Cambridge, MA: MIT Press. ♦
- Tresp, V., 2001, Scaling kernel-based systems to large data sets, *Data Mining Knowl. Discov.*, 5:197–211.
- Vapnik, V. N., 1995, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.
- Wahba, G., 1990, *Spline Models for Observational Data*, SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics. ♦
- Wahba, G., 1999, Support vector machines, reproducing kernel Hilbert

- spaces, and randomized GACV, in *Advances in Kernel Methods* (B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds.), Cambridge, MA: MIT Press, pp. 69–88.
- Williams, C. K. I., 1998, Prediction with Gaussian processes: From linear regression to linear prediction and beyond, in *Learning in Graphical Models* (M. I. Jordan, Ed.), Boston: Kluwer Academic, pp. 599–621. ♦
- Williams, C. K. I., and Barber, D., 1998, Bayesian classification with Gaussian processes, *IEEE Trans. Pattern Anal. Machine Intell.*, 20:1342–1351.

Generalization and Regularization in Nonlinear Learning Systems

Grace Wahba

Introduction

In this article we will describe generalization and regularization from the point of view of multivariate function estimation in a statistical context. Multivariate function estimation is not, in principle, distinguishable from supervised machine learning. However, until fairly recently, supervised machine learning and multivariate function estimation had fairly distinct groups of practitioners and little overlap in language, literature, and the kinds of practical problems under study.

In any case, we are given a *training set*, consisting of pairs of input (feature) vectors and associated outputs $\{\mathbf{t}(i), y_i\}$, for n training or example subjects, $i = 1, \dots, n$. From these data, it is desired to construct a map that *generalizes well*, that is, given a new value of \mathbf{t} , the map will provide a reasonable prediction for the unobserved output associated with this \mathbf{t} .

Most applications fall into one of two broad categories, which might be called nonparametric regression and classification. In *nonparametric regression*, y may be (any) real number or a vector of r real numbers. The desired algorithm will produce an estimate $\hat{f}(\mathbf{t})$ of the expected value of a (new) y to be associated with a (new) attribute vector \mathbf{t} . In the (two-class) *classification* problem, y_i will be an indicator, whether or not the example (subject) came from class \mathcal{A} . In some classification applications, the desired algorithm will return $\hat{f}(\mathbf{t})$, return an indicator that predicts whether or not an example with attribute vector \mathbf{t} comes from class \mathcal{A} (“hard”) classification. In other applications the desired algorithm will return $p(\mathbf{t})$, an estimate of the *probability* that the example with attribute vector \mathbf{t} is in class \mathcal{A} (“soft” classification). In some applications the feature vector \mathbf{t} of dimension d contains zeros and ones (for example, in a bitmap of handwriting); in others it may contain real numbers representing some physical quantities. Ordered or unordered category indicators are also possible, as in medical demographic studies. *Regularization*, loosely speaking, means that whereas the desired map is constructed to approximately send the observed feature vectors to the observed outputs, constraints are applied to the construction of the map, with the goal of reducing the generalization error (see also PROBABILISTIC REGULARIZATION METHODS FOR LOW-LEVEL VISION). In some applications, these constraints embody a priori information concerning the true relationship between input and output; alternatively, various ad hoc constraints have sometimes worked well in practice. Girosi, Jones, and Poggio (1995) give a wide-ranging review.

Generalization and Regularization in Nonparametric Regression

Single-Input Spline Smoothing

We will use Figure 1 to illustrate the ideas of generalization and regularization in the simplest possible nonparametric regression setup, that is, $d = 1$, $r = 1$, with $\mathbf{t} = t$ any real number in some interval of the real line. The circles (which are identical in each of the three panels of Figure 1) represent $n = 100$ (synthetically generated) input-output pairs $\{t(i), y_i\}$, generated according to the model

$$y_i = f_{\text{TRUE}}(t(i)) + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where $f_{\text{TRUE}}(t) = 4.26(e^{-t} - 4e^{-2t} + 3e^{-3t})$, and the ε_i came from a pseudo-random number generator for normally distributed random variables with mean 0 and standard deviation $\sigma = 0.2$. Given this training data $\{t(i), y_i, i = 1, \dots, n\}$, the learning problem is to create a map that, if given a new value of t , will predict the response $y(t)$. In this case, the data are noisy, so that even if the new t coincides with some predictor variable $t(i)$ in the training set, merely predicting y as the response y_i is not likely to be satisfactory. Also, this does not yet provide any ability to make predictions when t does not exactly match any predictor values in the training set. It is desired to generate a curve that will allow a reasonable prediction of the response for any t within a reasonable vicinity of the set of training predictors $\{t(i)\}$. The dashed line in each panel of Figure 1 is $f_{\text{TRUE}}(t)$; the three solid black lines in the three panels of Figure 1 are three solutions to the variational problem: find f in the (Hilbert) space W_2 of functions with continuous first derivatives and square integrable second derivatives that minimizes

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(t(i)))^2 + \lambda \int (f^{(2)}(u))^2 du \quad (2)$$

for three different values of λ . The parameter λ is known as the *regularization* or *smoothing parameter*. As $\lambda \rightarrow \infty$, f_λ tends to the least squares straight line best fitting the data, and as $\lambda \rightarrow 0$ the solution tends to that curve in W_2 that minimizes the penalty functional $J(f) = \int (f^{(2)}(u))^2 du$ subject to interpolating the data (provided the $\{t(i)\}$ are distinct). This latter interpolating curve is known as a *cubic interpolating spline*, and minimizers of Equation 2 are known as *smoothing splines*. We remark that, here as well as in the sequel, although a variational problem is being solved in

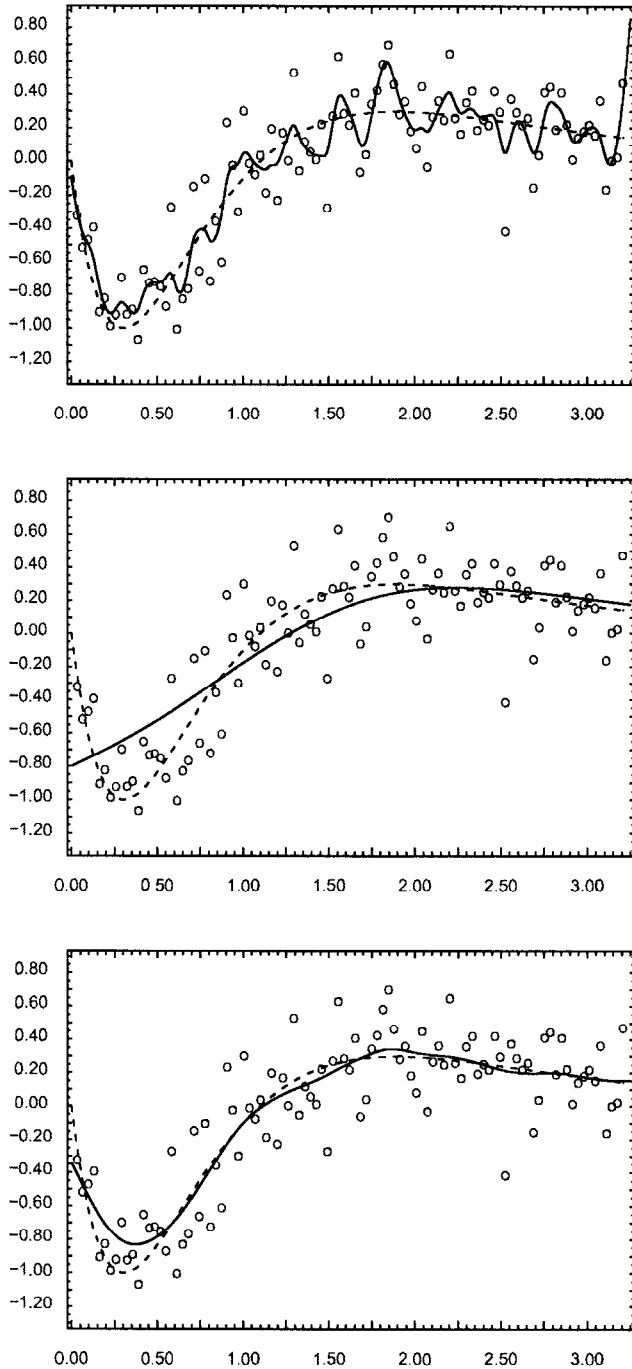


Figure 1. Training data (circles) have been generated by adding noise to $f_{TRUE}(t)$, shown by the dashed curve in each panel. All three panels have the same data. *Top*, Solid curve is fitted spline with λ too small. *Middle*, Solid curve is fitted spline with λ too large. *Bottom*, Solid curve is fitted spline with λ obtained by generalized cross-validation.

an infinite dimensional Hilbert space, the solution is in an n -dimensional subspace with a known spanning set. (See Wahba, 1990, and references cited there for further information concerning these and other properties of splines, and further references.)

In the top panel of Figure 1 λ has been chosen too small, and the wiggly solid line is attempting to fit the data too closely. It can

be seen that using the wiggly curve in the top panel is not likely to give a good prediction of y , assuming that future predictor-response data are generated by the same mechanism as the training data. In the middle panel, λ has been chosen too large; the curve has been forced to flatten out, and again it can be seen that the heavy line will not give a good prediction of y . In the bottom panel, λ has been chosen by generalized cross-validation (GCV). This is a method that behaves similarly to leaving-out-one in many cases, but with computational and theoretical advantages (see Li, 1986; Wahba, 1990, chap. 4; Girard, 1998). It can be seen that the λ obtained in this way does a good job of choosing the right amount of smoothing to best recover f_{TRUE} of Equation 1. The f_{TRUE} of Equation 1 would provide the best predictor of the response in an expected mean-square-error sense if future data were generated according to Equation 1. The curve in the bottom panel has a reasonable ability to *generalize*, that is, to predict the response given a new value t of the predictor variable, at least if t is not too far from the training predictor set $\{t(i)\}$.

For each positive λ , there exists a unique $\kappa = \kappa(\lambda)$ so that the minimizer f_λ of Equation 2 is also the solution to the problem: Find f in W_2 to minimize

$$L(y, f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(t(i)))^2 \quad (3)$$

subject to the condition

$$J(f) = \int (f^{(2)}(u))^2 du \leq \kappa \quad (4)$$

As λ becomes large, the associated $\kappa(\lambda)$ becomes small, and conversely. In general, the term *regularization* refers to solving some problem involving best fitting, subject to some constraints on the solution. These constraints may be of various forms. When they involve a quadratic penalty involving derivatives, like $J(f)$, the method is commonly referred to as *Tikhonov regularization*. The “tighter” the constraints (i.e., the smaller κ , equivalently the larger λ), the further away the solution f_λ will generally be from the training data; that is, L will be larger. As the constraints get weaker and weaker, ultimately (if there are enough degrees of freedom in the method) the solution will interpolate the data. However, as is clear from Figure 1, a curve that runs through all the data points is *not* a good solution.

A fundamental problem in machine learning with noisy and/or incomplete data is to balance the “tightness” of the constraints with the “goodness of fit” to the data, in such a way as to minimize the generalization error, that is, the ability to predict the unobserved response for new values of t (or \mathbf{t}). This trade-off is by now well known as the *bias-variance trade-off*, or, equivalently, the *goodness of fit-model complexity trade-off*. Methods abound in the statistical literature for univariate curve fitting, including Parzen kernel estimates, nearest neighbor estimates, orthogonal series estimates, least squares regression spline estimates, and, recently, wavelet estimates. Each method has one or more regularization parameters, whether they are kernel window widths, number of nearest neighbors included, number of terms in the orthogonal series expansion or regression basis, or factors or thresholds for shrinking or truncating wavelet coefficients, that control this trade-off. (See Ramsay and Silverman, 1997, and references therein.)

Multiple-Input, Single-Hidden-Layer Feedforward Neural Net

A multiple-input, single-hidden-layer feedforward neural net (NN) predictor for the learning problem described in the Introduction is typically of the form

$$f_{NN}(\mathbf{t}) = \sigma_0 \left(b_0 + \sum_{j=1}^N w_j \sigma_h(\mathbf{a}_j' \mathbf{t}(i) + b_j) \right) \quad (5)$$

where the \mathbf{a}_j and \mathbf{t} are d -vectors. The function σ_h is the so-called “activation function” of the hidden layer and σ_0 is the activation function for the output. σ_h is generally a sigmoidal function, for example, $\sigma_h(\tau) = e^\tau / (1 + e^\tau)$, while σ_0 may be linear, sigmoidal or a threshold unit. Here N is the number of hidden units, and the w_j , \mathbf{a}_j , and b_j are “learned” from the training data by some appropriate iterative descent algorithm that tries to steer these values toward minimizing some distance measure, typically $L(y, f_{NN}) = (1/n) \sum_{i=1}^n (y_i - f_{NN}(\mathbf{t}(i)))^2$. It is clear that if N is sufficiently large and the descent algorithm is run long enough, it should be possible to drive the L as close as one likes to zero. (In practice it is possible to get stuck in local minima.) However, it is also clear intuitively from Figure 1 that driving L all the way to zero is not a desirable thing to do. Regularization in this problem may be done by controlling the size of N , by imposing penalties on the w_j , by stopping the descent algorithm early (that is, by not driving down L as far as it can go), or by various combinations of these strategies. Each will influence how closely f_{NN} will fit the data, how “wiggly” it will be, and how well it will be able to predict unobserved data that are generated by a similar mechanism as the observed data.

Multiple-Input Radial Basis Function and Related Estimates

Radial basis functions are rapidly becoming a popular method for nonparametric regression (see RADIAL BASIS FUNCTION NETWORKS). We first describe a general form of nonparametric regression that will specialize to radial basis functions and other methods of interest. Let $R(\mathbf{s}, \mathbf{t})$ be any symmetric, strictly positive definite function on $E^d \times E^d$. Here, *strictly positive definite* means for any $K = 1, 2, \dots$, the $K \times K$ matrix with j, k th entry $R(\mathbf{s}(j), \mathbf{s}(k))$ is strictly positive definite whenever the $\mathbf{s}(1), \dots, \mathbf{s}(K)$ are distinct. (A symmetric $K \times K$ matrix M is said to be positive definite if for any K -dimensional column vector $x \neq 0$, $x'Mx$ is greater than or equal to 0, and is said to be strictly positive definite if $x'Mx$ is always strictly greater than 0.) Positive definiteness will play a key role in the discussion below because, among other reasons, any positive definite matrix can be the covariance matrix of a random vector and any positive definite function $R(\mathbf{s}, \mathbf{t})$ can be the covariance function of some stochastic process, $X(\mathbf{t})$. That is, there exists $X(\cdot)$ such that $\text{Cov } X(\mathbf{s})X(\mathbf{t}) = R(\mathbf{s}, \mathbf{t})$. Given training data $\{\mathbf{t}(i), y_i\}$, it is always possible in principle to obtain a (regularized) input-output map from this data by letting the model $f_{R,\lambda}$ be of the form

$$f_{R,\lambda}(\mathbf{t}) = \sum_{j=1}^N c_j R(\mathbf{t}, \mathbf{s}(j)) \quad (6)$$

where the $\mathbf{s}(j)$ are $N \leq n$ “centers” that are placed at distinct values of the $\{\mathbf{t}(i)\}$ and $c = (c_1, \dots, c_N)'$ is chosen to minimize $L(y, f) + \lambda J(f)$. Here

$$L(y, f_{R,\lambda}) = \frac{1}{n} \sum_{i=1}^n (y_i - f_{R,\lambda}(\mathbf{t}(i)))^2 \quad (7)$$

and the regularizing penalty $J(\cdot)$ is of the form

$$J(f_{R,\lambda}) = \sum_{j,k=1}^N c_j c_k J_{jk} \quad (8)$$

where J_{jk} are the entries of a non-negative definite quadratic form. The (strict) positive definiteness of R guarantees that

$$L(y, f_{R,\lambda}) + \lambda J(f_{R,\lambda}) \quad (9)$$

always has a unique minimizer in c , for any non-negative λ . This follows by substituting Equation 6 into Equation 9, and using the fact that the columns of the $n \times N$ matrix with ij entry $R(\mathbf{t}(i), \mathbf{s}(j))$ are linearly independent since they are just N columns of the $n \times n$ positive definite matrix with ij entry $R(\mathbf{t}(i), \mathbf{t}(j))$.

Radial basis function estimates are obtained for the special case where $R(\mathbf{s}, \mathbf{t})$ is of the special form

$$R(\mathbf{s}, \mathbf{t}) = r(\|W(\mathbf{s} - \mathbf{t})\|) \quad (10)$$

where W is some linear transformation on E^d and the norm is Euclidean distance. That is, $R(\mathbf{s}, \mathbf{t})$ depends only on some generalized distance in E_d between \mathbf{s} and \mathbf{t} . The regularization—that is, the effecting of the trade-off between goodness of fit to the data and “smoothness” of the solution—is performed by reducing N and/or increasing λ . The choice of W will also affect the “wiggleness” of $f_{R,\lambda}$ in the radial basis function case. Alternatively, a model can be obtained by choosing N small and minimizing $L(y, f)$. In that case, N and W are the smoothing parameters.

In the special case $N = n$, $\mathbf{s}(i) = \mathbf{t}(i)$, the $f_{R,\lambda}$ can (for any positive definite R) be shown to be Bayes estimates (see Kimeldorf and Wahba, 1970; Wahba, 1990). Arguments can be given to show that if n is large and $N < n$ is not too small, then they are good approximations to Bayes’s estimates (see Wahba, 1990, chap. 7). In the special case $J_{ij} = R(\mathbf{t}(i), \mathbf{t}(j))$, the Bayes model is easy to describe and we do it here; it is:

$$y_i = X(\mathbf{t}(i)) + \varepsilon_i \quad (11)$$

with $X(\mathbf{t})$ a zero-mean Gaussian process (see GAUSSIAN PROCESSES) with covariance $EX(\mathbf{s})X(\mathbf{t}) = bR(\mathbf{s}, \mathbf{t})$ and the ε_i independent zero-mean Gaussian random variables with common variance σ^2 , and independent of $X(\mathbf{t})$. In this case, the minimizer $f_{R,\lambda}$ of $L(y, f) + \lambda J(f)$, evaluated at \mathbf{t} , is the conditional expectation of $X(\mathbf{t})$, given y_1, \dots, y_n , provided that λ is chosen as σ^2/nb . In general, pretending that one has a prior and computing the posterior mean or mode will have a regularizing effect. The preceding discussion extends to symmetric positive definite functions on *arbitrary* domains for \mathbf{t} , including those mentioned in the Introduction of this article.

Thin plate splines in d variables (of order m) consist of radial basis functions plus polynomials of total degree less than m in d variables. ($2m - d > 0$ is required for technical reasons.) Letting $\mathbf{t} = (t_1, \dots, t_d)$, the thin plate splines are minimizers (in an appropriate function space) of

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{t}(i)))^2 + \lambda \sum_{\alpha_1 + \dots + \alpha_d = m} \frac{m!}{\alpha_1! \dots \alpha_d!} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(\frac{\partial^m f}{\partial t_1^{\alpha_1} \dots \partial t_d^{\alpha_d}} \right)^2 dt_1 \dots dt_d \quad (12)$$

Setting $d = 1$, $m = 2$ gives the cubic spline case discussed earlier. Note that there is no penalty on polynomials of total degree less than m ; the thin plate splines with a particular choice of λ are Bayes estimates with an improper prior (that is, infinite variance) on the polynomials of total degree less than m (see Wahba, 1990, and references cited therein).

Related variations on regularized estimates include additive smoothing splines, which are of the form

$$f(\mathbf{t}) = \mu + \sum_{\alpha=1}^d f_\alpha(t_\alpha) \quad (13)$$

where μ and the f_α are the solution to a variational problem of the form: Find μ and f_1, \dots, f_d in a certain function space to minimize

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{t}(i)))^2 + \sum_{\alpha=1}^d \lambda_\alpha J_\alpha(f_\alpha) \quad (14)$$

The J_α may be of the form of J in Equation 4. Here, there is a *regularization parameter* for each component (see Hastie and Tibshirani, 1990; Wahba, 1990). These additive models generalize to smoothing spline analysis of variance (SS-ANOVA) models. In the SS-ANOVA models, interaction terms of the form $f_{\alpha\beta}(t_\alpha, t_\beta), f_{\alpha\beta\gamma}(t_\alpha, t_\beta, t_\gamma)$, etc., which satisfy side conditions making them uniquely determined, are added to the representation in Equation 13, and corresponding penalty terms with regularization parameters are added in Equation 14. The f_α , etc., may be generalized to themselves, being radial basis functions. Behind these models are positive definite functions that are built up via tensor sums and products of positive definite functions (see Wahba, 1990; Wahba et al., 1995).

Regression spline ANOVA models may be obtained by setting the $f_\alpha, f_{\alpha\beta}$, etc. as linear combinations of a (relatively small) number of basis functions (usually splines). In this case the number of the basis functions is probably the most influential regularization parameter. These and similar methods again all have either explicit or implicit regularization parameters that govern the balance between the complexity of the model and the fit to the data—the bias-variance trade-off.

The usual criterion for the generalization error when the fit involves minimizing the observed residual sum of squares is the expected residual sum of squares for new data, $EL(y_{new}, f_\lambda) \equiv L(f_{TRUE}, f_\lambda) = (1/n) \sum_{i=1}^n (f_{TRUE}(\mathbf{t}(i)) - f_\lambda(\mathbf{t}(i)))^2$. Here the y_{new} are new observations at the original $\mathbf{t}(i)$. Leaving out one, leaving out 10%, leaving out a one-third representative sample (“tuning set”), and GCV (“in-sample tuning”) are popular methods for choosing the tuning parameters to minimize this criterion. Codes in Splus (smooth.spline()), SAS (tpspline), netlib (entire/gcv directory), Funfits (sreg, tps), R (smooth.Pspline, gss), and elsewhere are available for implementing the univariate spline, thin plate spline, and additive and interaction (ANOVA) splines with GCV to choose single or multiple smoothing parameters. Netlib (<http://www.netlib.org/>) and Funfits (<http://www.cgd.ucar.edu/stats/software.shtml>) are freeware. The smooth.Pspline code in R at <http://www.r-project.org> was used to generate Figure 1.

Generalization and Regularization in Soft Classification

Soft classification is a natural goal in certain kinds of demographic medical studies. For example, suppose a large training set is available from a demographic study, consisting of observations $\{\mathbf{t}(i), y_i\}$, where y_i is an indicator (1 or 0) of the presence or absence of some disease in subject i at the end of the study and $\mathbf{t}(i)$ is a vector of values of risk factors for this subject at the beginning of the study. With this kind of data, it is frequently of interest to make a “soft” classification, that is, to estimate the *probability* $p(\mathbf{t})$ that a new subject with predictor vector \mathbf{t} will contract the disease. A doctor, given this model, may advise new patients which risk factors are important for them to control to reduce the probability of their contracting the disease. A regularized (that is, “smooth”) estimate for $p(\mathbf{t})$ is desirable. Regularized estimates can be obtained as follows. First, define

$$f(\mathbf{t}) = \log[p(\mathbf{t})/(1 - p(\mathbf{t}))] \quad (15)$$

f is known in the statistics literature as the log odds ratio, or logit. Then $p(\mathbf{t})$ is a sigmoidal function of $f(\mathbf{t})$; that is, $p(\mathbf{t}) = e^{f(\mathbf{t})}/(1 + e^{f(\mathbf{t})})$. We will get a regularized estimate for f . $L(y, f)$ of Equation 3 will be replaced by an expression more suitable for 0–1 data, by using the likelihood for these data. To describe the likelihood, note that if y is a random variable with $\text{Prob}[y = 1] = p$ and $\text{Prob}[y = 0] = (1 - p)$, then the probability density (or likelihood) $P(y, p)$ for y , when p is true, is just $P(y, p) = p^y (1 - p)^{(1-y)}$. This merely says $P(1, p) = p$ and $P(0, p) = (1 - p)$. Thus, the likelihood for y_1, \dots, y_n (assuming that the y_i are independent) is

$$P(y_1, \dots, y_n; p(\mathbf{t}(1)), \dots, p(\mathbf{t}(n))) = \prod_{i=1}^n p(\mathbf{t}(i))^{y_i} (1 - p(\mathbf{t}(i))^{(1-y_i)}) \quad (16)$$

Substituting f for p in Equation 16 and taking the negative logarithm gives the negative log likelihood $L(y, f)$ in terms of f :

$$-\log P(y_1, \dots, y_n; f(\mathbf{t}(1)), \dots, f(\mathbf{t}(n))) = nL(y, f) = \sum_{i=1}^n [\log(1 + e^{f(\mathbf{t}(i))}) - y_i f(\mathbf{t}(i))] \quad (17)$$

It is natural for $L(y, f)$ to replace $L(y, f)$ in Equations 3, 7, and 14 when y_i is restricted to 0 or 1, since $L(y, f_{TRUE})$ is (a multiple of) the negative log likelihood for y generated by a model with Gaussian noise, as in Equation 1. A neural net implementation of soft classification would consist of finding $f_{NN}(\mathbf{t}) = \text{logit} p_{NN}(\mathbf{t})$ of the form of Equation 5 to minimize $L(y, f)$ of Equation 17. If N is large enough, then, in principle, f_{NN} may be driven so that $p_{NN}(\mathbf{t}(i))$ is close to 1 if y_i is 1, and is close to 0 if y_i is 0. Again, it is intuitively clear that this is not desirable. As before, a regularized or smooth f_{NN} can be obtained by controlling N , penalizing the w_i , stopping the iterative fitting early, or some combination of these actions.

Penalized likelihood estimates of f are obtained by minimizing $L(y, f) + J_\lambda(f)$, where $J_\lambda(f)$ is a penalty functional corresponding to those in Equations 2, 9, 12, or 14 and its generalizations. A popular definition for the generalization error is the (unobservable) comparative Kullback-Leibler distance of the estimate to the true probability distribution, which can be shown to be given by $EL(y_{new}, f_\lambda) = L(p_{TRUE}, f_\lambda)$. An estimate of λ that minimizes this criterion can be obtained by withholding a representative subset $y_{\text{left-out}}$ of the training set and choosing λ to minimize $L(y_{\text{left-out}}, f_\lambda)$. Leaving-out-one estimates are also possible but generally not feasible in this case. Generalized approximate cross-validation (GACV) is a feasible in-sample method of choosing λ ; based on a leaving-out-one argument, it has been shown in simulation studies to provide a good estimate of the minimizer of $L(p_{TRUE}, f_\lambda)$ (see Wahba et al., 1999).

Generalization and Regularization in Hard Classification

In the hard classification problem (here we will consider only two classes for simplicity), we are only interested in estimating whether an example with vector \mathbf{t} is in class \mathcal{A} or not. This is the typical situation in, for example, character recognition, voice recognition, and other situations where it is known that the \mathbf{t} 's from the two classes being examined are generally well separated. In that case (assuming, for simplicity, that the examples from the two classes are represented in the training set equally as is the future population of interest, and that costs of misclassification are the same for both classes), then the optimum classifier (to minimize the expected cost) would be \mathcal{A} if $p(\mathbf{t})$ is greater than one-half, and not \mathcal{A} otherwise. Equivalently, the same rule can be implemented by examining the sign of the logit $f(\mathbf{t})$. Here we are identifying \mathcal{A} with the 1s, and optimum is with respect to minimizing the expected cost of future misclassification. Unfortunately, in general it is neither desirable nor feasible to estimate the logit f directly by the methods described in the section on soft classification, because in the well-separated case, f takes on values near $\pm\infty$, and solving the penalized likelihood problem of that section is likely to be unstable. Recently, SUPPORT VECTOR MACHINES (q.v.) have been shown to provide an excellent method for classification in this situation (see Burges, 1998).

The support vector machine (SVM) is implemented coding the y_i as ± 1 according as the i th example is in \mathcal{A} or not. Given a

positive definite function $R(\mathbf{s}, \mathbf{t})$, we find a function f of the form $f(\mathbf{t}) = b + \sum_{i=1}^n c_i R(\mathbf{t}, \mathbf{t}(i))$ by finding b and $c = (c_1, \dots, c_n)$ to minimize

$$\frac{1}{n} \sum_{i=1}^n (1 - y_i f(\mathbf{t}(i))_+ + \lambda \sum_{i,j} c_i c_j R(\mathbf{t}(i), \mathbf{t}(j))) \quad (18)$$

where $(\tau)_+ = \tau$ for $\tau > 0$ and 0 otherwise. Letting f_λ be the minimizer of Equation 18, the classification algorithm is: for a new attribute vector \mathbf{t} , assign \mathcal{A} if $f(\mathbf{t}) > 0$ and not \mathcal{A} if $f(\mathbf{t}) < 0$. Lin (1999) has demonstrated the remarkable result that, under general circumstances with appropriately chosen λ , the SVM estimate f_λ tends almost everywhere to either 1 or -1 and is an estimate of $\text{sign} f_{\text{TRUE}} \equiv \text{sign}(p_{\text{TRUE}} - 1/2)$, which is exactly what is needed to carry out the optimum classification algorithm. It is interesting to note that if the data y_i in the penalized log likelihood estimate were recoded to $y_i = \pm 1$, then the i th term on the right of Equation 17 would become $\log(1 + e^{-y_i f(\mathbf{t}(i))})$, which is bounded below by $(\log 2)(1 - y_i f(\mathbf{t}(i)))_+$. A popular choice for $R(\mathbf{s}, \mathbf{t})$ is $R(\mathbf{s}, \mathbf{t}) = \exp(-(1/\sigma^2)\|\mathbf{s} - \mathbf{t}\|^2)$, where $\|\cdot\|$ is the Euclidean norm. In this choice of $R(\cdot, \cdot)$ the result may be sensitive to both σ and λ . As before, the λ and σ may be chosen by leaving out a representative subset of the observations and choosing λ and σ to minimize some measure of the generalization error. Here the natural choice for generalization error would be the misclassification rate. A version of GACV for SVMs, again based on a leaving-out-one argument, may be used as an in-sample method for choosing λ and σ (see Wahba, Lin, and Zhang, 2000). The generalization error target for the GACV is $E(1/n) \sum_{i=1}^n (1 - y_{\text{new}} f_\lambda(\mathbf{t}(i)))_+$. However, $(1/2)E(1/n) \sum_{i=1}^n (1 - y_{\text{new}} \text{sign}[f_\lambda(\mathbf{t}(i))])_+$ is the expected misclassification rate, so that to the extent that f_λ resembles $\text{sign} f_\lambda$, this criterion will be appropriate for the generalization error. There is a large literature on the multiclass case, generally involving repeated pairwise or one-versus-many comparisons. A multiclass SVM that deals with all classes simultaneously has recently been developed (Lee, Lin, and Wahba, 2001). This work also demonstrates how to modify the SVM to take into account nonrepresentative training sets and unequal misclassification costs.

Choosing How Much to Regularize

At the time of this writing, it is a matter of lively debate and much research how to choose the various regularization parameters. Leaving out a large fraction of the training sample for this purpose and tuning the regularization parameter(s) to best predict the left-out data (according to whatever criterion of best prediction is adopted) is conceptually simple, defensible, and widely used (this is called out-of-sample tuning). Successively leaving-out-one, successively leaving-out-10%, and the in-sample methods GCV and GACV are all popular. (See also Ye, 1998, who discusses in-sample tuning methods related to GCV in the Gaussian case that allows comparisons across different regularized estimates.) In the normally distributed observational error case, if the standard deviation of the observational error (σ in Equation 1) is known, then unbiased risk estimates become available (see Li, 1986; Wahba, 1990, and references therein). When there is a Bayesian model behind the regularization procedure, then maximum likelihood estimates may be derived (see Wahba, 1985), although in order for these and other Bayesian estimates to do a good job of minimizing the generalization error in practice, it is usually necessary that the priors on which they are based be realistic.

Which Method Is Best?

Feedforward neural nets, radial basis functions, and various forms of splines all provide regularized or regularizable methods for estimating smooth functions of several variables, given a training set

$\{\mathbf{t}(i), y_i\}$. Which approach is best? Unfortunately, there is no single answer to that question, nor is there likely to be one. The answer depends on the particular nature of the underlying but unknown "truth," the nature of any prior information that might be available about this truth, the nature of any noise in the data, the ability of the experimenter to choose the various smoothing or regularization parameters well, the size of the data set, the use to which the answer will be put, and the computational facilities available. From a mathematical point of view, the classes of functions well approximated by neural nets, radial basis functions, and additive and interaction splines (ANOVA splines) are not the same, although all of these methods have the capability of approximating large classes of functions. Of course, if a large enough data set is available, models utilizing all of these approaches can be built, tuned, and compared on data that have been set aside for this purpose. In-sample tuning methods for comparison across different regularized estimates in the hard and soft classification contexts are an area of active research.

Acknowledgments. This work was supported by NSF grant No. DMS 9704798 and NIH grant No. R01 EY09946.

Road Map: Learning in Artificial Networks

Related Reading: Learning and Statistical Inference; Stochastic Approximation and Efficient Learning

References

- Burges, C., 1998, A tutorial on support vector machines for pattern recognition, *Data Mining Knowledge Discovery*, 2:121–167. ♦
- Girard, D., 1998, Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression, *Ann. Statist.*, 126:315–334.
- Giroi, F., Jones, M., and Poggio, T., 1995, Regularization theory and neural networks architectures, *Neural Comput.*, 7:219–269.
- Hastie, T., and Tibshirani, R., 1990, *Generalized Additive Models*, New York: Chapman and Hall. ♦
- Kimeldorf, G., and Wahba, G., 1970, A correspondence between Bayesian estimation of stochastic processes and smoothing by splines, *Ann. Math. Statist.*, 41:495–502.
- Lee, Y., Lin, Y., and Wahba, G., 2001, *Multicategory Support Vector Machines*, Technical Report 1043, Madison, WI: University of Wisconsin, Department of Statistics; to appear in *Comput. Sci. Stat.*, 33.
- Li, K. C., 1986, Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing, *Ann. Statist.*, 14:1101–1112.
- Lin, Y., 1999, *Support Vector Machines and the Bayes rule in Classification*, Technical Report 1014, Madison WI: University of Wisconsin, Department of Statistics, in *Data Mining and Knowledge Discovery*, 6:259–275.
- Ramsay, J., and Silverman, B., 1997, *Functional Data Analysis*, New York: Springer-Verlag. ♦
- Wahba, G., 1985, A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, *Ann. Statist.*, 13:1378–1402.
- Wahba, G., 1990, *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, Philadelphia: Society for Industrial and Applied Mathematics. ♦
- Wahba, G., Lin, X., Gao, F., Xiang, D., Klein, R., and Klein, B., 1999, The bias-variance tradeoff and the randomized GACV, in *Advances in Information Processing Systems 11* (M. Kearns, S.olla, and D. Cohn, Eds.), Cambridge, MA: MIT Press, pp. 620–626.
- Wahba, G., Lin, Y., and Zhang, H., 2000, Generalized approximate cross validation for support vector machines, in *Advances in Large Margin Classifiers* (A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds.), Cambridge, MA: MIT Press, pp. 297–311.
- Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B., 1995, Neyman Lecture: Smoothing spline ANOVA for exponential families: With application to the Wisconsin Epidemiological Study of Diabetic Retinopathy, *Ann. Statist.*, 23:1865–1895.
- Ye, J., 1998, On measuring and correcting the effects of data mining and model selection, *J. Am. Statist. Assoc.*, 93:120–131.

GENESIS Simulation System

James M. Bower, David Beeman, and Michael Hucka

Introduction

GENESIS (the GEneral NEural SIMulation System) was developed as a research tool to provide a standard and flexible means for constructing structurally realistic models of biological neural systems. "Structurally realistic" simulations are computer-based implementations of models whose primary objective is to capture what is known of the anatomical structure and physiological characteristics of the neural system of interest. The GENESIS project is based on the belief that progress in understanding structure-function relationships in the nervous system specifically, or in biology in general, will increasingly require the development and use of structurally realistic models (Bower, 1995). It is our view that only through this type of modeling will general principles of neural or biological function emerge.

There is considerable debate within the computational neuroscience community concerning the appropriate level of modeling. As illustrated in other articles in the *Handbook*, many modeling efforts are currently focused on abstract "general" representations of neural function rather than on detailed, realistic models. However, the history of science clearly indicates that realistic models play an essential role in the development of quantitative understanding of physical systems. For example, philosophers and priests for thousands of years invented "models" to account for the motion of the planets in the night sky. These models, the most famous of which is probably the Ptolemaic system, "replicated the data" and made quantitative predictions. The structure of these planetary models, however, already assumed the general principles on which the universe was organized. It was not until the sixteenth and seventeenth centuries, when Kepler and, later, Newton constructed realistic models of the solar system, that general principles such as universal gravitation emerged. The inverse square law for gravitational attraction fell out of a model that Newton constructed of the moon's movement around Earth; it was not an apple-induced inspiration.

It is our view that neuroscience is not yet ready for its Newton. Instead, we are still in need of Kepler. Viewed most generally, GENESIS is intended to provide a framework for quantifying the physical description of the nervous system in a way that promotes common understanding of its physical structure. At the same time, this physical description also provides the base for simulations intent on understanding fundamental relationships between the structure of the brain and its measurable behavior. Again, looking back at the evolution of planetary science, Kepler's realization that the motion of the planets was elliptical came about as a result of his careful analysis of the detailed positions of the planets obtained by Tycho Brahe. Kepler's development of a mathematical formalism to describe elliptical motion provided a seminal framework for the work of later physicists, including Newton. Similarly, it is our hope and expectation that the formalism being developed within the GENESIS project and other simulation systems such as NEURON (see NEURON SIMULATION ENVIRONMENT) will provide a means for neurobiologists to collaboratively construct a physical description of the nervous system. We believe strongly that general principles of organization, function, and computation will only emerge once this description has been constructed.

GENESIS was designed from the beginning to allow the development of simulations at any level of complexity, from subcellular components and biochemical reactions to whole cells, networks of cells and systems-level models. The earliest GENESIS simulations were biologically realistic large-scale simulations of cortical networks (Wilson and Bower, 1992). The De Schutter and Bower

(1994a, 1994b) cerebellar Purkinje cell model is typical of a large, detailed single-cell model, with 4,550 compartments and 8,021 ionic conductances. GENESIS is now being used for large systems-level models of cerebellar pathways (Stricanne, Morissette, and Bower, 1998), and, at the other extreme, is increasingly being used to relate cellular and network properties to biochemical signaling pathways (Bhalla and Iyengar, 1999).

Although GENESIS continues to be widely used for single-cell modeling and for modeling small networks (see, e.g., HALF-CENTER OSCILLATORS UNDERLYING RHYTHMIC MOVEMENTS), we have seen a dramatic increase in the number of publications that report using GENESIS for large network models. We believe that this trend is largely due to the availability of our libraries of ion channels and complete cell models. A description of some notable large-scale network GENESIS simulations (many using parallel computers) that have been published recently can be found at <http://www.genesis-sim.org/GENESIS/research/genres.html>. This web page also contains links to a list of 170 papers based on research with GENESIS from groups outside of Caltech, and a summary of what various research groups are doing with GENESIS.

GENESIS is implemented in C, using the X Window System, and runs under most varieties of Unix, including Linux. There is also a parallel version of GENESIS (called PGENESIS) that runs on workstation networks, small-scale parallel computers, and large, massively parallel supercomputers. PGENESIS is being used for simulations that must be run many times independently (e.g., parameter searches), and for large-scale models (especially network models with thousands of neurons).

The GENESIS Design Philosophy

The objectives of this project were to reduce redundant software design efforts, establish standards for simulation technology, and provide a common base for the exchange of models and scientific information. The object-oriented nature of the software allows different modelers to easily exchange and reuse whole models or model components. GENESIS also includes a customizable user interface for use by modelers and educators. From the beginning, GENESIS was also designed to serve as an instructional tool, because our involvement in several educational projects had demonstrated that simulations could provide flexible and dynamic learning tools for neuroscience education (Bower and Beeman, 1998).

The design of the GENESIS simulator and interface is based on a building-block approach. Simulations are constructed from modules that receive inputs, perform calculations on them, and then generate outputs. Model neurons are constructed from these basic components, such as compartments (short sections of cellular membrane) and variable conductance ion channels. Compartments are linked to their channels and are then linked together to form multi-compartmental neurons of any desired level of complexity. Neurons may be linked together to form neural circuits. This object-oriented approach is central to the generality and flexibility of the system, as it allows modelers to easily exchange and reuse models or model components. In addition, it makes it possible to extend the functionality of GENESIS by adding new commands or simulation components to the simulator, without having to modify the GENESIS base code.

Neural systems are particularly amenable to this object-oriented approach because they typically consist of discrete components interacting in quite stereotyped ways, and because the different sim-

ulations tend to use similar neural components, display routines, numerical integration routines, and the like. This modularity means that it is possible to quickly construct a new simulation or to modify an existing simulation by changing modules that are chosen from a library or database of standard simulation components. Individual modules or linked assemblies of modules (such as compartments with channels, entire cells, or networks of cells) can be easily replicated.

Interacting with GENESIS

GENESIS uses a high-level simulation language to construct neurons and their networks. Commands may be issued either interactively to a command prompt, by use of simulation scripts, or through the graphical interface. A particular simulation is set up by writing a sequence of commands in the scripting language that creates the network itself and the graphical interface for a particular simulation. The scripting language and the modules are powerful enough that only a few lines of script are needed to specify a sophisticated simulation. The principal components of the simulation system and the various modes of interacting with GENESIS are shown in Figure 1.

The underlying level of the GENESIS user interface is the Script Language Interpreter (SLI). This is a command interpreter, similar to a Unix system shell, with an extensive set of commands related to building, monitoring, and controlling simulations. GENESIS simulation objects and graphical objects are linked together using the scripting language. The interpreter can read SLI commands either interactively from the keyboard (allowing interactive debugging, inspection, and control of the simulation) or from files containing simulation scripts.

The graphical user interface (GUI) is XODUS, the X-windows Output and Display Utility for Simulations. This provides a higher-level and user-friendly means for developing simulations and monitoring their execution. XODUS consists of a set of graphical

objects that are the same as the computational modules from the user's point of view, except that they perform graphical functions. As with the computational modules, XODUS modules can be set up in any manner that the user chooses to display or enter data. Furthermore, the graphical modules can call functions from the script language, so that the full power of the SLI is available through the graphical interface. This makes it possible to interactively change simulation parameters in real time to directly observe the effects of parameter variations. For example, the mouse can be used to plant recording or injection electrodes into a graphical representation of the cell. In addition to provisions for plotting the usual quantities of interest (membrane potentials, channel conductances, and so forth), XODUS has visualization features that permit such choices as using color to display the propagation of action potentials or other variables throughout a multicompartmental model, or to display connections and cell activity in a network model.

The GENESIS simulation engine (see Figure 1) consists of the simulator base code that provides the common control and support routines for the system, including those for input/output and for the numerical solution of the differential equations obeyed by the various neural simulation objects. GENESIS provides a choice of numerical integration methods, including highly accurate and stable implicit methods such as the Crank-Nicholson method (De Schutter and Beeman, 1998).

In addition to receiving commands from the SLI and the GUI, the simulation engine can construct simulations using information from data files and from the precompiled GENESIS object libraries. For example, the GENESIS "cell reader" allows one to build complex model neurons by reading their specifications from a data file instead of from a lengthy series of GENESIS commands delivered to the SLI. Similarly, network connection specifications may be read from a data file with the "fileconnect" command.

The GENESIS object libraries contain the building blocks from which many different simulations can be constructed. These in-

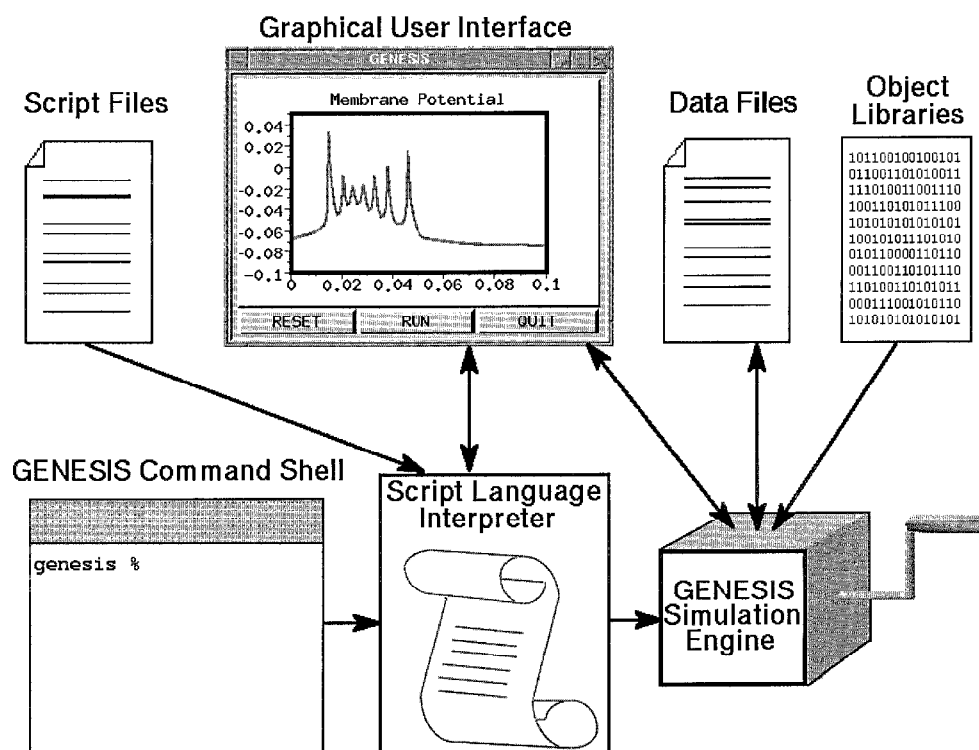


Figure 1. The components of GENESIS and modes of interaction. The Script Language Interpreter processes commands entered through the keyboard, script files, or the graphical user interface, and passes them to the GENESIS simulation engine. The simulation engine also loads compiled object libraries, reads and writes data files, and interacts with the GUI.

clude the spherical and cylindrical compartments from which the physical structure of neurons are constructed, voltage- and/or concentration-activated channels, dendrodendritic channels, and synaptically activated channels with synapses of several types, including Hebbian and facilitating synapses. In addition, there are objects for computing intracellular ionic concentrations from channel currents, for modeling the diffusion of ions within cells, and for allowing ligand gating of ion channels. There are also a number of "device objects" that may be interfaced to the simulation to provide various types of input to the simulation (e.g., pulse and spike generators, voltage clamp circuitry) or measurements (e.g., peristimulus and interspike interval histograms, spike frequency measurements, auto- and cross-correlation histograms).

The kinetics library supports kinetic-level modeling of biochemical pathways. This library currently includes objects for pools (molecular components), n -molecular reactions, enzymes, and channels to couple pools of different volume. The parameter search library provides a collection of objects and functions that automate the tedious process of adjusting model parameters to best reproduce experimental measurements carried out on the system being modeled.

GENESIS Script Libraries and Tools

In addition to the object libraries that are compiled into GENESIS, there are a number of libraries that are implemented as simulation scripts. These are available within the GENESIS distribution and in the archives of the GENESIS users group.

The channel library currently contains models for 39 different types of potassium channels (including several types of calcium-dependent channels), 24 types of sodium channels, and 14 types of calcium channels. The available single-cell models include cerebral cortical pyramidal cells, hippocampal pyramidal cells, cerebellar Purkinje cells, mitral, granule, and tufted cells from the olfactory bulb, a hippocampal granule cell model, a thalamic relay cell, and an *Aplysia* R15 bursting pacemaker cell.

GENESIS makes use of the GUI to provide other features to make the simulator more easily usable by people with limited programming experience. A set of kits, implemented as simulation scripts, has been provided to ease the modeling process. Neurokit provides an environment for building, modifying, and testing single-cell models without any programming on the part of the user. Kinetikit is a user-friendly, click-and-drag interface for modeling models of chemical reactions such as occur in biochemical signaling pathways. In addition to defining and running kinetic models, it is intended to facilitate managing kinetic data in these complex models (Bhalla, 1998; Bhalla and Iyengar, 1999).

GENESIS Documentation and Resources

GENESIS comes with extensive documentation. The GENESIS reference manual comes in three forms: a 566-page manual in Postscript format, and corresponding on-line help available either as plain text files viewable within the simulator or as hypertext help, which can be viewed with a web browser.

To complement the reference manual, we have published two editions of *The Book of GENESIS*. The most recent edition (Bower and Beeman, 1998) contains a CD-ROM with the GENESIS distribution, documentation, and files from the users group archives. It is widely used in both research and teaching, and consists of two parts serving complementary needs. Part I is designed to supplement instruction in neurobiology in upper division undergraduate and graduate neuroscience courses and includes chapters on various topics, each written by a known expert, to accompany a particular GENESIS tutorial. These interactive tutorial simulations are user-friendly, with on-line help, and may be used without any prior

knowledge of the GENESIS simulator or computer programming. Part II serves as a user's guide to GENESIS, complementing the GENESIS reference manual, by introducing the basic features of GENESIS as well as the process of creating GENESIS simulations, providing a starting point for the development of new simulations. (For further details, please see <http://www.genesis-sim.org/GENESIS/bog/bog.html>.)

The tutorials mentioned above are included in the GENESIS distribution, along with other tutorials and demonstrations designed to aid new users in building GENESIS simulations. These illustrate the use of advanced GENESIS features, including objects for modeling calcium diffusion; objects for spike analysis, recording, and generation; the use of facilitating and Hebbian synapses; objects for modeling stochastic ion channels; and the parameter search library. Several of the tutorials are based on significant published research simulations, including the piriform cortex model (Wilson and Bower, 1992), the hippocampal pyramidal cell model (Traub et al., 1991), and the detailed cerebellar Purkinje cell model (De Schutter and Bower, 1994a, 1994b). In addition to their educational purpose, these tutorial simulations provide examples of well-constructed GENESIS simulations, and they have been used by others as the basis for the construction of new published research simulations.

Individuals or research groups who make serious use of GENESIS are encouraged to join the GENESIS users group, BABEL. There are currently 347 BABEL memberships, representing approximately 700 users. Members of BABEL are entitled to access the BABEL directories and participate in the e-mail newsgroup. The directories are used as a repository for the latest contributions by GENESIS users and developers. Such contributions include new simulations, libraries of cells and channels, additional simulator components, new documentation and tutorials, bug reports and fixes, and the posting of questions and hints for setting up GENESIS simulations. As the results of GENESIS research simulations are published, many of these simulations are being made available through BABEL.

Use of GENESIS in Education

From its inception, GENESIS has had a strong educational component. We are currently aware of 49 institutions in 11 countries that have used GENESIS in teaching. Many of these institutions use GENESIS in association with *The Book of GENESIS* (Bower and Beeman, 1998).

GENESIS and the tutorials described in the previous sections are now being widely used in graduate and undergraduate instruction. Instructional options include full-semester courses in computational neuroscience or neural modeling, short intensive courses or workshops, course projects, and short units on computational neuroscience within courses on artificial neural nets. An example of the use of GENESIS tutorials as the basis for a short unit on neural modeling is available on the GENESIS web site as an HTML version of two lectures given at the University of Colorado.

GENESIS has also formed the basis for the laboratory section of the Methods in Computational Neuroscience course (1988–1996) at the Marine Biological Laboratory, a course in Mexico City in the summer of 1991, the Crete course in Computational Neuroscience (1996–1998), the EU Advanced Course in Computational Neuroscience (1999–2002), and the Computational Neuroscience course at the National Centre for Biological Sciences in Bangalore (1999–2000).

Discussion

In the many years since the GENESIS project began at Caltech, its development has spread to many other institutions, including the

University of Texas at San Antonio, the University of Antwerp, the National Centre for Biological Studies in Bangalore, the University of Colorado, the Pittsburgh Supercomputer Center, the San Diego Supercomputer Center, and Emory University. In addition, the tools available for GUIs and the decentralization of computational resources as a result of widespread use of the World Wide Web have dramatically changed the environment for computer-based education and research. In a collaborative effort involving many institutions, we are currently redesigning and reimplementing GENESIS in order to modernize the user interface and link the process of modeling with on-line databases of models and model components. Our efforts are taking place in the context of a new software framework called the Modeler's Workspace (Forss et al., 1999; Hucka et al., 2002) and a simulator-independent representation of neural models called NeuroML (Goddard et al., 2001).

Further information about GENESIS and PGENESIS, as well as instructions for downloading and installation, may be obtained from the GENESIS web site, <http://www.genesis-sim.org/GENESIS/>. Inquiries concerning GENESIS should be addressed to genesis@genesis-sim.org.

Road Map: Implementation and Analysis

Related Reading: Neurosimulation: Tools and Resources

References

- Bhalla, U. S., 1998, The network within: Signalling pathways, in *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System* (J.M. Bower and D. Beeman, Eds.), 2nd ed., New York: Springer-Verlag, pp. 169–191.
- Bhalla, U. S., and Iyengar, R., 1999, Emergent properties of networks of biological signaling pathways, *Science*, 283:381–387.
- Bower, J. M., 1995, Reverse engineering the nervous system: An in vivo, in vitro, and in computo approach to understanding the mammalian olfactory system, in *An Introduction to Neural and Electronic Networks*, (S. F. Zornetzer, J. L. Davis, and C. Lau, Eds.), 2nd ed., New York: Academic Press, pp. 3–28. ♦
- Bower, J. M., and Beeman, D., 1998, *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System*, 2nd ed., New York: Springer-Verlag. ♦
- De Schutter, E., and Beeman, D., 1998, Speeding up GENESIS simulations, in *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System* (J.M. Bower and D. Beeman, Eds.), 2nd ed., New York: Springer-Verlag, pp. 329–447.
- De Schutter, E., and Bower, J. M., 1994a, An active membrane model of the cerebellar Purkinje cell: I. Simulation of current clamps in slice, *J. Neurophysiol.*, 71:375–400.
- De Schutter, E., and Bower, J. M., 1994b, An active membrane model of the cerebellar Purkinje cell: II. Simulation of synaptic responses, *J. Neurophysiol.*, 71:401–419.
- Forss, J., Beeman, D., Eickler-West, R., and Bower, J. M., 1999, The Modeler's Workspace: A Distributed Digital Library for Neuroscience, *Future Generation Computer Systems*, vol. 16, pp. 111–121. ♦
- Goddard, N., Hucka, M., Howell, F., Cornelis, H., Shankar, K., and Beeman, D., 2001, Towards NeuroML: Model description methods for collaborative modelling in neuroscience, *Philos. Trans. R. Soc. Lond. B*, 356:1209–1228. ♦
- Hucka, M., Shankar, K., Beeman, D., and Bower, J. M., 2002, The Modeler's Workspace: Making model-based studies of the nervous system more accessible, in *Computational Neuroanatomy: Principles and Methods* (G. Ascoli, Ed.), Totowa, NJ: Humana Press. ♦
- Stricanne, B., Morissette, J., and Bower, J. M., 1998, Exploring the sources of cerebellar post-lesion plasticity with a network model of the somatosensory system, *Soc. Neurosci. Abs.*, 23:2364.
- Traub, R. D., Wong, R. K. S., Miles, R., and Michelson, H., 1991, A model of a CA3 hippocampal pyramidal neuron incorporating voltage-clamp data on intrinsic conductances, *J. Neurophysiol.*, 66:635–650.
- Wilson, M., and Bower, J. M., 1992, Simulating cerebral cortical networks: Oscillations and temporal interactions in a computer simulation of piriform (olfactory) cortex, *J. Neurophysiol.*, 67:981–995.

Geometrical Principles in Motor Control

Ferdinando A. Mussa-Ivaldi

Introduction

The central role played by geometry in the control of motor behaviors was recognized in the early years of the last century by Nikolai Bernstein. Using only the tool of logical reasoning applied to common observations, Bernstein reached the conclusion that “there exist in the higher levels of the CNS projections of space, and not projections of joint and muscles” (Bernstein, 1967). This intuition led others to consider motor planning and execution as separate processes.

The transition from the spatial representation of a motor goal to a set of neuromuscular commands is in many respects similar to a *coordinate transformation*. This analogy is the perspective of this article. We will begin by describing three types of coordinate systems, each one representing a particular point of view on motor behavior. Then, we will examine the geometrical rules that govern the transformations between these classes of coordinates. Finally, we will see how a proper representation of dynamics may greatly simplify the transformation of motor plans into actions.

Coordinate Systems for Motor Control

Endpoint Coordinates

Consider a monkey in the act of reaching for an apple with a wooden stick. The free extremity of the stick is the site at which the monkey interacts with its environment. We call such a site an *endpoint*. The position of the stick is fully determined by six coordinates. This is the smallest set of numbers needed to specify unambiguously the location and orientation of a rigid object in 3D space. The coordinates of the stick can be measured with respect to three orthogonal axes originating, for example, from the monkey's shoulder.

In our example a position in endpoint coordinates is a point

$$r = (x, y, z, \theta_x, \theta_y, \theta_z)$$

The coordinates, x , y , and z determine a translation with respect to the orthogonal axes. The angular coordinates, θ_x , θ_y , and θ_z , determine an orientation with respect to the same axes. Consistent with this notation, a *force* in endpoint coordinates is a vector with

three linear and three angular components:

$$F = (F_X, F_Y, F_Z, \tau_X, \tau_Y, \tau_Z)$$

Generalized Coordinates

A different way of describing the position of the monkey's arm is to provide the set of joint angles that define the orientation of each skeletal segment either with respect to fixed axes in space or with respect to the neighboring segments. Joint angles are a particular instance of *generalized coordinates*. According to the standard definitions of analytical mechanics, generalized coordinates are independent variables that are suitable for describing the dynamics of a system (Goldstein, 1980).

Once a set of generalized coordinates has been defined, one may also define a *generalized force*. For example, if one uses joint angles as generalized coordinates, the corresponding generalized forces are the torques measured at each joint. The dynamics of a mechanical system are described by differential equations relating the generalized coordinates to their first and second time derivatives and to the generalized forces.

In vector notation, the dynamics equations for the skeletal system of the monkey's arm can be written as

$$I(q)\ddot{q} + G(q, \dot{q}) = C(q, \dot{q}, u(t)) \quad (1)$$

where $q = (q_1, q_2, \dots, q_N)$ is the arm configuration in joint-angle coordinates, \dot{q} and \ddot{q} are, respectively, the first (velocity) and second (acceleration) time derivatives of q , I is an $N \times N$ matrix of inertia (that is configuration dependent) and $G(q, \dot{q})$ is a vector of centripetal and Coriolis torques (Sciavicco and Siciliano, 2000). The whole left-hand side of Equation 1 represents the torque due to the inertial properties of the arm. The term $C(\cdot)$ stands for the net torque generated nonlinearly by the muscles, by the environment (e.g., the gravitational torque), and by other dissipative elements, such as friction. The time-function $u(t)$ is a control vector—for example, a set of neural signals directed to the motoneurons or a representation of a desired limb position at time t (EQUILIBRIUM POINT HYPOTHESIS). Equation 1 may be regarded as a reformulation of Newton's law: $Ma = F$. The left side represents the passive dynamics associated with limb inertia. The right side is the applied force, which, in this case, is the output of a control process.

Actuator Coordinates

Actuator coordinates afford the most direct representation for the motor output of the central nervous system. A *position* in this coordinate system may be, for example, a collection of muscle lengths, $l = (l_1, l_2, \dots, l_M)$. Accordingly, a force in the same coordinate system is a collection of muscle tensions, $f = (f_1, f_2, \dots, f_M)$. The number of actuator coordinates depends on the level of detail of the model of control under consideration. *Unlike generalized coordinates, actuator coordinates do not constitute a system of mechanically independent variables*: one cannot set arbitrary values to l_i without eventually violating some kinematic constraint.

The Workspace and Its Transformations

Both the transformations from generalized coordinates to endpoint coordinates, and from generalized coordinates to actuator coordinates, are, in general, nonlinear mappings. In the case of the monkey's arm, the transformation from joint to hand coordinates is a nonlinear function

$$r = L(q) \quad (2)$$

where r indicates the position of the hand in endpoint coordinates and q is the joint configuration. The transformation from joint to

muscle coordinates is another nonlinear mapping

$$l = M(q) \quad (3)$$

We deal briefly with transformations between the different representations of the workspace in the framework of differential geometry. A modern tutorial on this subject can be found in Jose and Saitan (1998).

The Transformation of Vectors and Vector Fields

A function that associates a vector to each point of a multidimensional domain, M , is called a *vector field over M* . For example, one may rearrange the terms of the dynamics in Equation 1 so as to represent the arm's acceleration as a time-varying vector field over the state space described by q and \dot{q} :

$$\ddot{q} = I^{-1}(q)[C(q, \dot{q}, u(t)) - G(q, \dot{q})] \quad (4)$$

Another vector field describes the viscoelastic behavior of the arm muscles. This behavior can be measured by stimulating each muscle and recording the resulting tension at different muscle length rates of shortening and times. Then, the collective mechanical output of the skeletomotor system is summarized by a force field in muscle coordinates

$$f = \alpha(l, \dot{l}, t) \quad (5)$$

Vector fields such as those expressed in Equations 4 and 5 determine the way in which a system reacts to its environment on one hand and to its control signals on the other. To investigate the transformation from planning to control of actions, we must understand how such mechanical fields are affected by a change of coordinates.

Let us begin by considering the laws that govern the transformation of a point from a set of coordinates, x , into a new set of coordinates, \bar{x} . The coordinate transformation is a nonlinear function

$$\bar{x} = T(x) \quad (6)$$

We assume this function to be continuous and sufficiently differentiable (the existence and continuity of second partial derivatives is enough for most practical purposes.) However, we do not require the existence of an inverse mapping, $x = T^{-1}(\bar{x})$. In many biologically relevant cases the two coordinate systems have different dimensions and the inverse mapping is not defined uniquely or does not exist at all.

Next, consider how a vector field is related to the corresponding vector field in the new coordinate system. Let us begin by considering a field of velocity vectors, $\dot{x} = v(x)$, and apply the chain rule:

$$\dot{\bar{x}} = \frac{d\bar{x}}{dt} = \frac{\partial \bar{x}}{\partial x} \cdot \dot{x}$$

As we are dealing with multivariate functions, the expression $\partial \bar{x} / \partial x$ represents the functional derivative, or *Jacobian* of the transformation T . The Jacobian is a position-dependent matrix, $J(x)$, whose elements are

$$[J(x)]_{ij} = \frac{\partial \bar{x}_i}{\partial x_j}$$

and the transformation for the velocity vector can be rewritten as

$$\dot{\bar{x}} = J(x) \cdot \dot{x} \quad (7)$$

A vector that changes according to this law is said to be *contravariant*.

Does Equation 7 provide us with a rule for transforming the whole velocity field $v(x)$ into a new field $\bar{v}(\bar{x})$? The answer is generally negative. We may write \bar{x} as a function of \bar{x} :

$$\dot{\bar{x}} = J(T^{-1}(\bar{x}))v(T^{-1}(\bar{x})) = \bar{v}(\bar{x})$$

only if the mapping $T(x)$ can be inverted.

If a coordinate transformation cannot be inverted, we know how to transform a contravariant vector *at a given point* but we do not know how to transform a contravariant *field*.

The situation is quite different when dealing with vectors that in a change of coordinates transform like the gradient operator. Again, using the chain rule:

$$\frac{\partial}{\partial x} = \frac{\partial \bar{x}}{\partial x} \cdot \frac{\partial}{\partial \bar{x}}$$

that is,

$$\frac{\partial}{\partial x} = J^T(x) \cdot \frac{\partial}{\partial \bar{x}} \quad (8)$$

A vector following this type of transformation is said to be *covariant*. Note the dual or reciprocal nature of Equations 7 and 8. An infinitesimal displacement (or a velocity) is transformed by the Jacobian into the new coordinate system. The same Jacobian maps a covariant vector the other way around—from the new to the old coordinate system.

An example of a covariant vector is force, F . The covariance of force derives from the tensor invariance of work and power. *Work and power are indeed true scalar variables whose value is not modified by a change of coordinates.* In the original coordinate system, power is calculated as

$$F^T \dot{x}$$

and in the new coordinate system, as

$$\bar{F}^T \dot{\bar{x}}$$

By equating these two expressions and using Equation 7, we obtain

$$\bar{F}^T J(x) \dot{x} = F^T \dot{x}$$

Hence, the transformation of a force field, $\bar{F}(\bar{x})$, is

$$F(x) = J(x)^T \bar{F}(T(x)) \quad (9)$$

(compare with Equation 8).

Note that, here, the entire right side has been resolved in terms of x . We reach the important conclusion that unlike contravariant vectors, *covariant vectors transform globally, as fields*. No inverse transformation is required.

Transforming Plans into Action

When we plan a movement such as “trace the shape of a circle with the left hand,” we formulate the goal in terms of end point coordinates, without concern for the muscles that participate in the desired behavior. However, once we have decided to trace a circle, our brain must choose which muscles to activate and in which temporal sequence they should be activated. In carrying out this task, our brain must face the challenges associated with *kinematic redundancy*—the imbalance between the number of muscles, joints, and end point coordinates that is typical of any biological system. The issue of kinematic redundancy has attracted considerable attention both in robotics (Mussa-Ivaldi and Hogan, 1991) and in neural modeling (Hinton, 1984; Bullock, Grossberg, and Gunther, 1993).

The purpose of this section is to show how a proper representation of dynamics and of coordinate transformations leads to a simple solution for some problems associated with redundancy. The key for this solution lies in the representation of both the “high level” plans and the “low level” actuator actions as covariant fields of force.

The Transformation of Dynamics

The dynamics of a limb (Equation 1) are expressed as an equilibrium condition between the field of generalized forces, $D(q, \dot{q}, \ddot{q})$, which represent passive elements such as limb inertia and the field of time-varying generalized forces, and $C(q, \dot{q}, t)$, which are generated by the neuromuscular controller:

$$D(q, \dot{q}, \ddot{q}) = C(q, \dot{q}, t) \quad (10)$$

We formulate the problem of executing a motor plan as the problem of deriving a control field, $C(q, \dot{q}, t)$, that generates a desired end point behavior. *This task is carried out as an approximation problem, after transforming the covariant fields corresponding to the planned behavior and to the neuromuscular mechanics into their corresponding images in generalized coordinates.*

The first step consists of expressing the motor plans as fields of force in endpoint coordinates:

$$F = \pi(r, \dot{r}, t) \quad (12)$$

For example, a reaching movement of the hand may be planned, as proposed by Flash and Hogan (1985), by specifying a time-varying field whose equilibrium point moves along a smooth trajectory (EQUILIBRIUM POINT HYPOTHESIS). Following a similar approach, the movements of a robotic arm in an obstacle-ridden environment can be efficiently planned by associating a field of repulsive forces to each obstacle and a field of attractive forces to the target location (POTENTIAL FIELDS AND NEURAL NETWORKS).

From our earlier consideration of vector fields, we may conclude that the planned field, $\pi(r, \dot{r}, t)$, has a unique image, $C_\pi(q, \dot{q}, t)$, in configuration space. *This is true regardless of kinematic redundancy.* The only condition for C_π to be defined is that the kinematic transformation from generalized to endpoint coordinates be defined with its first partial derivatives. Operationally, we may construct C_π as a combination of three mappings:

$$C_\pi = l_2 \circ \pi \circ l_1 : (q, \dot{q}, t) \xrightarrow{l_1} (r, \dot{r}, t) \xrightarrow{\pi} F \xrightarrow{l_2} Q$$

The first mapping, l_1 , is the direct kinematics transformation ($q \rightarrow r$), its Jacobian (q, \dot{q}, \dot{r}), and the identity function ($t \rightarrow t$). The second mapping is the planned endpoint field, $\pi(r, \dot{r}, t)$. Finally, the third mapping, l_2 , is the transformation from endpoint to generalized force, which, again, is provided by the Jacobian of the direct kinematics.

As an example, consider the planning of an endpoint behavior, corresponding to a moving equilibrium position, $r_E(t)$, with linear stiffness and viscosity (that is a linear PD controller in endpoint coordinates):

$$\pi(r, \dot{r}, t) = K(r - r_E(t)) + B\dot{r}$$

The image of this endpoint field in joint coordinates is

$$C_\pi(q, \dot{q}, t) = J^T(q)K(L(q) - r_E(t)) + J^T(q)BJ(q)\dot{q}$$

Note that $r_E(t)$ is a time-varying input function that does not require a measure of the actual hand position. This derivation applies to kinematically redundant limbs, as only the direct transformations, L and J , are required.

On the other end of the planning/execution problem, one must deal with a number of actuators that, in any biological system exceeds the number of generalized coordinates. Microstimulation studies in frogs (Giszter, Mussa-Ivaldi, and Bizzi, 1993) and rats (Tresch and Bizzi, 1999) showed that the focal activation of inter-neuronal regions within the lumbar spinal cord impose a specific balance of muscle activations. This results in a force field

$$f = \alpha(l, \dot{l}, t) \sum_i \alpha_i \psi_i(l, \dot{l}, t) \quad (13)$$

This field has a well defined and measurable image in generalized coordinates, $\phi_\alpha(q, \dot{q}, t)$, which can be derived from the combination of three direct transformations:

$$\phi_\alpha = m_2 \circ \alpha \circ m_1 : (q, \dot{q}, t) \xrightarrow{m_1} (l, \dot{l}, t) \xrightarrow{\alpha} f \xrightarrow{m_2} Q$$

The first transformation, m_1 , is the actuator kinematics ($M : q \rightarrow l$) together with its Jacobian ($q, \dot{q} \rightarrow \dot{l}$) and the identity mapping (t). The second transformation is the actuator force field, $\alpha(l, \dot{l}, t)$, and the third transformation, m_2 , is again given by the Jacobian of M , $\mu(q) = \partial M(q)/\partial q$.

In generalized coordinates, the force field induced by a synergy of muscles is

$$\phi_\alpha(q, \dot{q}, t) = \sum_i \alpha_i \mu_i^T \psi_i(M(q), \mu(q)\dot{q}, t)$$

Once again, this derivation of the synergy image in generalized coordinates remains valid for redundant limb because no inverse transformation is involved.

Discussion

Microstimulation studies (Mussa-Ivaldi, Giszter, and Bizzi, 1994) as well as studies of reflex behaviors (Kargo and Giszter, 2000) have demonstrated that multiple muscle synergies can be combined by linear superposition to generate complex behaviors. In particular, Kargo and Giszter found that reflex responses to multiple cutaneous stimuli are accounted for by the linear superposition of the response fields triggered by each stimulus.

The finding of vector summation suggests that under descending supraspinal commands, the fields expressed by the spinal cord may form a broad repertoire:

$$\Gamma = \left\{ C_S(q, \dot{q}, t; c_\alpha) = \sum_\alpha c_\alpha \phi_\alpha(q, \dot{q}, t) \right\} \quad (14)$$

Each element of Γ is generated by the descending commands selecting a group of synergies through the weighting coefficients, c_α . Following this view, the neural control system may approximate a *target field* $C_\pi(q, \dot{q}, t)$ by finding the element of Γ that is closest to the target field. The approximating field may be obtained by least squares methods, that is by determining a set of coefficients, c_α , such that the norm

$$\|C_S(q, \dot{q}, t; c_\alpha) - C_\pi(q, \dot{q}, t)\|^2 \quad (15)$$

is at a minimum.

Field approximation has been directly applied to the generation of a desired trajectory, $q_D(t)$, in generalized coordinates (Mussa-Ivaldi and Bizzi, 2000). In this case, one may attempt to generate the appropriate controller by finding the parameters, c_α , which minimize the difference between passive dynamics and control field in Equation 10 that is by minimizing

$$\|D(q, \dot{q}, \ddot{q}) - C_S(q, \dot{q}, t; c_\alpha)\|^2 \quad (16)$$

along the desired trajectory. Since the parameters, c_α , appear linearly in C_S , this problem has a single global minimum at

$$c_\alpha = \sum_i [\Phi]_{\alpha,i}^{-1} \Lambda_i$$

with

$$\begin{cases} \Phi_{i,m} = \int \phi_i(q_D(t), \dot{q}_D(t), t) \cdot \phi_m(q_D(t), \dot{q}_D(t), t) dt \\ \Lambda_j = \int \phi_j(q_D(t), \dot{q}_D(t), t) \cdot D(q_D(t), \dot{q}_D(t), \ddot{q}_D(t)) dt \end{cases} \quad (17)$$

The symbol \cdot indicates the ordinary inner product.

The underlying idea of this method is that if the residual force error (16) could be reduced to zero, then the corresponding controller would produce exactly the desired trajectory. If, instead, there is a non-zero residual, then the problem of generating acceptable approximations becomes a problem of local stability: as residual forces may be regarded as a perturbation of the dynamics, one needs to insure that this perturbation does not lead to a motion that diverges from the desired trajectory. A study by Lohmiller and Slotine (1998) showed that the combination of control modules is stable if the modules are “contracting”—a condition germane to exponential stability.

If the modules corresponding to muscle synergies are stable, then the possibility of combining them provide the central nervous system with something equivalent to a movement’s representation. The movements of a limb can be considered as “points” in an abstract geometrical space, where the force fields produced by a set of modules play a role equivalent to that of coordinate axes and the selection parameters that generate a particular movement may be regarded as generalized projections of this movement along these axes.

The theoretical view of motor control as a form of function approximation has found support in recent studies of adaptive learning. In a set of elegant experiments, Thoroughman and Shadmehr (2000) have asked subjects to execute movements of the hand against a field of perturbing forces. Subjects learned gradually to compensate these forces, thus recovering the normal kinematics of reaching movements. However, if the forces were suddenly suppressed during “catch trials,” subjects showed a transient loss of learning that affected the following movements in variable amounts, depending on the angle between the movement in the catch trial and the following movement. Thoroughman and Shadmehr were able to reproduce the process of adaptation as well as the subtle effects of catch trials by assuming that the motor control system composed a representation of the disturbing field as a linear superposition of Gaussian primitives. These primitives encode the force generated in response to a velocity, with a narrow variance parameter (approximately 10 cm/sec). The analysis of motor learning, together with the study of motor primitives in the spinal cord has provided us with a strong support for a theoretical view based on the idea that the central nervous system generate and update a broad spectrum of behaviors by combining elementary building blocks. The mathematical language of force fields and of their transformations is the proper framework for relating this computational mechanism to its observable mechanical effects.

Road Map: Mammalian Motor Control

Related Reading: Equilibrium Point Hypothesis; Eye-Hand Coordination in Reaching Movements; Limb Geometry, Neural Control; Motor Primitives

References

- Bernstein, N., 1967, *The Coordination and Regulation of Movement*, Oxford, UK: Pergamon Press.
- Bullock, D., Grossberg, S., and Guenther, F. H., 1993, A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *J. Cognit. Neurosci.*, 5:408–435.
- Flash, T., and Hogan, N., 1985, The coordination of arm movements: An experimentally confirmed mathematical model, *J. Neurosci.*, 5:1688–1703.
- Giszter, S. F., Mussa-Ivaldi, F. A., and Bizzi, E., 1993, Convergent force fields organized in the frog’s spinal cord, *J. Neurosci.*, 13:467–491.
- Goldstein, H., 1980, *Classical Mechanics*, Reading, MA: Addison-Wesley.
- Hinton, G., 1984, Parallel computations for controlling an arm, *J. Motor Behav.*, 16:171–194.
- Jose, J. V., and Saletan, E. J., 1998, *Classical Dynamics: A Contemporary Approach*, Cambridge, UK: Cambridge University Press.

- Kargo, W. J., and Giszter, S. F., 2000, Rapid correction of aimed movements by summation of force-field primitives, *J. Neurosci.*, 20:409–426.
- Lohmiller, W., and Slotine, J.-J. E., 1998, On contraction analysis for nonlinear systems, *Automatica* 34:683–696. ♦
- Mussa-Ivaldi, F. A., and Hogan, N., 1991, Integrable solutions of kinematic redundancy via impedance control, *Int. J. Robotics Res.*, 10:481–491.
- Mussa-Ivaldi, F. A., Giszter, S. F., and Bizzi, E., 1994, Linear combinations of primitives in vertebrate motor control, *Proc. Natl. Acad. Sci. USA*, 91:7534–7538.
- Mussa-Ivaldi, F. A., and Bizzi, E., 2000, Motor learning through the combination of primitives, *Phil. Trans. Roy. Soc. Lond. B*, 355:1755–1769. ♦
- Sciavicco, L., and Siciliano, B., 2000, *Modeling and Control of Robot Manipulators*, New York: Springer Verlag.
- Thoroughman, K. A., and Shadmehr, R., 2000, Learning of action through adaptive combination of motor primitives, *Nature*, 407:742–747.
- Tresch, M. C., and Bizzi, E., 1999, Responses to spinal microstimulation in the chronically spinalized rat and their relationship to spinal systems activated by low threshold cutaneous stimulation, *Exp. Brain Res.*, 129:401–416.

Global Visual Pattern Extraction

Hugh R. Wilson and Frances Wilkinson

Introduction

Decades of research have established that visual pattern recognition begins with the extraction of local edge and contour information by orientation-selective simple cells in primary visual cortex (V1). At the highest levels of cortical form vision in inferior temporal cortex (IT), many neurons are sensitive to complex global patterns, including objects and faces (Desimone, 1991). This raises the key question “What processes occur at intervening stages of the form vision pathway to transform local V1 orientation information into global pattern representations?” It is known that the ventral form vision pathway includes at least areas V1, V2, V4, TEO, and TE (the highest level of IT), so there must be a sequence of transformations. Furthermore, mean receptive field size increases in diameter by roughly a factor of 2.5–2.7 from area to area in this processing hierarchy (Boussaoud, Desimone, and Ungerleider, 1991; Kobatake and Tanaka, 1994). Thus, a mean foveal receptive field diameter of about 0.4° in V1 is transformed into a mean of about 3.0° in V4 and about 15.0° – 20.0° in TE. Clearly, such large receptive fields must be combining information from many V1 neurons, but what sorts of combinations actually occur?

Essentially the same question may be posed in cortical motion processing along the dorsal pathway comprising V1, V2, MT, MST, and higher parietal areas. V1 neurons extract only local motion vectors perpendicular to moving edge segments, while MST neurons are sensitive to complex optic flow patterns, including expansion and rotation (Tanaka and Saito, 1989). In the dorsal pathway, receptive field diameter also grows by a factor of 2.5–2.7 from area to area. Thus analogous questions are raised about transitions from local to global processing in both motion and form vision. This article suggests answers to these questions at intermediate levels of these two pathways, primarily V4 and MST.

Global Processes in V4

Although primate V4 was originally believed to be a color vision area, more recent lesion studies have shown that it represents a major intermediate level of the cortical form vision system. Early physiological studies of V4 typically used the bar and grating stimuli that had proved so fruitful in elucidating orientation and spatial frequency selectivity in V1. Such stimuli, however, mainly revealed powerful end inhibition in V4. In a novel approach to V4, Gallant, Braun, and Van Essen (1993) used concentric, radial, and hyperbolic grating stimuli as well as conventional sinusoidal gratings. While many neurons responded well to all stimulus groups, two new groups were found: one responding optimally to concentric gratings and one responding optimally to radial or hyperbolic gratings. Very few neurons responded significantly better to con-

ventional gratings than to the other classes. Using a very different approach, Kobatake and Tanaka (1994) also found that many V4 neurons were selective for concentric, radial, or cross-shaped stimuli.

In an attempt to relate these results to human form vision, it was natural to study Glass patterns, which can convey concentric (Figure 1), radial, and other structures. These patterns are normally generated by randomly placing dot pairs of fixed separation on the pattern such that the orientation defined by the pair is tangent to contours of the desired global pattern (e.g., concentric circles in Figure 1). By randomizing some percentage of the dots, the global structure conveyed by Glass patterns can be degraded. When psychophysical thresholds were measured in this manner, it was discovered that humans are most sensitive to concentric structure, somewhat less sensitive to radial or hyperbolic structure, and least

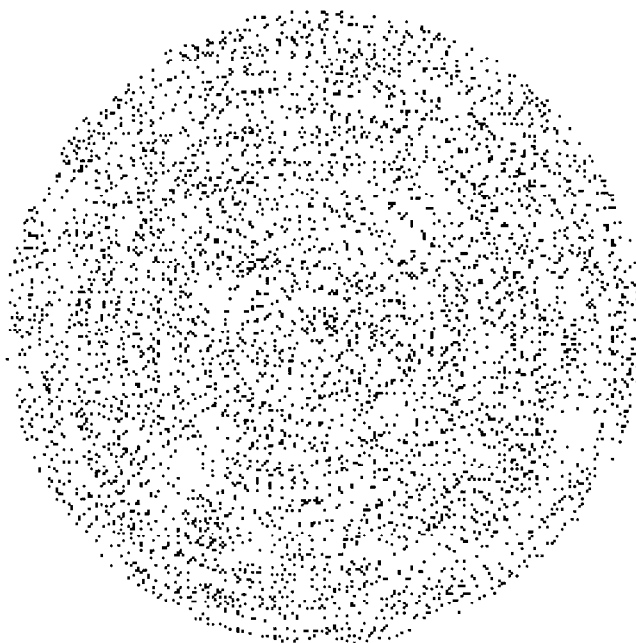


Figure 1. Concentric random dot Glass pattern. The global concentric structure in this pattern is produced by randomly positioning pairs of dots of fixed separation such that the orientation of the pair is locally tangent to a circle (not visible) concentric with the center of the pattern. Radial, hyperbolic, and parallel Glass patterns can be produced analogously.

sensitive to parallel or translational structure (Wilson, Wilkinson, and Asaad, 1997). Further experiments showed that the sensitivity to concentric structure resulted from linear pooling of concentric orientation information within an area estimated to be about 3.0° – 4.0° in diameter. Similar results were obtained in experiments with radial Glass patterns (Wilson, 1999b).

The psychophysical data are contrary to what would be expected for V1 but consistent with primate V4 physiology (Gallant et al., 1993; Kobatake and Tanaka, 1994). Accordingly, we asked whether human V4 might also show similar stimulus selectivity. After localizing V1, V4, and the fusiform face area using standard techniques, fMRI responses were measured while viewing concentric, radial, and sinusoidal gratings (Wilkinson et al., 2000). As predicted, the data showed that V1 activation was the same for all three stimulus patterns, while the fMRI signals in V4 were significantly higher for concentric and radial gratings than for sinusoidal gratings (Wilkinson et al., 2000). Interestingly, only concentric gratings and faces produced significant activation of the fusiform face area, suggesting that analysis of concentric structure represents one component of face perception.

Network Model

These data lead to the conclusion that human V4 is selectively sensitive to concentric and radial stimuli (and certainly other configural patterns), as is primate V4. The simple neural network diagrammed in Figure 2 describes a V4 concentric unit model consistent with these data (Wilson et al., 1997; Wilson, 1999b). Three stages make up the model, and these are hypothesized to represent V1, V2, and finally V4 processing. First, the stimulus is processed by oriented filters with properties of V1 simple cells (12 preferred orientations in 15° increments were used). Following this are contrast gain control and full-wave rectification operations for which there is experimental evidence. Next the responses are processed by larger second-stage filters oriented at right angles to their V1 inputs. This combination of filtering, rectification, and orthogonal filtering generates an end-stopped complex cell model that is sensitive to contour curvature (Wilson, 1999b). Finally, V2 responses that are concentric with the center of each V4 receptive field (gray circles) are summed linearly (Σ), and the result is passed through a threshold nonlinearity. An analogous model for V4 radial units has been constructed by simply changing the orientations of the

V1 filters to be parallel to the V2 filters. These V4 models are configural in the sense that they pool all V1 orientations, but each from pattern-specific subregions of the V4 receptive field.

This model offers an explanation for the 2.5- to 2.7-fold increase in receptive field sizes from V1 to V2 and from V2 to V4. The V2 filters must be about this much larger than V1 filters to effectively process curvature, while the V4 receptive fields must sum V2 responses over a diameter around 3.0 times the V2 filter diameter to extract circular or radial structure. Smaller increases in receptive field size from V1 to V4 simply cannot extract the relevant configural information.

The fMRI discovery that concentric gratings activate the fusiform face area is consistent with the hypothesis that V4 concentric units constitute an intermediate stage in face perception. To test this idea, the model in Figure 2 was applied to a variety of faces. For example, model convolution with the transparent face-house image in Figure 3A produced a single peak of activation centered at the black dot (Wilson, Krupa, and Wilkinson, 2000a). Furthermore, the response of this unit was proportional to the mean radius of the head, as indicated by the vertical arrow. This ability of model V4 units to *measure* concentric image structure is produced by the contrast gain control following the V1 filters. This causes the final summation stage Σ to add the number of units active around the pattern circumference, thus producing a signal proportional to radius and independent of contrast over a considerable range (Figure 3B). The model is also capable of encoding aspects of head shape, including axis of elongation and bilateral symmetry, as a sparse population code (Wilson et al., 2000b).

Global Unit Dynamics and Attention

There is considerable evidence that selective attention affects V4 neuron responses, and the evidence suggests that this results from biasing of competitive inhibitory networks (Reynolds, Chelazzi, and Desimone, 1999). Further evidence for such competitive inhibition among V4 concentric units has emerged from analysis of a visual illusion first introduced by Marroquin in 1978 (Wilson et al., 2000a). Marroquin patterns generate percepts of illusory circles appearing and vanishing at multiple locations within the pattern, much like the dynamic fluctuations in binocular rivalry. Psycho-physical measurement of the visibility times of illusory circles in

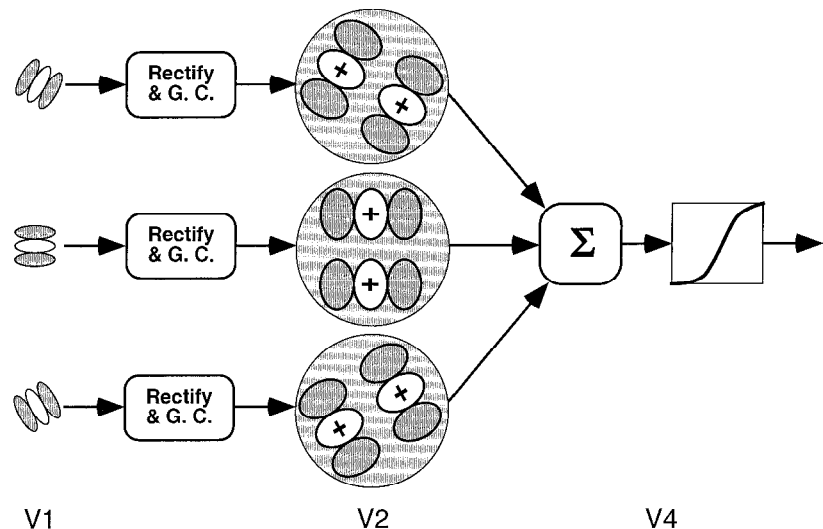


Figure 2. Global pooling model for a V4 concentric unit. Convolution of the stimulus with 12 different oriented simple cell filters (only three shown for clarity) having elongated excitatory (white) and inhibitory (gray) zones makes up the V1 stage. This is followed by full-wave rectification and a contrast gain control. V2 processing incorporates filtering by oriented filters that are 2.5–2.7 times the diameter of V1 filters. Finally, responses of concentrically arranged V2 filters are summed (Σ) and passed through a threshold nonlinearity to produce the simulated V4 response. This produces a V4 receptive field size (large gray circles) about 3.0 times that of V2 units, in agreement with cortical physiology.

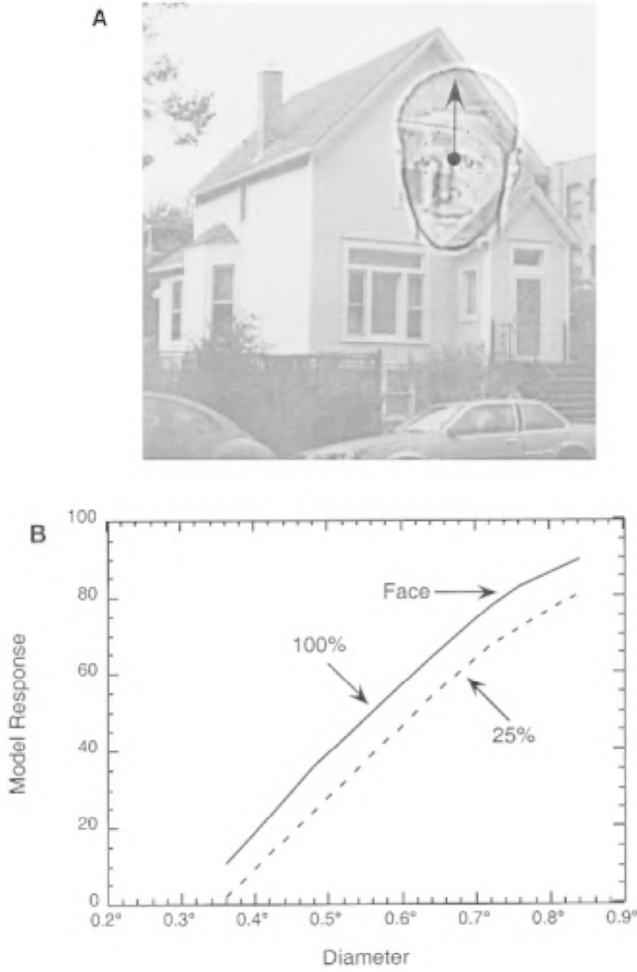


Figure 3. Image processing of a complex scene by the V4 model in Figure 2. *A*, Convolution of the model with a transparent face-house scene produces activation only at the center of the head (black dot), showing that the rectilinear house structure fails to affect model extraction of the ellipsoidal head shape. *B*, Response of maximally active V4 model unit in response to circles of different diameters. Owing to operation of the model contrast gain control, model responses are almost linear functions of diameter over more than a 2:1 size range despite contrast variations from 25% to 100%. The maximum model response to the head in *A* is indicated, and the radius estimated from the graph is plotted as an arrow in *A*.

the Marroquin pattern shows that they typically appear for a mean duration of 2.79 ± 2.75 s and are well described by a gamma distribution (Wilson et al., 2000a).

The V4 concentric unit model was extended to explain data on illusory circle visibility in Marroquin patterns by incorporating nonlinear dynamics and competitive inhibition. Individual neurons were simulated by using Wilson-Cowan-type spike rate equations in which a sigmoid function of postsynaptic potential P is approached exponentially in time (Wilson, 1999a):

$$\tau \frac{dE}{dt} = -E + \frac{MP_+^2}{\sigma^2 + P_+^2} \quad (1)$$

where the nonlinearity is a Naka-Rushton function and $P_+ = \max\{P, 0\}$. The maximum spike rate is conventionally $M = 100$, and σ is the semisaturation constant, that is, the value at which the Naka-Rushton function attains the value $M/2$.

Competitive inhibition was introduced by using the recurrent inhibitory spread function $\exp(-R^5/\sigma^5)$, where R is radius between competitors. This network implements a *spatially regional* winner-take-all competition. This means that the most strongly stimulated unit within each mutually inhibitory region will switch on while all others in that region are suppressed by inhibition. As the region is spatially limited, however, there will be multiple regional winners in the network.

Such networks can produce oscillations (rather than approaching a stable equilibrium) only if the regional winners adapt slowly until other winners emerge (Wilson, 1999a). Excitatory neocortical neurons are known to possess ion currents (typically Ca^{2+} mediated K^+ hyperpolarizations) that cause spike frequency adaptation. This reduces excitatory firing rates by a factor of about 3.0 within several hundred milliseconds following stimulus onset. The following equations describe activity in a network with such adaptation (Wilson, 1999a):

$$\tau_E \frac{dE_n}{dt} = -E_n + \frac{100P_+^2}{(10 + H_n)^2 + P_+^2}$$

where

$$P = S_{\text{Marroquin}} - 0.6 \sum_{k \neq n} I_k \exp\left(-\frac{R_{nk}^5}{\sigma^5}\right)$$

$$\tau_I \frac{dI_n}{dt} = -I_n + E_n$$

$$\tau_H \frac{dH_n}{dt} = -H_n + gE_n \quad (2)$$

E_n is the response of an excitatory neuron with a postsynaptic potential P that is the difference between stimulation S derived from the Marroquin pattern minus the spatially weighted inputs from inhibitory neurons I_k , where $n \neq k$, so there is no inhibition of a neuron by the inhibitory neuron it drives. The hyperpolarizing variable H_n produces spike frequency adaptation by increasing the semisaturation constant of the sigmoid nonlinearity in the first equation. Reasonable time constants are $\tau_E = 16$ ms, $\tau_I = 8$ ms, and $\tau_H = 400$ ms, the latter reflecting the much slower rate of spike frequency adaptation.

If Equation 1 is restricted to the case of two mutually inhibitory neurons, one can prove analytically that the system undergoes a Hopf bifurcation to a limit cycle oscillation at a critical value of the hyperpolarizing gain g (Wilson, 1999a). To simulate the illusory oscillating circles of the Marroquin illusion, the equations were extended to a 64×64 neuron array, and input was generated by applying the V4 concentric unit model in Figure 2 to a Marroquin pattern. The resulting model dynamics produced a gamma distribution of visibility durations with mean 2.24 ± 1.93 s (Wilson et al., 2000a), which agrees well with the human data. As Equation 1 represents a totally deterministic network without noise, it might be conjectured that the gamma distribution generated by the network reflects chaotic dynamics. However, the largest Lyapunov exponent was negative, so the dynamics are not chaotic; they apparently represent a very complex, long-period limit cycle.

It is gratifying that model V4 concentric units plus the regional competitive inhibition engendered in Equation 2 can predict an illusion that had remained unexplained for over 20 years. However, the significance of this network probably lies in its relationship to selective attention, which is evident in V4 and is thought to involve biasing of neural competition (Reynolds et al., 1999). Spike frequency adaptation in cortical neurons is controlled by modulatory neurotransmitters (serotonin, dopamine, and histamine), which *reduce* adaptation magnitude. Thus, modulatory transmitters function to tune network parameters, here the hyperpolarizing gain g in Equation 2. Reduction of g for a few excitatory neurons in the

network gives them an attentional advantage in subsequent competition. This suggests the hypothesis that biasing in V4 and other cortical areas may result from modulatory control of hyperpolarizing potentials in excitatory cells.

Global Processing in Motion Networks

The V1-V2-MT-MST dorsal motion pathway also shows progressively increasing receptive field size from area to area, with MST receptive fields averaging 50° in diameter (Tanaka and Saito, 1989). As in the ventral form vision pathway, this is indicative of progressively more global processing as information progresses through the hierarchy. Indeed, direction selective neurons in V1 have relatively small receptive fields and respond only to the component of motion perpendicular to local contour orientation. MT neurons pool over a larger visual area, combining V1 and V2 motion vectors over a range of about $\pm 90^\circ$ to determine the direction of pattern or object motion (Wilson, 1999b). (The restriction of vector pooling to $\pm 90^\circ$ reflects the ecological constraint that motion vectors in opposite directions cannot result from motion of a single rigid object.) Finally, neurons in MST combine MT responses over broad areas to extract expansion and rotation components of optic flow (Tanaka and Saito, 1989).

Psychophysical evidence that similar motion expansion and rotation units exist in human vision was provided in experiments using a motion analog of Glass patterns (Morrone, Burr, and Vaina, 1995). To generate a percept of rotary motion, for example, a circular patch of random dots is flashed on the screen and followed by a second flashed patch in which each dot is moved a fixed distance around a circular contour from its previous position. If dots are moved outward in the second frame relative to the first, motion expansion is perceived, and so on. Using this approach, Morrone et al. (1995) demonstrated that humans were extremely good at detecting rotary, expanding, and translational motion. Further experiments showed that the visual system globally summed motion vectors directed radially outward in detecting motion expansion. Similarly, clockwise or counterclockwise motion vectors were summed to detect clockwise or counterclockwise rotation, respectively.

These experiments demonstrate global, configural motion summation and are a direct motion analogue of global, configural orientation summation in the detection of concentric and radial Glass patterns (Wilson, 1999b). In consequence, a configural model analogous to the V4 configural model in Figure 2 may be applied to motion. The first stage of a configural motion model would incorporate V1 direction selective Reichardt or motion energy units rather than oriented simple cells. Following rectification, there would be pooling over larger areas to extract local object motions in MT. Finally, there would be configural summation of appropriate MT responses throughout large regions to extract expanding, radial, or translational optic flow.

Discussion

The evidence above indicates that similar global processes in higher cortical areas operate on local V1 orientation responses to extract shape information and on local V1 direction selective responses to detect optic flow patterns. Furthermore, the progressive enlargement of receptive field sizes in moving up either hierarchy is consistent with the requirements of global, configural processing. This leads to the natural question "Just how far do such global summation processes extend in cortical vision?" Certainly, receptive field size continues to increase from V4 to TEO and thence to TE (Boussaoud et al., 1991; Kobatake and Tanaka, 1994). Also, many TE receptive fields have been shown to respond to a complex object such as a face over a range of sizes and locations within visual space (Desimone, 1991). In principle, such size and position

invariance can be generated by obvious extensions of the V4 configural model in Figure 2. Given the evidence for regional competitive inhibition in V4, a larger receptive field sensitive to concentric shape independent of position can be produced by summing V4 concentric unit responses over an area similar to that of the competitive inhibition in Equation 2. Similarly, size invariance can be produced by replicating the V4 model at several different spatial scales, each about an octave apart, allowing regional competition within each scale, and then summing responses across spatial scales in an area beyond V4. Thus, both size and position invariance can be explained by a further stage of global processing with inhibition.

Units that are responsive to global forms such as human and monkey faces have been hypothesized to emerge from configural pooling of appropriate V4 unit responses (Kobatake and Tanaka, 1994). Support for this is provided by the fMRI finding that concentric patterns, which are very effective in activating human V4, are also effective stimuli in the fusiform face area (Wilkinson et al., 2000). Combination of V4 concentric responses with responses of units encoding, for example, the configuration of eyes and nose, could produce face-selective neurons in TE.

Cortical feedback between areas (e.g., V2 to V1, V4 to V2) poses an unsolved problem for models of global pooling, as the models presented here contain no such feedback. Lamme (1995) has shown that V1 responses are enhanced after a 30- to 40-ms latency period if the stimulus is inside an object rather than part of the background, and he has conjectured that this reflects extrastriate feedback. As model V4 concentric units code the location and radius of ellipsoidal regions (see Figure 3), V4 feedback could enhance activity in bounded V1 regions. However, excitatory feedback of this sort, if unchecked, can result in self-organization of network activity into a steady state indicating a hallucination. As hallucinations of circles and faces are common in Charles Bonnet syndrome (Schultz and Melzack, 1993), one can speculate that they result from feedback between cortical areas, including V4. Charles Bonnet syndrome may thus provide a glimpse into the interplay between cortical feedback and global pattern extraction.

Road Map: Vision

Related Reading: Cortical Population Dynamics and Psychophysics; Motion Perception, Elementary Mechanisms; Object Recognition, Neurophysiology

References

- Boussaoud, D., Desimone, R., and Ungerleider, L. G., 1991, Visual topography of area TEO in the macaque, *J. Comp. Neurol.*, 306:554–575.
- Desimone, R., 1991, Face selective cells in the temporal cortex of monkeys, *J. Cogn. Neurosci.*, 3:1–8. ◆
- Gallant, J. L., Braun, J., and Van Essen, D. C., 1993, Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex, *Science*, 259:100–103.
- Kobatake, E., and Tanaka, K., 1994, Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex, *J. Neurophys.*, 71:856–867.
- Lamme, V. A. F., 1995, The neurophysiology of figure-ground segregation in primary visual cortex, *J. Neurosci.*, 15:1605–1615.
- Morrone, M. C., Burr, D. C., and Vaina, L. M., 1995, Two stages of visual processing for radial and circular motion, *Nature*, 376:507–509.
- Reynolds, J. H., Chelazzi, L., and Desimone, R., 1999, Competitive mechanisms subserve attention in macaque areas V2 and V4, *J. Neurosci.*, 19:1736–1753.
- Schultz, G., and Melzack, R., 1993, Visual hallucinations and mental state: A study of 14 Charles Bonnet syndrome hallucinators, *J. Nerv. Ment. Dis.*, 181:639–643.
- Tanaka, K., and Saito, H., 1989, Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey, *J. Neurophysiol.*, 62:626–641.
- Wilkinson, F., James, T. W., Wilson, H. R., Gati, J. S., Menon, R. S., and Goodale, M. A., 2000, An fMRI study of the selective activation of

Graphical Models: Parameter Learning

Zoubin Ghahramani

Introduction

Graphical models combine graph theory and probability theory to provide a general framework for representing models in which a number of variables interact. Graphical models trace their origins to many different fields and have been applied in a wide variety of settings: for example, to develop probabilistic expert systems, to understand neural network models, to infer trait inheritance in genealogies, to model images, to correct errors in digital communication, and to solve complex decision problems. Remarkably, the same formalisms and algorithms can be applied to this wide range of problems.

Each node in the graph represents a random variable (or, more generally, a set of random variables). The pattern of edges in the graph represents the qualitative dependencies between the variables; the absence of an edge between two nodes means that any statistical dependency between these two variables is mediated via some other variable or set of variables. The quantitative dependencies between variables that are connected by edges are specified by means of parameterized conditional distributions, or, more generally, non-negative “potential functions.” The pattern of edges and the potential functions together specify a joint probability distribution over all the variables in the graph. We refer to the pattern of edges as the *structure* of the graph, while the parameters of the potential functions are simply called the *parameters* of the graph. In this article, we assume that the structure of the graph is given, and that our goal is to learn the parameters of the graph from data. Solutions to the problem of learning the graph structure from data are given in GRAPHICAL MODELS: STRUCTURE LEARNING (q.v.).

We briefly review some of the notation from GRAPHICAL MODELS: PROBABILISTIC INFERENCE (q.v.) that we will need to cover parameter learning in graphical models. More in-depth treatments of graphical models can be found in Pearl (1988), Heckerman (1996), Jordan (1999), and Cowell et al. (1999).

There are two main varieties of graphical model. *Directed graphical models*, also known as BAYESIAN NETWORKS (q.v.), represent the joint distribution of k random variables $\mathbf{X} = (X_1, \dots, X_k)$ by a directed acyclic graph in which each node i , representing variable X_i , receives directed edges from its set of parent nodes π_i . The semantics of a directed graphical model are that the joint distribution of \mathbf{X} can be factored into the product of conditional distributions of each variable given its parents. That is, for each setting \mathbf{x} of the variable \mathbf{X} ,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^k p(x_i|\mathbf{x}_{\pi_i}, \boldsymbol{\theta}_i) \quad (1)$$

This factorization formalizes the graphical intuition that X_i depends on its parents \mathbf{X}_{π_i} . Given its parents, X_i is statistically independent of all other variables that are not descendants of X_i . The set of parameters governing the conditional distribution that relates \mathbf{X}_{π_i} to X_i is denoted by $\boldsymbol{\theta}_i$, while the set of all parameters in the graphical model is denoted $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)$. Note that 1 is identical to Equation 1 in GRAPHICAL MODELS: PROBABILISTIC INFERENCE (q.v.), except that we have made explicit the dependence of the conditional distributions on the model parameters.

Undirected graphical models represent the joint distribution of a set of variables via a graph with undirected edges. Defining \mathcal{C} to be the set of maximal cliques (i.e., fully connected subgraphs) of this graph, an undirected graphical model corresponds to the statement that the joint distribution of \mathbf{X} can be factored into the product of functions over the variables in each clique:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C|\boldsymbol{\theta}_C) \quad (2)$$

where $\psi_C(\mathbf{x}_C|\boldsymbol{\theta}_C)$ is a potential function assigning a non-negative real number to each configuration \mathbf{x}_C of \mathbf{X}_C , and is parameterized by $\boldsymbol{\theta}_C$. An undirected graphical model corresponds to the graphical intuition that dependencies are transmitted via the edges in the graph: each variable is statistically independent of all other variables, given the set of variables it is connected to (i.e., its neighbors). Note again that we have reproduced Equation 2 from GRAPHICAL MODELS: PROBABILISTIC INFERENCE (q.v.), while making explicit the parameters of the potential functions.

The article is organized as follows. We start by concentrating on directed graphical models. In the next section, we discuss the problem of learning maximum likelihood (ML) parameters when all the variables are observed. The following section generalizes this problem to the case in which some of the variables are hidden or missing, and introduces the Expectation-Maximization (EM) algorithm. We then turn to learning parameters of undirected graphical models using both EM and IPF. Finally, we discuss the Bayesian approach, in which a posterior distribution over parameters is inferred from data.

Maximum Likelihood Learning from Complete Data

Assume we are given a data set \mathbf{d} of N independent and identically distributed observations of the settings of all the variables in our directed graphical model $\mathbf{d} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$, where $\mathbf{x}^{(n)} = (x_1^{(n)}, \dots, x_k^{(n)})$. The *likelihood* is a function of the parameters and is proportional to the probability of the observed data:

$$p(\mathbf{d}|\boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \quad (3)$$

We assume that the parameters are unknown and we wish to estimate them from data. We focus on the problem of estimating a single setting of the parameters that maximizes the likelihood formulated in Equation 3. (In contrast, the Bayesian approach to learning described in the last section starts with a prior distribution over the parameters $p(\boldsymbol{\theta})$ that is meant to capture background knowledge we may have about $\boldsymbol{\theta}$, and infers the posterior distribution over parameters given the data $p(\boldsymbol{\theta}|\mathbf{d})$, using Bayes's rule.) Equivalently, we can maximize the log likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{d}|\boldsymbol{\theta}) = \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \quad (4)$$

$$= \sum_{n=1}^N \sum_{i=1}^k \log p(x_i^{(n)}|\mathbf{x}_{\pi_i}^{(n)}, \boldsymbol{\theta}_i) \quad (5)$$

where the last equality makes use of the factorization (Equation 1) of joint distribution in the directed graphical model. If we assume that the parameters $\boldsymbol{\theta}_i$ governing the conditional probability distribution of X_i given its parents are distinct and functionally independent of the parameters governing the conditional probability distribution of other nodes in the graphical model, then the log likelihood decouples into a sum of local terms involving each node and its parents:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^k \mathcal{L}_i(\boldsymbol{\theta}_i) \quad (6)$$

where $\mathcal{L}_i(\theta_i) = \sum_n \log p(x_i^{(n)} | \mathbf{x}_{\pi_i}^{(n)}, \theta_i)$. Each \mathcal{L}_i can be maximized independently as a function of θ_i . For example, if the \mathbf{X} variables are discrete and θ_i is the conditional probability table for x_i given its parents, then the ML estimate of θ_i is simply a normalized table containing counts of each setting of X_i given each setting of its parents in the data set.

Maximum a posteriori (MAP) parameter estimation incorporates prior knowledge about the parameters in the form of a distribution $p(\theta)$. The goal of MAP estimation is to find the parameter setting that maximizes the posterior over parameters, $p(\theta | \mathbf{d})$, which is proportional to the prior times the likelihood. If the prior factorizes over the parameters governing each conditional probability distribution, i.e., $p(\theta) = \prod_i p(\theta_i)$, then MAP estimates can be found by maximizing

$$\mathcal{L}'(\theta) = \sum_{i=1}^k \mathcal{L}_i(\theta_i) + \log p(\theta) \quad (7)$$

The log prior can be seen as a regularizer, which can help reduce overfitting in situations where there are insufficient data for the parameters to be well-determined (see GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING SYSTEMS). Although ML estimation is invariant to reparameterization, since the location of the maximum of the likelihood function does not change if you apply a one-to-one transformation $f: \theta \rightarrow \phi$, MAP estimation is not. Indeed, for *any* $\tilde{\theta}$ one can always find a one-to-one mapping such that the MAP estimate of ϕ is $f(\tilde{\theta})$, as long as $p(\theta | \mathbf{d}) > 0$ in a small neighborhood around $\tilde{\theta}$. Thus, care should be taken in the choice of parameterization.

Maximum Likelihood Learning with Hidden Variables

The Expectation-Maximization (EM) Algorithm

Often, the observed data will not include the values of some of the variables in the graphical model. We refer to these variables as missing or hidden variables. With hidden variables, the log likelihood cannot be decomposed as in Equation 6. Rather, we find:

$$\mathcal{L}(\theta) = \log p(\mathbf{x} | \theta) = \log \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y} | \theta) \quad (8)$$

where \mathbf{x} denotes the setting of the observed variables, \mathbf{y} the setting of the hidden variables, and $\sum_{\mathbf{y}}$ is the sum (or integral) over \mathbf{Y} required to obtain the marginal probability of the observed data. (For notational convenience, we have dropped the superscript (n) in Equation 8 by evaluating the log likelihood for a single observation.) Maximizing Equation 8 directly is often difficult because the log of the sum can potentially couple all of the parameters of the model. We can simplify the problem of maximizing \mathcal{L} with respect to θ by making use of the following insight. Any distribution $q(\mathbf{Y})$ over the hidden variables defines a *lower bound* on \mathcal{L} :

$$\mathcal{L}(\theta) = \log \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y} | \theta) = \log \sum_{\mathbf{y}} q(\mathbf{y}) \frac{p(\mathbf{x}, \mathbf{y} | \theta)}{q(\mathbf{y})} \quad (9)$$

$$\geq \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y} | \theta)}{q(\mathbf{y})} \quad (10)$$

$$= \sum_{\mathbf{y}} q(\mathbf{y}) \log p(\mathbf{x}, \mathbf{y} | \theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log q(\mathbf{y}) \quad (11)$$

$$= \mathcal{F}(q, \theta) \quad (12)$$

where the inequality is known as Jensen's inequality and follows from the fact that the log function is concave. If we define the *energy* of a global configuration (\mathbf{x}, \mathbf{y}) to be $-\log p(\mathbf{x}, \mathbf{y} | \theta)$, then some readers may notice that the lower bound $\mathcal{F}(q, \theta) \leq \mathcal{L}(\theta)$ is the negative of a quantity known in statistical physics as the *free energy*: the expected energy under q minus the entropy of q (Neal

and Hinton in Jordan, 1999). The EM algorithm (Dempster, Laird, and Rubin, 1977) alternates between maximizing \mathcal{F} with respect to q and θ , respectively, holding the other fixed. Starting from some initial parameters θ_0 , the $\ell + 1$ st iteration of the algorithm consists of the following two steps:

$$\text{E step: } q_{[\ell+1]} \leftarrow \arg \max_q \mathcal{F}(q, \theta_{[\ell]}) \quad (13)$$

$$\text{M step: } \theta_{[\ell+1]} \leftarrow \arg \max_{\theta} \mathcal{F}(q_{[\ell+1]}, \theta) \quad (14)$$

It is easy to show that the maximum in the E step is obtained by setting $q_{[\ell+1]}(\mathbf{y}) = p(\mathbf{y} | \mathbf{x}, \theta_{[\ell]})$, at which point the bound becomes an equality: $\mathcal{F}(q_{[\ell+1]}, \theta_{[\ell]}) = \mathcal{L}(\theta_{[\ell]})$. This involves inferring the distribution over the hidden variables given the observed variables and the current settings of the parameters, $p(\mathbf{y} | \mathbf{x}, \theta_{[\ell]})$. Algorithms that solve this inference problem are presented in GRAPHICAL MODELS: PROBABILISTIC INFERENCE. These algorithms make use of the structure of the graphical model to compute the quantities of interest efficiently by passing local messages from each node to its neighbors. Exact inference results in the bound being satisfied, but is in general computationally intractable for multiply connected graphical structures. Even these "efficient" message-passing procedures can take exponential time to compute the exact solution in such cases. For such graphs, deterministic and Monte Carlo methods provide a tool for approximating the E step of EM. One deterministic approximation that can be used in intractable models is to increase but not fully maximize the functional with respect to q in the E step. In particular, if q is chosen to be in a tractable family of distributions \mathcal{Q} (i.e., a family of distributions for which the required expectations can be computed in polynomial time), then maximizing \mathcal{F} over this tractable family

$$\text{E step: } q_{[\ell+1]} \leftarrow \arg \max_{q \in \mathcal{Q}} \mathcal{F}(q, \theta_{[\ell]}) \quad (15)$$

is called a *variational approximation* to the EM algorithm (Jordan, Ghahramani, Jaakkola, and Saul in Jordan, 1999). This maximizes a lower bound to the likelihood rather than the likelihood itself.

The maximum in the M step is obtained by maximizing the first term in Equation 11, since the entropy of q does not depend on θ :

$$\text{M step: } \theta_{[\ell+1]} \leftarrow \arg \max_{\theta} \sum_{\mathbf{y}} p(\mathbf{y} | \mathbf{x}, \theta_{[\ell]}) \log p(\mathbf{y}, \mathbf{x} | \theta) \quad (16)$$

This is the expression most often associated with the EM algorithm (Dempster et al., 1977), but it obscures the elegant interpretation of EM as coordinate ascent in \mathcal{F} . Since $\mathcal{F} = \mathcal{L}$ at the beginning of each M step (following an exact E step), and since the E step does not change θ , we are guaranteed not to decrease the likelihood after each combined EM step.

It is usually not necessary to explicitly evaluate the entire posterior distribution $p(\mathbf{y} | \mathbf{x}, \theta_{[\ell]})$. Since $\log p(\mathbf{x}, \mathbf{y} | \theta)$ contains both hidden and observed variables in the network, it can be factored as before as the sum of log probabilities of each node given its parents (Equation 5). Consequently, the quantities required for the M step are the expected values, under the posterior distribution $p(\mathbf{y} | \mathbf{x}, \theta_{[\ell]})$, of the same quantities (namely the *sufficient statistics*) required for ML estimation in the complete data case.

Consider a directed graphical with discrete variables, some hidden and some observed. Each node is parameterized by a conditional probability table that relates its values to the values of its parents. For example, if node i has two parents, j and k , and each variable can take on L values, then θ_i is an $L \times L \times L$ table, with entries $\theta_{i,jst} = P(X_i = r | X_j = s, X_k = t)$. In the complete data setting where X_i, X_j, X_k are observed, the ML estimate is:

$$\hat{\theta}_{i,jst} = \frac{\#(X_i = r, X_j = s, X_k = t)}{\#(X_j = s, X_k = t)} \quad (17)$$

where $\#(\cdot)$ denotes the count (frequency) with which the bracketed expression occurs in the data. However, if all three variables were hidden, then one could use the EM algorithm for learning the directed graphical model (Lauritzen, 1995; Russell et al., 1995). The analogous M step for $\theta_{i,rst}$ would be:

$$\hat{\theta}_{i,rst} = \frac{\sum_n P(Y_i = r, Y_j = s, Y_k = t | \mathbf{X} = \mathbf{x}^{(n)})}{\sum_n P(Y_j = s, Y_k = t | \mathbf{X} = \mathbf{x}^{(n)})} \quad (18)$$

The sufficient statistics of the data required to estimate the parameters are the counts of the setting of each node and its parents; no other information in the data is relevant for ML parameter estimation. The *expectation* step of EM computes the expected value of these sufficient statistics.

The EM algorithm provides an intuitive way of dealing with hidden or missing data. The E step “fills in” the hidden variables with the distribution given by the current model. The M step then treats these filled-in values as if they had been observed, and reestimates the model parameters. It is pleasantly surprising that these steps result in a convergent procedure for finding the most likely parameters. Although EM is intuitive and often easy to implement, it is sometimes not the most efficient algorithm for finding ML parameters.

The EM procedure for learning directed graphical models with hidden variables can be applied to a wide variety of well-known models. Of particular note is the special case known as the Baum-Welch algorithm for training hidden markov models (Rabiner 1989; see HIDDEN MARKOV MODELS). In the E step it uses a local message-passing algorithm called the forward-backward algorithm to compute the required expected sufficient statistics. In the M step it uses a parameter reestimation equation based on expected counts, analogous to Equation 18. The EM algorithm can also be used to fit a variety of other models that have been studied in the machine learning, neural networks, statistics, and engineering literatures. These include linear dynamical systems, factor analysis, mixtures of Gaussians, and mixtures of experts (see Roweis and Ghahramani, 1999, for a review). It is straightforward to modify EM so that it maximizes the parameter posterior probability rather than the likelihood.

Parameter Learning in Undirected Graphical Models

When compared to learning the parameters of directed graphical models, undirected graphical models present an additional challenge: the partition function. Even if each clique has distinct and functionally independent parameters, the partition function from Equation 2,

$$Z(\theta) = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C | \theta_C) \quad (19)$$

couples all the parameters together. We examine the effect this coupling has in the context of an undirected graphical model that has had a great deal of impact in the neural networks field: the Boltzmann machine (Ackley, Hinton, and Sejnowski, 1985).

Boltzmann machines are undirected graphical models over a set of k binary variables $S_i \in \{0, 1\}$ (see also SIMULATED ANNEALING AND BOLTZMANN MACHINES). The probability distribution over the variables in a Boltzmann machine is given by

$$\begin{aligned} P(\mathbf{s} | W) &= \frac{1}{Z(W)} \exp \left\{ \frac{1}{2} \sum_{i=1}^k \sum_{j \in \text{ne}(i)} W_{ij} s_i s_j \right\} \\ &= \frac{1}{Z(W)} \prod_{(ij)} \exp \{ W_{ij} s_i s_j \} \end{aligned} \quad (20)$$

The first equation uses standard notation for Boltzmann machines, where W is the symmetric matrix of weights (i.e., model parameters)

and $\text{ne}(i)$ is the set of neighbors of node i in the Boltzmann machine. The second equation writes it as a product of clique potentials, where (ij) denotes the clique consisting of the pair of connected nodes i and j . (Actually, in Boltzmann machines, the maximal cliques in the graph may be very large, although the interactions are all pairwise; because of this pairwise constraint on interactions, we abuse terminology and consider the “cliques” to be the pairs, no matter what the graph connectivity is.)

Assuming that S_i and S_j are observed, taking derivatives of the log probability of the n th data point with respect to W_{ij} ,

$$\begin{aligned} \frac{\partial \log P(\mathbf{s}^{(n)} | W)}{\partial W_{ij}} &= s_i^{(n)} s_j^{(n)} - \sum_s s_i s_j P(\mathbf{s} | W) \\ &= \langle s_i s_j \rangle_n^+ - \langle s_i s_j \rangle^- \end{aligned} \quad (21)$$

we find that it is the difference between the correlation of S_i and S_j in the data and the correlation of S_i and S_j in the model (the $\langle \cdot \rangle$ notation means expectation). The standard ML gradient descent learning rule for Boltzmann machines therefore tries to make the model match the correlations in the data. The same learning rule applies if there are hidden variables.

The second term arises from the partition function. Note that even for fully observed data, although the first term can be computed directly from the observed data, the second term depends potentially on all the parameters, underlining the fact that the partition function couples the parameters in undirected models. Even for fully observed data, computing the second term is nontrivial; for fully connected Boltzmann machines it is intractable and needs to be approximated.

The IPF Algorithm

Consider the following problem: Given an undirected graphical model, and an initial set of clique potentials, we wish to find the clique potentials closest to the initial potentials that satisfy a certain set of consistent marginals. Closeness of probability distributions in this context is measured using the Kullback-Leibler divergence. The simplest example is a clique of two discrete variables, X_i and X_j , where the clique potential is proportional to the contingency table for the joint probability of these variables $P(X_i, X_j)$, and the marginal constraints are $P(x_i) = \hat{P}(x_i)$ and $P(x_j) = \hat{P}(x_j)$. These constraints could, for example, have come from observing data with these marginals. A very simple and intuitive iterative algorithm for trying to satisfy these constraints is to start from the initial table and satisfy each constraint in turn. For example, if we want to satisfy the marginal on X_i :

$$P(x_i, x_j) = P_{i-1}(x_i, x_j) \frac{\hat{P}(x_i)}{P_{i-1}(x_i)} \quad (22)$$

This has to be iterated over X_i and X_j , since satisfying one marginal can change the other marginal. The simple algorithm is known as Iterative Proportional Fitting (IPF), and it can be generalized in several ways (Darroch and Ratcliff, 1972).

More generally, we wish to find a distribution in the form

$$p(\mathbf{x}) = \frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c) \quad (23)$$

that minimizes the Kullback-Leibler divergence to some prior $p_0(\mathbf{x})$ and satisfies a set of constraints (the data) of the form:

$$\sum_{\mathbf{x}} a_r(\mathbf{x}) p(\mathbf{x}) = h_r \quad (24)$$

where r indexes the constraint. If the prior is set to the uniform distribution and the constraints are measured marginal distributions over all the variables in each of the cliques of the graph, then the

problem solved by IPF is equivalent to finding the maximum likelihood clique potentials given a complete data set of observations.

IPF can be used to train an ML Boltzmann machine if $\hat{P}(S_i, S_j)$ is given by the data set for all pairs of variables connected in the Boltzmann machine. The procedure is to start from the uniform distribution (i.e., all weights set to 0), then apply IPF steps to each clique potential until all marginals match those in the data. One can generalize this by starting from a nonuniform distribution, which would give a Boltzmann machine with minimum divergence from the starting distribution.

But what if some of the variables in the Boltzmann machine are hidden? Byrne (1992) presents an elegant solution to this problem using ideas from alternating minimization (AM) and information geometry (Csizár and Tusnády, 1984). One step of the alternating minimization computes the distribution over the hidden variables of the Boltzmann machine, given the observed variables. This is the E step of EM, and can also be interpreted within information geometry as finding the probability distribution that satisfies the marginal constraints and is closest to the space of probability distributions defined by the Boltzmann machines. The other step of the minimization starts from $W = 0$ (the maximum entropy distribution for a Boltzmann machine) and uses IPF to find the Boltzmann machine weights that satisfy all the marginals found in the E step, $\hat{P}_{ij} = \langle s_i s_j \rangle^+$. This is the M step of the algorithm; a single M step thus involves an entire IPF optimization. The update rule for each step of the IPF optimization for a particular weight is:

$$W_{ij} \leftarrow W_{ij} + \log \left[\frac{\hat{P}_{ij}}{\langle s_i s_j \rangle^-} \frac{(1 - \langle s_i s_j \rangle^-)}{(1 - \hat{P}_{ij})} \right] \quad (25)$$

This algorithm is conceptually interesting, as it presents an alternative method for fitting Boltzmann machines with ties to IPF and alternating minimization procedures. However, it is impractical for large, multiply connected Boltzmann machines, since computing the exact unclamped correlations $\langle s_i s_j \rangle^-$ can take exponential time.

Although we have presented this EM-IPF algorithm for the case of Boltzmann machines, it is widely applicable to learning many undirected graphical models. In particular, in the E step, marginal distributions over the set of variables in each clique in the graph are computed conditioned on the settings of the observed variables. A propagation algorithm such as the Junction Tree algorithm can be used for this step (see GRAPHICAL MODELS: PROBABILISTIC INFERENCE). In the M step the IPF procedure is run so as to satisfy all the marginals computed in the E step. There also exist Junction-Tree-style propagation algorithms that exploit the structure of the graphical model to solve the IPF problem efficiently (Jiroušek and Pfeučil, 1995; Teh and Welling, 2002).

Bayesian Learning of Parameters

A Bayesian approach to learning starts with some a priori knowledge about the model structure—the set of arcs in the Bayesian network—and model parameters. This initial knowledge is represented in the form of a prior probability distribution over model structures and parameters, and is updated using the data to obtain a posterior probability distribution over models and parameters. In this article we will assume that the model structure is given, and we focus on computing the posterior probability distribution over parameters (see GRAPHICAL MODELS: STRUCTURE LEARNING for solutions to the problem of inferring model structure; see also BAYESIAN METHODS AND NEURAL NETWORKS).

For a given model structure \mathbf{m} , we can compute the posterior distribution over the parameters:

$$p(\boldsymbol{\theta} | \mathbf{m}, \mathbf{d}) = \frac{p(\mathbf{d} | \boldsymbol{\theta}, \mathbf{m}) p(\boldsymbol{\theta} | \mathbf{m})}{p(\mathbf{d} | \mathbf{m})} \quad (26)$$

If the data set is $\mathbf{d} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ and we wish to predict the next observation, $\mathbf{x}^{(N+1)}$, based on our data and model, then the Bayesian prediction

$$p(\mathbf{x}^{(N+1)} | \mathbf{d}, \mathbf{m}) = \int p(\mathbf{x}^{(N+1)} | \boldsymbol{\theta}, \mathbf{m}, \mathbf{d}) p(\boldsymbol{\theta} | \mathbf{m}, \mathbf{d}) d\boldsymbol{\theta} \quad (27)$$

averages over the uncertainty in the model parameters. This is known as the *predictive distribution* for the model.

In the limit of a large data set, and as long as the prior over the parameters assigns non-zero probability in the region around the ML parameter values, the posterior $p(\boldsymbol{\theta} | \mathbf{m}, \mathbf{d})$ will be sharply peaked around the maxima of the likelihood, and therefore the predictions of a single ML model will be similar to those obtained by Bayesian integration over the parameters.

Often, models are fit with relatively small amounts of data, so asymptotic results are not applicable and the predictions of the ML estimate will differ significantly from those of Bayesian averaging. In such situations it is important to compute or approximate the averaging over parameters in Equation 27. For certain discrete models with Dirichlet distributed priors over the parameters, and for certain linear-Gaussian models, it is possible to compute these integrals exactly. Otherwise, approximations can be used. There are a large number of approximations to the integral over the parameter distribution that have been used in graphical models, including Laplace's approximation, variational approximations, and a variety of MCMC methods (see Neal, 1993, for a review).

Discussion

There are several key insights regarding parameter learning in graphical models. When there are no hidden variables in a directed graphical model, the graph structure determines the statistics of the data needed to learn the parameters: the joint distribution of each variable and its parents in the graph. Parameter estimation can then often occur independently for each node. The presence of hidden variables introduces dependencies between the parameters. However, the EM algorithm transforms the problem so that in each M step the parameters of the graphical model are again uncoupled. The E step of EM “fills in” the hidden variables with the distribution predicted by the model, thereby turning the hidden-data problem into a complete-data problem.

The intuitive appeal of EM has led to its widespread use in models of unsupervised learning where the goal is to learn a generative model of sensory data. The E step corresponds to perception or recognition: inferring the (hidden) state of the world from the sensory data; while the M step corresponds to learning: modifying the model that relates the actual world to the sensory data. It has been suggested that top-down and bottom-up connections in cortex play the roles of generative and recognition models (see HELMHOLTZ MACHINES AND SLEEP-WAKE LEARNING). From a graphical model perspective, the bottom-up recognition model in Helmholtz machines can be thought of as a graph that approximates the distribution of the hidden variables given the observed variables, in much the same way as the variational approximation approximates that same distribution.

Undirected graphical models pose additional challenges. The partition function is usually a function of the parameters, and can introduce dependencies between the parameters even in the case of complete data. The IPF algorithm can be used to fit undirected graphical models from complete data, and an IPF-EM algorithm can be used when there are hidden data as well.

Although ML learning is adequate when there are enough data, in general it is necessary to approximate the average over the parameters. Averaging avoids overfitting; it is hard to see how overfitting can occur when nothing is “fit” to the data. Non-Bayesian methods for avoiding overfitting often also involve averaging, for

example, via bootstrap resampling of the data. Even with parameter averaging, predictions can suffer in quality if the assumed structure of the model—the conditional independence relationships—is incorrect. To overcome this, it is necessary to generalize the approach presented in this article to learn the structure of the model as well as the parameters from data. This topic is covered in GRAPHICAL MODELS: STRUCTURE LEARNING (q.v.).

Road Maps: Artificial Intelligence; Learning in Artificial Networks

Background: Bayesian Networks

Related Reading: Graphical Models: Probabilistic Inference; Graphical Models: Structure Learning

References

- Ackley, D., Hinton, G., and Sejnowski, T., 1985, A learning algorithm for Boltzmann machines, *Cognit. Sci.*, 9:147–169.
- Byrne, W., 1992, Alternating minimization and Boltzmann machine learning, *IEEE Trans. Neural Netw.*, 3:612–620.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J., 1999, *Probabilistic Networks and Expert Systems*, New York: Springer-Verlag.
- Csiszár, I., and Tushnádý, G., 1984, Information geometry and alternating minimization procedures, in *Statistics and Decisions*, Supplementary Issue No. 1, (E. J. Dudewicz, D. Plachky, and P. K. Sen, Eds.), Munich: Oldenbourg Verlag, pp. 205–237.
- Darroch, J. N., and Ratcliff, D., 1972, Generalized iterative scaling for log-linear models, *Ann. Math. Statist.*, 43:1470–1480.
- Dempster, A., Laird, N., and Rubin, D., 1977, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B*, 39:1–38.
- Heckerman, D., 1996, *A Tutorial on Learning with Bayesian Networks*, Technical Report MSR-TR-95-06, Redmond, WA: Microsoft Research, available: <ftp://ftp.research.microsoft.com/pub/tr/TR-95-06.PS>. ♦
- Jiroušek, R., and Přeucil, S., 1995, On the effective implementation of the iterative proportional fitting procedure, *Computat. Statist. Data Anal.*, 19:177–189.
- Jordan, M. I., Ed., 1999, *Learning in Graphical Models*, Cambridge, MA: MIT Press. ♦
- Lauritzen, S. L., 1995, The EM algorithm for graphical association models with missing data, *Computat. Statist. Data Anal.*, 19:191–201.
- Neal, R. M., 1993, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.
- Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann. ♦
- Rabiner, L. R., 1989, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, 77:257–286.
- Roweis, S. T., and Ghahramani, Z., 1999, A unifying review of linear Gaussian models, *Neural Computat.*, 11:305–345.
- Russell, S. J., Binder, J., Koller, D., and Kanazawa, K., 1995, Local learning in probabilistic models with hidden variables, in *Proceedings of an International Joint Conference on Artificial Intelligence*, Montreal, Canada: Morgan Kaufmann, pp. 1146–1152.
- Teh, Y. W., and Welling, M., 2002, The unified propagation and scaling algorithm, *Adv. Neural Inf. Process. Syst.*, 14:1146–1152.

Graphical Models: Probabilistic Inference

Michael I. Jordan and Yair Weiss

Introduction

A *graphical model* is a type of probabilistic network that has roots in several different research communities, including artificial intelligence (Pearl, 1988), statistics (Lauritzen and Spiegelhalter, 1988), error-control coding (Gallager, 1963), and neural networks. The graphical models framework provides a clean mathematical formalism that has made it possible to understand the relationships among a wide variety of network-based approaches to computation, and in particular to understand many neural network algorithms and architectures as instances of a broader probabilistic methodology.

Graphical models use graphs to represent and manipulate joint probability distributions. The graph underlying a graphical model may be directed, in which case the model is often referred to as a *belief network* or a *Bayesian network* (see BAYESIAN NETWORKS), or the graph may be undirected, in which case the model is generally referred to as a *Markov random field*. A graphical model has both a structural component—encoded by the pattern of edges in the graph—and a parametric component—encoded by numerical “potentials” associated with sets of edges in the graph. The relationship between these components underlies the computational machinery associated with graphical models. In particular, general *inference algorithms* allow statistical quantities (such as likelihoods and conditional probabilities) and information-theoretic quantities (such as mutual information and conditional entropies) to be computed efficiently. These algorithms are the subject of the current article. *Learning algorithms* build on these inference algorithms and allow parameters and structures to be estimated from data (see GRAPHICAL MODELS: PARAMETER LEARNING and GRAPHICAL MODELS: STRUCTURE LEARNING).

Background

Directed and undirected graphical models differ in terms of their Markov properties (the relationship between graph separation and conditional independence) and their parameterization (the relationship between local numerical specifications and global joint probabilities). These differences are important in discussions of the family of joint probability distribution that a particular graph can represent. In the inference problem, however, we generally have a specific fixed joint probability distribution at hand, in which case the differences between directed and undirected graphical models are less important. Indeed, in the current article, we treat these classes of model together and emphasize their commonalities.

Let U denote a set of nodes of a graph (directed or undirected), and let X_i denote the random variable associated with node i , for $i \in U$. Let X_C denote the subset of random variables associated with a subset of nodes C , for any $C \subseteq U$, and let $X = X_U$ denote the collection of random variables associated with the graph.

The family of joint probability distributions associated with a given graph can be parameterized in terms of a product over *potential functions* associated with subsets of nodes in the graph. For directed graphs, the basic subset on which a potential is defined consists of a single node and its parents, and a potential turns out to be (necessarily) the conditional probability of the node given its parents. Thus, for a directed graph, we have the following representation for the joint probability:

$$p(x) = \prod_i p(x_i | x_{\pi_i}) \quad (1)$$

where $p(x_i | x_{\pi_i})$ is the *local conditional probability* associated with node i , and π_i is the set of indices labeling the parents of node i . For undirected graphs, the basic subsets are *cliques* of the graph—

subsets of nodes that are completely connected. For a given clique C , let $\psi_C(x_C)$ denote a general potential function—a function that assigns a positive real number to each configuration x_C . We have

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \quad (2)$$

where \mathcal{C} is the set of cliques associated with the graph and Z is an explicit normalizing factor, ensuring that $\sum_x p(x) = 1$. (We work with discrete random variables throughout for simplicity.)

Equation 1 can be viewed as a special case of Equation 2. Note in particular that we could have included a normalizing factor Z in Equation 1, but, as is easily verified, it is necessarily equal to 1. Second, note that $p(x_i|x_{\pi_i})$ is a perfectly good example of a potential function, except that the set of nodes that it is defined on—the collection $\{i \cup \pi_i\}$ —is not in general a clique (because the parents of a given node are not in general interconnected). Thus, to treat Equation 1 and Equation 2 on an equal footing, we find it convenient to define the so-called *moral graph* \mathcal{G}^m associated with a directed graph \mathcal{G} . The moral graph is an undirected graph obtained by connecting all of the parents of each node in \mathcal{G} , and removing the arrowheads. On the moral graph, a conditional probability $p(x_i|x_{\pi_i})$ is a potential function, and Equation 1 reduces to a special case of Equation 2.

Probabilistic Inference

Let (E, F) be a partitioning of the indices of the nodes in a graphical model into disjoint subsets such that (X_E, X_F) is a partitioning of the random variables. There are two basic kinds of inference problem that we wish to solve

- *Marginal probabilities:*

$$p(x_E) = \sum_{x_F} p(x_E, x_F)$$

- *Maximum a posteriori (MAP) probabilities:*

$$p^*(x_E) = \max_{x_F} p(x_E, x_F)$$

From these basic computations we can obtain other quantities of interest. In particular, the *conditional probability* $p(x_F|x_E)$ is equal to

$$p(x_F|x_E) = \frac{p(x_E, x_F)}{\sum_{x_F} p(x_E, x_F)}$$

and this is readily computed for any x_F once the denominator is computed—a marginalization computation. Moreover, we often wish to combine conditioning and marginalization, or conditioning, marginalization, and MAP computations. For example, letting (E, F, H) be a partitioning of the node indices, we may wish to compute

$$p(x_F|x_E) = \frac{p(x_E, x_F)}{\sum_{x_F} p(x_E, x_F)} = \frac{\sum_{x_H} p(x_E, x_F, x_H)}{\sum_{x_F} \sum_{x_H} p(x_E, x_F, x_H)}$$

We first perform the marginalization operation in the numerator and then perform a subsequent marginalization to obtain the denominator.

Elimination

In this section we introduce a basic algorithm for inference known as *elimination*. Although elimination applies to arbitrary graphs (as we will see), our focus in this section is on trees.

We proceed via an example. Referring to the tree in Figure 1A, let us calculate the marginal probability $p(x_5)$. We compute this probability by summing the joint probability with respect to $\{x_1, x_2, x_3, x_4\}$. We must pick an order over which to sum, and with some malice aforethought, let us choose the order $(1, 2, 4, 3)$. We have

$$\begin{aligned} p(x_5) &= \sum_{x_3} \sum_{x_4} \sum_{x_2} \sum_{x_1} p(x_1, x_2, x_3, x_4, x_5) \\ &= \sum_{x_3} \sum_{x_4} \sum_{x_2} \sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_3) \\ &= \sum_{x_3} p(x_5|x_3) \sum_{x_4} p(x_4|x_3) \sum_{x_2} p(x_3|x_2) \sum_{x_1} p(x_1)p(x_2|x_1) \\ &= \sum_{x_3} p(x_5|x_3) \sum_{x_4} p(x_4|x_3) \sum_{x_2} p(x_3|x_2)m_{12}(x_2) \end{aligned}$$

where we introduce the notation $m_{ij}(x_j)$ to refer to the intermediate terms that arise in performing the sum. The index i refers to the variable being summed over, and the index j refers to the other variable appearing in the summand (for trees, there will never be more than two variables appearing in any summand). The resulting term is a function of x_j . We continue the derivation:

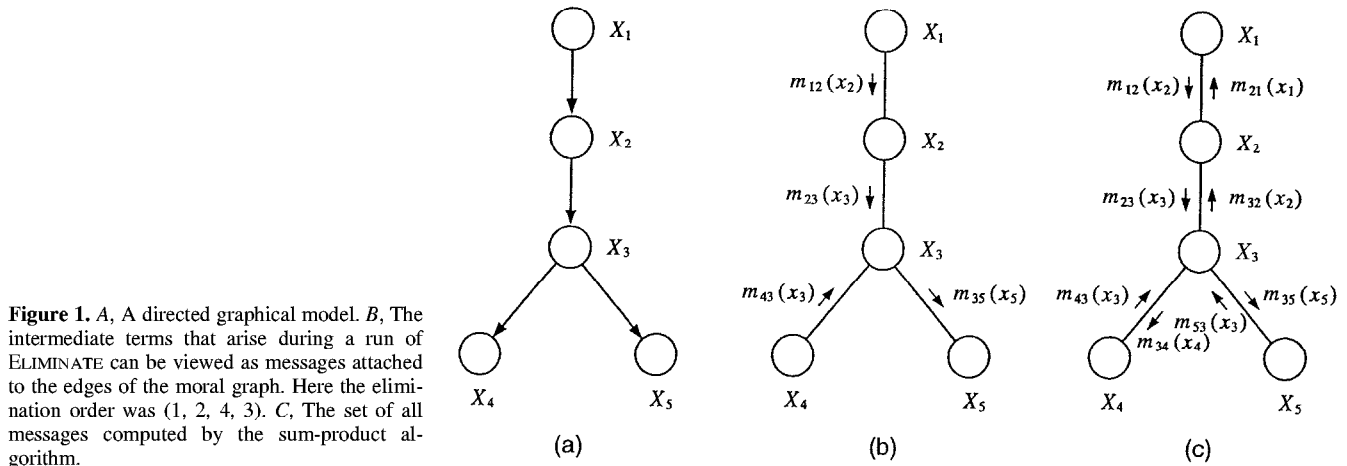


Figure 1. A, A directed graphical model. B, The intermediate terms that arise during a run of ELIMINATE can be viewed as messages attached to the edges of the moral graph. Here the elimination order was $(1, 2, 4, 3)$. C, The set of all messages computed by the sum-product algorithm.

$$\begin{aligned}
p(x_5) &= \sum_{x_3} p(x_5|x_3) \sum_{x_4} p(x_4|x_3) \sum_{x_2} p(x_3|x_2)m_{12}(x_2) \\
&= \sum_{x_3} p(x_5|x_3) \sum_{x_4} p(x_4|x_3)m_{23}(x_3) \\
&= \sum_{x_3} p(x_5|x_3)m_{23}(x_3) \sum_{x_4} p(x_4|x_3) \\
&= \sum_{x_3} p(x_5|x_3)m_{23}(x_3)m_{43}(x_3) \\
&= m_{35}(x_5)
\end{aligned}$$

The final expression is a function of x_5 only and is the desired marginal probability.

This computation is formally identically in the case of an undirected graph. In particular, an undirected version of the tree in Figure 1A has the parameterization

$$p(x) = \frac{1}{Z} \psi_{12}(x_1, x_2) \psi_{23}(x_2, x_3) \psi_{34}(x_3, x_4) \psi_{35}(x_3, x_5)$$

The first few steps of the computation of $p(x_5)$ are then as follows:

$$\begin{aligned}
p(x_5) &= \frac{1}{Z} \sum_{x_3} \psi_{35}(x_3, x_5) \sum_{x_4} \psi_{34}(x_3, x_4) \sum_{x_2} \psi_{23}(x_2, x_3) \\
&\quad \times \sum_{x_1} \psi_{12}(x_1, x_2) \\
&= \frac{1}{Z} \sum_{x_3} \psi_{35}(x_3, x_5) \sum_{x_4} \psi_{34}(x_3, x_4) \sum_{x_2} \psi_{23}(x_2, x_3) m_{12}(x_2)
\end{aligned}$$

and the remainder of the computation proceeds as before.

These algebraic manipulations can be summarized succinctly in terms of a general algorithm that we refer to here as **ELIMINATE** (Figure 2). The algorithm maintains an “active list” of potentials that, at the outset, represent the joint probability, and at the end represent the desired marginal probability. Nodes are removed from the graph according to an elimination ordering that must be specified. The algorithm applies to both directed and undirected graphs. Also, as we will see shortly, it is in fact a general algorithm, applying not only to trees but to general graphs.

Message-Passing Algorithms

In many problems we wish to obtain more than a single marginal probability. Thus, for example, we may wish to obtain both $p(x_4)$ and $p(x_5)$ in Figure 1A. Although we could compute each marginal with a separate run of **ELIMINATE**, this fails to exploit the fact that common intermediate terms appear in the different runs. We would like to develop an algebra of intermediate terms that allows them to be reused efficiently.

Suppose in particular that we wish to compute $p(x_4)$ in the example in Figure 1A. Using the elimination order (1, 2, 5, 3), it is easily verified that we generate the terms $m_{12}(x_2)$ and $m_{23}(x_3)$ as before, and also generate new terms $m_{53}(x_3)$ and $m_{34}(x_4)$.

```

ELIMINATE( $G$ )
  place all potentials  $\psi_c(x_c)$  on the active list
  choose an ordering  $I$  of the indices  $F$ 
  for each  $X_i$  in  $I$ 
    find all potentials on the active list that reference  $X_i$ 
    and remove them from the active list
    define a new potential as the sum (with respect to  $x_i$ ) of the
    product of these potentials
    place the new potential on the active list
  end
  return the product of the remaining potentials

```

Figure 2. A simple elimination algorithm for marginalization in graphical models.

As suggested by Figure 1B, the intermediate terms that arise during elimination can be viewed as “messages” attached to edges in the moral graph. Rather than viewing inference as an elimination process, based on a global ordering, we instead view inference in terms of local computation and routing of messages. The key operation of summing a product can be written as follows:

$$m_{ij}(x_j) = \sum_{x_i} \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i) \quad (3)$$

where $N(i)$ is the set of neighbors of node i . Thus, summing over x_i creates a message $m_{ij}(x_j)$ that is sent to the node j . The reader can verify that each step in our earlier computation of $p(x_5)$ has this form.

A node can send a message to a neighboring node once it has received messages from all of its other neighbors. As in our example, a message arriving at a leaf node is necessarily a marginal probability. In general, the marginal probability at a node is given by the product of all incoming messages:

$$p(x_i) \propto \prod_{k \in N(i)} m_{ki}(x_i) \quad (4)$$

The pair of equations given by Equations 3 and 4 defines an algorithm known as *sum-product algorithm* or the *belief propagation algorithm*. It is not difficult to prove that this algorithm is correct for trees.

The set of messages needed to compute all of the individual marginal probabilities for the graph in Figure 1A is shown in Figure 1C. Note that a pair of messages is sent along each edge, one message in each direction.

Neural networks also involve message-passing algorithms and local numerical operations. An important difference, however, is that in the neural network setting, each node generally has a single “activation” value that it passes to all of its neighbors. In the sum-product algorithm, on the other hand, individual messages are prepared for each neighbor. Moreover, the message $m_{ij}(x_j)$ from i to j is not included in the product that node j forms in computing a message to send back to node i . The sum-product algorithm avoids double-counting.

Maximum a posteriori (MAP) Probabilities

Referring again to Figure 1A, let us suppose that we wish to compute $p^*(x_5)$, the maximum probability configuration of the variables (X_1, X_2, X_3, X_4), for a given value of X_5 . Again choosing a particular ordering of the variables, we compute

$$\begin{aligned}
p^*(x_5) &= \max_{x_3} \max_{x_4} \max_{x_2} \max_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_3) \\
&= \max_{x_3} p(x_5|x_3) \max_{x_4} p(x_4|x_3) \max_{x_2} p(x_3|x_2) \\
&\quad \times \max_{x_1} p(x_1)p(x_2|x_1)
\end{aligned}$$

and the remaining computation proceeds as before. We see that the algebraic operations involved in performing the MAP computation are isomorphic to those in the earlier marginalization computation. Indeed, both the elimination algorithm and the sum-product algorithm extend immediately to MAP computation; we simply replace “sum” with “max” throughout in both cases. The underlying justification is that “max” commutes with products just as “sum” does.

General Graphs

Our goal in this section is to describe the *junction tree algorithm*, a generalization of the sum-product algorithm that is correct for arbitrary graphs. We derive the junction tree algorithm by returning to the elimination algorithm.

The first point to note is that **ELIMINATE** is correct for arbitrary graphs—the algorithm simply describes the creation of intermedi-

ate terms in a chain of summations that compute a marginal probability. Thus the algorithm is correct, but it is limited to the computation of a single marginal probability.

To show how to generalize the elimination algorithm to allow all individual marginals to be computed, we again proceed by example. Referring to the graph in Figure 3A, suppose that we wish to calculate the conditional probability $p(x_1)$. Let us use the elimination ordering (5, 4, 3, 2). At the first step, in which we sum over x_5 , we remove the potentials $\psi_{35}(x_3, x_5)$ and $\psi_{45}(x_4, x_5)$ from the active list and form the sum

$$m_{32}(x_3, x_4) = \sum_{x_5} \psi_{35}(x_3, x_5) \psi_{45}(x_4, x_5)$$

where the intermediate term, which is clearly a function of x_3 and x_4 , is denoted $m_{32}(x_3, x_4)$. (We explain the subscripts below.) The elimination of x_5 has created an intermediate term that effectively links x_3 and x_4 , variables that were not linked in the original graph. Similarly, at the following step, we eliminate x_4 :

$$m_{21}(x_2, x_3) = \sum_{x_4} \psi_{24}(x_2, x_4) m_{32}(x_3, x_4)$$

and obtain a term that links x_2 and x_3 , variables that were not linked in the original graph.

A graphical record of the dependencies induced during the run of ELIMINATE is shown in Figure 3B. We could also have created this graph according to a simple graph-theoretic algorithm in which nodes are removed in order from a graph where, when a node is removed, its remaining neighbors are linked. Thus, for example, when node 5 is removed, nodes 3 and 4 are linked. When node 4 is removed, nodes 2 and 3 are linked. Let us refer to this algorithm as GRAPH ELIMINATE.

We can also obtain the desired marginal $p(x_1)$ by working with the “filled-in” graph in Figure 3B from the outset. Noting that the cliques in this graph are $C_1 = \{x_1, x_2, x_3\}$, $C_2 = \{x_2, x_3, x_4\}$, and $C_3 = \{x_3, x_4, x_5\}$, and defining the potentials:

$$\begin{aligned} \psi_{C_1}(x_1, x_2, x_3) &= \psi_{12}(x_1, x_2) \psi_{13}(x_1, x_3) \\ \psi_{C_2}(x_2, x_3, x_4) &= \psi_{24}(x_2, x_4) \\ \psi_{C_3}(x_3, x_4, x_5) &= \psi_{35}(x_3, x_5) \psi_{45}(x_4, x_5) \end{aligned}$$

we obtain exactly the same product of potentials as before. Thus we have

$$\begin{aligned} p(x) &= \frac{1}{Z} \psi_{12}(x_1, x_2) \psi_{13}(x_1, x_3) \psi_{24}(x_2, x_4) \psi_{35}(x_3, x_5) \psi_{45}(x_4, x_5) \\ &= \frac{1}{Z} \psi_{C_1}(x_1, x_2, x_3) \psi_{C_2}(x_2, x_3, x_4) \psi_{C_3}(x_3, x_4, x_5) \end{aligned}$$

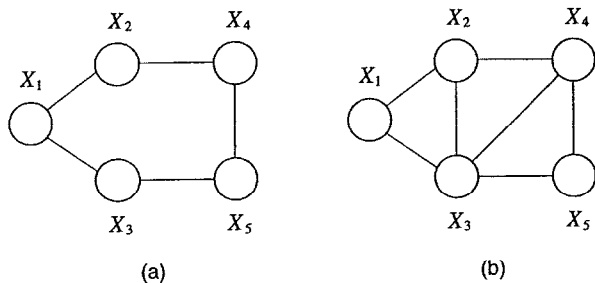


Figure 3. A, An undirected graphical model. B, The same model, with additional edges that reflect the dependencies created by the elimination algorithm.

We have essentially transferred the joint probability distribution from Figure 3A to Figure 3B. Moreover, the steps of the elimination algorithm applied to Figure 3B are exactly the same as before, and we obtain the same marginal. An important difference, however, is that in the case of Figure 3B all of the intermediate potentials created during the run of the algorithm are also supported by cliques in the graph.

Graphs created by GRAPH ELIMINATE are known as *triangulated graphs*, and they have a number of special properties. In particular, they allow the creation of a data structure known as a *junction tree* on which a generalized message-passing algorithm can be defined. A junction tree is a tree in which each node is a clique from the original graph. Messages, which correspond to intermediate terms in ELIMINATE, pass between these cliques.

Although a full discussion of the construction of junction trees is beyond the scope of the article, it is worth noting that a junction tree is not just any tree of cliques from a triangulated graph. Rather, it is a maximal spanning tree (of cliques), with weights given by the cardinalities of the intersections between cliques.

Given a triangulated graph, with cliques $C_i \in \mathcal{C}$ and potentials $\psi_{C_i}(x_{C_i})$, and given a corresponding junction tree (which defines links between the cliques), we send the following “message” from clique C_i to clique C_j :

$$m_{ij}(x_{S_{ij}}) = \sum_{C_i \setminus S_{ij}} \psi_{C_i}(x_{C_i}) \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_{S_{ki}}) \quad (5)$$

where $S_{ij} = C_i \cap C_j$, and where $\mathcal{N}(i)$ are the neighbors of clique C_i in the junction tree. Moreover, it is possible to prove that we obtain marginal probabilities as products of messages. Thus

$$p(x_{C_i}) \propto \prod_{k \in \mathcal{N}(i)} m_{ki}(x_{S_{ki}}) \quad (6)$$

is the marginal probability for clique C_i . (Marginals for single nodes can be obtained via further marginalization: i.e., $p(x_i) = \sum_{C_i \setminus \{x_i\}} p(x_C)$, for $i \in C$.)

The junction tree corresponding to the triangulated graph in Figure 3B is shown in Figure 4, where the corresponding messages are also shown. The reader can verify that the leftward-going messages are identical to the intermediate terms created during the run of ELIMINATE. The junction tree algorithm differs from ELIMINATE, however, in that messages pass in all directions, and the algorithm yields all clique marginals, not merely those corresponding to a single clique.

The sum-product algorithm described earlier in Equations 3 and 4 is a special case of Equations 5 and 6, obtained by noting that the original tree in Figure 1A is already triangulated and has pairs of nodes as cliques. In this case, the “separator sets” S_{ij} are singleton nodes.

Once again, the problem of computing MAP probabilities can be solved with a minor change to the basic algorithm. In particular, the “sum” in Equation 5 is changed to a “max.”

There are many variations on exact inference algorithms, but all of them are either special cases of the junction tree algorithm or are close cousins. The basic message from the research literature on exact inference is that the operations of triangulating a graph

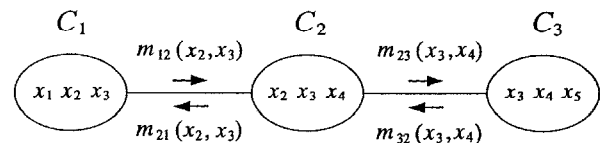


Figure 4. The junction tree corresponding to the triangulated graph in Figure 3B.

and passing messages on the resulting junction tree capture in a succinct way the basic algebraic structure of probabilistic inference.

Cutset Conditioning

In the method of *cutset conditioning*, we identify a “cutset”—defined (in the graphical model literature) as a set of nodes whose removal from the graph yields a tree. For example, in Figure 3A, any single node is a cutset. Denoting the indices of the cutset by Q , we loop over all instantiations x_Q , computing the conditional marginals $p(x_i|x_Q)$. The unconditional marginals are then given by $p(x_i) = \sum_{x_Q} p(x_i|x_Q)p(x_Q)$.

By considering an augmented graph in which edges are drawn from the nodes in the cutset to all other nodes in the graph, we can view cutset conditioning as a special case of the junction tree algorithm (Shachter, Andersen, and Szolovits, 1994). Note, however, that an implementation of cutset conditioning method involves operating on a single tree at a time, with cliques that are never larger than two nodes. Essentially, cutset conditioning involves implementing the junction tree algorithm in a way that trades time for space.

Computational Complexity

The computational complexity of the junction tree algorithm is a function of the size of the cliques upon which message-passing operations are performed. In particular, summing a clique potential is exponential in the number of nodes in the clique.

The problem of finding the optimal triangulation—the triangulation yielding the smallest maximal clique—turns out to be NP-hard. Clearly, if we had to search over all possible elimination orderings, the search would take exponential time. Triangulation can also be defined in other ways, however, and practical triangulation algorithms need not search over orderings. But the problem is still intractable, and can be a practical computational bottleneck.

An even more serious problem is that in practical graphical models, the original graph may have large cliques, or long loops, and even the optimal triangulation would yield unacceptable complexity. This problem is particularly serious because it arises not during the “compile time” operation of triangulation, but during the “run time” operation of message-passing. Problems in error-control coding and image processing are particularly noteworthy for yielding such graphs, as are discretizations of continuous-time problems and layered graphs of the kinds studied in the neural network field. To address these problems, we turn to the topic of approximate probabilistic inference.

Approximate Inference

The junction tree algorithm focuses on the algebraic structure of probabilistic inference, exploiting the conditional independencies present in a joint probability distribution, as encoded in the pattern of (missing) edges in the graph. There is another form of structure in probability theory, however, that is not exploited in the junction tree framework, and which leads us to hope that successful approximate inference algorithms can be developed. In particular, laws of large numbers and other concentration theorems in probability theory show that sums and products of large numbers of terms can behave in simple, predictable ways, despite the apparent combinatorial complexity of these operations. Approximate algorithms attempt to exploit these numerical aspects of probability theory.

We discuss two large classes of approximate inference algorithms in this section—Monte Carlo algorithms and variational algorithms. Although these classes do not exhaust all of the approximation techniques that have been studied, they capture the most widely used examples.

Monte Carlo Algorithms

Monte Carlo algorithms are based on the fact that while it may not be feasible to compute expectations under $p(x)$, it may be possible to obtain samples from $p(x)$, or from a closely related distribution, such that marginals and other expectations can be approximated using sample-based averages. We discuss three examples of Monte Carlo algorithms that are commonly used in the graphical model setting—Gibbs sampling, the Metropolis-Hastings algorithm, and importance sampling (for a comprehensive presentation of these methods and others, see Andrieu et al., 2003).

Gibbs sampling is an example of a Markov chain Monte Carlo (MCMC) algorithm. In an MCMC algorithm, samples are obtained via a Markov chain whose stationary distribution is the desired $p(x)$. The state of the Markov chain is a set of assignments of values to each of the variables, and, after a suitable “burn-in” period so that the chain approaches its stationary distribution, these states are used as samples.

The Markov chain for the Gibbs sampler is constructed in a straightforward way: (1) at each step one of the variables X_i is selected (at random or according to some fixed sequence), (2) the conditional distribution $p(x_i|x_{\setminus i})$ is computed, (3) a value x_i is sampled from this distribution, and (4) the sample x_i replaces the previous value of the i th variable.

The implementation of Gibbs sampling thus reduces to the computation of the conditional distributions of individual variables given all of the other variables. For graphical models, these conditionals take the following form:

$$p(x_i|x_{\setminus i}) = \frac{\prod_{C \in \mathcal{C}_i} \psi_C(x_C)}{\sum_{x_i} \prod_{C \in \mathcal{C}_i} \psi_C(x_C)} = \frac{\prod_{C \in \mathcal{C}_i} \psi_C(x_C)}{\sum_{x_i} \prod_{C \in \mathcal{C}_i} \psi_C(x_C)} \quad (7)$$

where \mathcal{C}_i denotes the set of cliques that contain index i . This set is often much smaller than the set \mathcal{C} of all cliques, and in such cases each step of the Gibbs sampler can be implemented efficiently. Indeed, the conditional of node i depends only on the neighbors of node i in the graph, and thus the computation of the conditionals often takes the form of a simple message-passing algorithm that is reminiscent of the sum-product algorithm.

A simple example of a Gibbs sampler is provided by the *Boltzmann machine*, an undirected graphical model in which the potentials are defined on pairwise cliques. Gibbs sampling is often used for inference in the Boltzmann machine, and the algorithm in Equation 7 takes the form of the classical computation of the logistic function of a weighted sum of the values of neighboring nodes.

When the computation in Equation 7 is overly complex, the *Metropolis-Hastings algorithm* can provide an effective alternative. The Metropolis-Hastings algorithm is an MCMC algorithm that is not based on conditional probabilities and thus does not require normalization. Given the current state x of the algorithm, Metropolis-Hastings chooses a new state \tilde{x} from a “proposal distribution” $q(\tilde{x}|x)$, which often simply involves picking a variable X_i at random and choosing a new value for that variable, again at random. The algorithm then computes the “acceptance probability”:

$$\alpha = \min \left(1, \frac{q(x|\tilde{x}) \prod_{C \in \mathcal{C}_i} \psi_C(\tilde{x}_C)}{q(\tilde{x}|x) \prod_{C \in \mathcal{C}_i} \psi_C(x_C)} \right)$$

With probability α the algorithm accepts the proposal and moves to \tilde{x} , and with probability $1 - \alpha$ the algorithm remains in state x . For graphical models, this computation also turns out to often take the form of a simple message-passing algorithm.

While Gibbs sampling and Metropolis-Hastings aim at sampling from $p(x)$, *importance sampling* is a Monte Carlo technique for

computing expectations in which samples are chosen from a simpler distribution $q(x)$, and these samples are reweighted appropriately. In particular, we approximate the expectation of a function $f(x)$ as follows:

$$\begin{aligned} E[f(x)] &= \sum_x p(x)f(x) \\ &= \sum_x q(x) \left(\frac{p(x)}{q(x)} f(x) \right) \\ &\approx \frac{1}{N} \sum_{i=1}^N \frac{p(x^{(i)})}{q(x^{(i)})} f(x^{(i)}) \end{aligned}$$

where the values $x^{(i)}$ are samples from $q(x)$. The choice of $q(x)$ is in the hands of the designer, and the idea is that $q(x)$ should be chosen to be relatively simple to sample from, while reasonably close to $p(x)$ so that the weight $p(x^{(i)})/q(x^{(i)})$ is reasonably large. In the graphical model setting, natural choices of $q(x)$ are often provided by simplifying the graph underlying $p(x)$ in some way, in particular by deleting edges.

The principal advantages of Monte Carlo algorithms are their simplicity of implementation and their generality. Under weak conditions, the algorithms are guaranteed to converge. A problem with the Monte Carlo approach, however, is that convergence times can be long, and it can be difficult to diagnose convergence.

We might hope to be able to improve on Monte Carlo methods in situations in which laws of large numbers are operative. Consider, for example, the case in which a node i has many neighbors, such that the conditional $p(x_i | x_{\setminus i})$ has a single, sharply determined maximum for most configurations of the neighbors. In this case, it would seem wasteful to continue to sample from this distribution; rather, we would like to be able to compute the maximizing value directly in some way. This way of thinking leads to the variational approach to approximate inference.

Variational Methods

The key to the variational approach lies in converting the probabilistic inference problem into an optimization problem, such that the standard tools of constrained optimization can be exploited. The basic approach has a similar flavor to importance sampling, but instead of choosing a single $q(x)$ a priori, a family of approximating distributions $\{q(x)\}$ is used, and the optimization machinery chooses a particular member from this family.

We begin by showing that the joint probability $p(x)$ can be viewed as the solution to an optimization problem. In particular, define the *energy* of a configuration x by $E(x) = -\log p(x) - \log Z$, and define the *variational free energy* as follows:

$$\begin{aligned} F(q) &= \sum_x q(x)E(x) + \sum_x q(x) \log q(x) \\ &= -\sum_x q(x) \log p(x) + \sum_x q(x) \log q(x) - \log Z \end{aligned}$$

The variational free energy is equal (up to an additive constant) to the Kullback-Leibler divergence between $q(x)$ and $p(x)$. It is therefore minimized when $q(x) = p(x)$ and attains the value of $-\log Z$ at the minimum. We have thus characterized $p(x)$ variationally.

Minimizing F is as difficult as doing exact inference, and much effort has been invested in finding approximate forms of F that are easier to minimize. Each approximate version of F gives an approximate variational inference algorithm.

For example, the simplest variational algorithm is the *mean field* approximation, in which $\{q(x)\}$ is restricted to the family of factorized distributions: $q(x) = \prod_i q_i(x_i)$. In this case F simplifies to

$$\begin{aligned} F_{MF}(q) &= -\sum_C \sum_{x_C} \log \psi_C(x_C) \prod_{i \in C} q_i(x_i) \\ &\quad + \sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i) \end{aligned}$$

subject to the constraint $\sum_{x_i} q_i(x_i) = 1$.

Setting the derivative with respect to $q_i(x_i)$ equal to zero gives

$$q_i(x_i) = \alpha \exp \left(\sum_{C \in \mathcal{C}_i} \sum_{x_{C \setminus i}} \log \psi_C(x_C) \prod_{j \in C, j \neq i} q_j(x_j) \right) \quad (8)$$

where α is a normalization constant chosen so that $\sum_{x_i} q_i(x_i) = 1$. The sum over cliques C_i is a sum over all cliques that node i belongs to.

Equation 8 defines an approximate inference algorithm. We initialize approximate marginals $q_i(x_i)$ for all nodes in the graph and then update the approximate marginal at one node based on those at neighboring nodes (note that the right-hand side of Equation 8 depends only on cliques that node i belongs to). This yields a message-passing algorithm that is similar to neural network algorithms; in particular, the value $q_i(x_i)$ can be viewed as the “activation” of node i .

More elaborate approximations to the free energy give better approximate marginal probabilities. While the mean field free energy depends only on approximate marginals at single nodes, the *Bethe free energy* depends on approximate marginals at single nodes $q_i(x_i)$ as well as on approximate marginals on cliques $q_C(x_C)$:

$$\begin{aligned} F_B(q) &= \sum_C \sum_{x_C} q_C(x_C) \log \frac{q_C(x_C)}{\psi_C(x_C)} \\ &\quad - \sum_i (d_i - 1) \sum_{x_i} q_i(x_i) \log q_i(x_i) \end{aligned}$$

where $d_i - 1$ denotes the number of cliques that node i belongs to.

The approximate clique marginals and the approximate singleton marginals must satisfy a simple marginalization constraint: $\sum_{x_C} q_C(x_C) = q_i(x_i)$. When we add Lagrange multipliers and differentiate the Lagrangian, we obtain a set of fixed point equations. Surprisingly, these equations end up being equivalent to the “sum-product” algorithm for trees in Equation 3. The messages $m_{ij}(x_j)$ are simply exponentiated Lagrange multipliers. Thus the Bethe approximation is equivalent to applying the local message-passing scheme developed for trees to graphs that have loops (see Yedidia, Freeman, and Weiss, 2001). This approach to approximate inference has been very successful in the domain of error-control coding, allowing practical codes based on graphical models to nearly reach the Shannon limit.

Discussion

The unified perspective on inference algorithms that we have presented in this article has arisen through several different historical strands. We briefly summarize these strands here and note some of the linkages with developments in the neural network field.

The elimination algorithm has had a long history. The “peeling” algorithm developed by geneticists is an early example (Cannings, Thompson, and Skolnick, 1978), as are the “decimation” and “transfer matrix” procedures in statistical physics (Itzykson and Drouffe, 1991). For a recent discussion of elimination algorithms, including more efficient algorithms than the simple ELIMINATE algorithm presented here, see Dechter (1999).

Belief propagation has also had a long history. An early version of the sum-product algorithm was studied by Gallager (1963) in the context of error-control codes (see Kschischang, Frey, and Loeliger (2001) for a recent perspective). Well-known special cases of sum-product include the forward-backward algorithm for hidden Markov models (see HIDDEN MARKOV MODELS), and the Kalman filtering/smoothing algorithms for state-space models. A seminal presentation of the sum-product algorithm was provided by Pearl (1988).

The variant of the junction tree algorithm that we have defined is due to Shafer and Shenoy (1990), and has also been called the *generalized distributive law* by Aji and McEliece (2000). A closely related variant known as the *Hugin algorithm* arose from the work of Lauritzen and Spiegelhalter (1988); it is described by Jensen (2001).

Many neural network architectures are special cases of general graphical model formalism, both representationally and algorithmically. Special cases of graphical models include essentially all of the models developed under the rubric of unsupervised learning (see UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS, INDEPENDENT COMPONENT ANALYSIS, and HELMHOLTZ MACHINES AND SLEEP-WAKE LEARNING), as well as Boltzmann machines (see SIMULATED ANNEALING AND BOLTZMANN MACHINES), mixtures of experts (see MODULAR AND HIERARCHICAL LEARNING SYSTEMS), and radial basis function networks (see RADIAL BASIS FUNCTION NETWORKS). Many other neural networks, including the classical multilayer perceptron (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION) can be profitably analyzed from the point of view of graphical models. For more discussion of these links, see the articles in Jordan (1999).

Road Map: Artificial Intelligence; Learning in Artificial Networks

Related Reading: Bayesian Networks; Graphical Models: Parameter Learning; Graphical Models: Structure Learning; Markov Random Field Models in Image Processing

References

- Aji, S. M., and McEliece, R. J., 2000, The generalized distributive law, *IEEE Trans. Inform. Theory*, 46:325–343.
- Andrieu, C., De Freitas, J., Doucet, A., and Jordan, M. I., 2003, An introduction to MCMC for machine learning, *Machine Learn* (in press). ♦
- Cannings, C., Thompson, E. A., and Skolnick, M. H., 1978, Probability functions on complex pedigrees, *Adv. Appl. Probab.*, 10:26–91.
- Dechter, R., 1999, Bucket elimination: A unifying framework for probabilistic inference, in *Learning in Graphical Models* (M. I. Jordan, Ed.), Cambridge, MA: MIT Press.
- Gallager, R. G., 1963, *Low-Density Parity Check Codes*, Cambridge, MA: MIT Press.
- Itzykson, C., and Drouffe, J., 1991, *Statistical Field Theory*, Cambridge, Engl.: Cambridge University Press.
- Jensen, F. V., 2001, *Bayesian Networks and Decision Graphs*, New York: Springer-Verlag. ♦
- Jordan, M. I., Ed., 1999, *Learning in Graphical Models*, Cambridge, MA: MIT Press. ♦
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K., 1999, An introduction to variational methods for graphical models, *Machine Learn.*, 37:183–233. ♦
- Kschischang, F. R., Frey, B. J., and Loeliger, H.-A., 2001, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory*, 47:498–519.
- Lauritzen, S. L., and Spiegelhalter, D., 1988, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. R. Statist. Soc. B*, 50:157–224.
- Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann. ♦
- Shachter, R., Andersen, S., and Szolovits, P., 1994, Global conditioning for probabilistic inference in belief networks, in *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, pp. 514–522.
- Shafer, G. R., and Shenoy, P. P., 1990, Probability propagation, *Ann. Math. Artif. Intell.*, 2:327–352.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y., 2001, *Bethe Free Energies, Kikuchi Approximations, and Belief Propagation Algorithms.*, MERL Technical Report 2001–16.

Graphical Models: Structure Learning

David Heckerman

Introduction

The article GRAPHICAL MODELS: PARAMETER LEARNING (q.v.) discussed the learning of parameters for a fixed graphical model. In this article, we discuss the simultaneous learning of parameters and structure. Real-world applications of such learning abound and can be found in, for example, *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (1991 and after). An index to software for parameter and structure learning can be found at <http://www.cs.berkeley.edu/murphyk/Bayes/bnsoft.html>.

For simplicity, we concentrate on directed-acyclic graphical models (DAG models), but the basic principles described here can be applied more generally. We describe the Bayesian approach in detail and mention several common non-Bayesian approaches.

We use notation that is slightly different from that used in the article on parameter learning. In particular, we use $\mathbf{X} = (X_1, \dots, X_n)$ to denote the n variables that we are modeling, \mathbf{x} to denote a configuration or observation of \mathbf{X} , and $\mathbf{d} = (\mathbf{x}^1, \dots, \mathbf{x}^N)$ to denote a random sample of N observations of \mathbf{X} . In addition, we use \mathbf{Pa}_i to denote the variables corresponding to the parents of X_i in a DAG model and \mathbf{pa}_i to denote a configuration of those variables. Finally, we shall use the terms *model* and *structure* interchangeably. In particular, a DAG model (and hence its structure) is described by (1) its nodes and arcs, and (2) the distribution class of each of its local distributions $p(x_i|\mathbf{pa}_i)$.

The Bayesian Approach

When we learn a model and its parameters, we presumably are uncertain about their identity. When following the Bayesian approach—in which all uncertainty is encoded as (subjective) probability—we encode this uncertainty as prior distributions over random variables corresponding to structure and parameters. In particular, let \mathbf{m} be a random variable having states $\mathbf{m}^1, \dots, \mathbf{m}^M$ corresponding to the possible models. (Note that we are assuming the models are mutually exclusive.) In addition, let $\theta^1, \dots, \theta^M$ be random variables corresponding to the unknown parameters of each of the M possible models. Then we express our uncertainty prior to learning as the prior distributions $p(\mathbf{m})$, and $p(\theta^1), \dots, p(\theta^M)$.

Given data \mathbf{d} , a random sample from the true but unknown joint distribution for \mathbf{X} , we compute the posterior distributions for each \mathbf{m} and θ^m using Bayes's rule:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{\sum_{\mathbf{m}'} p(\mathbf{m}')p(\mathbf{d}|\mathbf{m}')} \quad (1)$$

$$p(\theta^m|\mathbf{d}, \mathbf{m}) = \frac{p(\theta^m|\mathbf{m})p(\mathbf{d}|\theta^m, \mathbf{m})}{p(\mathbf{d}|\mathbf{m})} \quad (2)$$

where

$$p(\mathbf{d}|\mathbf{m}) = \int p(\mathbf{d}|\boldsymbol{\theta}^m, \mathbf{m})p(\boldsymbol{\theta}^m|\mathbf{m})d\boldsymbol{\theta}^m \quad (3)$$

is called the *marginal likelihood*. Given some hypothesis of interest, h , we determine the probability that h is true given data \mathbf{d} by averaging over all possible models and their parameters:

$$p(h|\mathbf{d}) = \sum_m p(\mathbf{m}|\mathbf{d})p(h|\mathbf{d}, \mathbf{m}) \quad (4)$$

$$p(h|\mathbf{d}, \mathbf{m}) = \int p(h|\boldsymbol{\theta}^m, \mathbf{m})p(\boldsymbol{\theta}^m|\mathbf{d}, \mathbf{m})d\boldsymbol{\theta}^m \quad (5)$$

For example, h may be the event that the next case \mathbf{x}^{N+1} is observed in configuration \mathbf{x}^{N+1} . In this situation, we obtain

$$p(\mathbf{x}^{N+1}|\mathbf{d}) = \sum_m p(\mathbf{m}|\mathbf{d}) \int p(\mathbf{x}^{N+1}|\boldsymbol{\theta}^m, \mathbf{m})p(\boldsymbol{\theta}^m|\mathbf{d}, \mathbf{m})d\boldsymbol{\theta}^m \quad (6)$$

where $p(\mathbf{x}^{N+1}|\boldsymbol{\theta}^m, \mathbf{m})$ is the likelihood for the model. It is important to note that, in the Bayesian approach, no single model is learned. Instead, data is used to update the probability that each possible model is the correct one.

Unfortunately, this approach, sometimes called *Bayesian model averaging* or the *full Bayesian approach*, is often impractical. For example, the number of different DAG models for a domain containing n variables grows superexponentially with n . Thus, the approach can only be applied in those few settings where one has strong prior knowledge that can eliminate almost all possible models.

Statisticians, who have been confronted by this problem for decades in the context of other types of models, use two approximations to address this problem: *Bayesian model selection* and *selective Bayesian model averaging*. The former approach is to select a likely model from among all possible models and use it as if it were the correct model. For example, to predict the next case, we use

$$\begin{aligned} p(\mathbf{x}^{N+1}|\mathbf{d}) &\cong p(\mathbf{x}^{N+1}|\mathbf{m}, \mathbf{d}) \\ &= \int p(\mathbf{x}^{N+1}|\boldsymbol{\theta}^m, \mathbf{m})p(\boldsymbol{\theta}^m|\mathbf{d}, \mathbf{m})d\boldsymbol{\theta}^m \end{aligned} \quad (7)$$

where \mathbf{m} is the selected model. The latter approach is to select a manageable number of good models from among all possible models and pretend that these models are exhaustive. In either approach, we need only the *relative* model posterior— $p(\mathbf{m})p(\mathbf{d}|\mathbf{m})$ —to select likely models.

Both approaches can be characterized as *search-and-score* techniques. That is, in these approaches, we search among a large set of models looking for those with good scores. The use of these approximate methods raise several important questions. Do they yield accurate results when applied to graphical model learning? If so, can we compute the model posteriors and perform a search efficiently?

The question of accuracy is difficult to answer in theory. Nonetheless, several researchers have shown experimentally that the selection of a single good hypothesis often yields accurate predictions (e.g., Cooper and Herskovits, 1992; Heckerman, Geiger, and Chickering, 1995) and that selective model averaging using Monte Carlo methods can sometimes be efficient and yield even better predictions (Madigan et al., 1996). These results, which are somewhat surprising, are largely responsible for the considerable interest in learning graphical models.

In the remainder of this section, we address computational efficiency. In particular, we consider situations in which (relative) model posteriors can be computed efficiently as well as efficient search procedures.

We note that model averaging, model selection, and selective model averaging all help avoid overfitting—situations where mod-

els perform well on training data and poorly on new data. In particular, the marginal likelihood balances the fit of the model structure to data with the complexity of the model. One way to understand this fact is to note that, when the number of cases N is large and other conditions hold, the marginal likelihood can be approximated as follows:

$$p(\mathbf{d}|\mathbf{m}) \cong p(\mathbf{d}|\hat{\boldsymbol{\theta}}, \mathbf{m}) - \frac{|\boldsymbol{\theta}|}{2} \log N$$

where $\hat{\boldsymbol{\theta}}$ is the maximum-likelihood estimator of the data (e.g., Kass and Raftery, 1995). The first quantity in this expression represents the degree to which the model fits the data, which increases as the model complexity increases. The second quantity, in contrast, penalizes model complexity.

Computation of the Marginal Likelihood

Under certain conditions, the marginal likelihood of a graphical model—and hence its relative posterior—can be computed efficiently. In this section, we examine a particular set of these conditions for structure learning of DAG models. We note that a similar set of conditions holds for the learning of decomposable UG models. For details, see Lauritzen (1996).

Given any DAG model \mathbf{m} , we can factor the likelihood of a single sample as follows:

$$p(\mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) = \prod_{i=1}^n p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m}) \quad (8)$$

We shall refer to each term $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m})$ in this equation as the *local likelihood* for X_i . Also, in this equation, $\boldsymbol{\theta}_i$ denotes the set of parameters associated with the local likelihood for variable X_i .

The first condition in our set of sufficient conditions yielding efficient computation is that each local likelihood is in the exponential family. One example of such a factorization occurs when each variable $X_i \in \mathbf{X}$ is finite, having r_i possible values $x_i^1, \dots, x_i^{r_i}$, and each local likelihood is a collection of multinomial distributions, one distribution for each configuration of \mathbf{Pa}_i —that is,

$$p(x_i^k|\mathbf{pa}_i^j, \boldsymbol{\theta}_i, \mathbf{m}) = \theta_{ijk} > 0 \quad (9)$$

where $\mathbf{pa}_i^1, \dots, \mathbf{pa}_i^{q_i}$ ($q_i = \prod_{X_j \in \mathbf{Pa}_i} r_j$) denotes the configurations of \mathbf{Pa}_i , and $\boldsymbol{\theta}_i = ((\theta_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$ denotes the parameters. The parameter θ_{ij1} is given by $1 - \sum_{k=2}^{r_i} \theta_{ijk}$. We shall use this example to illustrate many of the concepts in this article. For convenience, we define the vector of parameters

$$\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \dots, \theta_{ijr_i})$$

for all i and j . Examples of other exponential families can be found in Bernardo and Smith (1994).

The second assumption for efficient computation is one of parameter independence. In our multinomial example, we assume that the parameter vectors $\boldsymbol{\theta}_{ij}$ are mutually independent. Note that, when this independence holds and we are given a random sample \mathbf{d} that contains no missing observations, the parameters remain independent:

$$p(\boldsymbol{\theta}_m|\mathbf{d}, \mathbf{m}) = \prod_{i=1}^n \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|\mathbf{d}, \mathbf{m}) \quad (10)$$

Thus, we can update each vector of parameters $\boldsymbol{\theta}_{ij}$ independently.

The third assumption is that each independent parameter set has a conjugate prior (e.g., Bernardo and Smith, 1994). In our multinomial example, we assume that each $\boldsymbol{\theta}_{ij}$ has a Dirichlet prior $\text{Dir}(\boldsymbol{\theta}_{ij}|\alpha_{ij1}, \dots, \alpha_{ijr_i})$. In this case, we obtain

$$p(\theta_{ij}|\mathbf{d}, \mathbf{m}) = \text{Dir}(\theta_{ij}|\alpha_{ij1} + N_{ij1}, \dots, \alpha_{ijr_i} + N_{ijr_i}) \quad (11)$$

where N_{ijk} is the number of cases in \mathbf{d} in which $X_i = x_i^k$ and $\mathbf{Pa}_i = \mathbf{pa}_i^j$. Note that the collection of counts N_{ijk} are sufficient statistics of the data for the model \mathbf{m} .

Under these conditions, we can compute the marginal likelihood efficiently and in closed form. For our multinomial example (as first derived in Cooper and Herskovits, 1992), we obtain

$$p(\mathbf{d}|\mathbf{m}) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (12)$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

Under these same conditions, the integral in Equation 7 also can be computed efficiently. In our example, suppose that, for a given outcome \mathbf{x}_{N+1} of \mathbf{X}_{N+1} , the value of X_i is x_i^k and the configuration of \mathbf{Pa}_i is \mathbf{pa}_i^j , where k and j depend on i . Using Equations 4, 8, and 9, we obtain

$$p(\mathbf{x}_{N+1}|\mathbf{d}, \mathbf{m}) = \int \left(\prod_{i=1}^n \theta_{ijk} \right) p(\theta_m|\mathbf{d}, \mathbf{m}) d\theta_m$$

Because parameters remain independent given \mathbf{d} , we get

$$p(\mathbf{x}_{N+1}|\mathbf{d}, \mathbf{m}) = \prod_{i=1}^n \int \theta_{ijk} p(\theta_{ij}|\mathbf{d}, \mathbf{m}) d\theta_{ij}$$

Finally, because each integral in this product is the expectation of a Dirichlet distribution, we have

$$p(\mathbf{x}_{N+1}|\mathbf{d}, \mathbf{m}) = \prod_{i=1}^n \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \quad (13)$$

To compute the relative posterior probability of a model, we must assess the structure prior $p(\mathbf{m})$ and the parameter priors $p(\theta^m|\mathbf{m})$. Unfortunately, when many models are possible, the assessment process will be intractable. Nonetheless, under certain assumptions, we can derive the structure and parameter priors for many models from a manageable number of direct assessments. Several authors have discussed such assumptions and corresponding methods for deriving priors (e.g., Buntine, 1991; Cooper and Herskovits, 1992; Heckerman et al., 1995; Cowell et al., 1999). In the following two sections, we examine some of these approaches.

Priors for Model Parameters

First, let us consider the assessment of priors for the parameters of DAG models. We consider the approach of Heckerman, Geiger, and Chickering (1995)—herein, HGC—who address the case for \mathbf{X} where the local likelihoods are multinomial distributions. A similar approach exists for situations where the local likelihoods are linear regressions (Heckerman and Geiger, 1995).

Their approach is based on two key concepts: Markov equivalence and distribution equivalence. We say that two models for \mathbf{X} are *Markov equivalent* if they represent the same set of conditional-independence assertions for \mathbf{X} . For example, given $\mathbf{X} = \{X, Y, Z\}$, the models $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, and $X \leftarrow Y \leftarrow Z$ represent only the independence assertion that X and Z are conditionally independent given Y . Consequently, these models are equivalent. Another example of Markov equivalence is the set of *complete models* on \mathbf{X} . A complete model is one that has no missing edge and that encodes no assertion of conditional independence. When \mathbf{d} contains n variables, there are $n!$ possible complete models, one model structure for every possible ordering of the variables. All complete models for \mathbf{X} are Markov equivalent. In general, two models are Markov equivalent if and only if they have the same structure, ignoring arc directions, and the same v -structures. A v -structure is an ordered tuple (X, Y, Z) such that there is an arc from X to Y and from Z to Y , but no arc between X and Z .

The concept of distribution equivalence is closely related to that of Markov equivalence. Suppose that all models for \mathbf{X} under consideration have local likelihoods in the family \mathcal{F} . This is not a restriction per se, because \mathcal{F} can be a large family. We say that two model structures \mathbf{m}_1 and \mathbf{m}_2 for \mathbf{X} are *distribution equivalent with respect to \mathcal{F}* if they can represent the same joint probability distributions for \mathbf{X} , that is, if, for every θ_{m_1} , there exists a θ_{m_2} such that $p(\mathbf{x}|\theta_{m_1}, \mathbf{m}_1) = p(\mathbf{x}|\theta_{m_2}, \mathbf{m}_2)$, and vice versa.

Distribution equivalence with respect to some \mathcal{F} implies Markov equivalence, but the converse does not hold. For example, when \mathcal{F} is the family of generalized linear regression models, the complete model structures for $n \geq 3$ variables do not represent the same sets of distributions. Nonetheless, there are families \mathcal{F} —for example, multinomial distributions and linear regression models with Gaussian noise—where Markov equivalence implies distribution equivalence with respect to \mathcal{F} (see HGC). The notion of distribution equivalence is important, because if two model structures \mathbf{m}_1 and \mathbf{m}_2 are distribution equivalent with respect to a given \mathcal{F} , then it is often reasonable to expect that data cannot help to discriminate them. That is, we expect $p(\mathbf{d}|\mathbf{m}_1) = p(\mathbf{d}|\mathbf{m}_2)$ for any data set \mathbf{d} . HGC call this property *likelihood equivalence*.

Now let us return to the main issue of this section: the derivation of parameter priors from a manageable number of assessments. HGC show that the assumption of likelihood equivalence combined with the assumption that the θ_{ij} are mutually independent imply that the parameters for any *complete* model \mathbf{m}_c must have a Dirichlet distribution with constraints on the hyperparameters given by

$$\alpha_{ijk} = \alpha p(x_i^k, \mathbf{pa}_i^j|\mathbf{m}_c) \quad (14)$$

where α is the user's equivalent sample size and $p(x_i^k, \mathbf{pa}_i^j|\mathbf{m}_c)$ is computed from the user's joint probability distribution $p(\mathbf{d}|\mathbf{m}_c)$ (discussions of equivalent sample size can be found in, e.g., Heckerman et al., 1995). Note that this result is rather surprising, as the two assumptions leading to the constrained Dirichlet solution are qualitative.

To determine the priors for parameters of *incomplete* models, HGC use the assumption of *parameter modularity*, which says that if X_i has the same parents in models \mathbf{m}_1 and \mathbf{m}_2 , then

$$p(\theta_{ij}|\mathbf{m}_1) = p(\theta_{ij}|\mathbf{m}_2)$$

for $j = 1, \dots, q_i$. They call this property *parameter modularity*, because it says that the distributions for parameters θ_{ij} depend only on a portion of the graph structure, namely, X_i and its parents.

Given the assumptions of parameter modularity and parameter independence, it is a simple matter to construct priors for the parameters of an arbitrary model given the priors on complete models. In particular, given parameter independence, we construct the priors for the parameters of each node separately. Furthermore, if node X_i has parents \mathbf{Pa}_i in the given model, then we identify a complete model structure where X_i has these parents, and use Equation 14 and parameter modularity to determine the priors for this node. The result is that all terms α_{ijk} for all model structures are determined by Equation 14. Thus, from the assessments α and $p(\mathbf{d}|\mathbf{m}_c)$, we can derive the parameter priors for all possible model structures. We can assess $p(\mathbf{d}|\mathbf{m}_c)$ by constructing a parameterized model, called a *prior network*, that encodes this joint distribution.

Priors for Model Structures

Now let us consider the assessment of priors on structure. The simplest approach for assigning priors to models is to assume that every model is equally likely. Of course, this assumption is typically inaccurate and is used only for the sake of convenience. A simple refinement of this approach is to ask the user to exclude various structures (perhaps based on judgments of cause and ef-

fect), and then impose a uniform prior on the remaining structures. We use this approach in an example described later.

Buntine (1991) describes a set of assumptions that leads to a richer yet efficient approach for assigning priors. The first assumption is that the variables can be ordered (e.g., through a knowledge of time precedence). The second assumption is that the presence or absence of possible arcs are mutually independent. Given these assumptions, $n(n-1)/2$ probability assessments (one for each possible arc in an ordering) determines the prior probability of every possible model. One extension to this approach is to allow for multiple possible orderings. One simplification is to assume that the probability that an arc is absent or present is independent of the specific arc in question. In this case, only one probability assessment is required.

An alternative approach, described by Heckerman et al. (1995), uses the prior network described in the previous section. The basic idea is to penalize the prior probability of any structure according to some measure of deviation between that structure and the prior network. Heckerman et al. (1995) suggest one reasonable measure of deviation.

Search Methods

In this section, we examine search methods for identifying DAG models with high scores. Consider the problem of finding the best DAG model from the set of all DAG models in which each node has no more than k parents. Unfortunately, the problem for $k > 1$ is NP-hard even when we use the restrictive prior given by Equation 14 (Chickering, 1996). Thus, researchers have used heuristic search algorithms, including greedy search, greedy search with restarts, best-first search, and Monte Carlo methods.

One consolation is that these search methods can be made more computationally efficient when the model score is factorable. Given a DAG model for domain \mathbf{X} , we say that a score for that model $S(\mathbf{m}, \mathbf{d})$ is *factorable* if it can be written as a product of variable-specific scores:

$$S(\mathbf{m}, \mathbf{d}) = \prod_{i=1}^n s(X_i, \mathbf{Pa}_i, \mathbf{d}_i) \quad (15)$$

where \mathbf{d}_i is the data restricted to the variables X_i and \mathbf{Pa}_i . An example of a factorable score is Equation 12 used in conjunction with any of the structure priors described previously.

Most of the commonly used search methods for DAG models also make successive arc changes to the graph structure, and employ the property of factorability to evaluate the merit of each change. One commonly used set of arc changes is as follows. For any pair of variables, if there is an arc connecting them, then this arc can either be reversed or removed. If there is no arc connecting them, then an arc can be added in either direction. All changes are subject to the constraint that the resulting DAG contains no directed cycles. We use E to denote the set of eligible changes to a graph, and $\Delta(e)$ to denote the change in $\log p(\mathbf{d}|\mathbf{m})p(\mathbf{m})$ resulting from the modification $e \in E$. Given a factorable score, if an arc to X_i is added or deleted, only $c(X_i, \mathbf{Pa}_i, \mathbf{d}_i)$ need be evaluated to determine $\Delta(e)$. If an arc between X_i and X_j is reversed, then only $c(X_i, \mathbf{Pa}_i, \mathbf{d}_i)$ and $c(X_j, \mathbf{Pa}_j, \mathbf{d}_j)$ need be evaluated.

One simple heuristic search algorithm is greedy hill climbing. We begin with some DAG model. Then, we evaluate $\Delta(e)$ for all $e \in E$, and make the change e for which $\Delta(e)$ is a maximum, provided it is positive. We terminate search when there is no e with a positive value for $\Delta(e)$. Candidates for the initial model include the empty graph, a random graph, and the prior network used for the assessment of parameter and structure priors.

A potential problem with any local-search method is getting stuck at a local maximum. One method for escaping local maxima

is greedy search with random restarts. In this approach, we apply greedy search until we hit a local maximum. Then we randomly perturb the structure, and repeat the process for some manageable number of iterations. Another method for escaping local maxima is simulated annealing. In this approach, we initialize the system at some temperature T_0 . Then we pick some eligible change e at random, and evaluate the expression $p = \exp(\Delta(e)/T_0)$. If $p > 1$, then we make the change e ; otherwise, we make the change with probability p . We repeat this selection and evaluation process α times or until we make β changes. If we make no changes in α repetitions, then we stop searching. Otherwise, we lower the temperature by multiplying the current temperature T_0 by a decay factor $0 < \gamma < 1$, and continue the search process. We stop searching if we have lowered the temperature more than δ times. Thus, this algorithm is controlled by five parameters: T_0 , α , β , γ , and δ . To initialize this algorithm, we can start with the empty graph, and make T_0 large enough so that almost every eligible change is made, thus creating a random graph. Alternatively, we may start with a lower temperature, and use one of the initialization methods described for local search.

Another method for escaping local maxima is best-first search. In this approach, the space of all models is searched systematically using a heuristic measure that determines the next best structure to examine. Experiments (e.g., Heckerman et al., 1995) have shown that, for a fixed amount of computation time, greedy search with random restarts produces better models than does best-first search.

One important consideration for any search algorithm is the search space. The methods that we have described search through the space of DAG models. Nonetheless, when likelihood equivalence is assumed, one can search through the space of model equivalence classes. One benefit of the latter approach is that the search space is smaller. One drawback of the latter approach is that it takes longer to move from one element in the search space to another. Experiments have shown that the two effects roughly cancel.

Example: College Plans

In this section, we consider an analysis of data, obtained by Sewell and Shah (1968), regarding factors that influence the intention of high school students to attend college. This analysis was given previously by Heckerman in Jordan (1999).

Sewell and Shah (1968) measured the following variables for 10,318 Wisconsin high school seniors: *sex* (SEX): male, female; *socioeconomic status* (SES): low, lower middle, upper middle, high; *intelligence quotient* (IQ): low, lower middle, upper middle, high; *parental encouragement* (PE): low, high; and *college plans* (CP): yes, no. Our goal in this analysis is to understand the relationships among these variables.

The data are (completely) described by the counts in Table 1. Each entry denotes the number of cases in which the five variables take on some particular configuration. The first entry corresponds to the configuration SEX = male, SES = low, IQ = low, PE =

Table 1. Sufficient Statistics for the Sewall and Shah (1968) Study

4	349	13	64	9	207	33	72	12	126	38	54	10	67	49	43
2	232	27	84	7	201	64	95	12	115	93	92	17	79	119	59
8	166	47	91	6	120	74	110	17	92	148	100	6	42	198	73
4	48	39	57	5	47	123	90	9	41	224	65	8	17	414	54
5	454	9	44	5	312	14	47	8	216	20	35	13	96	28	24
11	285	29	61	19	236	47	88	12	164	62	85	15	113	72	50
7	163	36	72	13	193	75	90	12	174	91	100	20	81	142	77
6	50	36	58	5	70	110	76	12	48	230	81	13	49	360	98

Reproduced by permission of the University of Chicago Press. © 1968 by The University of Chicago. All rights reserved.

low, and CP = yes. The remaining entries correspond to configurations obtained by cycling through the states of each variable such that the last variable (CP) varies most quickly. Thus, for example, the upper (lower) half of the table corresponds to male (female) students.

To generate priors for model parameters, we used the method described earlier in this section with an equivalent sample size of five and a prior network describing a uniform distribution over \mathbf{X} . (The results we report remain qualitatively the same for equivalent sample sizes ranging from 3 to 40.) For structure priors, we assumed that all models were equally likely, except that we excluded structures (based on causal considerations) where SEX and/or SES had parents, and/or CP had children. We used Equation 12 to compute the marginal likelihoods of the models. The two most likely models that we found after an exhaustive search over all structures are shown in Figure 1. Note that the most likely model has a posterior probability that is extremely close to 1. Both models show a reasonable result: that CP and SEX are independent, given the remaining variables.

Methods for Incomplete Data

Among the assumptions that yield an efficient method for computing the marginal likelihood, the one that is most often violated is the assumption that all variables are observed in every case. In many situations, some variables will be hidden (i.e., never observed) or will be observed for only a subset of the data samples. There are a variety of methods for handling such situations—at greater computational cost—including Monte Carlo (MC) approaches (e.g., DiCiccio et al., 1995), large-sample approximations (e.g., Kass and Raftery, 1995), and variational approximations (e.g., Jordan et al. in Jordan, 1999).

In this section, we examine a simple MC approach called *Gibbs sampling* (e.g., MacKay in Jordan, 1999). In general, given variables $\mathbf{X} = \{X_1, \dots, X_n\}$ with some joint distribution $p(x)$, we can use a Gibbs sampler to approximate the expectation of a function $f(x)$ with respect to $p(x)$. This approximation is made as follows. First, we choose an initial state for each of the variables in \mathbf{X} somehow (e.g., at random). Next, we pick some variable X_i , unassign its current state, and compute its probability distribution given the states of the other $n - 1$ variables. Then, we sample a state for X_i based on this probability distribution, and compute $f(x)$. Finally, we iterate the previous two steps, keeping track of the average value of $f(x)$. In the limit, as the number of cases approach infinity, this average is equal to $E_{p(x)}(f(x))$ provided two conditions are met. First, the Gibbs sampler must be *irreducible*. That is, the probability distribution $p(x)$ must be such that we can eventually sample any possible configuration of \mathbf{X} given any possible initial configuration of \mathbf{X} . For example, if $p(x)$ contains no zero probabilities, then the Gibbs sampler will be irreducible. Second, each X_i must be chosen infinitely often. In practice, an algorithm for deterministically rotating through the variables is typically used. An intro-

duction to Gibbs sampling and other Monte Carlo methods, including methods for initialization and a discussion of convergence, is given by Neal (1993).

To illustrate Gibbs sampling, consider again the case where every variable in \mathbf{X} is finite, the parameters θ_{ij} for a given DAG model \mathbf{m} are mutually independent, and each θ_{ij} has a Dirichlet prior. In this situation, let us approximate the probability density $p(\theta_m | \mathbf{d}, \mathbf{m})$ for some particular configuration of θ_m , given an incomplete data set \mathbf{d} . First, we initialize the states of the unobserved variables in each case somehow. As a result, we have a complete random sample \mathbf{d}_c . Second, we choose some variable X_{ij} (variable X_i in case I) that is not observed in the original random sample D , and reassign its state according to the probability distribution

$$p(x'_{ij} | \mathbf{d}_c \setminus x_{ij}, \mathbf{m}) = \frac{p(x'_{ij}, \mathbf{d}_c \setminus x_{ij} | \mathbf{m})}{\sum_{x_{ij}} p(x''_{ij}, \mathbf{d}_c \setminus x_{ij} | \mathbf{m})}$$

where $\mathbf{d}_c \setminus x_{ij}$ denotes the data set \mathbf{d}_c with observation x_{ij} removed, and the sum in the denominator runs over all states of variable X_{ij} . As we have seen, the terms in the numerator and denominator can be computed efficiently (see Equation 12). Third, we repeat this reassignment for all unobserved variables in \mathbf{d} , producing a new complete random sample \mathbf{d}'_c . Fourth, we compute the posterior density $p(\theta_m | \mathbf{d}'_c, \mathbf{m})$ as described in Equations 10 and 11. Finally, we iterate the previous three steps, and use the average of $p(\theta_m | \mathbf{d}'_c, \mathbf{m})$ as our approximation.

Monte Carlo approximations are also useful for computing the marginal likelihood given incomplete data. One Monte Carlo approach uses Bayes's theorem:

$$p(\mathbf{d} | \mathbf{m}) = \frac{p(\theta_m | \mathbf{m}) p(\mathbf{d} | \theta_m, \mathbf{m})}{p(\theta_m | \mathbf{d}, \mathbf{m})} \quad (16)$$

For any configuration of θ_m , the prior term in the numerator can be evaluated directly. In addition, the likelihood term in the numerator can be computed using DAG-model inference (e.g., Kjærulff in Jordan, 1999). Finally, the posterior term in the denominator can be computed using Gibbs sampling, as we have just described.

Non-Bayesian Approaches

In this section, we consider several commonly used alternatives to the Bayesian approach for structure learning.

One such class of algorithms mimic the search-and-score approach of Bayesian model selection but incorporate a non-Bayesian score. Alternative scores include (1) prediction accuracy on new data, (2) prediction accuracy over cross-validated data sets, and (3) non-Bayesian information criteria such as AIC.

Another class of algorithms for structure learning is the *constraint-based* approach, described by Pearl (2000) and Spirtes, Glymour, and Scheines (2001). In this set of algorithms, statistical tests are performed on the data to determine independence and dependence relationships among the variables. Then, search methods are used to identify one or more models that are consistent with those relationships.

To illustrate this approach, suppose we seek to learn one or more DAG models given data for three finite variables (X_1, X_2, X_3). Assuming each local likelihood is a collection of multinomial distributions, there are 11 possible DAG models that are distinct: (1) a complete model, (2) $X_1 \rightarrow X_2 \rightarrow X_3$, (3) $X_1 \rightarrow X_3 \rightarrow X_2$, (4) $X_2 \rightarrow X_1 \rightarrow X_3$, (5) $X_1 \rightarrow X_2 \leftarrow X_3$, (6) $X_1 \rightarrow X_3 \leftarrow X_2$, (7) $X_2 \rightarrow X_1 \leftarrow X_3$, (8) $X_1 \rightarrow X_2 X_3$, (9) $X_1 \rightarrow X_3 X_2$, (10) $X_2 \rightarrow X_3 X_1$, and (11) $X_1 X_2 X_3$, where $X_i X_j$ means there is no arc between X_i and X_j . There are other possible models that are not listed, but each such model represents a set of distributions that is equivalent to one of the other models

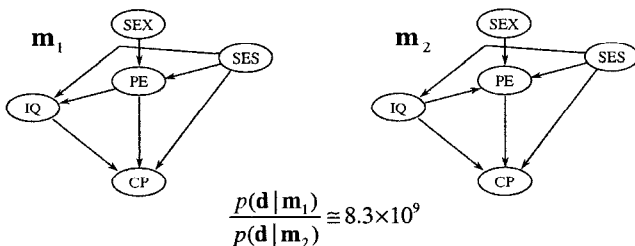


Figure 1. The a posteriori most likely models.

above. For example, $X_3 \rightarrow X_2 \rightarrow X_1$ and model 2 are distribution equivalent.

Now, suppose that statistical tests applied to the data reveal that the *only* independence relationship is that X_1 and X_3 are independent. Only models 1 and 5 can exhibit only this independence. Furthermore, if we use parameter prior assignments of the form described earlier in this section, then model 1 will exhibit this independence with probability zero. Consequently, we conclude that model 5 is correct (with probability one).

One drawback of the constraint-based approach is that any statistical test will be an approximation for finite data, and errors in the tests may lead the search mechanism to (1) conclude that the found relationships are inconsistent or (2) return erroneous models. One advantage of the approach over most search-and-score methods is that more structures can be considered for a fixed amount of computation, because the results of some statistical tests can greatly constrain model search.

Road Maps: Artificial Intelligence; Learning in Artificial Networks

Background: Graphical Models: Probabilistic Inference

Related Reading: Graphical Models: Parameter Learning

References

- Bernardo, J., and Smith, A., 1994, *Bayesian Theory*, New York: Wiley. ♦
- Buntine, W., 1991, Theory refinement on Bayesian networks, in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, pp. 52–60.
- Chickering, D., 1996, Learning Bayesian networks is NP-complete, in *Learning from Data* (D. Fisher and H. Lenz, Eds.), New York: Springer-Verlag, pp. 121–130.
- Cooper, G., and Herskovits, E., 1992, A Bayesian method for the induction of probabilistic networks from data, *Machine Learn.*, 9:309–347.
- Cowell, R., Dawid, A. P., Lauritzen, S., and Spiegelhalter, D., 1999, *Probabilistic Networks and Expert Systems (Statistics for Engineering and Information Science)*, New York: Springer-Verlag. ♦
- DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L., 1995, *Computing Bayes Factors by Combining Simulation and Asymptotic Approximations*, Technical Report 630, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Heckerman, D., and Geiger, D., 1995, Learning Bayesian networks: A unification for discrete and Gaussian domains, in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, pp. 274–284.
- Heckerman, D., Geiger, D., and Chickering, D., 1995, Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learn.*, 20:197–243.
- Jordan, M., Ed., 1999, *Learning in Graphical Models*, Cambridge, MA: MIT Press. ♦
- Kass, R., and Raftery, A., 1995, Bayes factors, *J. Am. Statist. Assoc.*, 90:773–795. ♦
- Lauritzen, S., 1996, *Graphical Models*, Oxford, Engl.: Clarendon Press. ♦
- Madigan, D., Raftery, A., Volinsky, C., and Hoeting, J., 1996, Bayesian model averaging, in *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, Portland, OR. ♦
- Neal, R., 1993, Probabilistic inference using Markov chain Monte Carlo Methods, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.
- Pearl, J., Ed., 2000, *Causality: Models, Reasoning, and Inference*, Cambridge, Engl.: Cambridge University Press. ♦
- Sewell, W., and Shah, V., 1968, Social class, parental encouragement, and educational aspirations, *Am. J. Sociol.*, 73:559–572.
- Spirtes, P., Glymour, C., and Scheines, R., 2001, *Causation, Prediction, and Search*, 2nd ed., Cambridge, MA: MIT Press. ♦

Grasping Movements: Visuomotor Transformations

Giuseppe Rizzolatti and Giacomo Luppino

Introduction

When one attempts to pick up an object, one executes two distinct motor operations. One—reaching—consists of bringing the hand toward an object's location in space, the other—grasping—consists of shaping the hand and fingers in anticipation of the object's size, shape, and orientation (Arbib, 1981; Jeannerod, 1988). This article focuses on grasping. Its aim is to examine where in the cerebral cortex visual information on intrinsic properties of objects is transformed into hand movements and how this transformation occurs.

Motor Areas for Grasping

The agranular frontal cortex of primates consists of several distinct motor areas (Rizzolatti and Luppino, 2001; Picard and Strick, 2001). Their location in the monkey cerebral cortex is shown in Figure 1. Recent data showed that many of them contain distal movement representations (Rizzolatti and Luppino, 2001).

Since the early electrical cortical stimulation experiments, it has been known that the largest and most detailed representation of distal movements is that of the primary motor cortex (F1 or area 4; see Porter and Lemon, 1993). Following lesion of this area, grasping movements, especially those demanding a subtle control of fingers, are lost. The deficit is characterized by a dramatic decrease in force and a loss of the capacity to control individual fingers, but does not affect the mechanisms underlying visuomotor transformation for grasping movements. This view is confirmed by

neurophysiological findings showing that the visual properties (brisk responses to abrupt stimulus presentation) of the few neurons that respond to visual stimuli, do not match those necessary for grip formation.

This last finding raises the problem of which of the areas that have access to the F1 neural machinery also have the visual properties required for organizing grasping movements. Anatomical data and recording studies have shown that this area is F5. This area is richly connected with F1 and many of its neurons respond to the presentation specific visual stimuli (Rizzolatti and Luppino, 2001).

Area F5

Area F5 forms the rostral part of inferior area 6. A fundamental property of F5 is that the discharge of most of its neurons correlates with specific actions (or fragments of a specific action) much better than with elementary movements. A clear instance of this behavior is represented by F5 neurons that discharge when the monkey grasps an object with its right hand, with its left hand, or with its mouth. It is obvious that in this case a description of neuron behavior in terms of elementary movements makes little sense (Rizzolatti et al., 1988).

Using the effective action as classification criterion, F5 neurons were subdivided into various classes. Among them the most represented are “grasping” neurons, “holding” neurons, “tearing” neurons, and “manipulating” neurons. The largest F5 class is related to grasping.

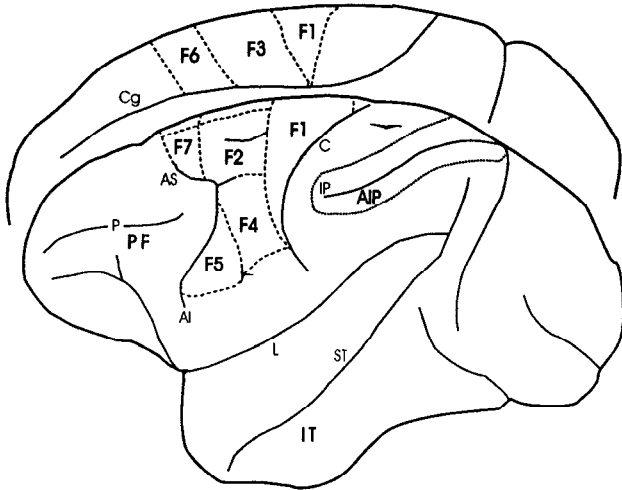


Figure 1. Lateral and mesial views of the monkey cerebral cortex showing the location of the agranular frontal areas and of other cortical areas or sectors referred to in the text. Motor areas in which distal movements are represented are F1, F3, F2, and F5. The intraparietal sulcus (IP) is opened. AI, inferior arcuate sulcus; AS, superior arcuate sulcus; C, central sulcus; Cg, cingulate sulcus; IT, inferotemporal cortex; L, lateral fissure; P, principal sulcus; PF, prefrontal cortex; ST, superior temporal sulcus.

Typically, grasping neurons discharge for actions made with the contralateral as well as the ipsilateral hand. Many of them code specific hand grips. Three basic grip types are extensively represented: precision grip, finger prehension, and whole-hand prehension. Most neurons are selective for one of these types of grasping, the precision grip being the most represented. The temporal relations between neuron discharge and grasping movements vary among neurons. Some of them fire only during the last part of grasping, i.e., during finger flexion. Others start to fire with finger extension and continue to fire during finger flexion. Finally, others are activated in advance of the movement initiation and often cease to discharge only when the object is grasped.

Rizzolatti and co-workers proposed that F5 contains a “vocabulary” of motor acts related to prehension. The “words” of the vocabulary are comprised of populations of neurons related to different motor actions. There are various categories of “words.” Some indicate very general commands, e.g., “grasp,” “hold,” and “tear.” Others indicate how the objects have to be grasped, held, or torn (e.g., by precision grip, finger prehension, whole-hand prehension, or their subtypes). Finally, a third group of “words” is concerned with the temporal segments of the actions (e.g., hand aperture, hand closure).

The presence in F5 of a store of “words” or motor schemas (Arbib, 1981) has two important consequences. First, since information is concentrated in relatively few abstract elements, the number of variables to be controlled is much less than it would be if the movements were described in terms of motor neurons or muscles. This solution to the problem of controlling the large number of hand degrees of freedom is remarkably similar to that proposed theoretically (Arbib, Iberall, and Lyons, 1985). Second, the retrieval of the appropriate movement is simplified. Both for internally generated actions and for those emitted in response to an external stimulus, a schema, or a small ensemble of schemas, must be selected. In particular, the retrieval of a movement in response to a visual object is reduced to the task of matching its size and orientation with the appropriate schema.

How can the motor vocabulary of F5 be addressed? The simplest way to examine this issue is to present different types of stimuli—

for example, different objects—and to establish whether the recorded neuron respond to them. Using this approach, “visual” responses are observed in about 20%–30% of F5 neurons. According to the type of stimuli that is effective, two separate classes of neurons have been distinguished. Neurons of the first class (*canonical* neurons) respond to the presentation of graspable objects (Rizzolatti et al., 1988). Neurons of the second class (*mirror* neurons) respond when the monkey sees object-directed actions (Rizzolatti and Luppino, 2001).

The functional properties of canonical neurons and inactivation experiments showed that this class of F5 visuomotor neurons is crucially involved in visuomotor transformations for grasping objects.

The properties of canonical neurons were recently studied in a task in which monkeys were trained to fixate objects of different size and shape and, after a go signal, to grasp them (Murata et al., 1997). The timing of the different phases of the task was such as to allow one to identify the neuron’s activity related to object presentation, the preparatory phase preceding the movement, and movement execution. Most neurons were further tested in a task in which monkeys had to fixate an object, but not to grasp it. At the go signal they had only to release a switch. The main result of the experiment was that canonical neurons discharge in response to object presentation even in the absence of any subsequent movement directed toward it. In the majority of neurons the visual discharge was evoked either exclusively by one object or by a small set of objects. Furthermore, the visual responses were evoked only by objects whose size and shape were congruent with the grip coded by the neurons (i.e., neurons visually activated by small objects discharged also selectively during precision grip, while neurons visually activated by large objects, discharged during whole hand prehension).

Recently, a paradigm similar to that described previously was employed to study the effects of inactivation of F5 with muscimol, a GABAergic agonist (Gallese et al., 1997). Small and large inactivations were performed. As a control, in separate sessions, the hand field of F1 was inactivated. The most interesting result was that following F5 inactivations (bank sector of F5) the monkeys were unable to shape the hand according to the intrinsic visual properties of the object. The deficit was particularly evident for small objects, where it was also present following limited inactivations. After large inactivations, however, the prehension of large objects was also affected. An important observation was that the monkeys, in spite of their visuomotor deficit, were still able to grasp and manipulate objects after touching them without any apparent skill impairment. Thus, the execution of individual finger movements was preserved. In contrast, after inactivation of F1 hand field, the monkeys showed a hypotonic paralysis and lack of individual finger movements (Schieber and Poliakov, 1998). Another important difference between F1 and F5 inactivations was that, following F5 inactivation, both hands were affected, not just the hand contralateral to the lesion.

Anterior Intraparietal Area

What is the origin of F5 visual information? Injection of neural tracers in F5 showed that this area receives a strong input from the inferior parietal lobule and, in particular, from an area located in the rostral part of the lateral bank of the intraparietal sulcus—area AIP (Rizzolatti and Luppino, 2001). The functional properties of area AIP have been extensively studied by Sakata and his co-workers (Murata et al., 2000).

The paradigms used by Sakata and colleagues were basically similar to those described for F5. In their initial experiments, monkeys were trained to manipulate four different types of switches, each of which required a peculiar type of grasping. The movements were performed under visual guidance and in the dark. According

to each neuron's behavior in the task, neurons were subdivided into three main classes. Neurons of the first class, "motor dominant" neurons, discharged equally well during movements performed in light and in dark. These neurons represented about one-third of the task-related neurons. Neurons of the second class, "visual dominant" neurons, were activated only when the task was performed in the light. They represented about 25% of the studied neurons. Neurons of the third class, "visual and motor" neurons, were less active during movement performed in the dark than in the light. They represented about 40% of the studied neurons.

Neurons belonging to "visual dominant" and "visual and motor" classes typically responded as soon as the object to be grasped was presented. To better study the visual properties of AIP neurons, another series of experiments was carried out by the same authors using a variety of 3D objects. They included spheres, cubes, cones, cylinders, rings, and plates of different sizes and orientations. Furthermore, the neurons were also tested in a condition in which there was no request to act on the presented objects. The results showed that most of the recorded neurons were selective for object shape. Of these, one-third were highly selective, being activated only by one of the six shapes used in the experiment, while half were moderately selective, being activated by two or three different shapes. Many shape-selective neurons were also selective for object size. Finally, in "visual and motor" neurons, a clear congruence was observed between object and motor selectivity.

It is clear from this overview of the functional properties of AIP that the neurons of this area share many common features with F5 neurons. There are also, however, some important differences. First, there are no "visual dominant" neurons in F5. Second, purely motor neurons are much more frequent in F5 than in AIP. Third, most AIP neurons discharge during the whole action leading to the grasping of the objects, often remaining active during the object holding period. In contrast, F5 grasping neurons are typically active only during some of the phases of the grasping/holding action.

Taken together, these data strongly suggest that AIP and F5 form a parieto-frontal circuit devoted to the visuomotor transformation for hand-object interactions (Jeannerod et al., 1995). They predict also that an inactivation of AIP should disrupt the monkey's capacity to preshape the hand and fingers in anticipation of object grasping. This prediction was fully confirmed by experiments in which different sectors of AIP were reversibly inactivated using muscimol microinjections. After inactivation, monkeys showed marked deficits of contralateral hand preshaping without any deficit in arm reaching (Gallese et al., 1997).

Circuit for Grasping in Humans

Earlier brain imaging experiments failed to convincingly demonstrate the existence of a cortical circuit for grasping movements in humans. Recent data by Binkofski et al. (1999) clearly demonstrated that a circuit specifically involved in object manipulation exists also in humans. By using fMRI, they showed that the manipulation of complex objects results in an activation of ventral premotor cortex (BA 44) and of a region in the intraparietal sulcus. If one considers the anatomical location of these areas, and, for area 44, the cytoarchitectonics similarities with F5, it is very likely that this circuit is the human homologue of the monkey circuit formed by F5-AIP.

Discussion

The data described previously indicate that areas AIP and F5 form the key elements in a circuit that transforms visual information on intrinsic properties of objects into grasping movements. Various attempts have been made to explain how AIP and F5 neurons per-

form this transformation (Sakata et al., 1992; Gallese et al., 1997; Fagg and Arbib, 1998).

Common to all these proposals is the idea that "visual dominant" neurons of AIP code the object's intrinsic properties and, then, either directly or via other AIP elements, send this information to specific sets of F5 neurons. F5 neurons transform the received information into patterns of hand movements that are appropriate to the size and shape of the objects to be grasped. Finally, the F5 grasping pattern recruits specific F1 neurons that command grasping execution.

According to Sakata and co-workers, the AIP discharge associated with movements represents a corollary discharge originating in the premotor cortex. Its function is that of creating a reverberatory activity that keeps active AIP neurons. AIP "visual-and-motor" neurons are, therefore, a kind of "memory" that keeps the representation of the object active during the entire movement execution. In this way the representation of the object remains present even when vision of the object is obstructed by the hand movements. Finally, the "motor dominant" neurons would represent an intermediate stage in the transmission of F5 corollary discharge to the AIP "visual-and-motor" neurons.

Sakata's model did not take into consideration the fact that an object may be grasped in several ways and that the chosen grip depends, in addition to the object visual properties, on object semantics and on what the individual who grasps the object wants to do with it. Let us imagine a mug. Once the mug is recognized as a mug, it is grasped by the handle, if one wants to drink from it. However, if one wants to throw the mug at somebody or simply to move it, one will take it by its body or by its upper edge. These possible ways of grasping depend on (1) a preliminary object recognition, and (2) on motor decisions, and obviously not on the visual intrinsic properties of the object. Thus, a more complete model of how F5-AIP circuit works, also requires information (1) from circuits that code object semantics (inferotemporal lobe, IT) and (2) from circuits where decision are taken on what to do. To solve these problems, Fagg and Arbib (1998) proposed that AIP provides F5 not with a single visual description of the object, but with multiple descriptions of the possible way in which a given object may be grasped (affordances). This multitude of possibilities is sent to F5, where the selection of the desired grip is made on the basis of prefrontal inputs that signal the nature of the object as well as the current goals of the organism. Finally, the control exerted by F6 (pre-SMA) on F5 will determine whether the external contingencies allow the action execution.

Fagg and Arbib's model (1998, FARS) is not only physiologically plausible, but has also been computationally implemented. Anatomically, FARS relies heavily on connections between prefrontal cortex and F5. There is evidence, however, that although these connections are very modest, rich connections exist between prefrontal cortex and AIP. Furthermore, AIP, unlike F5, receives a direct input from IT. These findings suggest an alternative possibility, namely that information on object semantics and the goals of the individuals influence AIP rather than F5 neurons. Thus, the fundamental process for selecting an appropriate grip occurs in AIP by biasing those affordances that will lead to the grip appropriate to the individual intention. According to this view, AIP describes several affordances, but information on that bias is able to influence F5. This affordance then activates the F5 neurons for the appropriate grip. The selected action remains a potential action until an appropriate signal comes from F6. A version of the FARS model, modified according to these considerations, is shown in Figure 2.

Road Maps: Mammalian Brain Regions; Mammalian Motor Control

Related Reading: Action Monitoring and Forward Control of Movements; Arm and Hand Movement Control; Eye-Hand Coordination in Reaching Movements; Language Evolution: The Mirror System Hypothesis; Prefrontal Cortex in Temporal Organization of Action

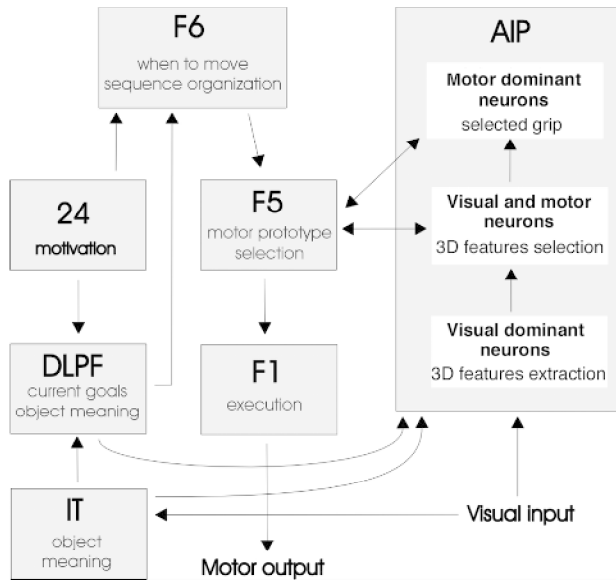


Figure 2. Schematic model of visuomotor transformations for grasping. (From Rizzolatti and Luppino, 2001.)

References

- Arbib, M. A., 1981, Perceptual structures and distributed motor control, in *Handbook of Physiology*, sect. 1, vol. 2, part 2 (V. B. Brooks, Ed.), Bethesda, MD: American Physiological Society, pp. 1449–1480.
- Arbib, M. A., Iberall, T., and Lyons, D., 1985, Coordinated control programs for movements of the hand, in *Hand Function and the Neocortex* (A. W. Goodman and I. Darian-Smith, Eds.), *Exp. Brain Res. Suppl* 10. Berlin: Springer-Verlag, pp. 111–129.
- Binkofski, F., Buccino, G., Posse, S., Seitz, R. J., Rizzolatti, G., and Freund, H.-J., 1999, A fronto-parietal circuit for object manipulation in man: Evidence from an fMRI-study, *Eur. J. Neurosci.*, 11:3276–3286.
- Fagg, A. H., and Arbib, M. A., 1998, Modeling parietal-premotor interactions in primate control of grasping, *Neural Networks*, 11:1277–1303.
- Gallese, V., Fadiga, L., Fogassi, L., Luppino, G., and Murata, A., 1997, A parietal-frontal circuit for hand grasping movements in the monkey: Evidence from reversible inactivation experiments, in *Parietal lobe contributions to orientation in 3D space* (P. Thier and H.-O. Karnath, Eds.), Heidelberg: Springer, pp. 255–270.
- Jeannerod, M., 1988, *The Neural and Behavioral Organization of Goal-Directed Movements*, Oxford: Clarendon.
- Jeannerod, M., Arbib, M. A., Rizzolatti, G., and Sakata, H., 1995, Grasping objects: The cortical mechanisms of visuomotor transformation, *Trends Neurosci.*, 18:314–320.
- Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., and Rizzolatti, G., 1997, Object representation in the ventral premotor cortex (area F5) of the monkey, *J. Neurophysiol.*, 78:2226–2230.
- Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H., 2000, Selectivity for the shape, size and orientation of objects in the hand-manipulation-related neurons in the anterior intraparietal (AIP) area of the macaque, *J. Neurophysiol.*, 83:2580–2601.
- Picard, N., and Strick, P. L., 2001, Imaging the premotor areas, *Curr. Opin. Neurobiol.*, 11:663–672. ♦
- Porter, R., and Lemon, R., 1993, *Corticospinal function and voluntary movement*, Clarendon Press, Oxford, p. 427.
- Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., and Matelli, M., 1988, Functional organization of inferior area 6 in the macaque monkey, II: Area F5 and the control of distal movements, *Exp. Brain Res.*, 71:491–507.
- Rizzolatti, G., and Luppino, G., 2001, The cortical motor system, *Neuron*, 31:889–901. ♦
- Sakata, H., Taira, M., Mine, S., and Murata, A., 1992, Hand-movement related neurons in the posterior parietal cortex of the monkey: Their role in visual guidance of hand movements, in *Control of Arm Movement in Space* (R. Caminiti, P. B. Johnson, and Y. Burnod, Eds.), *Exp. Brain Res. Suppl.* 22, Berlin: Springer-Verlag, pp. 185–198.
- Schieber, M. H., and Poliakov, A. V., 1998, Partial inactivation of the primary motor cortex hand area: Effects of individuated finger movements, *J. Neurosci.*, 18:9038–9054.

Habituation

Adam S. Bristol, Angela L. Purcell, and Thomas J. Carew

Introduction

Habituation, one of the simplest and most common forms of learning, is typically defined as the progressive decrement in a behavioral response with repeated presentations of the eliciting stimulus. Despite its apparent simplicity, a complete understanding of the underlying neural mechanisms of habituation has not been achieved in any preparation. In this article, we briefly review the fundamental characteristics of habituation and describe several experimental preparations in which the neural basis of habituation has been examined. We conclude by describing attempts to model the habituation process in computational terms, the success and shortcomings of these attempts, and directions for future research.

Characteristics of Habituation

Historically, the unambiguous identification of habituation as a form of learning has been complicated by the necessity of distinguishing it from other causes of response decrement, such as sensory adaptation and motor fatigue. A landmark publication by

Thompson and Spencer (1966) presented nine behavioral criteria for identifying habituation. These criteria have been successfully applied to a wide variety of systems and continue to be used today. They are:

1. Repeated stimuli result in diminishing response (within-session, short-term habituation).
2. After a series of stimuli and a period of rest, the response recovers spontaneously.
3. During repeated series of stimuli separated by spontaneous recovery, habituation occurs more rapidly within each series (savings).
4. The rate and/or magnitude of habituation increases with increases in stimulation frequency.
5. The rate and/or magnitude of habituation decreases with increases in stimulus intensity.
6. When a steady level of habituation has been reached during a series, additional stimuli prolong the time until spontaneous recovery ("subzero" habituation).

7. Habituation to one type of stimulus may result in habituation to other similar stimuli (generalization).
8. The presentation of a strong stimulus different from the habituating stimulus leads to a recovery of the habituated response (dishabituation).
9. Repetition of the dishabituating stimulus leads to successively less dishabituation (habituation of dishabituation).

Two important features of habituation, stimulus generalization and dishabituation (Criteria 7 and 8), are important indicators that the decrease in responding is not due to sensory or effector fatigue. Additionally, the finding that spontaneous recovery occurs more quickly with shorter interstimulus interval (ISI) training (Groves and Thompson, 1970) would not be predicted if sensory adaptation or muscle fatigue were the cause of the response decrement.

Thompson and Spencer's (1966) criteria have been highly influential in the study of habituation, but they have not gone without criticism (for a thorough evaluation, see Hinde, 1970). Moreover, subsequent studies have warranted amendments to these guidelines. First, Davis (1970) questioned the notion that habituation increases with shorter ISIs (Criterion 4) because, within a single training session, training and testing intervals are confounded. That is, when comparing different ISIs (e.g., 2 s and 10 s), both the training and testing intervals for each ISI are different because the response to each stimulus serves as both a training trial and a measure of behavioral habituation. When he controlled for differences in testing conditions, Davis (1970) found that *longer* ISIs resulted in greater and longer-lasting retention, although the rate of habituation was faster with short ISI. This is consistent with the finding that spontaneous recovery occurs more rapidly after short ISIs than long ISIs (Groves and Thompson, 1970) and with the general notion that massed training produces inferior learning relative to spaced training.

Second, the Thompson and Spencer criteria did not distinguish between short-term and long-term forms of habituation. Criterion 3 states that the rate of habituation increases across repeated training sessions, a form of "savings." However, the criteria do not reflect another common feature of multiple habituation trials, namely, that the first response of each repeated session is often progressively diminished, indicating a longer-lasting retention of habituation across sessions. Importantly, short- and long-term forms of habituation have been experimentally dissociated in a number of preparations, including *Aplysia*, *C. elegans*, crab, and rat.

The third amendment to the Thompson and Spencer criteria comes from work by Christoffersen (1997), who recently conducted a comparative analysis of the kinetics of short-term habituation across a wide variety of species and behaviors. He noted that, despite variability in response types, there is a characteristic learning curve for within-session habituation consisting of a rapid and pronounced early phase of habituation resulting from only a few response activations (from 5 to 15) and a second, slowly declining phase. However, he found that some behaviors do not seem to have a rapid, early phase and that many more stimuli (from 50 to several hundreds) are required to achieve the same absolute level of response decrement. The kinetics of the learning curves could be fitted by the equation:

$$H_{n+1} = H_n(1 + P_n) \quad (1)$$

in which H_n is the degree of habituation of a normalized reflex before the n th stimulus and P_n is a factor that determines the change in H from stimulus n to $n + 1$. The factor P_n decreases as habituation increases, such that

$$P_n = P_1(H_{\max} - H_n) \quad (2)$$

where H_{\max} is the maximum amount of habituation. Interestingly, when Christoffersen (1997) solved for the plasticity parameter, P_1 , for numerous cases of habituation, he found that it was many times smaller for slowly habituating reflexes than for rapidly habituating reflexes, suggesting two distinct types of habituation. Thus, this computational approach to examining habituation extends Thompson and Spencer's criteria by indicating that not all reflexes habituate with the same rapidity.

Experimental Preparations and Neural Mechanisms Underlying Habituation

Habituation has been extensively studied in invertebrate preparations, in particular the defensive reflexes of the mollusk *Aplysia* and crayfish, as their relatively simple and accessible nervous systems provide the opportunity to study the neural basis of behavior in considerable detail (see INVERTEBRATE MODELS OF LEARNING: *APLYSIA* AND *HERMISSENDA*). Importantly, these preparations meet many of the criteria put forth by Thompson and Spencer and show short- and long-term forms of habituation. Reflex habituation has also been studied in vertebrate preparations, first in the scratch reflex of the dog and later in the vestibulo-ocular reflex (VOR) in goldfish, and the prey-catching orienting response in frog. (See SCRATCH REFLEX, VESTIBULO-OCULAR REFLEX, and VISUOMOTOR COORDINATION IN FROG AND TOAD for more discussion.)

Neurophysiological studies in invertebrates examining the mechanisms underlying behavioral habituation have led to the widespread notion that habituation results from homosynaptic depression (the activity-dependent decrement in synaptic efficacy) of one or more synapses in a reflex circuit (i.e., an intrinsic mechanism). There are numerous examples showing that synaptic depression occurs during or as a result of a habituation process and that the kinetics of synaptic depression are similar to those of behavioral habituation (for a review, see Christoffersen, 1997). However, even in *Aplysia*, where synaptic depression has been firmly linked to habituation, there is evidence suggesting that other mechanisms may contribute. For example, Stopfer and Carew (1996) found that *facilitation* of the tail sensorimotor synapses accompanied habituation of the tail-elicited siphon withdrawal reflex, suggesting that plasticity at interneuronal sites underlies habituation of this reflex.

Such an "intrinsic" mechanism, where plasticity occurs within the circuit mediating the behavior, is the most popular model, but cases in which habituation also involves modulation from cells extrinsic to the circuit (e.g., heterosynaptic depression) have also been reported. For example, whereas homosynaptic depression of presynaptic sensory cells has been found to contribute to habituation of the crayfish tail-flip reflex, more recent investigations have implicated an increase in tonic descending inhibition as an additional mechanism of habituation in this system (Krasne and Teshiba, 1995). Thus, even within the same animal, homosynaptic and heterosynaptic mechanisms may contribute to habituation in different response systems.

Additional evidence of "top-down" processing in habituation has come from studies of higher vertebrate systems. Frogs with lesions of the medial pallidum, a homolog of the mammalian hippocampus, show no habituation of the orienting response to repeated visual stimuli (Finkenstadt, 1989). In a different behavioral paradigm, decerebrate rats failed to show long-term habituation of acoustic startle despite exhibiting short-term habituation (Leaton, Casella, and Borszcz, 1985). These data suggest that short-term and long-term forms of habituation may develop by separate mechanisms, with short-term habituation occurring within the circuit mediating the reflex (perhaps by homosynaptic depression) and long-term habituation occurring via descending modulatory input (perhaps by heterosynaptic depression or tonic synaptic inhibition). Experiments in *Aplysia* have indicated a morphological correlate of long-term

habituation: a decreased number of synaptic endings (Bailey and Chen, 1983).

The experiments of Sokolov (1960) on the orienting reflex in humans demonstrate the complex “top-down” control of habituation. For instance, he showed that habituation of the orienting response did not generalize to other stimuli but was stimulus specific, such that any change in the stimulus parameters (e.g., intensity, duration) resulted in a reinstatement of the response (dishabituation). His observation that even *decreases* in stimulus intensities resulted in dishabituation argued against homosynaptic depression of sensory processing. In an incredible demonstration of top-down processing, he showed that habituated responses spontaneously recovered during sleep, became resistant to rehabituation during sleep, and returned to the habituated state when the subject reawakened. These data led him to propose a “comparator theory” of habituation, in which an internal representation, or “template,” is created in the brain, to which subsequent external stimuli are compared. Stimuli that “match” the internal representation are habituated, whereas stimuli that are “mismatches” evoke a response (i.e., dishabituation). Sokolov (1960) proposed a neural model of habituation based on his theory in which the cortex, the site of the comparator mechanism, inhibits the reticular formation and, hence, the physiological correlates of the orienting response.

In summary, habituation has been linked to homosynaptic depression within the reflex circuit in a variety of systems. Accordingly, homosynaptic depression continues to be widely regarded as the cellular mechanism underlying habituation, especially short-term forms. However, considerable evidence demonstrating the importance of extrinsic modulatory processes, such as descending cortical input in mammals and tonic inhibition in crayfish, suggests that some forms of habituation may be due to mechanisms both intrinsic and extrinsic to the reflex circuit.

Computational Models of Habituation

Habituation is particularly well-suited for a computational analysis. First, the habituated response is a repeatable behavior, thus allowing for a detailed analysis of the kinetics of learning (i.e., ISI function, trials to criterion, etc.). Second, the habituated behavior is, in many cases, an evoked reflex that relies on a relatively simple neural circuit; for example, in invertebrate preparations such as *Aplysia* and crayfish, many of the individual neural elements in the reflex circuit have been identified. Third, the general underlying assumption guiding computational modeling—namely, that network output is linked to circuit architecture, cellular properties, and synaptic plasticity—also guides physiological studies of behavior, thus allowing the results of computational and biological analyses of habituation to inform each other.

Most attempts at modeling habituation have been top-down and have focused either implicitly or explicitly on short-term, within-session habituation typically involving synaptic plasticity-like learning mechanisms. For example, Horn (1967) put forth a remarkably modern theoretical network model based on synaptic mechanisms very similar to homosynaptic depression and heterosynaptic facilitation. In his model, various features of habituation and dishabituation are accounted for by independent decrementing and facilitating processes, foreshadowing the dual process theory of Groves and Thompson (1970). Stanley (1976) took a similar approach, but added mathematical complexity to his modeling of habituation of the hindlimb flexion reflex in spinal cats. He posited a circuit architecture consisting of two independent pathways, (1) a direct pathway between input and output capable of use-dependent decrement and (2) an indirect pathway with an additional intercalated element between input and output capable of use-dependent enhancement. This simple organization captured the essence of the dual process theory: repetitive low-intensity stimulation of

the direct pathway resulted in synaptic decrement, whereas moderate- and high-intensity stimulation increasingly recruited the indirect, facilitatory pathway, leading to dishabituation (and sensitization). Synaptic change was simulated using a first-order differential equation:

$$\tau \frac{dy(t)}{dt} = \alpha(y_0 - y(t)) - S(t) \quad (3)$$

where y_0 is the initial synaptic strength, $S(t)$ is the effect of external stimulation (which could be habituating or sensitizing), τ is a time constant governing the rate of habituation, and α determines the rate of recovery. This model yielded exponential learning curves typical of the short-term habituation data with varying intensity stimulation but could not account for long-term habituation.

Wang and Arbib (1992) modified the traditional first-order differential equation to include an activity-gated input and included a second equation that, by regulating the rate of recovery, captured the transition from short-term to long-term habituation in their model of habituation of prey-catching behavior in frogs. These two equations took the form:

$$\tau \frac{dy(t)}{dt} = \alpha z(y_0 - y(t)) - \beta y(t)S(t) \quad (4)$$

$$\frac{dz(t)}{dt} = \gamma z(t)(z(t) - 1)S(t) \quad (5)$$

where the second term in Equation 4 constitutes activity-gated input with β as a rate parameter. In Equation 5, which regulates recovery, variable $z(t)$ generates an inverse S-shaped curve under constant stimulation $S(t)$. When $z(t)$ is large after few trials, recovery is rapid. In contrast, when $z(t)$ is small, as it is after many trials, recovery is slow. Thus, as habituation progresses, recovery becomes more prolonged, producing longer-lasting habituation. This approach yielded simulated results that resembled experimental data for short recovery times (minutes) but underestimated forgetting for longer recovery times (hours).

Recently, Anastasio (2001) presented a computational model of habituation of the goldfish VOR to sinusoidal stimulation that is conceptually similar to the comparator theory of Sokolov (1960). Structurally, the model is based on a simplified view of vestibular neuroanatomy: a direct excitatory path through the brainstem (BS) to the output of the circuit, the vestibular nucleus (VN), and an inhibitory and plastic indirect path through the vestibulocerebellum (VC) (Figure 1). Behavioral VOR habituation shows frequency-specificity and nonlinearity; the most pronounced habituation occurs at the peak stimulus amplitude and does not occur at the same rate for the two directions of sinusoidal head rotations. These features suggested that the habituation mechanism does not treat the stimulus as a continuous function but as discrete units that habituate independently. Thus, a primary component of the model is that the sinusoidal stimulus is partitioned into segments or patterns. In the model, habituation occurs because of a pattern correlation mechanism in which the VC compares the similarity (the correlation) of the current stimulus (the pattern) with the representation of previous stimuli (a history vector stored in the VC) and weights the inhibition of the VN according to the maximum correlation using the equation:

$$y(t) = x(t) - w_k r_k(t) \quad (6)$$

where $y(t)$ is the output of the VC at time step t , $x(t)$ is the value of the input pattern, and w_k is the weight value corresponding to the pattern having the maximum correlation, r_k , with the history vector. After filtering and dithering the simulation output to mimic neuromuscular dynamics, Anastasio (2001) showed that the model can accurately account for the features of behavioral VOR habituation, thus providing a computational mechanism for producing

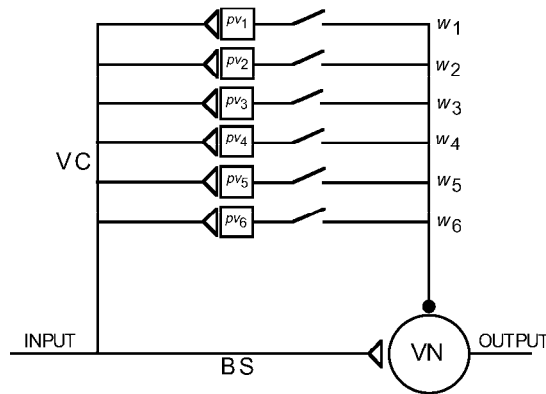


Figure 1. Anastasio's (2001) pattern correlation model of VOR habituation. Open triangles represent excitatory connections. Solid circles represent inhibitory connections. BS, brainstem; VC, vestibulocerebellar nucleus; VN, vestibular nucleus. Pattern vectors and their respective weights are represented by pv_i and w_i ($i = 1, 2, \dots, 6$). (Adapted from Anastasio, T. J., 2001, A pattern correlation model of vestibulo-ocular reflex habituation, *Neural Netw.*, 14:1–22. Reproduced with permission.)

discontinuous, nonlinear output plasticity from a continuous, elementary function input. Moreover, the pattern correlation model achieves computational instantiations of a Sokolov-like comparator mechanism by storing and matching a history vector to subsequently present stimuli.

Discussion

Habituation continues to be an attractive paradigm in the study of the neural basis of learning and memory. Experimental studies have identified at least two important neural mechanisms of habituation, homosynaptic depression within the reflex circuit and extrinsic descending modulatory input. Although the computational models of habituation described above have achieved success in accounting for important aspects of behavioral habituation in some systems, much work remains to be done in simulating processing in known biological circuitry. There are several well-studied model systems conducive to a biologically realistic computational analysis, such as the tail-flip escape reflex in crayfish and the tap-elicited withdrawal reflex in the nematode. In each of these cases, much of the underlying neural circuitry is known. Habituation in the crayfish tail-flip reflex, due to both afferent depression as well as descending inhibition, is attractive because it offers the opportunity to analyze the interaction and cooperativity of mechanisms intrinsic and extrinsic to the reflex circuit. The nematode *C. elegans* is attractive because of the possibility of a genetic analysis of habituation (see

Rose and Rankin, 2001). Progress in understanding the computational processes of habituation in complex vertebrates and mammals will require greater knowledge of underlying circuitry and neurophysiology. The goal of future work, both experimental and computational, will be to account for the intriguing complexity of this not-so-simple form of learning.

Road Map: Neural Plasticity

Related Reading: Invertebrate Models of Learning: *Aplysia* and *Hermisenda*; Visuomotor Coordination in Frog and Toad

References

- Anastasio, T. J., 2001, A pattern correlation model of vestibulo-ocular reflex habituation, *Neural Netw.*, 14:1–22.
- Bailey, C. H., and Chen, M., 1983, Morphological basis of long-term habituation and sensitization in *Aplysia*, *Science*, 220:91–93.
- Christoffersen, G. R. J., 1997, Habituation: Events in the history of its characterization and linkage to synaptic depression. A new proposed kinetic criterion for its identification, *Prog. Neurobiol.*, 53:45–66. ♦
- Davis, M., 1970, Effects of interstimulus interval length and variability on startle-response habituation in the rat, *J. Comp. Physiol. Psychol.*, 72:177–192.
- Finkenstadt, T., 1989, Stimulus-specific habituation in toads: 2DG and lesion studies, in *Visuomotor Coordination: Amphibians, Comparisons, Models, and Robots* (J.-P. Ewert and M. A. Arbib, Eds.), New York: Plenum Press, pp. 767–797.
- Groves, P. M., and Thompson, R. F., 1970, Habituation: A dual process theory, *Psychol. Rev.*, 77:419–450.
- Hinde, R. A., 1970, Behavioral habituation, in *Short-Term Changes in Neural Activity and Behavior* (G. Horn and R. A. Hinde, Eds.), Cambridge, Engl.: Cambridge University Press, pp. 3–40. ♦
- Horn, G., 1967, Neuronal mechanisms of habituation, *Nature*, 215:707–711.
- Krasne, F. B., and Teshiba, T. M., 1995, Habituation of an invertebrate escape reflex due to modulation by higher centers rather than local events, *Proc. Natl. Acad. Sci. USA*, 92:3362–3366.
- Leaton, R. N., Casella, J. V., and Borszcz, G. S., 1985, Short-term and long-term habituation of the acoustic startle response in chronic decerebrate rats, *Behav. Neurosci.*, 99:901–912.
- Rose, J. K., and Rankin, C. H., 2001, Analyses of habituation in *Caenorhabditis elegans*, *Learn. Mem.*, 8:63–69.
- Sokolov, E. N., 1960, Neuronal models and the orienting reflex, in *The Central Nervous System and Behavior* (M. A. B. Brazier, Ed.), Madison, NJ: Madison, pp. 187–276. ♦
- Stanley, J. C., 1976, Computer simulation of a model of habituation, *Nature*, 261:146–148.
- Stopfer, M., and Carew, T. J., 1996, Heterosynaptic facilitation of tail sensory neuron synaptic transmission during habituation in induced tail and siphon withdrawal reflexes of *Aplysia*, *J. Neurosci.*, 16:4933–4948.
- Thompson, R. F., and Spencer, W. A., 1966, Habituation: A model phenomenon for the study of neuronal substrates of behavior, *Psychol. Rev.*, 73:16–43. ♦
- Wang, D., and Arbib, M. A., 1992, Modeling the dishabituation hierarchy: The role of the primordial hippocampus, *Biol. Cybern.*, 67:535–544.

Half-Center Oscillators Underlying Rhythmic Movements

Andrew A. V. Hill, Stephen D. Van Hooser, and Ronald L. Calabrese

Introduction

The half-center oscillator model was first proposed by T. Graham Brown (1914) to account for the observation that spinal cats could

produce stepping movements even when all dorsal roots were severed, thereby eliminating sensory feedback from the animals' motion. He envisioned pools of interneurons controlling flexor and extensor motor neurons (the half-centers) that had reciprocal inhib-

itory connections, and that were capable of sustaining alternating oscillatory activity if properly activated. In the model, he assumed that the duration of reciprocal inhibition was limited by some intrinsic factor, e.g., synaptic fatigue, and that the neuron pools (half-centers) showed rebound excitation. In the intervening years we have learned that almost all rhythmic movements of animals are programmed in part by central pattern-generating networks that comprise neural oscillators (MOTOR PATTERN GENERATION; SPINAL CORD OF LAMPREY: GENERATION OF LOCOMOTOR PATTERNS; CRUSTACEAN STOMATOGASTRIC SYSTEM; and LOCOMOTION, VERTEBRATE). In many of these motor pattern-generating networks, in both vertebrates and invertebrates, reciprocal inhibitory synaptic interactions between neurons or groups of neurons are found (Calabrese, 1995). We are beginning to understand how the intrinsic membrane properties of the component neurons interact with reciprocal inhibition to initiate and sustain oscillation in these networks.

Theoretical Framework

A theoretical framework for understanding how reciprocally inhibitory neurons oscillate (i.e. how half-center oscillators work) was developed by Wang and Rinzel (1992) (OSCILLATORY AND BURSTING PROPERTIES OF NEURONS). Their model neurons are minimal. Each contains a synaptic conductance that is a sigmoidal function of presynaptic membrane potential with a set threshold and instantaneous kinetics, a constant leak conductance, and a voltage-gated postinhibitory rebound current, I_{pir} . I_{pir} was originally envisioned to be a T-like Ca^{2+} current (low-threshold, inactivating), but its expression in the model can also accommodate an h current (hyperpolarization activated inward current) (Wang and Rinzel, 1992) or a Ca^{2+} dependent K^+ current (Grillner et al., 2000). Two different modes of oscillation appear in the model, "release" and "escape" (Wang and Rinzel, 1992). For the release mode to occur, the synaptic threshold must be above the steady state membrane potential of the uninhibited neurons. In the release mode, the inactivation of I_{pir} erodes the depolarized or active phase of a neuron so that it falls below threshold for synaptic transmission. Consequently, its partner is released from inhibition and rebounds into the active depolarized state. For the escape mode to occur the synaptic threshold must be below the steady-state voltage of the neurons when uninhibited. This condition can be accomplished simply by increasing g_{pir} . In the escape mode, once inactivation of I_{pir} is removed by the hyperpolarization associated with inhibition, it activates and overcomes the maintained synaptic current so that the neuron escapes into the active phase and thus inhibits its partner.

Skinner, Kopell, and Marder (1994) have extended this analysis using similar model neurons based on the Morris-Lecar equations (low-threshold noninactivating inward current and delayed rectifier current) with a synaptic conductance, which is a steep sigmoidal function of presynaptic membrane potential with a set threshold and instantaneous kinetics. Such model neurons, like the Wang and Rinzel neurons, oscillate between a depolarized plateau and a sustained inhibitory trough. Each of the two modes of oscillation can be further differentiated depending on whether the escape or release is intrinsic or synaptic. If the release is due to a cessation of synaptic transmission (crossing synaptic threshold), it is synaptic release, but if it is due to termination (deactivation of the inward current, activation of the delayed rectifier, or both) of the depolarized plateau, it is intrinsic release. If the escape is due to the commencement of synaptic transmission (crossing synaptic threshold), it is synaptic escape, but if it is due to expression of the depolarized phase (crossing plateau threshold), it is intrinsic escape. Varying the synaptic threshold causes transitions between the modes.

These theoretical studies have been corroborated by hybrid systems studies in which artificial reciprocal inhibitory synapses were introduced between crustacean stomatogastric neurons that are nor-

mally unconnected by using dynamic clamp (Sharp, Skinner, and Marder, 1996). To obtain robust oscillations an artificial h current was also added. By adjusting the synaptic threshold of the artificial synapses it was possible to capture all of the richness of the dynamic systems analysis. More recently, a theoretical analysis, based on work in crustacean stomatogastric networks, has shown that even passive neurons employing graded synapses can generate oscillation in the half-center configuration (Manor et al., 1999). Although the models used in the theoretical analyses are simplistic, the major insights that they impart can be transferred to biological neurons, which are connected by simplistic artificial synapses. Real neurons display more complicated intrinsic membrane properties and plastic synaptic interactions that may blur the conclusions of these analyses, but nevertheless they serve as a useful organizing point for the exploration of richer biological systems.

Leech Heartbeat

This review will now focus on the motor pattern generating network that controls heartbeat in the leech. Progress has been made in understanding the role reciprocal inhibitory synaptic interactions and membrane properties play in generating oscillations by combining experimental analyses with realistic modeling.

A network of seven bilateral pairs of segmental heart (HN) interneurons produces rhythmic activity (at about 0.1 Hz) that paces segmental heart motor neurons, which in turn drive the two hearts. The synaptic connections among the interneurons and from the interneurons to the motor neurons are inhibitory. The first four pairs of heart interneurons control the timing of the network. The timing oscillation is dominated by the activity of the third and fourth pairs of heart interneurons. Reciprocally inhibitory synapses between these bilateral pairs of oscillator interneurons, combined with their ability to escape from inhibition and begin firing, pace the oscillation (Figure 1A). Thus, each of these two reciprocally inhibitory heart interneuron pairs can each be considered an elemental half-center oscillator. The first two pairs of heart interneurons act as coordinating fibers, serving to link these two elemental oscillators (Figure 1B).

Several ionic currents have been identified in single electrode voltage-clamp studies that contribute to the activity of oscillator heart interneurons (Calabrese, Nadim, and Olsen, 1995). These include, in addition to the fast Na^+ current that mediates spikes, two low-threshold Ca^{2+} currents [one rapidly inactivating (I_{CaF}) and one slowly inactivating (I_{CaS})], three outward currents [a fast transient K^+ current (I_A) and two delayed rectifier-like K^+ currents, one inactivating (I_{K1}) and one persistent (I_{K2})], a hyperpolarization-activated inward current (I_h) (mixed Na^+/K^+ , $E_{rev} = -20$ mV), a low-threshold persistent Na^+ current (I_p) and a leakage current (I_L). The inhibition between oscillator interneurons consists of a graded component that is associated with the low-threshold Ca^{2+} currents and a spike-mediated component that appears to be mediated by a high-threshold Ca^{2+} current. Spike-mediated transmission varies in amplitude throughout a burst according to the baseline level of depolarization (Olsen and Calabrese, 1996). Graded transmission wanes during a burst owing to the inactivation of low-threshold Ca^{2+} currents.

Much of this biophysical data has been incorporated into a detailed conductance-based model of an elemental (two-cell) oscillator (Nadim et al., 1995; Hill et al., 2001). This model uses standard Hodgkin-Huxley representations of each voltage-gated current. Synaptic transmission in the model is complex. A spike-triggered alpha-function is used to describe the postsynaptic conductance associated each action potential and the maximal conductance reached is a function of the past membrane potential to reflect the fact that spike-mediated transmission varies in amplitude throughout a burst according to the baseline level of depolarization. Graded

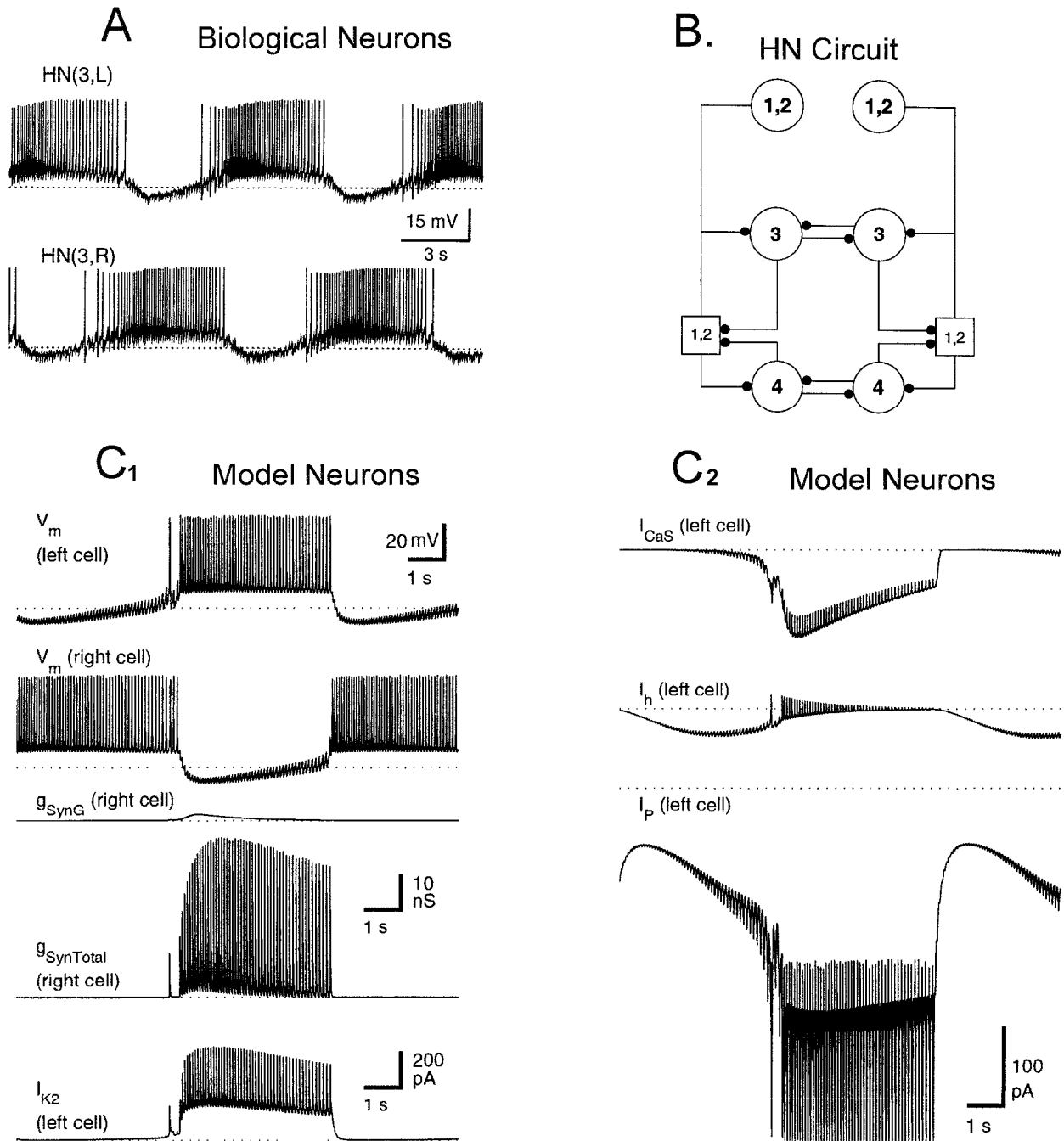


Figure 1. *A*, Simultaneous intracellular recordings showing the normal rhythmic activity of two reciprocally inhibitory heart (HN) interneurons that compose a half-center oscillator in an isolated ganglion preparation. Heart interneurons are indexed by body side (R, L) and ganglion number. *B*, Circuit diagram showing inhibitory synaptic connections among the HN interneurons of the timing network. Coordinating neurons HN(1) and HN(2) are functionally equivalent and are lumped together in the diagram. The HN(1) and HN(2) neurons receive synaptic inputs, initiate action potentials, and make synaptic outputs at sites located in the third and fourth ganglia (open squares). *C*₁–*C*₂, Synaptic conductances and some major

intrinsic currents that are active during a single cycle of the third-generation model of a two-cell HN interneuron oscillator (half-center). The graded synaptic conductance (g_{SynG}) is shown at the same scale as the total synaptic conductance (g_{SynTotal}), which is the sum of the graded and spike-mediated conductances. The slow calcium current (I_{CaS}), the hyperpolarization-activated current (I_h), and the persistent sodium current (I_p) are shown to the same scale. Note that I_p is active throughout the entire cycle period (Hill and Calabrese, unpublished). In *A* and *C* dashed lines indicate -50 mV in voltage traces, 0 nA in current traces, and 0 nS in conductance traces.

synaptic transmission is represented by a synaptic transfer function, which relates postsynaptic conductance (the result of transmitter release) to presynaptic Ca^{2+} build-up and decline, via low-threshold Ca^{2+} currents and a Ca^{2+} removal mechanism, respectively. The model is now in its fourth generation (Hill et al., 2001), having been upgraded each time by the incorporation of new data from experiments suggested by the previous generation of the model (Olsen and Calabrese, 1996). Free parameters in the model are the maximal conductance ($g_{\text{max}_{ion}}$) for each current (voltage-gated or synaptic). The $g_{\text{max}_{ion}}$ s were adjusted to be close to the average observed experimentally. The reversal potential, E_{ion} , for each current was determined experimentally and they were considered fixed. Final selection of parameters to form a canonical model was dictated by model behavior under control conditions, passive response of the model to hyperpolarizing current pulses, and reaction of the model to current perturbations. The model cells were tuned by adjusting leak parameters (E_{leak} and $g_{\text{max}_{leak}}$) so that they fire tonically when all inhibition between them was blocked. This tuning was chosen because under conditions of our experiments, which use sharp microelectrodes for recordings, the neurons fire tonically when synaptically isolated with bicuculline (Schmidt and Calabrese, 1992). Recent unpublished work (Gaudry, Cymbalyuk, and Calabrese, Department of Biology, Emory University) suggest that oscillator heart interneurons burst endogenously when isolated with bicuculline. Because the cells must be recorded extracellularly to reveal this bursting, we hypothesize that their bursting behavior is very sensitive to leak parameters that are altered with intracellular recording methods.

The canonical model generates activity, which closely approximates that observed for an elemental half-center oscillator (Figure 1C). Analysis of current flows during this activity (Figure 1C) indicates that graded transmission occurs only at the beginning of the inhibitory period due to inactivation of the low-threshold Ca^{2+} currents that mediate this inhibition. Thus, graded inhibition helps turn off the antagonist neuron, but sustained inhibition of the antagonist neuron is all spike-mediated. The inward currents in the model neurons act to overcome this inhibition and force a transition to burst phase of oscillation. I_p is active throughout the activity cycle, providing a persistent excitatory drive to the system. I_h is slowly activated by the hyperpolarization-associated inhibition, adding a delayed inward current that drives further activation of I_p and eventually the low-threshold Ca^{2+} currents (I_{CaS} and I_{CaF}). These regenerative currents support burst formation. I_p , because it does not inactivate, provides steady depolarization to sustain spiking, while the low-threshold Ca^{2+} currents help force the transition to the burst phase but inactivate as the burst proceeds, thus spike frequency slowly wanes during the burst. Outward currents also play important roles, especially the I_{KS} , I_{K2} , which activates and deactivates relatively slowly and does not inactivate, regulates the amplitude of the depolarization that underlies the burst, while I_{K1} , which activates and deactivates relatively quickly and inactivates, controls spike frequency.

Increasing $g_{\text{max}_{\text{SynS}}}$ (the maximal spike-mediated synaptic conductance) in the model slows the oscillation while reducing $g_{\text{max}_{\text{SynS}}}$ speeds the oscillation. Under canonical conditions graded transmission is suppressed (Nadim et al., 1995). Analysis of state variables (m and h) for low-threshold Ca^{2+} currents indicates that deinactivation of these currents is not effective during the inhibitory period. In the canonical model cells, as in the real cells, the potential for prolonged and intense graded transmission is revealed on rebound from a hyperpolarizing pulse.

The period of the oscillation is sensitive to the level of g_{max_h} , as would be predicted from its key role in forcing the transition from the inhibitory phase to the burst phase. Decreasing g_{max_h} from canonical levels slows the oscillation proportionately, while increasing it speeds the oscillations. In contrast, increasing g_{max} of I_p , the other inward current active during the inhibited phase,

slows the oscillation and decreasing g_{max_p} speeds the oscillation. These observations indicate that the predominate effect of I_p is to prolong the burst phase and concomitant inhibition of the antagonist heart interneuron, rather than to promote escape during the inhibited phase. Addition of a slowly activating and deactivating outward current (I_{KF}), which is induced by the endogenous neuropeptide FMRFamide, speeds the cycling of the elemental oscillator model as it does in the biological oscillator interneurons (Nadim and Calabrese, 1997).

It appears that in the heart interneuron half-center oscillators forces that promote both escape and release are at work. Spike-mediated transmission gradually wanes during a burst because of the slowly declining envelope of depolarization during the burst phase, which slows spike frequency and down modulates IPSP amplitude (Figure 1C). This decline in the inhibition of the inhibited cell represents a release. Indeed, if this decline is eliminated in the model by eliminating inactivation from I_{CaS} , then oscillations cease (Hill et al., 2001). Nevertheless, whenever I_h is sufficiently activated to overcome the waning synaptic current, a transition from the inactive state to the active state occurs, and the "trigger point" for this release is determined by g_{max_h} , which is consistent with an escape mechanism (Hill et al., 2001). Thus this half-center oscillator is not easily categorized, but the theoretical comparisons have been illuminating.

To fully explore the leech heart interneurons half-center model and a model of the entire heartbeat timing network, they can be downloaded from <http://calabreseix.biology.emory.edu> by anonymous ftp. Various UNIX-based operating systems, including LINUX, are supported.

Conclusions

The half-center-based pattern generators controlling swimming in the pelagic mollusk *Clione* (Arshavsky et al., 1993), in tadpoles (*Xenopus*) (Roberts et al., 1998), and in lampreys (Grillner et al., 2000), which have also been extensively analyzed and modeled, employ neurons that have pacemaking or bursting properties. Both the *Clione* and the *Xenopus* tadpole swim oscillators appear to operate in the intrinsic release mode (Arshavsky et al., 1993). "Spike" (active state) termination terminates inhibition and allows the antagonist cell to rebound into the active state. Perhaps this mode is more suited to the operational frequency range of these oscillators, which are some ten times faster (about 1 Hz) than the leech heartbeat oscillator.

Acknowledgments. Research in the authors' lab was supported by NIH grant NS24072.

Road Map: Motor Pattern Generators

Related Reading: Chains of Oscillators in Motor and Sensory Systems; Crustacean Stomatogastric System; Locomotion, Invertebrate; Locomotion, Vertebrate; Locust Flight: Components and Mechanisms in the Motor; Oscillatory and Bursting Properties of Neurons; Sensorimotor Interactions and Central Pattern Generators

References

- Arshavsky, Y. I., Orlovsky, G. N., Panchin, Y. V., Roberts, A., and Soffe, S. R., 1993, Neuronal control of swimming locomotion: Analysis of the pteropod mollusk *Clione* and embryos of the amphibian *Xenopus*, *TINS*, 16:227-233. ♦
- Brown, T. G., 1914, On the nature of the fundamental activity of the nervous centres; together with an analysis of the conditioning of rhythmic activity in progression, and a theory of the evolution of function in the nervous system, *J. Physiol. (Lond.)*, 48:18-46.
- Calabrese, R. L., 1995, Oscillation in motor pattern generating networks, *Curr. Opin. Neurobio.*, 5:816-823.
- Calabrese, R. L., Nadim, F., and Olsen Ø. H., 1995, Heartbeat control in the medicinal leech: a model system for understanding the origin, co-

- ordination, and modulation of rhythmic motor patterns, *J. Neurobiol.*, 27:390–402.
- Grillner, S., Cangiano, L., Hu, G.-Y., Thompson, R., Hill, R., and Wallen, P., 2000, The intrinsic function of a motor system—from ion channels to networks and behavior, *Brain Res.*, 886:224–236. ♦
- Hill, A. A. V., Lu, J., Masino, M. A., Olsen, Ø. H., and Calabrese, R. L., 2001, A model of a segmental oscillator in the leech heartbeat neuronal network, *J. Compu. Neurosci.*, 10:281–302.
- Manor, Y., Nadim, F., Epstein, S., Ritt, J., Marder, E., and Kopell, N., 1999, Network oscillations generated by balancing graded asymmetric reciprocal inhibition in passive neurons, *J. Neurosci.*, 19:2765–2779.
- Nadim, F., Olsen, Ø. H., De Schutter, E., and Calabrese, R. L., 1995, Modeling the leech heartbeat elemental oscillator: I. Interactions of intrinsic and synaptic currents, *J. Compu. Neurosci.*, 2:215–235. ♦
- Nadim, F., and Calabrese, R. L., 1997, A slow outward current activated by FMRFamide in heart interneurons of the medicinal leech, *J. Neurosci.*, 17:4461–4472.
- Olsen, Ø. H., and Calabrese, R. L., 1996, Activation of intrinsic and synaptic currents in leech heart interneurons by realistic waveforms, *J. Neurosci.*, 16:4958–4970.
- Roberts, A., Soffe, S. R., Wolf, E. S., Yoshida, M. and Zhao, R. L., 1998, Central circuits controlling locomotion in young frog tadpoles, *Ann. NY Acad. Sci.*, 860:19–34. ♦
- Schmidt, J., and Calabrese, R. L., 1992, Evidence that acetylcholine is an inhibitory transmitter of heart interneurons in the leech, *J. Exp. Biol.*, 171:339–347.
- Sharp, A. A., Skinner, F. K., and Marder, E., 1996, Mechanisms of oscillation in dynamic clamp constructed two-cell half-center circuits, *J. Neurophysiol.*, 76:867–883.
- Skinner, F. K., Kopell, N., and Marder, E., 1994, Mechanisms for oscillation and frequency control in reciprocally inhibitory model neural networks, *J. Compu. Neurosci.*, 1:69–87.
- Wang, X.-J., and Rinzal, J., 1992, Alternating and synchronous rhythms in reciprocally inhibitory model neurons, *J. Neural Comp.*, 4:84–97. ♦

Hebbian Learning and Neuronal Regulation

Gal Chechik, David Horn, and Eytan Ruppin

Introduction

Since its conception half a century ago, Hebbian learning has become a fundamental paradigm in the neurosciences. The idea that neurons that fire together wire together has become fairly well understood, as in the case of NMDA-dependent long-term potentiation in the hippocampus (Bliss and Collingridge, 1993). However, for both computational and biological reasons, this type of plasticity has to be accompanied by synaptic changes that are not synapse specific but neuron specific; i.e., they involve many synapses of the same neuron. Biologically, such interactions are inevitable as synapses compete for finite resources and are subject to common processes of the same neuron to which they all belong. Computationally, neuron-specific modifications of synaptic efficacies are required in order to obtain efficient learning, or to faithfully model biological systems. Hence, *neuronal regulation*, defined here as a process modulating all synapses of a postsynaptic neuron, is a general phenomenon that complements Hebbian learning.

There exists evidence for cellular mechanisms resulting in normalization of synaptic efficacies, some of which operate to maintain total synaptic strength and others to regulate mean postsynaptic activity (Miller, 1996). Among these mechanisms are cellular regulation of the number of synapses or of trophic factors, competition between synapses for some finite resources, changes in presynaptic and postsynaptic learning thresholds, or activity-dependent regulation of conductances. Normalization of synaptic efficacies is also induced by certain types of plasticity as an emergent phenomenon, for example in the case of spike-time-dependent plasticity (Song, Miller, and Abbott, 2000). Of particular interest are the findings by Turrigiano et al. (1998), who studied cultures of pyramidal neurons of postnatal rats. They observed slow postsynaptic up- or down-regulation of adenosine monophosphate (AMPA)-mediated synaptic currents in a way that maintained the mean firing activity of the neuron. This scaling resulted in overall synaptic normalization through a multiplicative factor that is inversely related to the neuron's activity.

What are the computational consequences of such neuronal-level processes? It turns out that learning through Hebbian learning alone raises many theoretical difficulties and questions, such as: What stops the positive feedback loop of Hebbian learning and guarantees some normalization of the synaptic efficacies of a neuron?

How can a neuron acquire specificity to particular inputs without being prewired? How can memories be maintained throughout life, while synapses suffer degradation due to metabolic turnover? As we will see, neuronal regulation provides a possible answer to all of the above.

We can divide the computational problems to be looked at according to the traditional dichotomy of supervised and unsupervised learning. In unsupervised learning, the important role of neuronal regulation is to allow for *competition* between the various synapses, leading to *normalization* of the synaptic efficacies. This role will be further explained in the next section, where we review some basic learning paradigms, discuss the difference between multiplicative ($\Delta \mathbf{w} \propto \mathbf{w}$) and additive ($\Delta \mathbf{w} = \text{const}$) scaling, and identify some applications to biological systems. We will then turn to supervised learning paradigms, and show that neuronal regulation improves the *capacity* of associative memory models and can be used to guarantee the *maintenance* of biological memory systems.

Unsupervised Learning

When Hebbian plasticity operates in a network in an unsupervised manner, a positive feedback loop is created. To illustrate the problem, think of a presynaptic cell *A* that causes the firing of a postsynaptic cell *B*. Because of Hebbian plasticity, the efficacy of the synapse w_{BA} from *A* to *B* is strengthened. This leads to an increase in the ability of cell *A* to activate cell *B*, which in turn leads to strengthening of the same synapse again. When this positive feedback is unconstrained, it leads to synaptic runaway, i.e., the divergence of synaptic efficacies. It seems reasonable to assume that synaptic values are limited by some upper bound that stops this process. Even then a problem emerges: different afferents will activate different synapses of the target neuron *B*. When all of them get saturated at their upper bounds, the neuron will not have any discrimination ability. This problem may be solved by introducing constraints, such as limiting the total synaptic strength of a neuron, $\sum_i w_i^2 = \text{const}$. This results in competition between synapses: the increased strength of one synapse causes a decrease in the strength of another, preventing saturation of all synapses.

Multiplicative Versus Additive Constraints

Normalization prevents synaptic divergence; thus, the combined operation of Hebbian learning and normalization induces new dynamics of synaptic efficacies. This combined dynamics was described by Oja (1982), showing that multiplicative weight normalization of a neuron with real-valued stochastic input extracts the first principal component of the input distribution (also known as PRINCIPAL COMPONENT ANALYSIS [p.c.a.] or Karunen Leove feature extraction).

Whereas PCA extraction follows for multiplicative normalization, the results change if other types of constraints, such as additive normalization, are imposed (Miller and MacKay, 1994). Although multiplicative normalization leads to a graded weight vector that represents even weak correlations in the input (upper plot of Figure 1B), additive normalization yields a sharpened receptive field where weights saturate at their lower and upper bounds in a way that reflects only the maximally correlated inputs (Figure 1B, bottom). Similar results were obtained for competitive learning, where only the weights of the winning unit are changed (Goodhill and Barrow, 1994).

Neuronal Regulation and Synaptic Normalization

What is the relation between synaptic normalization and neuronal regulation? Normalization of synaptic efficacies involves all synapses of a postsynaptic neuron, and thus requires neuronal-level computation. Moreover, in some cases synaptic normalization is an emergent result of synaptic changes that depend on neuronal activity. We approach this idea by discussing two learning models: the Oja learning rule mentioned earlier and the Bienenstock, Cooper, and Munro (BCM) model.

Using the linear perceptron $V = \sum_i w_i x_i$, Oja's learning rule can be implemented by $\Delta w_i = \eta V(x_i - Vw_i)$. Here, neuronal regulation is explicitly manifested by the second term, which provides a multiplicative correction that is independent of the specific input x_i but is determined by the neuronal output V . Interestingly, this neuronal regulation term guarantees $\sum_i w_i^2 = 1$.

The BCM model (Bienenstock, Cooper, and Munro, 1982) is

another example of complex interplay of neuronal regulation and synaptic competition. In the BCM approach, both Hebbian potentiation and depression are used in defining the synaptic learning rule: synapses are potentiated when the presynaptic neuron fires frequently, and depressed otherwise. The boundary between potentiation and depression is determined by the activity of the postsynaptic neuron, which is where neuronal regulation comes in. This component of neuronal regulation leads to competition between synapses and introduces statistical correlations (Intrator and Cooper, 1992) that are higher than the second order used in PCA. Thus, the BCM model captures high-order statistical structures in the input and tunes the efficacies of incoming synapses accordingly.

Biological Models

The BCM model was developed to describe the emergence of orientation-selective cells and ocular dominance in the visual cortex and their dependence on the stimuli that the visual system receives during its critical developmental stages. A study that discusses the same issues with particular emphasis on the dynamics of synaptic efficacies under additive and multiplicative normalization schemes is that of Miller and MacKay (1994). They show that when the inputs to a neuron have positive correlations only, additive normalization leads to the convergence of weights to an on-center-off-surround receptive field, or to a bilobed receptive field, depending on the parameter regime. When the cell receives inputs from both eyes, additive normalization leads to ocular dominance through the sharpening of receptive fields. Multiplicative normalization can lead in this case to ocular dominance only if the inputs from both eyes are negatively correlated.

A convenient system for the study of neuronal regulation is the vertebrate neuromuscular junction (NMJ). Its development is characterized by an initial stage of superinnervation (each muscle fiber is innervated by several motor neurons) followed by withdrawal of axon terminals until a state of single innervation is reached. Modeling the NMJ, Willshaw (1993) has shown that competition for postsynaptic resources explains the decrease in innervation, yet fails to account for other experimental findings such as incomplete innervation after artificial partial denervation during development.

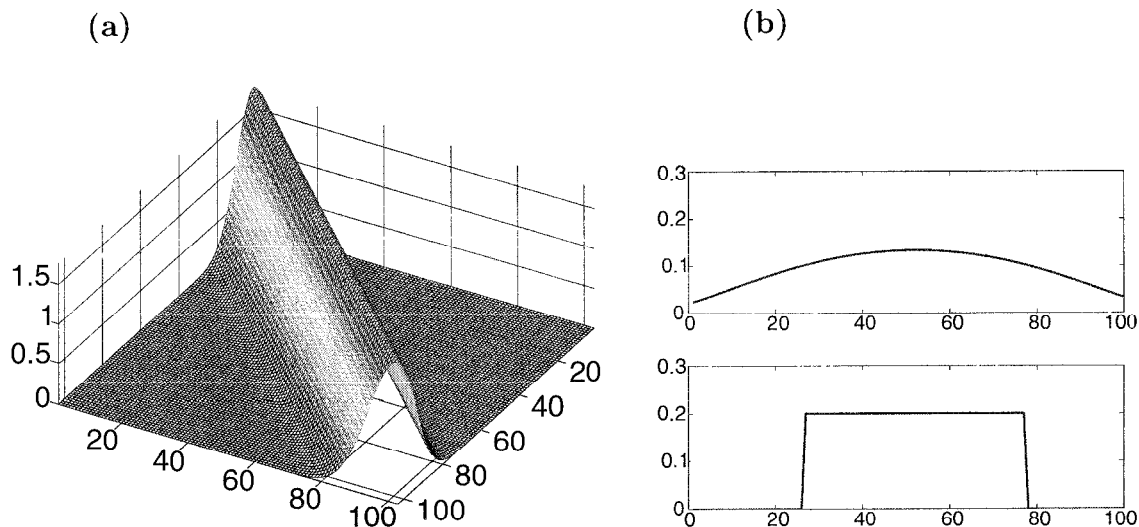


Figure 1. Steady states of synaptic weights under multiplicative and additive normalization in a toy example of a neuron with 100 inputs. *A*, The correlation matrix of all inputs. *B*, Under multiplicative constraints, the

weights converge onto the first principal component (upper plot), while an additive rule leads to a binary separation of weights that reach their extremal values (bottom plot).

However, a combination of post- and presynaptic competition provides a good account of the data. Indeed, the efficacy of the neuromuscular synapse during the period preceding axonal withdrawal was traced by Colman, Nabekura, and Lichtman (1997), who found changes in quantal content and efficacy. These changes led to continuous strengthening of some synapses with a parallel weakening of the rest, suggesting the operation of a cascade of pre- and postsynaptic processes regulating synaptic efficacies. This is therefore an interesting example in which both pre- and postsynaptic normalization cooperate during early development.

Finally, we wish to point out that competition may arise also through effects other than neuronal regulation. An example is the case of spike-time-dependent synaptic plasticity, in which potentiation of excitatory synapses occurs when the presynaptic spike shortly precedes the postsynaptic spike, and depression occurs when the opposite temporal order holds. Song et al. (2000) assumed that spike-dependent synaptic potentiation is weaker than depression and showed that in a neuron that is driven by net positive input, the excitatory synapses will be weakened. This eventually leads to a balanced input in which the more relevant excitatory synapses are strengthened. Thus, effective competition between synapses may occur even in the absence of an explicit neuronal regulatory term, one that depends on the postsynaptic neuron only.

Supervised Learning

We saw that in unsupervised learning, neuronal regulation solves the problem of synaptic runaway and guarantees specificity of neuronal response through synaptic competition. In supervised learning, normalization constraints introduced by neuronal regulation provide both maintenance of memory systems and high memory capacity.

Memory Maintenance

The concept of neuronal regulation was introduced by Horn, Levy, and Ruppin (1998) in the context of associative memory networks, while these authors were developing a model that could account

for the stability of memory systems in the face of continuous metabolic turnover of synapses. This repetitive process of synaptic degeneration and buildup occurs on a time scale of few days. Under these conditions, one wonders how memories can be stored in synaptic connections for prolonged periods. It turns out that neuronal regulation may play an important role in bringing about the necessary homeostasis of this system, i.e., account for its ability to continue to both learn and retrieve memories (Horn et al., 1998).

To understand this issue, let us consider an associative memory system that is tested through activation by random inputs. Neurons that belong to memories with large basins of attraction will be much more active than those that participate only in memories with small basins of attraction. At this point, we introduce neuronal regulation through multiplicative synaptic corrections that are inversely proportional to the activity of the postsynaptic neuron. This will up-regulate weak memories and downregulate strong ones. The multiplicative nature of the correction guarantees that the relative weights of different memories on the same neuron are maintained. The result of this procedure is depicted in Figure 2. As can be seen, repeated synaptic degradation and neuronal regulation leads to normalization of the basins of attraction.

The homeostasis strategy suggested by Horn et al. (1998) involves repeated sessions of random activation, synaptic degradation, and neuronal regulation that provide the required maintenance of the network after it goes through some period of Hebbian learning. It can be shown that the combined effect of synaptic degradation and neuronal regulation also results in the removal of weak synapses, owing to emerging synaptic competition (Chechik, Meilijson, and Ruppin, 1999). This can provide insight into the phenomenon of synaptic pruning that is believed to occur during early development in mammals (see Quartz and Sejnowski, 1997, for a review of the constructive versus selectionist approaches to brain development).

Learning Capacity

Normalization of synaptic efficacies plays a crucial role in producing effective Hebbian learning: Without normalization, Hebbian

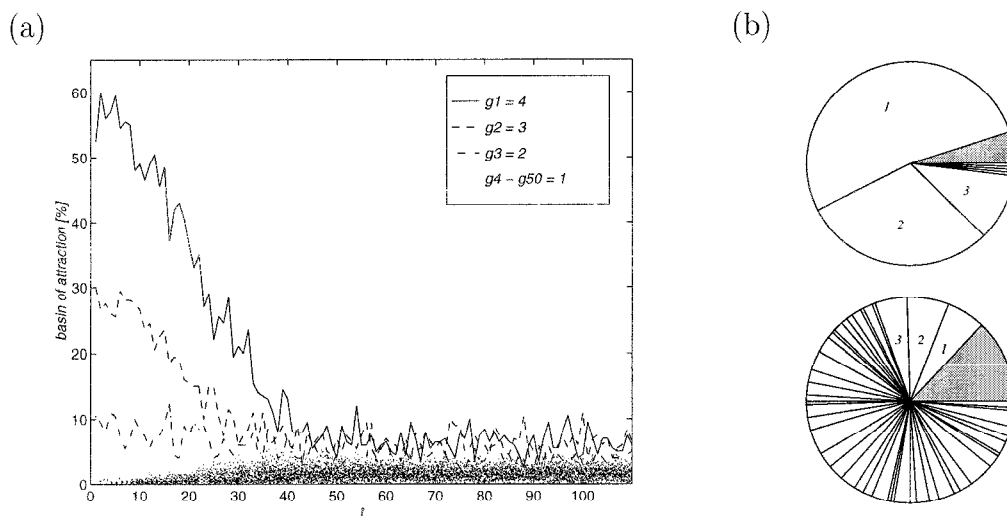


Figure 2. A, Size of basins of attraction as measured by the percentage of retrievals of specific memories. Fifty memories are stored in a system of a thousand neurons. Three of the memories are stronger (parametrized by g) than the rest, overshadowing all others before the corrective dynamic action of NR is introduced. B, Shares of memory space (relative sizes of basins of attraction) at the beginning (upper figure) and the end (lower figure) of

the simulation that consists of repeated cycles of synaptic degradation and neuronal regulation. Random inputs lead either to encoded memories or to the null attractor (gray shading) in which all activity stops. (From Horn, D., Levy, N., and Ruppin, E., 1998, Synaptic maintenance via neuronal regulation, *Neural Computat.*, 10:1–18. Reprinted with permission.)

learning leads to poor associative memory capacity that does not grow with the size of the network.

Several authors (e.g., Dayan and Willshaw, 1991; Chechik, Meilijson, and Ruppin, 2001) have studied the space of additive Hebbian learning rules for associative memory networks with low-activity patterns (i.e., patterns where only a low fraction of the neurons fire). Such learning rules determine the changes in synaptic efficacy when a memory pattern is stored, and may be formally written as $\Delta w_{ij} = aS_iS_j + bS_i + cS_j + d$, where $S_i \in \{0, 1\}$ is the activity of the i th neuron of the stored pattern. Analyzing the associative memory capacity of such learning rules shows that only a constrained subspace of learning rules leads to effective memory storage. Figure 3 illustrates this phenomenon, showing the capacity resulting from such rules within a subspace of two parameters. Most learning rules are ineffective and lead to low memory capacity because they create correlations between synaptic weights even when the stored memory patterns are uncorrelated. Moreover, all effective learning rules fulfill a constraint that depends on the fraction of firing neurons within the stored memory patterns (a global network parameter). Unfortunately, small perturbations in the learning rule parameters lead to violation of this constraint, and consequently to memory capacity collapse.

Interestingly, learning with effective learning rules lead to a vanishing sum of synaptic efficacies for each neuron. This is true, for example, for the learning rules on the ridge in Figure 3. More important, the converse also holds: a vanishing synaptic sum guarantees effective learning. Thus, enforcing through neuronal regulation the condition that the sum of synaptic efficacies vanishes yields high memory capacity, irrespective of the generalized Hebbian rule one starts with (Chechik et al., 2001).

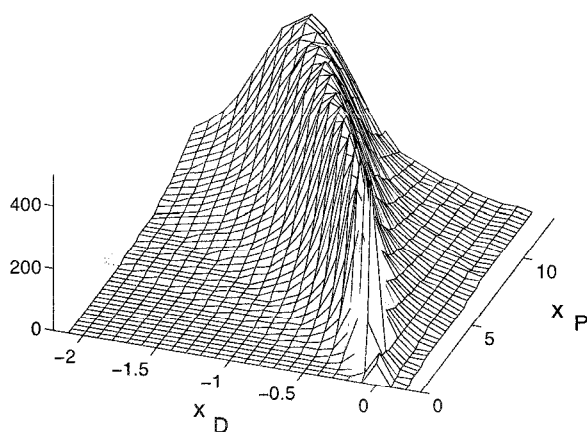


Figure 3. The memory capacity of an associative memory network for various learning rules. The number of memories that can be stored in a 1,000-neurons network and later retrieved from distorted cue is plotted as a function of two parameters: the strength of synaptic potentiation (x_P) and heterosynaptic depression (x_D). These two parameters span the two-dimensional subspace of learning rules $\Delta w_{ij} = x_P S_i S_j + x_D S_i (1 - S_j)$, where $S_i \in \{0, 1\}$ is the activity of the i th neuron. Apparently, only a one-dimensional set of learning rules provides effective learning. (From Chechik, G., Meilijson, I., and Ruppin, E., 2001, Effective learning with ineffective Hebbian learning rules, *Neural Computat.*, 13:817–840. Reprinted with permission.)

Discussion

Hebbian mechanisms per se fail to provide robust and effective learning, both in supervised and in unsupervised scenarios. Although some synapse-specific mechanisms may provide partial remedies for these problems, the current article focused on neuronal regulation of synaptic efficacies and its role in complementing Hebbian learning. Experimental evidence exists for cellular mechanisms that regulate synapses to maintain global constraints on activity or total synaptic strengths (Turrigiano et al., 1998). This evidence suggests that neuronal regulation and Hebbian learning are distinct mechanisms: they are mediated through different receptors (NMDA versus AMPA) and operate on different time scales. From a computational standpoint, the combined operation of Hebbian learning and neuronal regulation provides powerful learning capabilities, ranging from PCA and ICA extraction to robust associative memory learning. We conclude that the functional interplay between synaptic and neuronal mechanisms plays a fundamental role in biological neural networks.

Road Map: Neural Plasticity

Background: Hebbian Synaptic Plasticity

Related Reading: Post-Hebbian Learning Algorithms

References

- Bienenstock, E. L., Cooper, L. N., and Munro, P. W., 1982, Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex, *J. Neurosci.*, 2:32–48.
- Bliss, T. V. P., and Collingridge, G. L., 1993, Synaptic model of memory: Long-term potentiation in the hippocampus, *Nature*, 361:31–39.
- Chechik, G., Meilijson, I., and Ruppin, E., 1999, Neuronal regulation: A mechanism for synaptic pruning during brain maturation, *Neural Computat.*, 11:2061–2080.
- Chechik, G., Meilijson, I., and Ruppin, E., 2001, Effective neuronal learning with ineffective Hebbian learning rules, *Neural Computat.*, 13:817–840.
- Colman, H., Nabekura, J., and Lichtman, J. W., 1997, Alterations in synaptic strength preceding axon withdrawal, *Science*, 275:356–361.
- Dayan, P., and Willshaw, D. J., 1991, Optimizing synaptic learning rules in linear associative memories, *Biol. Cybern.*, 65:253.
- Goodhill, G. J., and Barrow, H. G., 1994, The role of weight normalization in competitive learning, *Neural Computat.*, 6:255–269.
- Horn, D., Levy, N., and Ruppin, E., 1998, Synaptic maintenance via neuronal regulation, *Neural Computat.*, 10:1–18.
- Intrator, N., and Cooper, L. N., 1992, Objective function formulation theory of visual cortical plasticity: Statistical connections, stability conditions, *Neural Netw.*, 5:3–17.
- Miller, K. D., 1996, Synaptic economics: Competition and cooperation in synaptic plasticity, *Neuron*, 17:371–374.
- Miller, K. D., and Mackay, D. J. C., 1994, The role of constraints in Hebbian learning, *Neural Computat.*, 6:100–126. ♦
- Oja, E., 1982, A simplified neuron model as a principal component analyzer, *J. Math. Biol.*, 15:267–273.
- Quartz, S. R., and Sejnowski, T. J., 1997, The neural basis of cognitive development: A constructivist manifesto, *Behav. Brain Sci.*, 20:537–556.
- Song, S., Miller, K. D., and Abbott, L. F., 2000, Competitive Hebbian learning through spike-timing dependent synaptic plasticity, *Nature Neurosci.*, 3:919–926.
- Turrigiano, G. G., Leslie, K., Desai, N., and Nelson, S. B., 1998, Activity dependent scaling of quantal amplitude in neocortical pyramidal neurons, *Nature*, 391:892–896.
- Willshaw, D. J., 1993, Presynaptic and postsynaptic competition in models for the development of neuromuscular connections, *Biol. Cybern.*, 61:85–93.

Hebbian Synaptic Plasticity

Yves Frégnac

Introduction

Appropriate levels of description must be chosen in order to analyze dynamic changes in brain function during development, learning, and perception. One approach is to go from simple phenomenological rules to complex mechanistic scenarios of synaptic plasticity and to evaluate progressively how each level of complexity affects the processing and adaptive capacities of the overall network. This article briefly summarizes the historical foundations and subsequent elaboration by theoreticians and experimenters of a simple activity-dependent algorithm of synaptic plasticity proposed by Donald Hebb in 1949. The predictions derived from Hebb's postulate can be generalized for different levels of integration (i.e., synaptic efficacy, functional coupling, adaptive change in behavior) simply by adjusting the variables derived from various measures of neural activity and the time scale over which each operates. It is thus interesting to consider to what degree this association law may be independent of the biological substrate that is considered and should be viewed as one of the most general computational principles in brain dynamics.

Five major questions will be addressed:

1. Should the definition of Hebbian plasticity refer to a simple positive correlational rule of learning, or are there biological justifications for including additional pseudo-Hebbian terms (such as synaptic depression due to disuse or competition) in a generalized phenomenological algorithm?
2. What are the spatiotemporal constraints (e.g., input specificity, temporal associativity) that characterize the induction process? In particular, should the Hebbian postulate be interpreted as a co-activity rule, or should it incorporate a causal temporal asymmetry, where presynaptic activity precedes postsynaptic activity by a few milliseconds in order to induce synaptic potentiation?
3. Do the predictions of Hebbian-based algorithms account for most forms of activity-dependent dynamics in synaptic transmission throughout phylogenesis? How do the predictions depend on the complexity of the considered neural network (e.g., direct sensory motor connections in *Aplysia* versus associative networks in neocortex)?
4. On which time scales (perception, learning, epigenesis) and at which stage of development of the organism (embryonic life, critical postnatal developmental periods, adult age) are activity-dependent changes in functional links predicted by Hebb's rule?
5. Are there examples of correlation-based plasticity that contradict the predictions of Hebb's postulate (i.e., those termed anti-Hebbian modifications)?

The Conceptual Framework of Cell Assemblies

Pre-Hebbian Theories

Long before our current knowledge of the synapse-neuron-based structure of the brain, philosophers of antiquity had already theorized how causal relations between external events could be established by the human mind and had pointed out the necessity of repeating sequences of activation in order to link mental representations (Aristotle, ca. 350 B.C.). The application of association theories to the brain can be traced back to as early as 1890: according to William James, the adaptive capacities of our brain depend on mechanistic laws of association that operate under the guidance of central neural structures such as cerebral cortex in higher vertebrates: "When two elementary brain processes have been active

together or in immediate succession, one of them, on re-occurring tends to propagate its excitement into the other" (James, 1890). These concepts of association are immediately applicable to the understanding of behavioral learning, and the first extension of classical conditioning in cellular terms was proposed by Jerzy Konorski, a contemporary of Donald Hebb, who assumed that "when the excitation of a given center is synchronous with the rise of excitation in another center, conditioned excitatory connexions are formed from the first of these centers to the latter" (Konorski, 1948).

A second field of application of association theories is the transient formation of mental representations during perception or dreams. In contrast to the previous proposals, the effect produced by repeated associations is no longer restricted to sequences of external events but extends to autonomous activity of the brain. Rather than being transformed under a long-lasting "mnestic" form, the trace of the association is seen here as a reversible facilitation, promoting neural links over the time required for the establishment of the percept (a few hundred milliseconds). In his seminal essay *Le rêve*, Yves Delage hypothesized that each cortical neuron exhibits an intrinsic characteristic periodicity in its activity (Delage, 1919). The relative diversity of the intrinsic frequencies adopted by possible future functional partners (heterochrony) would be dynamically restructured during perception and give place to transient and highly synchronized states (parachrony) among the activated members of the functional assembly.

A Neurophysiological Postulate to Build Assemblies

The physiological association principle proposed by Hebb was in fact just one of several keystones incorporated in a multilevel model of cerebral functioning during perceptual and learning processes. The main concept of a *cellular assembly*, pivotal to his theory, designated an activity process that reverberates in a *set of closed pathways*. The neurophysiological postulate of Hebbian synapses was introduced as one possible way to reinforce functional coupling between co-active cells and thus of growing assemblies. Similar hypotheses were developed at a higher hierarchical level of organization that allowed the linkage between cognitive events and their recall in the form of a temporally organized series of activations of assemblies. Donald Hebb referred to this final binding process as a *phase sequence*. Thus, Hebb's postulate appears simply as a putative low-level biophysical mechanism for establishing the perseverance of activity among assemblies: "When an axon of cell A is near enough to excite cell B, and repeatedly or consistently takes part in firing it, some growth process or metabolic change takes part in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949).

The formulation of Hebb's postulate requires not only close temporal coincidence but also the spatial convergence of one neuron onto another, supporting a causal relationship between the afferent activity and the postsynaptic spike. It provides a specific prediction: a period of maintained positive temporal correlation between pre- and postsynaptic activity will lead to an increase in the efficacy of synaptic transmission. Hebb did not elaborate on whether the modifications responsible for this decrease in *synaptic resistance* were presynaptic, postsynaptic, or both. Neither did he describe the biophysical substrate responsible for the modification, leaving the choice open between "metabolic change" and "oriented growth." Both of these options turned out to be true. Historically, Hebb's postulate referred exclusively to excitatory synapses. A symmetric,

if not synergetic, version of Hebb's postulate was proposed much later for the case of inhibitory synapses (Stent, 1973), in which functional coupling can be increased by reducing the strength of inhibitory synapses activated at the same time that action potentials are fired in the postsynaptic cell. Some models introduced inhibitory plasticity well before it was proved analytically that the use of negative weights is required for endowing associative memories with an optimal mapping and memory capacity. Thus, the effective gain between input and output, defined at an ideal Hebbian synapse, should be envisioned as a dynamic variable between positive and negative boundaries, depending on whether the net effect induced by the input is excitatory or inhibitory (Figures 1A and 1B). Some theoretical studies also proposed that synaptic gain could change sign during development (Bienenstock, Cooper, and Munro, 1982), a suggestion later found to apply to inhibitory neocortical or hippocampal circuits.

Post-Hebbian Theories: Dynamic Binding of Assemblies

Peter Milner was probably the first theoretician in the fifties to propose explicit rules for the compositionality of assemblies. The repeated activation of a given cell assembly would reinforce syn-

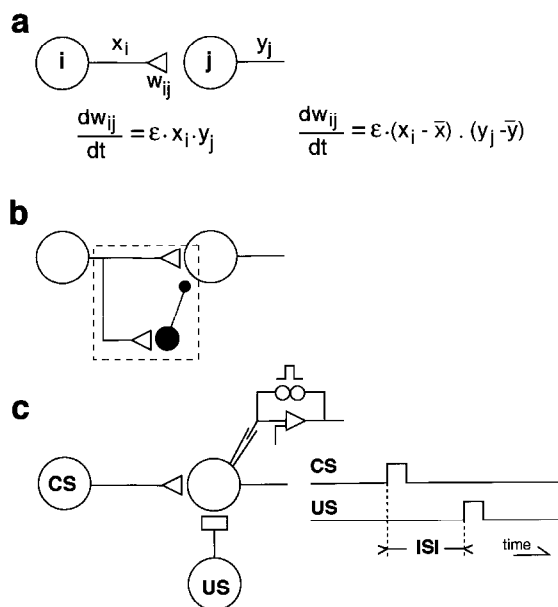


Figure 1. The multiple identities of a Hebbian synapse. **A**, Monosynaptic connection between a presynaptic axon (*i*) and a postsynaptic target cell (*j*). Left, the Hebbian algorithm posits that the change in synaptic efficacy $\Delta\omega_{ij}$ is given by the product (or logical AND) of presynaptic (x_i) and postsynaptic (y_j) activities at any point in time. Right, the “covariance rule” replaces the pre- and postsynaptic terms by the departure of instantaneous pre- and postsynaptic activities from their respective mean values ($\langle x \rangle$) and ($\langle y \rangle$) averaged over a certain time window. **B**, A dual excitatory/inhibitory circuit equivalent to an ideal Hebbian synapse, the gain of which varies between negative and positive boundary values. Open circle indicates an excitatory cell; open triangles indicate synapses. Solid symbols indicate inhibitory cells and synapses. **C**, A cellular analogue of classical conditioning: the Hebbian synapse transmits the neural information fed by the conditioned stimulus (CS). The unconditioned stimulus (US) activates the postsynaptic cell in an all-or-none fashion through a nonmodifiable synapse or through a depolarizing current injection directly applied by the experimenter to the postsynaptic cell. The conditioned response stems from the repeated association of the CS input followed by the US input. The phase of the association (ISI) is defined by the temporal lag between the CS and the US inputs.

aptic links within this assembly, and in addition would “prime” a restricted number of cells, allowing future binding and thus composition with other associative processes. The latent labeled synapses would remain transiently eligible for further potentiation by the contiguous firing of other assemblies, a “tagging” concept that would find its biological counterpart only much later. Repeated sequential activation would reinforce the primed connections so that they became an integral part of the next active assembly, thereby resulting in second-order associations: their firing would thus allow the recall of a complete *phase sequence*. Similar ideas were reworded with the addition of a glue mechanism, namely temporal synchrony, acting on a much faster time scale than that initially proposed by Hebb such as to bind elementary representations into a cognitive whole (Milner, 1974) (see SYNCHRONIZATION, BINDING AND EXPECTANCY). The theoretical work fostered by von der Malsburg and colleagues showed that the combined use of “fast” and reversible Hebbian-like synaptic changes and “slow” Hebbian plasticity provides interesting compositional properties when applied to a specific type of assembly called synfire chains (see SYNFIRES CHAINS). Simultaneous recordings of cortical single-unit activity in behaving monkeys have since shown that a significant fraction of activity correlated with a specific cognitive task can be described as waves of synchrony relayed between sets of co-active neurons, at various delays ranging from a few to several hundred milliseconds. A Lego-like interlacing of dynamic assemblies can be further used to give a neuronal embodiment to abstract compositional models of cognition based on dynamical binding operations between symbols that sit at various levels of a representational hierarchy (reviewed in Frégnac and Bienenstock, 1998; see DYNAMIC LINK ARCHITECTURE).

Theoretical Predictions and Neurobiological Tests of Hebb's Postulate

Ten years after the publication of his *The Organization of Behavior*, Donald Hebb was doubtful whether his theory was definite enough to be testable, not so much because of technical constraints but because of conceptual ones. Most of the premises were based on a number of postulates, each addressing a different level of integration, and failures of tests restricted to the biophysical level of the synapse would not threaten or contradict the theory in itself.

Hebbian Analogs of Cellular Learning

Hebb's postulate was initially applied by cyberneticians and electrophysiologists in the context of supervised learning. In a similar way to the external teacher of the gamma-perceptron, a classifier machine developed by cyberneticians at the end of the sixties that imposes an increase in the gains of active synapses that participate in the “correct” answer (see PERCEPTRONS, ADALINES, AND BACK-PROPAGATION), electrophysiological tests of Hebbian synapses impose depolarization of the postsynaptic element concomitantly with afferent activity. The exogenous control of postsynaptic activity has been achieved using various technical means (electrical stimulation of an unmodifiable pathway, iontophoresis of excitatory neurotransmitters, intracellular current injection, uncaging of calcium at a dendritic spot) in order to elicit positive reinforcement of the modifiable test response (Figure 1C). This strategy has been attempted at a variety of sites in the central nervous system ranging from molluscan neuronal ganglia to the mammalian forebrain (reviewed in Brown et al., 1990; Frégnac and Shulz, 1994; Bi and Poo, 2001). Most success in demonstrating the role of postsynaptic factors has been observed in the vertebrate cortex (hippocampus, neocortex, motor cortex), where synaptic potentiation of various durations can be induced under the cooperative influence of other inputs (as is the case during high-frequency tetanus of afferent path-

ways) or by forcing the postsynaptic cell to an artificially high level of activity. As we will discuss later, results opposite to the Hebbian prediction have been found in the striatum, in the cerebellum, and in related structures in electric fish (reviewed in Bell, 2001; see CEREBELLUM: NEURAL PLASTICITY).

In spite of some earlier reports that invertebrate synapses possess the capacity for long-term potentiation (LTP), until recently it was assumed that nonassociative and associative forms of behavioral plasticity in *Aplysia* resulted from an activity-dependent presynaptic modulation of the efficacy of sensorimotor synapses (see INVERTEBRATE MODELS OF LEARNING: *APLYSIA* AND *HERMISSENDA*). However, later studies provided strong evidence for a postsynaptic regulation of *Aplysia* sensorimotor synapses in dissociated cell culture. These studies showed a specific influence of the appropriate motor cell in inducing spatial competition and segregation of the locus of termination of the afferent axons corresponding to different presynaptic axons. It also demonstrated that Hebbian pairing protocols induced potentiation of identified sensorimotor synapses.

By which cellular machinery does co-activity exert control over synaptic gain? In vertebrate hippocampus as well as in *Aplysia* sensorimotor co-cultures, evidence implicates the NMDA receptor and its invertebrate homologous form, respectively (see NMDA RECEPTORS: SYNAPTIC, CELLULAR, AND NETWORK MODELS). These receptors are ideally suited to operate conjunctive mechanisms, since they require concomitant presynaptic activation and depolarization of the postsynaptic neuron above a critical level to free their embedded ionophore channel from the magnesium block. However, this mechanism is certainly not unique, since Hebbian forms of plasticity can be observed even during the pharmacological APV blockade of NMDA receptors.

The use of multiple simultaneous whole-cell recordings in vitro has recently improved control of the respective timing of pre- and postsynaptic activity and the ability to patch at different distances from the soma. These experiments suggested that action potentials propagating back into dendrites serve to modify single active synaptic connections (see BACKPROPAGATION: GENERAL PRINCIPLES). Backpropagation can contribute to the induction of synaptic plasticity through three distinct mechanisms. The simplest one relies on the voltage dependency requirements of NMDA receptor activation and remains input specific. The second one, involving voltage-gated calcium channels, will apply to all parts of the dendrite in which the spike is efficiently propagated. A third mechanism involves a chain reaction and depends on the initiation of calcium spikes by otherwise subthreshold distal inputs when the EPSP follows by a 5-ms delay the invasion of the dendrite by a backpropagating action potential (reviewed in Frégnac, 1999). In all cases the backpropagating spike can be seen as the “binding signal” emitted by the soma to differentially modify synapses that are active within a precise temporal window.

A Theoretical Need for Synaptic Depression and Competition: Pseudo-Hebbian Rules

Most algorithms of synaptic plasticity use rules of normalization that require depression of certain synapses in addition to Hebbian reinforcement of active connections (see POST-HEBBIAN LEARNING ALGORITHMS). The “divergence” problem caused by a straightforward application of Hebb’s principle was solved by the first modelers using various ad hoc hypotheses: upper bound values (saturation) for individual synaptic weights, forgetting mechanisms slowly activated by disuse, and complementary plasticity rules that operated at the level of synapses fed both by active pathways (associative depression) and by neighboring inactive afferents (heterosynaptic depression). In the latter case, different weight normalization procedures were proposed that gave rise to what is called *competitive learning* in the modeling literature (see COM-

PETITIVE LEARNING). The additional decay terms introduced in the plasticity rule can depend only on local variables, or it can operate as a global constraint that maintains constant the sum (or the sum of the squares) of all the synaptic weights converging onto the same neuron. Gunther Stent proposed a biophysical mechanism inducing a selective decrease in the synaptic efficacy of afferent fibers that were inactive at the time the postsynaptic neuron was discharging under the influence of other inputs (Stent, 1973). He was probably the first theoretician to introduce the concept of a threshold in synaptic plasticity, linked to the local postsynaptic membrane potential, below which synaptic depression occurs. The prediction of this postulate found strong support from later cross-depression studies in visual pathways and from the observation of heterosynaptic depression in the CA1 field and the dentate area of the hippocampus.

The assumption of global constraints in maintaining the total synaptic weight constant onto the recipient neuron has been also made on the basis of more biological grounds, such as the theory of selective stabilization. Correlates to this model were found in simple in vitro systems that allow the culture of specified numbers of pre- and postsynaptic partners, suggesting that the total capacity of the target neuron for synaptic interaction is fixed and divided among the different input lines. Related arguments can also be found during synaptogenesis of neuromuscular junctions and of vertebrate visual pathways.

In spite of their diversity, these different rules have a common implication: they predict spatial and temporal competition between active fibers that impinge on a common target cell; they are referred to as being *pseudo-Hebbian*.

Experimental Support for a Generalized Hebbian Algorithm

Most Hebbian algorithms that were used to model synaptic plasticity in self-organizing networks or behavioral learning, before the concept of spike-timing-dependent plasticity (STDP) arose, are surprisingly uniform and are based on co-activity of pre- and postsynaptic neurons. They may be summarized by the same general equation in which the change of synaptic efficacy with respect to time is equal to the product of a presynaptic term and a postsynaptic term (reviewed in Frégnac and Shulz, 1994). The so-called covariance hypothesis (in visual cortex: Bienenstock et al., 1982) replaces the pre- and postsynaptic terms by the departure of instantaneous pre- and postsynaptic activities from their respective mean values averaged over a certain time window (see Figure 1A). Average values in the covariance product constitute pre- and postsynaptic plasticity thresholds that determine the sign of the modification. They can be replaced by nonlinear functions of past activity (power function with an exponent greater than 1; see Bienenstock et al., 1982). Because of the particular choice of these nonlinearities, synaptic depression will be more readily induced by regimes of high activity during which the plasticity threshold increases faster than the mean postsynaptic activity. Conversely, synaptic potentiation will be promoted following low-activity regimes during which the plasticity threshold decreases slower than mean postsynaptic activity. The “floating threshold” hypothesis predictions agree with the observation of an increased rate of cortical specification in previously deprived animals that are re-exposed to a visually structured environment, when compared with the normal process observed in nondeprived animals. Although most theoretical studies have been concerned with postsynaptic thresholds, experimental evidence (priming protocols, mostly tested in hippocampal slices) suggests as well the existence of a presynaptic averaging mechanism.

The BCM covariance rule in its most general form does account for spatial and temporal competition. In addition to the straightforward Hebbian condition (positive covariance induces an in-

crease in synaptic gain), the covariance hypothesis predicts two forms of depression. The first one is an associative heterosynaptic depression at the level of synapses whose activity was uncorrelated with that of the tetanized pathway (Levy and Steward, 1983). The second form is a homosynaptic depression, when presynaptic activity is associated with repeated failure in synaptic transmission (Frégnac et al., 1988; Frégnac and Shulz, 1999).

Constraints on Spatial and Temporal Specificity

Input Specificity and Cooperativity

A first limitation in the locality of the changes produced by learning depends on the minimal neuroanatomical convergence of input that is necessary to induce a functional change. In the case in which the activity of one afferent alone would be sufficient to induce synaptic change, convergence should be considered to be related to the density and the spatial distribution of boutons made by a single presynaptic axon onto the same target cell. Some studies based on simultaneous dual intracellular recordings failed to observe significant potentiation of individual synaptic connections, whereas their compound activation revealed an increased postsynaptic response. These results suggest the existence of a postsynaptic threshold mechanism controlling the *expression* of LTP: a critical level of depolarization has to be achieved so that the enhancement at the “primed” synapse would be revealed in response to the test input. This nonlinear behavior in the input/output curve of the postsynaptic neuron would have a major consequence: it would greatly increase the spatial input selectivity of LTP by making it conditional on the strength and the convergence of multiple inputs. It could thus prevent temporally unstructured or spatially dispersed afferent information from benefiting from the potentiation.

Volume Plasticity

The input specificity of Hebbian schemes of plasticity—i.e., their restriction to active synapses—might suffer strong limitations when the release of retrograde factors and the spatial diffusion of second messengers are considered. Since quantal analysis studies have implicated presynaptic factors in the maintenance of LTP, it is admitted that some feedback signal indicates to the presynaptic terminal that the correlation operation has been accomplished and that potentiation is authorized. Various retrograde messengers have been proposed, including arachidonic acid, nitric oxide, carbon monoxide, and platelet-activating factor. Taking into account the diffusion of messengers in the extracellular medium, the correlation between high levels of the released molecule and active axon terminals could then become the key factor controlling which synapses should be potentiated. This scheme accounts for the observed generalization of potentiation to neighboring synapses belonging to the axon that has initiated the retrograde messenger process, independently of the target neuron. The consequence of this “volume plasticity” is that correlation will be reinforced between elements that are co-active within a given time window and are within some critical volume without being necessarily physically connected. Reasonable estimates of the space constant on which retrograde messenger-induced changes occur are in the order of 50–150 μm , based on dual intracellular recordings of a conditioned cell and a neighbor in vitro, both of which receive parent branches from the same input fiber. More surprisingly, the postsynaptic modification also seems to spread to different presynaptic axons that have or have not been implicated in the induction of the LTP process, regardless of their own history of activation.

More advanced studies have recently been achieved at the level of identified neurons and synapses in low-density hippocampal cultures and have revealed extensive but selective spread of both LTP

and long-term depression (LTD) from the site of induction to other synapses in the network (reviewed in Bi and Poo, 2001). LTD induced at synapses between two glutamatergic neurons can spread to other synapses made by divergent outputs of the same presynaptic neuron (*presynaptic lateral propagation*) or to synapses made by other convergent inputs on the same postsynaptic cell (*postsynaptic lateral propagation*). Furthermore, LTD can spread in a retrograde direction to depress synapses afferent to the presynaptic neuron (*backpropagation*). In contrast, LTP can exhibit only lateral propagation and backpropagation to the synapses associated with the presynaptic neuron. If output synapses of the paired presynaptic neuron undergo LTP/LTD, then the input synapses undergo similar changes. It is interesting to observe that the backpropagation of LTP/LTD observed in cell cultures appears to fit qualitatively the requirement for backpropagation algorithms in multilayer networks. Similar functional effects could operate on a larger scale in vivo if a permanent imbalance in activity is introduced between competing axons, for example altering the spatial grouping of bands of ocular dominance driven by the open eye in monocular deprived visual cortex.

Breaking the Timing Symmetry: Causality Rather than Synchrony

Most applications of Hebbian theories based on pairing protocols indeed stressed the importance of co-activity, ignoring the few-milliseconds step that separates a presynaptic spike from the triggering of postsynaptic activation. The temporal contiguity requirement of Hebbian potentiation in cortex was first estimated in the ± 50 ms range, both in vivo and in vitro, and no temporal ordering was required between pre- and postsynaptic activation. However, the wording of Hebb’s principle dictates that presynaptic activity should precede the spike initiation in the postsynaptic element to which it contributes in a causal way (“A’s efficiency, as one of the cells firing B”). Temporal asymmetric Hebbian and anti-Hebbian rules (Figure 2B) that have been introduced only in recent years agree in this respect with the original concept.

An overlooked consequence of additional pseudo-Hebbian rules is that their interplay with a purely Hebbian scheme already predicts a loss of symmetry in the temporal domain and a possible narrowing of the critical interval of association. In vitro studies of heterosynaptic plasticity in cocultures of embryonic spinal neurons and myotomal muscles show that synchronous activation of two presynaptic pathways protect them from depression, whereas a delay as short as 100 ms is sufficient to depress one or both pathways. Associative forms of LTD have been observed during contiguous dual-pathway stimulation paradigms or when the test input follows a postsynaptic depolarization induced by a brief current pulse. The exact temporal window during which a recurrent input remains eligible for potentiation depends, however, on the strength of the last unconditioned activation of the cell. These results agree partially with the temporal order requirement in associative heterosynaptic depression reported 20 years ago in the study of the crossed (weak) and uncrossed (strong) entorhinal cortex projection to the dentate gyrus, which still described much more shorter association intervals (20 ms), enabling associative LTP (Levy and Steward, 1983).

Recent work, based in most cases on dual patch recordings in vitro, has been realized in preparations as diverse as cultured hippocampal networks, the developing retinotectal system of the frog, the adult electrosensory lobe of electric fish, and the sensory neocortex of the rat. Results suggest an even tighter temporal contingency rule (10-ms range), where the temporal order of the onset of the postsynaptic subthreshold potential reflecting the arrival of the presynaptic spike and the postsynaptic spike backpropagating in the dendrite decides whether potentiation or depression occurs (re-

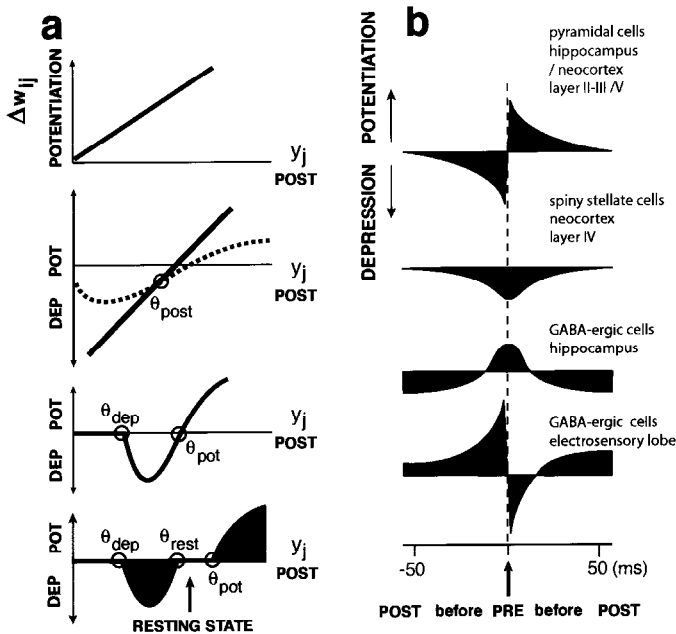


Figure 2. Hebbian synapses and spike-timing-dependent plasticity. *A*, A representation of Hebb's rule (top) and the most commonly observed rules of associative plasticity. Each graph expresses the relationship between the induced synaptic change Δw_{ij} (positive ordinates for potentiation, negative for depression) and postsynaptic activity (y_j on abscissa) at the time of the association defined by the occurrence of the presynaptic spike. The slope of Hebb's rule is proportional to presynaptic activity. The postsynaptic term (POST) has been equated successively with postsynaptic firing frequency, postsynaptic membrane potential, or calcium influx in the postsynaptic cell. In addition, many authors have confounded, rightly or wrongly, the postsynaptic term with the presynaptic stimulus frequency applied during conditioning, which made these models more manageable in predicting the effects of various input frequencies used more specifically to induce LTD or LTP. From top to bottom: The simple Hebbian rule (top) predicts potentiation only. The covariance rule or BCM rule (second graph from the top) and ABS rule (third graph from the top) predict both depression and potentiation, respectively, with one (θ_{post}) or two postsynaptic plasticity thresholds (θ_{dep} , θ_{pot}). The lower graph is derived from *in vivo* experiments in which the resting state of the synapse corresponds to a dead zone between depression and potentiation where information is reliably transmitted (Frégnac et al., 1988). *B*, Different forms of spike-timing-dependent plasticity rules established *in vitro* in cocultures and acute slices. The induced synaptic change is expressed as a function of the temporal delay separating postsynaptic firing from presynaptic firing (taken here as the zero-delay reference), imposed during the pairing protocol. From top to bottom: pyramidal cells in hippocampus or in nongranular layers in neocortex, granular spiny stellate cells in neocortex, GABAergic neurons in hippocampal cultures, GABAergic medium ganglionic layer cells in the electrosensory lobe of the electric fish. Positive ordinates indicate synaptic potentiation and negative ordinates indicate depression.

viewed in Bi and Poo, 2001; Abbott and Nelson, 2000; see Figure 2*B*). An unexpected twist in the story is the claim that the outcome of the rule itself depends on the frequency and the strength of the presynaptic inputs, which could have nontrivial consequences *in vivo*.

The obvious consequence is that models that incorporate STDP rules account most accurately for the emergence of causal chains within neuronal assemblies and best support phase sequence learning. It is nevertheless possible that both synchrony and causality are required to promote reverberating assemblies rather than the simple build-up of open-ended chains of feedforward activity, as

exemplified by the concept of synfire chains. It is highly plausible that STDP reinforces the progressive establishment of a transmission mode through "pulse packets," since the asymmetric nature of this plasticity rule will indirectly control the jitter that could be observed in the timing spread within each pulse packet.

The most obvious cases of associative learning that depart from Hebbian co-activity or STDP rules are those requiring associations between neural events separated by long delays (see CEREBELLUM AND CONDITIONING), such as the optimal interstimulus intervals (in the range of seconds) of association in classical conditioning (see Figure 1*C*), or between the sample and choice periods in delayed-matching-to-sample comparison tasks (see CORTICAL HEBBIAN MODULES). Predictive rules have been modeled with ad hoc phase-lagged correlation functions (see REINFORCEMENT LEARNING). However, no experimental evidence has yet been obtained to account for the build-up of optimal interstimulus intervals through Hebbian mechanisms. This lack of evidence does not negate the implication of ionic mechanisms responsible for a delayed excitability change, or the slow build-up of a second-messenger-mediated intracellular response.

From Hebbian Synapses to Behavioral Learning

This section addresses the functional and behavioral consequences that Hebbian rules of plasticity induce in biological self-organizing systems.

Unsupervised Learning

Four possible applications of Hebbian processes acting on a long-term scale can be found in the early development or in the forced oriented growth of retinofugal pathways in lower vertebrates, mammals, and primates. These applications exemplify unsupervised learning, or self-organization.

1. The intrinsic synchronous bursting activity ("dark discharge") that arises prenatally from the retina before rods and cones are even formed exerts a structuring influence on the developing retinofugal pathway. The correlated firing among neighboring retinal ganglion cells within one eye and the lack of synchronous firing between ganglion cells of each eye are conditions that allow competition between geniculate afferents according to their ocularity. This correlated input is present in a still stronger way very early in prenatal life, taking the form of spatially organized waves of activity that spread intermittently in random directions across the whole retina. These synchronizing waves could provide the local correlations necessary for the sorting out and topographic refinement of retinal projections onto the lateral geniculate nucleus and, if still present after filtering through the thalamic relay, could be instrumental in the segregation of geniculate afferents in the recipient layer of visual cortex. After the first week of postnatal life in the cat, this intrinsic pattern-generating mechanism will give way to correlated inputs under the guidance of vision (see OCULAR DOMINANCE AND ORIENTATION COLUMNS).

2. Similar activity-dependent rules might apply to the development of intracortical connectivity: the validity of correlational Hebbian rules seems to hold throughout ontogenesis, if one does not restrict the choice of the postsynaptic control variable to spike activity. Indeed, free calcium activity and electrical and glial coupling could act prenatally as a substitute for synaptic transmission to ensure assembly formation in the absence of conventional fast Na^+ action potentials. A more classical form of Hebbian plasticity responsible for the progressive maturation of the horizontal intracortical network occurs during a few weeks following birth in the cat; the process results in a selective activity-dependent pruning and stabilization of horizontal connectivity.

3. Evidence has been obtained for the implication of Hebbian mechanisms during the functional reorganization of cortical processing following anomalous visuomotor behavior. A neuroanatomical and electrophysiological study in divergent strabismic kittens, which compensate misalignment of their eyes by alternate fixation, showed that only territories with the same ocular dominance are linked by tangential intracortical connections, and that synchronized activity is achieved only between cell groups dominated by the same eye. Furthermore, in the case of convergent strabismus, which results behaviorally in a loss of acuity through the eye that is not used for fixation, neurons dominated by the “amblyopic” eye exhibited much weaker synchronized activity than cells driven by the “good” eye. The observed correspondence between alterations in intracortical horizontal connectivity topology, the selective impairment of response synchronization, and the perceptual deficit constitutes probably the best evidence so far for a role of temporal correlation in the functional organization of cortical domains (see SYNCHRONIZATION, BINDING AND EXPECTANCY).

4. A last example illustrates the case of oriented growth of retinal axons that are artificially forced to connect the auditory thalamus after chronic deafferentation from its normal input at an early stage of development (see AUDITORY CORTEX). A combined electrophysiological and optical imaging study in the cortex, which normally should have become a primary auditory area, shows that the thalamic rewiring procedure induces the emergence of an anatomofunctional architecture similar to that observed in normal visual cortex (Sharma, Angelucci, and Sur, 2000): a normal retinotopic projection is formed, an orientation preference map is found with classical pinwheel organization, and, furthermore, the intrinsic intracortical connections show the distinctive patchy pattern along a mediolateral axis that is specific to a V1 area. Sur and colleagues further demonstrated that this rewired cortex can mediate adequate visual behavioral responses in response to visual stimuli. These findings suggest that the specificity found in the columnar organization of a given cortical area and its intrinsic horizontal connectivity are shaped by the particular temporal structure of the sensory input experienced during a critical postnatal period.

In summary, the grouping and sorting out of fibers afferent to cortex, the morphological tuning in the spatial distribution of the terminal boutons of intrinsic and extrinsic axons, the functional expression and possibly silencing of synapses, and the setting of a columnar architecture specific to the sensory modality to be processed could all, at some stage of postnatal development, be under the influence of temporal correlation between presynaptic fibers converging onto the same target, or between pre- and postsynaptic partners. This essential role of co-activity in self-organization was foreseen by theoreticians such as Linsker and Miller, who proposed a unifying role for activity, whether triggered endogenously by the nervous system or evoked by interaction with the environment (see OCULAR DOMINANCE AND ORIENTATION COLUMNS).

Supervised Learning and Cellular Analogues of Visual Cortical Epigenesis

Support for a functional implication of Hebb's postulate has also been found in studies on the neuroanatomical and physiological effects of forced patterns of activity, which simulate the functional effects of anomalous visual experience during critical periods of development (epigenesis, in Frégnac et al., 1988; Frégnac and Shulz, 1999). A differential supervised association protocol was used in vivo to test specific predictions of the covariance rule by imposing opposite changes in the temporal correlation between two test parameters characterizing afferent visual activity and the output signal of the cell. Here, an external supervisor (i.e., the experimenter) helped the cell to respond to one input and blocked the

cell's response to another, different input. The common outcome was that the relative preference between the two test stimulus characteristics was generally displaced toward that which had been paired with imposed increased visual responsiveness.

These pairing-induced modifications of specificity of the visual response have been considered as cellular analogues of epigenesis since they reproduce functional changes occurring during development or following early manipulation of the visual environment (monocular deprivation, rearing in an orientation-biased environment, or optically induced interocular orientation disparity). Surprisingly, the probability of inducing functional changes was found to be comparable in the kitten during the critical period and in older kittens and adults, suggesting that the cellular potential for plasticity might extend well beyond the classical extent of the critical period. Local supervised learning procedures, applied at the cellular level, might bypass the systemic control that normally blocks the expression of plasticity in the mature brain.

Functional Consequences of Spike-Timing-Dependent Plasticity

It has been suggested that, with repeated experience of a sequence of sensory events, spike-timing-dependent plasticity will promote the learning of the sequence and anticipation of future events from past stimuli. Recent tests of this hypothesis have been engineered in vivo in cat primary visual cortex: repetitive sequential presentation of two stimuli, one being the orientation preferred by the recorded cell and the other being either a suboptimal orientation or an electrical stimulation of the cortex, resulted in a compensatory reorganization of orientation tuning, with the direction of the shift in orientation preference depending on the temporal order of the stimulus pair. Furthermore, similar conditioning in human subjects induced a similar shift in perceived orientation, thus mirroring the plasticity described at the single-cell level in cat visual cortex. Thus the relative timing of visual stimulation and cortical activity plays a critical role in dynamic modulations of adult cortical function. In addition, these various adaptive protocols suggest that the susceptibility to adaptive changes is not uniform throughout cortex, and that orientation pinwheel centers may obey plasticity rules differently from their surround.

Gating Signals and Attentional Processes

Both Hebb and Milner were aware of the fact that the expression of synaptic changes could largely depend on the level of pre-activation of nonspecific projection systems and arousal. These factors could influence the likelihood of summation at the synapse and thus could affect the amount of correlated input needed to induce synaptic changes. Because of methodological and technical difficulties, the role of the *behavioral context* (i.e., attention, reinforcement) has been often ignored in the study of the synaptic mechanisms underlying learning in mammals. Ahissar and collaborators applied cross-correlation techniques to study the plasticity of “functional connectivity” between simultaneously recorded pairs of neurons in the auditory cortex in awake monkeys performing a sensory discrimination task (reviewed in Ahissar et al., 1998). In order to control the correlation of activity between cells, the auditory stimulus preferred by the presumed postsynaptic cell was applied every time (and immediately after) the presynaptic cell fired spontaneously. The tone used to control the activity of the postsynaptic cell also signaled the reward occurrence. Under these Hebbian conditions, reversible changes in functional coupling could be induced only when the animal was attentive to the tone. These changes lasted for a few minutes and followed the covariance hypothesis predictions. The results indicate that Hebb's requirement is necessary but not sufficient for cortical plasticity in the adult monkey

to occur: internal signals indicating the behavioral relevance are also required. More recently, cellular analogues of state-dependent learning have been proposed that show that the recall of a learned association in adult sensory cortex requires the application of the same neuromodulatory signals (ACh) as those present during conditioning in order for the functional changes to be expressed (Shulz et al., 2000).

Anti-Hebbian Forms of Learning

Depending on the neural structure under study (i.e., cerebellum versus cortex) or the time course of the functional effect looked for (i.e., sensory adaptation versus learning), forms of plasticity have been observed that are contrary to the predictions of Hebb's postulate. Such changes are called anti-Hebbian (or reverse Hebbian) and should be unambiguously distinguished from pseudo-Hebbian modifications (see POST-HEBBIAN LEARNING ALGORITHMS). The best-known example of anti-Hebbian plasticity is cerebellar long-term depression (see CEREBELLUM: NEURAL PLASTICITY), which, by its time course and induction requirements, appears similar to Hebbian associative potentiation. The trigger mechanism appears to be the same in both cases: free calcium entry into the postsynaptic cell. In order to explain why in the cerebellar case depression occurs, whereas potentiation is predicted by Hebbian schemes of plasticity, it could be assumed that the sign of the change of the synaptic modification depends on the type of neurotransmission (excitatory/inhibitory) that the postsynaptic neuron will exert on other neuronal targets. Purkinje cells for which Hebbian protocols induce depression are the inhibitory output neurons of the cerebellar cortex. Although no biological basis has been found to support the hypothesis that excitatory and inhibitory cells undergo Hebbian potentiation and depression, respectively, the implications of the hypothesis are very attractive in terms of systems theory: forced co-activity would produce the same type of global positive gain control whatever the neuronal structure under study, either by increasing the transmission of the selected input through a purely excitatory loop or by reducing the excitation fed into the inhibitory efferent pathway.

Other examples of anti-Hebbian plasticity link both cerebellar LTD and fast adaptation processes in perception. In the teleost electric fishes (Mormyridae and Gymnotidae), the cerebellar-like structure that receives primary electrosensory projections is the electrosensory lobe (ELL) (see ELECTROLOCATION). The feedforward sensory input informs the fish of its electrical environment and also provides a reafferent response, owing to the sensory effect of the fish's own electric discharge (EOD). The principal cells of the ELL, like Purkinje cells in cerebellum, receive in addition a diversity of contextual inputs conveyed through parallel fibers. The context can provide a copy of the motor command responsible for the electric discharge ("efferent copy" or "corollary discharge"), proprioceptive cues about movements from the fins or the whole body, or control signals descending from higher integrative areas. The pairing of an electrosensory stimulus at a fixed phase or delay after the EOD motor command, or after the passive bend of the body tail, results, when the contextual signal is reapplied alone, in the recall of a sign-inverted image of the firing pattern evoked by the paired electrosensory stimulus. The plastic changes at the parallel fiber-principal cell synapse can be long-lasting, but the extinction of the effect is an active process depending on the frequency of the recall process. Thus, in the mormyrid ELL, the modifiable corollary discharge elicits the transient storage of a negative image of the temporal and spatial pattern of sensory input that has followed the motor command. Bell and co-workers demonstrated the synaptic nature of the change and its reversibility in GABAergic medium ganglionic cells of the ELL by replacing the sensory input (which affects the whole structure) with an intracel-

lular current pulse affecting only the cell under study (Han, Grant, and Bell, 2000; reviewed in Bell, 2001). Similar findings have been observed with other contextual signals in other cerebellar-like structures, such as the ventilatory motor commands of the fish in the dorsal octaval nucleus of elasmobranchs (sharks and rays).

The application of sign-inverted Hebbian rules to excitatory networks has by itself a straightforward prediction: the output of the association neurons will tend to be reduced in response to input patterns to which the neural system is exposed frequently. Evidence in the visual system has been found for gain control processes acting in the range of hundreds of milliseconds. It is known, for instance, that sensory adaptation in forward masking (the fact that the first presentation of a stimulus can bias the perceived features of a second stimulus or alter its visibility) occurs on a time range too long to be accounted for by direct inhibitory action and too short to be compared with the effects of behavioral learning. Anti-Hebbian plasticity could potentially act in the vertebrate brain to filter out modification of sensory input resulting from motor exploration of the environment, and thus could optimize the detection of new events.

Discussion

The study of memory formation benefits from the use of simple putative elementary principles of plasticity operating at a local level (the synapse) and uniformly across the cell assembly. The large number of experimental attempts to demonstrate the validity of Hebb's postulate prediction during the last 50 years should have inevitably narrowed its fields of application. Surprisingly, Hebbian schemes have survived to become the symbol of an ever-renewed concept of synaptic plasticity, open for more generalization. A variety of experimental networks, ranging from the abdominal ganglion in the invertebrate *Aplysia* to the hippocampus and visual cortex, offer converging validation of the prediction of Hebb's postulate. In these networks, similar algorithms of potentiation can be implemented using different cascades of second messengers triggered by activation of synaptic and/or voltage-dependent conductances. Classes of processes occurring on different time scales—development or epigenesis (days and weeks), learning (minutes or hours), and even perception (milliseconds)—could have similar phenomenological outcomes.

When followed literally, Hebb's postulate refers to the set of direct excitatory synaptic contacts, originating from one presynaptic neuron, onto a postsynaptic neuron that may participate in triggering its activity, and to the correlational rules that predict an increase in efficacy in synaptic transmission. Modelers have often simplified this view to the extreme, using ideal connections between pairs of neurons, and ignoring much of the complexity of different biological implementations of the so-called Hebbian synapse in invertebrates and vertebrates. Most cellular data supporting Hebb's predictions have been derived from electrophysiological measurements of composite postsynaptic potentials or synaptic currents, or from short-latency peaks in cross-correlograms, which cannot always be interpreted simply at the synaptic level. The basic conclusion of these experiments is that covariance between pre- and postsynaptic activity up- and downregulates the "effective" connectivity between pairs of functionally coupled cells.

It may be concluded that what changes according to a correlational rule is not so much the efficacy of transmission at a given synapse but rather a more general coupling term mixing the influence of polysynaptic excitatory and inhibitory circuits linking the two cells, modulated by the diffuse network background activation. Replacing this composite interaction by a single coupling term defines an ideal Hebbian synapse and has the additional interest for the modeler of providing a weighting function of the input, which can even change sign when inhibition overcomes excitation.

Road Maps: Grounding Models of Neurons; Neural Plasticity

Related Reading: Cerebellum: Neural Plasticity; Dendritic Learning; Hebbian Learning and Neuronal Regulation; NMDA Receptors: Synaptic, Cellular, and Network Models; Post-Hebbian Learning Algorithms

References

- Abbott, L. F., and Nelson, S. B., 2000, Synaptic plasticity: Taming the beast, *Nature Neurosci.*, 3(Suppl.):1178–1183. ♦
- Ahissar, E., Abeles, M., Ahissar, M., Haidarliu, S., and Vaadia, E., 1998, Hebbian-like functional plasticity in the auditory cortex of the behaving monkey, *Neuropharmacology*, 37:633–655.
- Bell, C. C., 2001, Memory-based expectations in electrosensory systems, *Curr. Opin. Neurobiol.*, 11:481–487. ♦
- Bi, G., and Poo, M., 2001, Synaptic modification by correlated activity: Hebb's postulate revisited, *Annu. Rev. Neurosci.*, 24:139–166. ♦
- Bienenstock, E., Cooper, L. N., and Munro, P., 1982, Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex, *J. Neurosci.*, 2:32–48.
- Brown, T. H., Ganong, A. H., Kairiss, E. W., and Keenan, C. L., 1990, Hebbian synapses: Biophysical mechanisms and algorithms, *Annu. Rev. Neurosci.*, 13:475–511. ♦
- Delage, Y., 1919, *Le Rêve: Etude psychologique, philosophique et littéraire*, Paris: Presses Universitaires de France.
- Frégnac, Y., 1999, A tale of two spikes, *Nature Neurosci.*, 2:299–301.
- Frégnac, Y., and Bienenstock E., 1998, Correlational models of synaptic plasticity: Development, learning and cortical dynamics of mental representation, in *Mechanistic Relationships between Development and Learning: Beyond Metaphor* (T. Carew, R. Menzel, and C. J. Shatz, Eds.), Chichester: Wiley, pp. 113–148. ♦
- Frégnac, Y., and Shulz, D., 1994, Models of synaptic plasticity and cellular analogs of learning in the developing and adult vertebrate visual cortex, in *Advances in Neural and Behavioral Development* (V. Casagrande and P. Shinkman, Eds.), Norwood, NJ: Ablex, pp. 149–235.
- Frégnac, Y., and Shulz, D., 1999, Activity-dependent regulation of receptive field properties of cat area 17 by supervised Hebbian learning, *J. Neurobiol.*, 41:69–82. ♦
- Frégnac, Y., Shulz, D., Thorpe, S., and Bienenstock, E., 1988, A cellular analogue of visual cortical plasticity, *Nature*, 333:367–370.
- Han, V. Z., Grant, K., and Bell, C. C., 2000, Reversible associative depression and non-associative potentiation at a parallel fiber synapse, *Neuron*, 27:611–622.
- Hebb, D. O., 1949, *The Organization of Behavior*, New York: Wiley. ♦
- James, W., 1890, *Psychology: Briefer Course*, Cambridge, MA: Harvard University Press. ♦
- Konorski, J., 1948, *Conditioned Reflexes and Neuron Organization*, London: Cambridge University Press.
- Levy, W. B., and Steward, O., 1983, Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus, *Neuroscience*, 8:791–797.
- Milner, P. M., 1974, A model for visual shape recognition. *Psychol. Rev.*, 81:521–535.
- Sharma, J., Angelucci, A., and Sur, M., 2000, Induction of visual orientation modules in auditory cortex, *Nature*, 404:841–847. ♦
- Shulz, D. E., Sosnik, R., Ego, V., Haidarliu, S., and Ahissar, E., 2000, A neuronal analogue of state-dependent learning, *Nature*, 403:549–553.
- Stent, G., 1973, A physiological mechanism for Hebb's postulate of learning, *Proc. Natl. Acad. Sci. USA*, 70:997–1001.

Helmholtz Machines and Sleep-Wake Learning

Peter Dayan

Introduction

Unsupervised learning is largely concerned with finding structure among sets of input patterns such as visual scenes. An important example of structure occurs when the input patterns are generated or caused in a systematic way, for instance when objects with different shapes, surface properties, and positions are illuminated by lights of different characters and viewed by an observer with a digital camera at a particular relative location. Here, the inputs can be seen as living on a manifold that has many fewer dimensions than the space of all possible activation patterns over the pixels of the camera; otherwise, random visual noise in the camera would appear to be a normal visual scene. The manifold should correctly be parameterized by the generators themselves (i.e., the objects, the lights, etc.) (see Hinton and Ghahramani, 1997).

The Helmholtz machine (Dayan et al., 1995; Hinton et al., 1995) is an example of an approach to unsupervised learning called *analysis by synthesis* (see, e.g., Neisser, 1967). Imagine that we have a perfect computer graphics model that indicates how objects appear to observers. We can use this model to synthesize input patterns that look just like the input patterns the observer would normally receive, with the crucial difference that, since we synthesized them, we know in detail how the images were generated. We can then use these pairs of images and generators to train a model that analyzes new images to find out how they too were generated—that is, a model that represents them according to which particular generators underlie them. Conversely, if we have a perfect analysis model that indicates the generators underlying any image, then it is straightforward to use the paired images and generators to train a graphics model. In the Helmholtz machine, we attempt to have an imperfect graphics or generative model train a better analysis or

recognition model, and an imperfect recognition model train a better generative model.

There are three key issues for an analysis by synthesis model. First is the nature of the synthetic or generative model. For the Helmholtz machine, this is a structured belief network (Jordan, 1998; see also BAYESIAN NETWORKS) that is a model for hierarchical top-down connections in the cortex. This model has an overall structure (the hierarchically organized layers, units within a layer, and so on) and a set of generative parameters that determine the probability distribution the model expresses. The units in the lowest layer of the network are observable, in the sense that it is on them that the inputs are presented; units higher up in the network are latent, since they are not directly observable from inputs.

The second issue for an analysis by synthesis model is how new inputs are analyzed or recognized in light of this generative model, i.e., how the states of the latent units are determined so that the input is represented in terms of the way it would be generated by the generative model (see GRAPHICAL MODELS: PROBABILISTIC INFERENCE). For the Helmholtz machine, this is done in an approximate fashion using a second structured belief network (called the recognition model) over the latent units, whose parameters are also learned. The recognition network is a model for the standard, bottom-up connections in cortex.

The third issue is the way that the generative and recognition models are learned from data (see GRAPHICAL MODELS: PARAMETER LEARNING). For the sleep-wake learning algorithm for the stochastic Helmholtz machine, this happens in two phases. In the wake phase, the recognition model is used to estimate the underlying generators (i.e., the states of the latent units) for a particular input pattern, and then the generative model is altered so that those

generators are more likely to have produced the input that is actually observed. In the sleep phase, the generative model fantasizes inputs by choosing particular generators stochastically, and then the recognition model is altered so that it is more likely to report those particular generators, if the fantasized input were actually to be observed.

The Generative Model

Figure 1 shows an example Helmholtz machine, involving (for the sake of simplicity) three layers (\mathbf{x} , \mathbf{y} , \mathbf{d}) of binary stochastic units. The generative model uses top-down biases and weights $\mathcal{G} = \{g^x, g^y, g^d, G^{xy}, G^{yd}\}$ to parameterize a probability distribution over the input units $\mathbf{d} = (d_1, d_2, \dots)$. Consider an example in which the inputs are binary, pixelated, handwritten versions of the digit 9. In this case, we might contrive that x_i represent some relatively abstract features of the handwritten fonts (such as high curvature for circular portions or abnormal lengths for straight portions), and y_j represent some more concrete aspects of a sample character, such as tight corners at the top and bottom of the loops or elongated stems and tails.

In the standard Helmholtz machine model, the units *within* each layer are conditionally independent given the binary states of the layer above (this is called a *factorial* property). In our contrived example, across fonts, the existence of tight curvature is independent of long lengths, i.e.,

$$\rho[\mathbf{x}; \mathcal{G}] = \prod_i \rho[x_i; \mathcal{G}] \quad (1)$$

Next, given the particular abstract properties of the font (i.e., given the state of \mathbf{x}), the precise concrete realizations (i.e., the y_j) are individually independent:

$$\rho[\mathbf{y}|\mathbf{x}; \mathcal{G}] = \prod_j \rho[y_j|\mathbf{x}; \mathcal{G}] \quad (2)$$

Finally, given these concrete properties, pixels are independently inked:

$$\rho[\mathbf{d}|\mathbf{y}; \mathcal{G}] = \prod_k \rho[d_k|\mathbf{y}; \mathcal{G}] \quad (3)$$

In sum, by marginalizing, one can write

$$\rho[\mathbf{d}; \mathcal{G}] = \sum_{\mathbf{x}, \mathbf{y}} \rho[\mathbf{x}; \mathcal{G}] \rho[\mathbf{y}|\mathbf{x}; \mathcal{G}] \rho[\mathbf{d}|\mathbf{y}; \mathcal{G}] \quad (4)$$

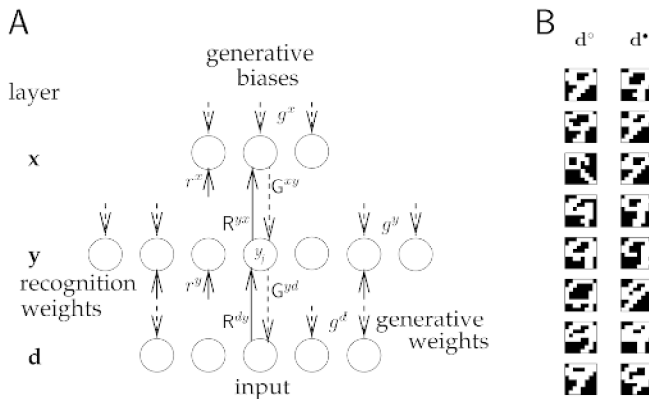


Figure 1. Helmholtz machine. A, Structure of a three-layer Helmholtz machine, with generative weights and biases \mathcal{G} (dashed) and recognition weights \mathcal{R} (solid). B, Handwritten digit example. The left column shows eight samples from a training set of binarized, 8×8 , handwritten 9s; the right column shows eight samples produced by the generative model after training. The training set is as described in Hinton et al. (1995).

For binary stochastic units,

$$\begin{aligned} \rho[x_i; g_i^x] &= \sigma(g_i^x) \\ \rho[y_j|\mathbf{x}; g_j^y, G^{xy}] &= \sigma\left(g_j^y + \sum_i G_{ji}^{xy} x_i\right) \equiv \hat{y}_j(\mathbf{x}) \end{aligned} \quad (5)$$

where $\sigma(u) = 1/(1 + \exp(-u))$ is the standard sigmoid function. Although the units \mathbf{y} are conditionally independent given the states of the units \mathbf{x} in the layer above, they are not marginally independent across all the patterns that can be generated. That is, \mathbf{x} can capture statistical structure (i.e., correlations) in the states of \mathbf{y} . So, for instance, the abstract curvature property captures the correlation that a tight corner at the bottom of the loop of a 9 is typically associated with a tight corner at the top of the loop. Similarly, \mathbf{y} captures correlations in the states \mathbf{d} . The font example is only for illustration. We actually expect the unsupervised learning algorithm itself to find the representations \mathbf{y} and \mathbf{x} that collectively best capture the statistical structure of \mathbf{d} .

This top-down generative model is a simple example of a sigmoid belief net (or BAYESIAN NETWORKS). The conditional independence within a layer makes it very straightforward to generate a sample \mathbf{d}^* from $\rho[\mathbf{d}; \mathcal{G}]$ by fantasizing a sample \mathbf{x}^* , then \mathbf{y}^* given \mathbf{x}^* (both as in Equation 5), and then \mathbf{d}^* given \mathbf{y}^* (using a similar expression).

The Recognition Model

When the generative model is used to create such a complete fantasy, we consider \mathbf{x}^* and \mathbf{y}^* as the *generators* of \mathbf{d}^* . The task for the recognition model is to take a new example \mathbf{d} and report the state(s) of \mathbf{x} and \mathbf{y} that might have generated it. In terms of our example, this implies finding a representation for a new image of a handwritten 9 in terms of the settings of the more concrete and more abstract parameters that could have generated it. Using Bayes's theorem, we know that

$$\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{G}] = \rho[\mathbf{d}|\mathbf{x}, \mathbf{y}; \mathcal{G}] \frac{\rho[\mathbf{x}; \mathcal{G}] \rho[\mathbf{y}|\mathbf{x}; \mathcal{G}]}{\rho[\mathbf{d}; \mathcal{G}]} \quad (6)$$

It is straightforward to calculate all the terms on the right-hand side *except* for the denominator $\rho[\mathbf{d}; \mathcal{G}]$, which involves a sum over all the possible states of \mathbf{x} and \mathbf{y} (a set that grows exponentially large as the number of elements in \mathbf{x} and \mathbf{y} grows). Thus, an approximation to $\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{G}]$ is usually required. The stochastic version of the Helmholtz machine (the only version we discuss here) uses for approximate recognition a bottom-up belief network (see Figure 1) over exactly the same units. This network instantiates a probability distribution $\mathcal{Z}[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{R}] = \mathcal{Z}[\mathbf{y}|\mathbf{d}; \mathcal{R}] \mathcal{Z}[\mathbf{x}|\mathbf{y}; \mathcal{R}]$ using a separate set of parameters, the bottom-up biases and weights $\mathcal{R} = \{r^x, r^y, R^{dy}, R^{yx}\}$. A critical approximation is that the recognition model is assumed to be factorial in the bottom-up direction, i.e., y_1 is independent of y_2 given \mathbf{d} , and so forth. Over the course of learning, it is intended that $\mathcal{Z}[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{R}]$ should come to be as close to $\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{G}]$ as possible, subject to this approximation. On account of this approximation, it is as easy to generate a sample bottom-up from the recognition model, i.e., to recognize the input in terms of its generators, as it is to generate a sample top-down from the generative model, i.e., to create a fantasy.

Sleep-Wake Learning

As for many unsupervised learning methods (see UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS), the underlying goal of sleep-wake learning is to perform maximum likelihood density estimation by maximizing the log probability of the observed

data $\mathcal{D} = \{\mathbf{d}(1), \mathbf{d}(2), \dots\}$ under the generative model, that is, $E(\mathcal{Q}) = \sum_t \log \rho[\mathbf{d}(t)|\mathcal{Q}]$. One key idea, due to Neal and Hinton (1998) and Zemel (1994) (see MINIMUM DESCRIPTION LENGTH ANALYSIS), is to take the logarithm of both sides of Equation 4, though using the exact recognition distribution of Equation 6 to swap the order of the log and the sum:

$$\log \rho[\mathbf{d}; \mathcal{Q}] = \sum_{\mathbf{x}, \mathbf{y}} \rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}] \log \rho[\mathbf{x}, \mathbf{y}, \mathbf{d}; \mathcal{Q}] + \mathcal{H}[\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}]] \quad (7)$$

and then to introduce an approximate recognition distribution $\mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]$, with a bounded effect on the expression

$$\log \rho[\mathbf{d}; \mathcal{Q}] \geq \sum_{\mathbf{x}, \mathbf{y}} \mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}] \log \rho[\mathbf{x}, \mathbf{y}, \mathbf{d}; \mathcal{Q}] + \mathcal{H}[\mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]] \quad (8)$$

that can be quantified using a measure of the discrepancy between approximate and exact recognition distributions:

$$= \log \rho[\mathbf{d}; \mathcal{Q}] - \text{KL}[\mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}], \rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}]] \quad (9)$$

$$\equiv -\mathcal{F}[\mathbf{d}; \mathcal{R}, \mathcal{Q}] \quad (10)$$

Here, $\mathcal{H}[\mathcal{A}] = -\sum_{\mathbf{a}} \mathcal{A}[\mathbf{a}] \log \mathcal{A}[\mathbf{a}]$ is the entropy of probability distribution \mathcal{A} , and, in Equation 9, $\text{KL}[\mathcal{A}, \mathcal{B}] = \sum_{\mathbf{a}} \mathcal{A}[\mathbf{a}] \log \mathcal{A}[\mathbf{a}] / \mathcal{B}[\mathbf{a}]$ is the Kullback-Liebler (KL) divergence between two distributions \mathcal{A} and \mathcal{B} . This KL divergence is greater than or equal to 0, with equality when \mathcal{A} and \mathcal{B} are essentially equal. Thus, Inequality 8 holds for an arbitrary distribution $\mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]$ over \mathbf{x}, \mathbf{y} , with equality if $\mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]$ is the true analytical distribution $\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}]$. Expression $\mathcal{F}[\mathbf{d}; \mathcal{R}, \mathcal{Q}]$ can be seen as a Helmholtz free energy; hence the name of the machine.

During the *wake* phase, a single pattern \mathbf{d}° is sampled from \mathcal{D} and presented to the recognition model. This is executed bottom-up to produce a single sample \mathbf{y}° given \mathbf{d} and \mathbf{x}° given \mathbf{y}° . Then, the parameters \mathcal{Q} of the generative model are changed using stochastic gradient ascent of the lower bound to the log probability, i.e., proportionally to

$$\nabla_{\mathcal{Q}} \log \rho[\mathbf{x}^\circ, \mathbf{y}^\circ, \mathbf{d}; \mathcal{Q}] = \nabla_{\mathcal{Q}} \{\log \rho[\mathbf{x}^\circ; \mathcal{Q}] + \log \rho[\mathbf{y}^\circ|\mathbf{x}^\circ; \mathcal{Q}] + \log \rho[\mathbf{d}|\mathbf{y}^\circ; \mathcal{Q}]\}$$

For activation functions such as those in Equation 5, this leads to particularly simple “delta” learning rules such as

$$\Delta g_j^y \propto (y_j^\circ - \hat{y}_j(\mathbf{x}^\circ)) \quad \Delta G_{ij}^{xy} \propto (y_j^\circ - \hat{y}_j(\mathbf{x}^\circ)) x_i^\circ$$

in which the output of the recognition model is used as the *target* for the generative model instead of vice versa.

The ideal for the *sleep* phase would be to change the recognition weights \mathcal{R} using stochastic gradient descent also of the lower bound in Equation 9. Unfortunately, this procedure is not generally computationally tractable. The second key idea in sleep-wake learning is to attempt during sleep to minimize $\text{KL}[\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}]; \mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]]$ instead. This is not the same, since the KL divergence is not symmetric, although they are equal at their joint minimum where $\rho[\mathbf{x}, \mathbf{y}|\mathbf{d}; \mathcal{Q}] = \mathcal{Z}[\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathcal{R}]$. The KL divergence the wrong way around can be minimized by fantasizing sample \mathbf{x}^* , \mathbf{y}^* , and \mathbf{d}^* from the generative model (in the way described at the end of the section on that model), and then changing the recognition weights accordingly.

$$\nabla_{\mathcal{R}} \log \mathcal{Z}[\mathbf{x}^*, \mathbf{y}^*, \mathbf{d}^*, \mathcal{R}] = \nabla_{\mathcal{R}} \{\log \mathcal{Z}[\mathbf{y}^*|\mathbf{d}^*; \mathcal{R}] + \log \mathcal{Z}[\mathbf{x}^*|\mathbf{y}^*; \mathcal{R}]\}$$

For activation functions such as those in Equation 5, this leads to the same simple delta learning rules as for the generative model, except that the output of the generative model is used as the target for the recognition model.

Since sleep learning involves an approximation, it is only in very special cases (see Neal and Dayan, 1997) that it is possible to prove even that it is appropriately stable. Nevertheless, the model has

been shown to work quite well in practice. Figure 1B shows the result of applying sleep-wake learning to a set of input patterns (left column) that are binary images of the handwritten digit 9. The right column shows fantasized samples following learning, and these can be seen to be generated by a distribution close to that in the training distribution.

Discussion

As a directed belief network for analysis by synthesis that is trained according to maximum likelihood density estimation, the Helmholtz machine lives in what has become a rather crowded space (see UNSUPERVISED LEARNING WITH GLOBAL OBJECTIVE FUNCTIONS). In this context, the key property of the Helmholtz machine is that it uses an explicit recognition model that has its own parameters rather than performing recognition by an iterative process involving only the parameters of the generative model. In some ways this is an advantage; in particular, recognition can occur swiftly in a single pass. Learning during the sleep phase can be considered as a way of caching knowledge about how to do recognition effectively. In other ways it is a disadvantage, since the recognition model introduces an extra set of parameters that need to be learned and since, unlike the iterative mean field recognition methods that underlie most of the architectures mentioned earlier, the approximation involved in the recognition model in the Helmholtz machine cannot be tailored on-line to the particular input pattern that is presented. Another key feature is that, unlike many of these methods, the Helmholtz machine is explicitly designed to be hierarchical: units in one layer capture (i.e., both represent and generate) correlations in the layer below. Unlike INDEPENDENT COMPONENT ANALYSIS (q.v.), for instance, units within a layer are not forced to be marginally independent in the generative model, only conditionally independent *given* the activities in the layer above. This potentially allows it a much richer representation of the inputs. Also, the recognition model in the Helmholtz machine allows at least some correlations among the states of the hierarchical generators, a feature denied to mean field methods.

The Helmholtz machine also bears an interesting relationship to the Boltzmann machine (Hinton and Sejnowski, 1986; see also SIMULATED ANNEALING AND BOLTZMANN MACHINES), which can be seen as an undirected belief net. In the Boltzmann machine, which also lacks an explicit recognition model, a potentially drawn-out process of Gibbs sampling is used to recognize and generate inputs, since there is nothing like the simple, one-pass, directed recognition and generative belief networks of the Helmholtz machine. Also, the Boltzmann machine learning rule performs true stochastic gradient ascent of the log likelihood using a contrastive procedure, which, confusingly, involves wake and sleep phases that are quite different from the wake and the sleep phases of the sleep-wake algorithm. The two phases of the Boltzmann machine contrast the statistics of the activations of the network when input patterns are presented with the statistics of the activations of the network when it is running “free.” This contrastive procedure involves substantial noise and is therefore slow (a problem rectified in Hinton’s (2000) new approximate contrastive divergence learning rule). In the sleep-wake learning procedure for the Helmholtz machine, the wake and sleep phases are not contrastive. Rather, the recognition and generative models are forced to chase each other.

The most important open issue for the Helmholtz machine as a model of top-down and bottom-up connections in the cortex is how to weaken the approximation that the recognition and generative models are factorial within layers, without destroying the simplicity of sampling from and learning the models.

Road Map: Learning in Artificial Networks

Related Reading: Ying-Yang Learning

References

- Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S., 1995, The Helmholtz machine, *Neural Computat.*, 7:889–904.
- Hinton, G. E., 2000, *Training Products of Experts by Minimizing Contrastive Divergence*, Gatsby Computational Neuroscience Unit Technical Report TR 2000–004, London: Alexandra House.
- Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M., 1995, The wake-sleep algorithm for unsupervised neural networks, *Science*, 268:1158–1160. ♦
- Hinton, G. E., and Ghahramani, Z., 1997, Generative models for discovering sparse distributed representations, *Philos. Trans. R. Soc. B*, 352:1177–1190. ♦
- Hinton, G. E., and Sejnowski, T. J., 1986, Learning and relearning in Boltzmann machines, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, Foundations* (D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Eds.), Cambridge, MA: MIT Press, pp. 282–317.
- Jordan, M. I., Ed., 1998, *Learning in Graphical Models*, Dordrecht: Kluwer. ♦
- Neal, R. M., and Dayan, P., 1997, Factor analysis using delta-rule wake-sleep learning, *Neural Computat.*, 9:1781–1803.
- Neal, R. M., and Hinton, G. E., 1998, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in Graphical Models* (M. I. Jordan, Ed.), Dordrecht: Kluwer, pp. 355–368.
- Neisser, U., 1967, *Cognitive Psychology*, New York: Appleton-Century-Crofts.
- Zemel, R. S., 1994, A minimum description length framework for unsupervised learning, Ph.D. diss., Toronto: University of Toronto, Computer Science Department.

Hemispheric Interactions and Specialization

James A. Reggia and Svetlana Levitan

Introduction

Currently recognized *hemispheric specializations*, where one cerebral hemisphere performs a task better than the other, include language, handedness, visuospatial processing, emotion and its facial expression, olfaction, and attention (Hellige, 1993). For example, in roughly 95% of people the left cerebral hemisphere is dominant for language, so language is said to be *lateralized* to the left hemisphere in such individuals. Behavioral lateralization in areas such as vocalization and motor preferences has been demonstrated not only in people but also in rodents, birds, primates, and other animals.

The underlying causes of hemispheric specialization/lateralization are not well understood at present. The many anatomical, biochemical, and physiological asymmetries that exist in the brain include a larger left temporal plane in the majority of subjects, greater dendritic branching in speech areas of the left hemisphere, different distributions of important neurotransmitters such as dopamine and norepinephrine between the hemispheres, and a lower left hemisphere threshold for motor-evoked potentials. Understanding which, if any, of these asymmetries actually contribute to hemispheric specialization remains an important problem in neuropsychology and is an instance of the more general issue of how functional modularity arises in the brain (Jacobs, 1997).

Besides the underlying hemispheric asymmetries listed above, another potential factor in function lateralization is hemispheric interactions via pathways connecting the hemispheres, such as the corpus callosum. Callosal fibers are largely but not exclusively homotopic: roughly mirror-symmetric points in each hemisphere are connected to each other. Callosal connections between the hemispheres are excitatory. This, as well as clinical data and split-brain experiments, suggests that transcallosal hemispheric interactions are mainly excitatory in nature, but this hypothesis has long been quite controversial. Transcallosal monosynaptic excitatory effects are subthreshold and followed by stronger, more prolonged inhibition, suggesting to some that transcallosal inhibitory interactions are much more important (Cook, 1986). The case for interhemispheric inhibition/competition has been strengthened recently by transcranial magnetic stimulation studies indicating that activation of one primary motor cortex inhibits the opposite one.

Neural modeling provides a useful way to investigate the implications of hypotheses that complement more traditional methods. In this article, we first consider models of hemispheric interactions that do not incorporate hemispheric differences and, conversely,

models examining the effects of hemispheric differences that do not incorporate hemispheric interactions. We then look in more detail at some examples of recent work that studies both hemispheric interactions and differences in the same model, demonstrating how these factors influence the emergence of lateralization in models in which lateralization is not initially present. Finally, we briefly summarize insights gained from simulated damage in these models.

Modeling Hemispheric Interactions/Differences

A number of neural models have examined issues involving hemispheric interactions. In one group of studies, several models were developed representing homotopic left and right hemispheric regions connected via a simulated corpus callosum. For example, one early study demonstrated that oscillatory activity in a simulated hemisphere could be transferred to the other hemisphere via interhemispheric connections (Anninos and Cook, 1988). This was true regardless of assumptions about the excitatory/inhibitory nature of callosal connections, although learning was more rapid when callosal connections were excitatory. Another model, using paired neural networks representing left and right cortical regions, showed that homotopic inhibitory callosal connections produce complementary activity patterns but not lateralization in the two simulated hemispheric regions (Cook, 1986). This was done without postulating intrinsic differences between the cerebral hemispheres. These and other earlier neural models involved simulating hemispheric interactions, but none of them considered underlying asymmetric hemispheric regions or emergent lateralization.

In contrast, other models have simulated single hemispheric regions under differing conditions that are motivated by known asymmetries in the left and right cerebral hemispheres. Although these models have neither paired left and right cortical regions nor callosal connections like those above, they do examine issues related to lateralization.

One study focused on modeling hemispheric differences observed during semantic priming experiments (Burgess and Lund, 1998). Such experiments examine how the occurrence of one word facilitates the subsequent recall from memory of words with similar meanings. This model assumed that each hemisphere has the same representation of semantic information about words, but that the right hemisphere processes this information more slowly. The speed of activation of a target word for a hemisphere was also

assumed to be based on a function of the semantic distance between that word and a priming word, and on the word's rate of activation decay. These word-specific quantities were derived from a large corpus of written text. Separate parameterized functions for the rate of activation of a word were determined for left and right hemispheres by evolving the parameters in these functions using a genetic algorithm. The resultant activation functions for each hemisphere, each function having different parameters, were then used successfully to predict the time course of activating word meaning in a number of priming experiments in which stimuli were presented separately to each hemisphere. These results implicitly provide a theory about the underlying differences in left and right hemisphere processing of semantic information.

Another study examining issues related to cerebral specialization without actually simulating paired hemispheric regions, corpus callosum, or the emergence of lateralization focused on the processing of spatial relations (Kosslyn et al., 1992). This investigation hypothesized that receptive field size asymmetries involving low-level visual neurons were responsible for experimental observations that the right cerebral hemisphere is faster at computing coordinate spatial relations involving precise metric information, such as judging distances, while the left hemisphere is faster in evaluating some categorical relations, such as above versus below. The coarse coding provided by error backpropagation neural networks with larger, overlapping receptive fields assumed for the right hemisphere was found to be superior for distance judgments, whereas networks with smaller, nonoverlapping receptive fields assumed for the left hemisphere were found to be superior in judging categorical relations. Conversely, training a network with adaptable receptive fields to do distance judgments led to larger receptive fields than training a network to judge categorical relations (Jacobs and Kosslyn, 1994). A separate model focusing on receptive field asymmetries from a different perspective has also been used to explain several experimentally observed visual processing asymmetries (Ivry and Robertson, 1998).

Simulating the Emergence of Lateralization

A number of neural models of paired left and right cortical regions have been used to examine conditions under which lateralization that was not present initially might emerge during learning. These models incorporate both hemispheric interactions and hemispheric differences. Three examples are given here. The first two incorporate a simulated corpus callosum, and represent end points on a spectrum from supervised to unsupervised learning within which other models fall (Cook, 1999; Shevtsova and Reggia, 1999). The third example illustrates how noncallosal hemispheric interactions can influence lateralization.

Phoneme Sequence Generation

A phoneme sequence generation model, trained using recurrent error backpropagation, was created that takes three-letter words as input and produces the correct temporal sequence of phonemes for the pronunciation of each word as output (Reggia, Goodall, and Shkuro, 1998). Figure 1a schematically summarizes the network architecture, where input elements are fully connected to two sets of neural elements representing corresponding regions of the left and right hemisphere cortex. These regions are connected to each other via a simulated corpus callosum and to output elements representing individual phonemes.

The effects of different hemispheric asymmetries (relative size, excitability, learning rate parameter, etc.) on the emergence of lateralization were examined one at a time. For each hemispheric asymmetry, and for a symmetric control version of the model, the uniform value of callosal connection influences was varied over

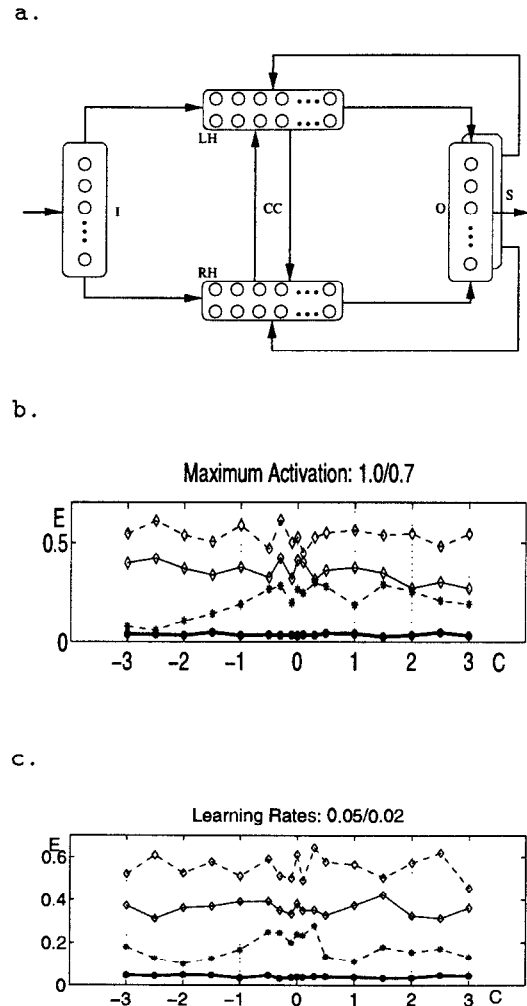


Figure 1. a, Network architecture for phoneme sequence generation. I, input elements; LH (RH), left (right) hemisphere cortex region; CC, corpus callosum; O, output elements; S, state elements. In the graphs, error (E) is plotted versus callosal strength (C) for model versions where the left hemisphere (b) is more excitable or (c') has more potent synaptic plasticity. In each case the upper dashed curve shows pretraining error and the lowest (thick solid) curve shows post-training error for the full model's output. The two middle curves show post-training output error when the left hemisphere alone (dashed line) or the right hemisphere alone (solid line) controls output.

several excitatory and inhibitory values. Lateralization was measured based on the difference between the output error when the left hemispheric region alone controlled the output versus when the right hemispheric region alone did so. These simulations showed that, within the limitations of the model, it is easy to produce lateralization. For example, lateralization occurred toward the side with higher excitability (Figure 1b) or a higher learning rate parameter (Figure 1c), depending on callosal strength. Lateralization tended to occur most readily and intensely when callosal connections exerted predominantly an inhibitory influence (Figure 1b), but with some asymmetries significant lateralization occurred for all callosal strengths (Figure 1c). In this specific model, the results could be interpreted as a competitive "race-to-learn" involving the two hemispheric regions, with the "winner" (dominant side) determined when the model as a whole acquired the input-output map-

ping and learning largely ceased. These results suggest that lateralization to a cerebral region can, in some cases, be associated with increased synaptic plasticity in that region relative to its mirror-image region in the opposite hemisphere, a testable prediction.

Self-Organizing Topographic Maps

Topographic maps representing various aspects of the environment are found in primary sensory regions of cortex. Maps in mirror-image regions of sensory cortex exhibit a rich range of patterns of asymmetries and lateralization (Bianki, 1993). A model of corresponding left and right hemispheric regions receiving sensory input has been used to study emergent map asymmetries and lateralization (Levitan and Reggia, 1999). Unlike the phoneme sequence generation model, purely unsupervised competitive learning was used. Map formation was examined while varying the underlying cortical region asymmetries and the assumed excitatory/inhibitory nature of callosal connections.

Figure 2 illustrates pairs of cortical maps appearing in this model where the centers of receptive fields of the cortical elements are plotted in the space of the sensory surface (i.e., these are *not* pictures of the cortical regions involved). Lines between plotted points indicate adjacent cortical elements. In all simulations, cortical maps were initially highly disorganized before learning, owing to randomly assigned initial synaptic strengths (Figure 2a). For excitatory, absent, or weakly inhibitory callosal interactions, complete

and symmetric mirror-image maps appeared after learning in both hemispheric regions (Figure 2b). In contrast, with stronger inhibitory callosal interactions, after learning, map lateralization tended to occur (Figure 2c; left more organized than right), or the maps became complementary (Figure 2d), reminiscent of “mosaic patterns” described experimentally (Bianki, 1993). Lateralization occurred readily toward the side having higher excitability or a larger cortical region. Unlike with the phoneme sequence generation model, asymmetric plasticity had only a transitory effect on lateralization, indicating that the effects of this factor may differ substantially depending on whether supervised or unsupervised learning is used. In this model, a “phase transition” in behavior occurs at a specific inhibitory callosal strength: above this value, bilateral symmetric maps occur; below it, lateralization and complementary maps occur.

Spatial Relations

Asymmetries in perceptual abilities are sometimes hypothesized to be due to an underlying asymmetry in receptive field sizes in the early visual system: smaller visual receptive fields on the left than on the right. A neural model of two hemispheric regions, represented as multilayer feedforward networks having different afferent pathway receptive field sizes, was implemented to examine the plausibility of this hypothesis (Jacobs and Kosslyn, 1994). This model did not include a simulated corpus callosum but instead had an extra module called a “gating network” that determined the extent to which each hemispheric region controlled the model’s output. Each hemispheric region competed to learn each output pattern, and the region whose output most closely matched the correct output pattern “won” and learned more.

In simulations performed with this model, the network was trained to perform both categorical and coordinate tasks, while the left input layer was forced to have smaller receptive field sizes. When the difference in left and right receptive field sizes was sufficiently large, the left hemispheric region became superior at the categorical task, while the right hemispheric region became superior with the coordinate task. These results are consistent with the hypothesis that asymmetric receptive field sizes contribute to hemispheric specialization in processing spatial relations.

Effects of Simulated Damage

Several versions of the above models of interacting left and right hemispheric regions have been subjected to simulated focal damage (Levitan and Reggia, 1999; Shkuro, Glezer, and Reggia, 2000). For example, with the phoneme sequence generation model, an area of focal damage was introduced into an intact model by making part of one simulated hemispheric region nonfunctional. Performance errors of the full model and each hemisphere alone were measured immediately after the damage and then after further training restored the model’s performance to normal. During the recovery period, in which performance eventually returned to predamage levels, the undamaged hemispheric region very often participated in and contributed to recovery, more so as the amount of damage increased. These results support the controversial hypothesis that the intact cerebral hemisphere plays a role in adult recovery from damage in the opposite hemisphere. Further, in these simulations of focal damage, when callosal influences were excitatory, the undamaged hemispheric region often had a drop in mean activation and exhibited impaired performance, representing the analog of transcallosal diaschisis (a fall in regional cerebral blood flow and glucose metabolism in the intact cerebral hemisphere following unilateral brain damage). These and other observations following model damage support the view that hemispheric interactions via callosal connections are excitatory.

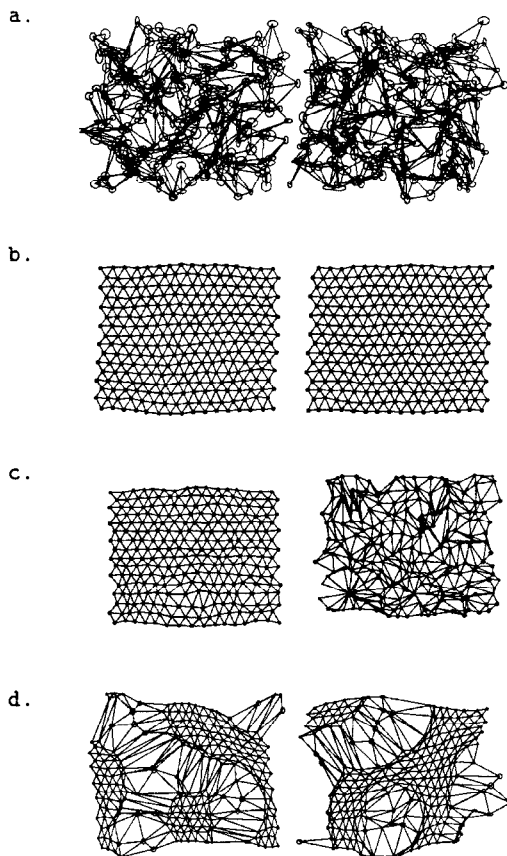


Figure 2. *a*, Unorganized maps bilaterally (e.g., prelearning). *b*, Bilaterally organized maps (e.g., post-training with weak excitatory callosal connections). *c*, Left map more organized than right. *d*, Complementary mosaic maps.

Discussion

Neural models have recently been used to investigate a variety of issues concerning whether underlying asymmetries can lead to or can explain lateralization and how assumptions about callosal influences affect lateralization. Although these models have been limited in size and scope, and although they are greatly simplified from neurobiological and behavioral reality, they have demonstrated a number of results that are relevant to current issues in brain theory.

First, any one of a variety of underlying hemispheric asymmetries can lead to hemispheric specialization in these models, including asymmetric size, excitability, receptive field sizes, and synaptic plasticity. Such a finding supports past arguments that a single underlying hemispheric asymmetry is unlikely to account for language and other behavioral lateralizations (Hellige, 1993).

Second, resolving the issue of the inhibitory versus excitatory nature of callosal influences remains elusive. Lateralization generally appeared most intensely in models with inhibitory inter-hemispheric interactions, lending support to past arguments that whatever the actual neurophysiological nature of callosal synaptic connections, callosal influences are apparently inhibitory in nature, producing competition between the two hemispheres. However, lateralization also occurred in some models with excitatory callosal influences and asymmetric hemispheric regions, and focal damage in these models caused a decrease in activation and sometimes performance in the opposite intact hemispheric region. These latter changes resemble those seen experimentally in stroke patients, supporting past arguments that callosal influences are predominantly excitatory. This conflicting evidence concerning the excitatory versus inhibitory nature of callosal influences may be referred to as the *callosal dilemma*. A possible resolution of this dilemma, excitatory transcallosal influences in the context of subcortical cross-midline inhibitory connections, was recently studied computationally (Reggia et al., 2001). Further examination of this issue would be a fertile topic for future research.

Finally, during the recovery period following focal damage to models, the opposite intact hemispheric region generally contributed to recovery, supporting the controversial hypothesis that the undamaged cerebral hemisphere often contributes to language recovery following unilateral brain damage in adults. This effect increased with increasing damage, suggesting that relevant future ex-

perimental studies of recovery from stroke should carefully control for this factor in interpreting data.

Road Map: Cognitive Neuroscience

Related Reading: Lesioned Networks as Models of Neuropsychological Deficits; Neuropsychological Impairments

References

- Anninos, P., and Cook, N., 1988, Neural net simulation of the corpus callosum, *Int. J. Neurosci.*, 38:381–391.
- Bianki, V., 1993, *The Mechanism of Brain Lateralization*, Newark, NJ: Gordon & Breach.
- Burgess, C., and Lund, K., 1998, Modeling cerebral asymmetries in high-dimensional semantic space, in *Right Hemisphere Language Comprehension* (M. Beeman and C. Chiarello, Eds.), Mahwah, NJ: Erlbaum, pp. 215–244.
- Cook, N., 1986, *The Brain Code*, New York: Methuen. ♦
- Cook, N., 1999, Simulating consciousness in a bilateral neural network, *Consciousness Cognit.*, 8:62–93.
- Hellige, J., 1993, *Hemispheric Asymmetry*, Cambridge, MA: Harvard University Press. ♦
- Ivry, R., and Robertson, L., 1998, *The Two Sides of Perception*, Cambridge, MA: MIT Press, pp. 225–255.
- Jacobs, R., 1997, Nature, nurture and the development of functional specialization, *Psychonom. Bull. Rev.*, 4:299–309.
- Jacobs, R., and Kosslyn, S., 1994, Encoding shape and spatial relations, *Cognit. Sci.*, 18:361–386.
- Kosslyn, S., Chabris, C., Marsolek, C., and Koenig, O., 1992, Categorical vs. coordinate spatial relations, *J. Exper. Psychol. Hum. Percept. Performance*, 18:562–577.
- Leviton, S., and Reggia, J., 1999, Interhemispheric effects on map organization following simulated cortical lesions, *Artific. Intell. Med.*, 17:59–85.
- Reggia, J., Goodall, S., and Shkuro, Y., 1998, Computational studies of lateralization of phoneme sequence generation, *Neural Computat.*, 10:1277–1297.
- Reggia, J., Goodall, S., Shkuro, Y., and Glezer, M., 2001, The callosal dilemma, *Neurol. Res.*, 23:465–471.
- Shevtsova, N., and Reggia, J., 1999, A neural network model of lateralization during letter identification, *J. Cognit. Neurosci.*, 11:167–181.
- Shkuro, Y., Glezer, M., and Reggia, J., 2000, Interhemispheric effects of simulated lesions in a neural model of single-word reading, *Brain Lang.*, 72:343–374.

Hidden Markov Models

Hervé Bourlard and Samy Bengio

Introduction

Over the past 20 years, finite-state automata (FSA), and more particularly stochastic finite-state automata (SFSA) and different variants of hidden Markov models (HMMs), have been used successfully to address several complex sequential pattern recognition problems, among them continuous speech recognition, cursive (handwritten) text recognition, time series prediction, and biological sequence analysis.

FSAs allow complex learning problems to be solved by assuming that the sequential pattern can be decomposed into piecewise stationary segments, encoded through the topology of the FSA. Each stationary segment can be parameterized in terms of a deterministic or a stochastic function. In the latter case, it may also be possible that the SFSA state sequence is not observed directly but is a probabilistic function of the underlying finite-state Markov

chain. This leads to the definition of the powerful HMMs, involving two concurrent stochastic processes: the sequence of HMM states modeling the sequential structure of the data, and a set of state output processes modeling the (local) stationary character of the data. The Markov model is called hidden because there is an underlying stochastic process (i.e., the sequence of states) that is not observable but that affects the observed sequence of events.

Furthermore, depending on the way the SFSA is parameterized and trained, FSAs (and HMMs in particular) can be used either as a *production model*, in which the observation sequence is considered to be an output signal produced by the model, or as a *recognition model* (acceptor), in which the observation sequence is considered as being accepted by the model. Finally, it may also be the case that the HMM is used to explicitly model the stochastic relationship between two (input and output) event sequences, resulting in a model usually referred to as an input-output HMM.

The parameters of these models can be trained by different variants of the powerful Expectation-Maximization (EM) algorithm (Baum and Petrie, 1966; Liporace, 1982), which, depending on the criterion being used, is referred to as maximum likelihood (ML) or maximum a posteriori (MAP) training. However, even though they belong to the same family, all of these models exhibit different properties. This article compares some of the variants of these powerful SFSA and HMM models currently used for sequence processing.

Finite-State Automata

In its more general form (Hopcroft, Motwani, and Ullman, 2000) and as summarized in Table 1, an FSA, which will be denoted M in this paper, is defined as an abstract machine consisting of the following:

- A set of states $\mathcal{Q} = \{I, 1, \dots, k, \dots, K, F\}$, including the initial state I and final state F , also referred to as accepting state (in the case of recognizers). Variants of this machine include machines having multiple initial states and multiple accepting states. In this article, a specific state visited at time t will be denoted q_t .
- A set \mathcal{Y} of (discrete or continuous) input symbols or vectors. A particular sequence of size T of input symbols/vectors will be denoted $Y = \{y_1, y_2, \dots, y_T\} = y_1^T$, where y_t represents the input symbol/vector at time t .
- A set \mathcal{Z} of (continuous or discrete) output symbols or vectors. A particular sequence of size T of output symbols/vectors will be denoted $Z = z_1^T$, where z_t represents the output symbol/vector at time t .
- A *state transition function* $q_t = f(y_t, q_{t-1})$, which takes the current input event and the previous state q_{t-1} and returns the next state q_t .
- An *emission function* $z_t = g(q_t, q_{t-1})$, which takes the current state q_t and the previous state q_{t-1} and returns an output event z_t . This automaton is usually known as a *Mealy FSA*, i.e., one producing an output for each input-dependent *transition*. As a variant of this, the emission function of a *Moore FSA* depends only on the current state, i.e., $z_t = g(q_t)$, thus producing an output for each visited state. There is, however, a homomorphic equiv-

alence between Mealy and Moore automata, given an increase and renaming of the states.

Finally, in the case of sequential pattern processing, the processed sequence is often represented as an *observed sequence* of symbols or vectors which, depending on the type of automata and optimization criterion, will sometimes be considered input events and at other times output events. To accommodate this flexibility, we also define the *observed sequence* of size T as $X = x_1^T$, where x_t is the observed event/vector at time t . For example, in the case of speech recognition, x_t would be the acoustic vector resulting from the spectral analysis of the signal at time t , and is equivalent to z_t (since in that case the observations are the outputs of the FSA).

A *deterministic FSA* is one in which the transition and emission functions $f(\cdot)$ and $g(\cdot)$ are deterministic, implying that the output event and next state are uniquely determined by a single input event (i.e., there is exactly one transition for each given input event and state). In comparison, a nondeterministic FSA is one in which the next state is not uniquely determined by the current previous state and input event. However, it is often possible to transform a nondeterministic FSA into a deterministic FSA, at the cost of a significant increase in the possible number of input symbols.

We will not further discuss deterministic FSAs, which have been largely used in language theory (Hopcroft et al., 2000), where FSAs are often used to accept or reject a language (i.e., certain sequences of input events).

Stochastic Finite-State Automata

A *stochastic FSA* (SFSA) is an FSA in which the transition and/or emission functions are probabilistic functions. In the case of Markov models, there is a one-to-one relationship between the observation and the state, and the transition function is probabilistic. In the case of HMMs, the emission function is also probabilistic, and the states are no longer directly observable through the input events. Instead, each state produces one of the possible output events with a certain probability.

Depending on their structure (discussed below), transition and emission (probability density) functions are represented in terms of a set of parameters Θ , which will have to be estimated on repre-

Table 1. Deterministic and Stochastic Finite-State Automata

	Deterministic Finite-State Automata	Stochastic Finite-State Automata			
		Markov Model	HMM	HMM/ANN	IOHMM
States	$k \in \mathcal{Q}$	$x_t = k \in \mathcal{Q}$	$k \in \mathcal{Q}$	$k \in \mathcal{Q}$	$k \in \mathcal{Q}$
Input symbols	$y_t \in \mathcal{Y}$	—	—	$x_t = y_t \in \mathcal{Y}$	$x_t = y_t \in \mathcal{Y}$
Output symbols	$z_t \in \mathcal{Z}$	—	$x_t = z_t \in \mathcal{Z}$	—	$z_t \in \mathcal{Z}$
Transition law	$q_t = f(y_t, q_{t-1})$	Transition probabilities $p(x_t q_{t-1})$	Transition probabilities $p(q_t q_{t-1})$	Conditional transition probabilities $p(q_t x_t, q_{t-1})$	Conditional transition probabilities $p(q_t x_t, q_{t-1})$
Emission law					
Mealy	$z_t = g(q_t, q_{t-1})$	—	Emission on transition $x_t = g(q_t, q_{t-1})$ $p(x_t q_{t-1}, q_t)$	—	$p(z_t q_t, q_{t-1}, x_t, q_{t-1})$
Moore	$z_t = g(q_t)$	—	Emission on state $x_t = g(q_t) p(x_t q_t)$	—	$p(z_t q_t, x_t)$
Training methodology	Many, often based on heuristics	Relative counts (including smoothing)	EM, Viterbi recurrence	REMAP (GEM)	GEM, EM, GD, Viterbi recurrence
Training criterion	Deterministic	Maximum likelihood	Maximum likelihood	Maximum a posteriori	Maximum a posteriori

Abbreviations: HMM, hidden Markov model; ANN, artificial neural network; IOHMM, input-output HMM; EM, Expectation-Maximization algorithm; GEM, Generalized Expectation-Maximization algorithm; REMAP, recurrent estimation and maximization of a posteriori probabilities

sentative training data. If X represents the whole sequence of training data and M its associated SFSA, the estimation of optimal parameter set Θ^* is usually achieved by optimizing a *maximum likelihood criterion*:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} p(X | M, \Theta) \quad (1)$$

or a *maximum a posteriori* criterion, which could be either

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} p(M | X, \Theta) = \underset{\Theta}{\operatorname{argmax}} p(X | M, \Theta)p(M | \Theta) \quad (2)$$

or

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} p(M, \Theta | X) = \underset{\Theta}{\operatorname{argmax}} p(X | M, \Theta)p(M, \Theta) \quad (3)$$

In the first case, we take into account the prior distribution of the model M ; in the second case, we take into account the prior distribution of the model M as well as the parameters Θ .

Markov Models

The simplest form of SFSA is a Markov model in which states are directly associated with observations (see the second column in Table 1). We are interested in modeling

$$p(X) = p(F | x_1^T)p(x_1 | I) \prod_{t=2}^T p(x_t | x_{t-1}^{t-1}, I)$$

which can be simplified, using the k th-order Markov assumption, by

$$p(X) = p(F | x_{T-k+1}^T)p(x_1 | I) \prod_{t=2}^T p(x_t | x_{t-k}^{t-1})$$

which leads in the simplest case to the first-order Markov model,

$$p(X) = p(F | x_T)p(x_1 | I) \prod_{t=2}^T p(x_t | x_{t-1})$$

where $p(x_1 | I)$ is the initial state probability and the other terms can be seen as transition probabilities. Note that any k th-order Markov model can be expressed as a first-order Markov model, at the cost of possibly exponentially more states. Note also that the transition probabilities are time invariant, i.e., $p(x_t = \ell | x_{t-1} = k)$ is fixed for all t .

The set of parameters, represented by the $(K \times K)$ -transition probability matrix, i.e.,

$$\Theta = p(x_t = \ell | x_{t-1} = k), \text{ for all } \ell, k \in \mathcal{Q}$$

is then directly estimated on a large number of possible observation (and, thus, state) sequences such that

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} p(X | M, \Theta)$$

and simply amounts to estimating the relative counts of observed transitions, possibly smoothed in the case of undersampled training data, i.e.,

$$p(x_t = \ell | x_{t-1} = k) = \frac{n_{k\ell}}{n_k}$$

where $n_{k\ell}$ stands for the number of times a transition from state k to state ℓ was observed, while n_k represents the number of times state k was visited.

It is sometimes desirable to compute the probability of going from the initial state I to the final state F in exactly T steps, which could naively be estimated by summing path likelihoods over all possible paths of length T in model M , i.e.,

$$p(F | I) = p(x_1 | I) \sum_{\text{paths}} p(F | x_T) \prod_{t=2}^T p(x_t | x_{t-1})$$

although there is a possibly exponential number of paths to explore. Fortunately, a more tractable solution exists, using the intermediate variable

$$\alpha_t(\ell) = p(x_t = \ell, x_1^{t-1}, I)$$

which can be computed using the *forward recurrence*:

$$\alpha_t(\ell) = \sum_k \alpha_{t-1}(k)p(x_t = \ell | x_{t-1} = k) \quad (4)$$

and can be used as follows:

$$p(F | I) = \alpha_{T+1}(F)$$

Replacing the sum operator in Equation 4 by the max operator is equivalent to finding the most probable path of length T between I and F .

Although quite simple, Markov models have many uses. For example, they are used in all state-of-the-art continuous speech recognition systems to represent statistical grammars (Jelinek, 1998), usually referred to as N -grams, and for estimating the probability of a sequence of K words

$$p(w_1^K) \approx \prod_{k=N+1}^K p(w_k | w_{k-N}^{k-1})$$

which is equivalent to assuming that possible word sequences can be modeled by a Markov model of order N .

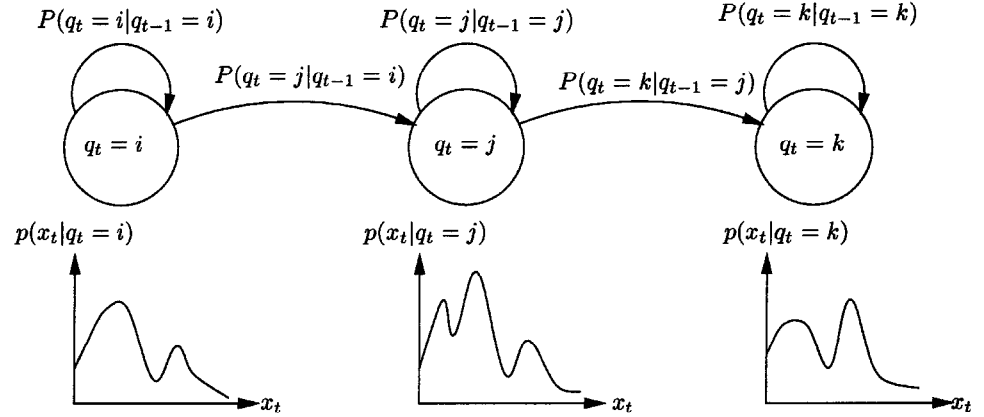
Hidden Markov Models

In many sequential pattern processing or classification problems (such as speech recognition and cursive handwriting recognition), one of the greatest difficulties is to simultaneously model the inherent statistical variations in sequential rates and feature characteristics. In this respect, HMMs have been one of the most successful approaches used so far. As shown in Table 1, an HMM is a particular form of SFSA in which Markov models (modeling the sequential properties of the data) are complemented by a second stochastic process modeling the local properties of the data. The Markov model is called hidden because there is an underlying stochastic process (i.e., the sequence of states) that is not observable but that affects the observed sequence of events.

Although sequential signals, such as speech and handwriting, are nonstationary processes, HMMs assume that the sequence of observation vectors is a *piecewise stationary* process. That is, a sequence $X = x_1^T$ is modeled as a succession of discrete stationary states $Q = \{1, \dots, k, \dots, K\}$, with instantaneous transitions between these states. In this case, an HMM is defined as a stochastic FSA with a particular (generally strictly left-to-right for speech data) topology. An example of a simple HMM is given in Figure 1. In speech recognition, this could be the model of a word or phoneme that is assumed to be composed of three stationary parts. In cursive handwriting recognition, this could be the model of a letter.

Once the topology of the HMM has been defined (usually arbitrarily), the main criterion used for training and decoding is based on the likelihood $p(X | M, \Theta)$, i.e., the probability that the observed vector sequence X was produced by Markov model M . In this case, the HMM is thus considered a *production model*, and the observation vectors x_t are considered to be output variables z_t of the HMM. It can be shown that, provided several assumptions are met (Bourlard and Morgan, 1993), the likelihood $p(X | M, \Theta)$ can be expressed and computed in terms of *transition probabilities* $p(q_t = \ell | q_{t-1} = k, \Theta)$ and *emission probabilities*, which can be of the Mealy type (emission on transitions), $p(x_t | q_t, q_{t-1}, \Theta)$, or

Figure 1. A three-state, left-to-right hidden Markov model.



of the Moore type (emission on states), $p(x_t | q_t, \Theta)$. In the case of multivariate continuous observations, these emission probabilities are estimated by assuming that they follow a particular functional distribution, usually (mixtures of) multivariate Gaussian densities. In this case, the set of parameters Θ comprises all the Gaussian means and variances, mixing coefficients, and transition probabilities. These parameters are then usually trained according to the maximum likelihood criterion (Equation 1), resulting in the efficient EM algorithm (Liporace, 1982; Gold and Morgan, 2000).

Given this formalism, the likelihood of an observation sequence X given the model M can be calculated by extending the forward recurrence (Equation 4) defined for Markov models to also include the emission probabilities. Assuming a Moore automaton (emission on states), we thus have the *forward recurrence*

$$\begin{aligned} \alpha_t(\ell) &= p(x_t^1, q_t = \ell) \\ &= p(x_t | q_t = \ell) \sum_k \alpha_{t-1}(k) p(q_t = \ell | q_{t-1} = k) \end{aligned} \quad (5)$$

which will be applied over all possible t , and where \sum_k is applied over all possible predecessor states of ℓ , thus resulting in

$$p(X | M, \Theta) = \alpha_{T+1}(F)$$

Replacing the sum operator in Equation 5 by the max operator is equivalent to finding the most probable path of length T generating the sequence X , and then yields the well-known dynamic programming recurrence, also referred to as the *Viterbi recurrence* in the case of HMMs:

$$\begin{aligned} \bar{p}(x_t^1, q_t = \ell) &= p(x_t | q_t = \ell) \\ &\times \max_{\{k\}} \{ \bar{p}(x_{t-1}^1, q_{t-1} = k) p(q_t = \ell | q_{t-1} = k) \} \end{aligned} \quad (6)$$

where $\bar{p}(x_t^1, q_t = \ell)$ represents the probability of having produced the partial observation sequence x_t^1 while being in state ℓ at time t and having followed the most probable path, $\{k\}$ represents the set of possible predecessor states of ℓ (given by the topology of the HMM), and $\bar{p}(X | M, \Theta)$ represents the likelihood that the most probable path is obtained at the end of the sequence and is equal to $\bar{p}(x_T^1, F)$.

During training, the HMM parameters Θ are optimized to maximize the likelihood of a set of training utterances given their associated (and known during training) HMM model, according to Equation 1, where $p(X | M, \Theta)$ is computed by taking all possible paths into account (forward recurrence) or only the most probable path (Viterbi recurrence). Powerful iterative training procedures based on the EM algorithm exist for both criteria and have been proved to converge to a local optimum. At each iteration of the EM algorithm, the E step estimates the most probable segmentation or

the best state posterior distribution (referred to as hidden variables) based on the current values of the parameters, while the M step reestimates the optimal value of these parameters assuming that the current estimate of the hidden variables is correct.

For further reading about HMM training and decoding algorithms, see Bourlard and Morgan (1993), Deller, Proakis, and Hansen (1993), Gold and Morgan (2000), and Jelinek (1998).

HMM Advantages and Drawbacks

The most successful application of HMMs is in speech recognition. Given a sequence of acoustic signals, the goal is to produce a sequence of associated phoneme or word transcriptions. To solve such a problem, one usually associates one HMM per different phoneme (or word). During training, a new HMM is created for each training sentence as the concatenation of the corresponding target phoneme models, and its parameters are maximized. Over the last few years, a number of laboratories have demonstrated large-vocabulary (at least 1,000 words), speaker-independent, continuous speech recognition systems based on HMMs.

HMMs can deal efficiently with the temporal aspect of speech (time warping) as well as with frequency distortion. They also benefit from powerful and efficient training and decoding algorithms. For training, only the transcription in terms of the speech units that are trained is necessary, and no explicit segmentation of the training material is required. Also, HMMs can easily be extended to include phonological and syntactical rules (at least when these rules use the same statistical formalism).

However, the assumptions that make the efficiency of these models and their optimization possible limit their generality. As a consequence, they also suffer from several drawbacks, including the following:

- Poor discrimination as a result of the training algorithm, which maximizes likelihoods instead of a posteriori probabilities $p(M|X)$ (i.e., the HMM associated with each speech unit is trained independently of the other models).
- A priori choice of model topology and statistical distributions, e.g., assuming that the probability density functions associated with the HMM state can be described as (mixtures of) multivariate Gaussian densities, each with a diagonal-only covariance matrix (i.e., the possible correlation between the components of the acoustic vectors is disregarded).
- Assumption that the state sequences are first-order Markov chains.
- Assumption that the input observations are not correlated over time. Thus, apart from the HMM topology, the possible temporal

correlation across features associated with the same HMM state is simply disregarded.

In order to overcome some of these problems, many researchers have concentrated on integrating artificial neural networks (ANNs) into the formalism of HMMs. In the next section we discuss some of the most promising approaches.

ANN-Based Stochastic Finite-State Automata

The idea of combining HMMs and ANNs was motivated by the observation that HMMs and ANNs have complementary properties. HMMs are clearly dynamic and very well suited to sequential data, but several assumptions limit their generality, whereas ANNs can approximate any kind of nonlinear discriminant functions, are very flexible, and do not need strong assumptions about the distribution of the input data, but they cannot properly handle time sequences (although recurrent neural networks can indeed handle time, they are known to be difficult to train long-term dependencies, and cannot easily incorporate knowledge in their structure, as is the case for HMMs). Therefore, a number of hybrid models have been proposed in the literature.

Hybrid HMM/ANN Systems

HMMs are based on a strict probabilistic formalism, making them difficult to interface with other modules in a heterogeneous system. However, it has indeed been shown (Richard and Lippmann, 1991; Bourlard and Morgan, 1993) that if each output unit of an ANN (typically a multilayer perceptron) is associated with a state k of the set of states $\mathcal{Q} = \{1, 2, \dots, K\}$ on which the SFSA's are defined, it is possible to train the ANN (e.g., according to the usual least-mean-square or relative entropy criteria) to generate good estimates of a posteriori probabilities of the output classes conditioned on the input. In other words, if $g_k(x_t|\Theta)$ represents the output function observed on the k th ANN output unit when the ANN is presented with the input observation vector x_t , we will have

$$g_k(x_t | \Theta^*) \approx p(q_t = k | x_t) \quad (7)$$

where Θ^* represents the optimal set of ANN parameters.

When using these posterior probabilities (instead of local likelihoods) in SFSA's, the model becomes a recognition model (sometimes referred to as a stochastic finite-state acceptor) where the observation sequence is an *input* to the system and where all local and global measures are based on a posteriori probabilities. It became necessary to revisit the SFSA basis to accommodate this formalism. In Bourlard and Morgan (1993) and Bourlard, Konig, and Morgan (1996), it is shown that $p(M|X, \Theta)$ can be expressed in terms of *conditional transition probabilities* $p(q_t | x_t, q_{t-1})$ and that it is possible to train the optimum ANN parameter set Θ according to the MAP criterion (Equation 2). The resulting training algorithm (Bourlard, Konig, and Morgan, 1994), referred to as REMAP (recursive estimation and maximization of a posteriori probabilities) is a particular form of EM training, directly involving posteriors, in which the M step involves the (gradient-based) training of the ANN and the desired target distribution (required to train the ANN) has been estimated in the previous E step. Since this EM version includes an iterative M step, it is also sometimes referred to as Generalized EM (GEM). As for standard HMMs, there is a full likelihood version (taking all possible paths into account) as well as a Viterbi version of the training procedure.

Another popular solution in using hybrid HMM/ANN as a sequence recognizer is to turn the local posterior probabilities $p(q_t = k | x_t)$ into *scaled likelihoods* by dividing these by the estimated value of the class priors as observed on the training data, i.e.,

$$\frac{p(q_t = k | x_t)}{p(q_t = k)} = \frac{p(x_t | q_t = k)}{p(x_t)} \quad (8)$$

These scaled likelihoods are trained discriminatively (using the discriminant properties of ANN). During decoding, though, the denominator of the resulting scaled likelihoods $p(x_t | q_t = k)/p(x_t)$ is independent of the class and simply appears as a normalization constant. The scaled likelihoods can thus be simply used in a regular Viterbi or forward recurrence to yield an estimator of the global scaled likelihood (Hennebert et al., 1997):

$$\frac{p(X | M, \Theta)}{p(X)} = \sum_{\text{paths}} \prod_{t=1}^T \frac{p(x_t | q_t)}{p(x_t)} p(q_t | q_{t-1}) \quad (9)$$

where the sum extends over all possible paths of length T in model M .

These hybrid HMM/ANN approaches provide more discriminant estimates of the emission probabilities needed for HMMs without requiring strong hypotheses about the statistical distribution of the data. Since this result still holds with modified ANN architectures, the approach has been extended in a number of ways, including the following:

- Extending the input field to accommodate not only the current input vector but also its right and left contexts, leading to HMM systems that take into account the correlation between acoustic vectors (Bourlard and Morgan, 1993).
- Partially recurrent ANNs (Robinson, Hochberg, and Renals, 1996) feeding back previous activation vectors on the hidden or output units, leading to some kind of higher-order HMM.

Input-Output HMMs

Input-output HMMs (IOHMMs) (Bengio and Frasconi, 1995) are an extension of classical HMMs in which the emission and transition probability distributions are conditioned on another sequence, called the *input sequence*, and notated as $Y = y_t^T$. The emitted sequence is now called the *output sequence*, notated $Z = z_t^T$. Hence, in the simplest case of the Moore model (see Table 1), the emission distribution now models $p(z_t | q_t, y_t)$ while the transition distribution models $p(q_t | q_{t-1}, y_t)$.

Although this looks like an apparently simple modification, it has a structural impact on the resulting model and hence on the hypotheses of the problems to solve. For instance, whereas in classical HMMs the emission and transition distributions do not depend on t (we say that HMMs are homogeneous), this is not the case for IOHMMs, which therefore are called inhomogeneous, as the distributions are now conditioned on y_t , which changes with time t .

Applications of IOHMMs range from speech processing to sequence classification tasks and include time-series prediction and robot navigation. For example, for economic time series, the input sequence could represent different economic indicators while the output sequence could be, for example, the future values of some target assets or the evolution of a given portfolio.

In order to train IOHMMs, an EM algorithm has been developed (Bengio and Frasconi, 1995) that looks very much like the classical EM algorithm used for HMMs, except that all distributions and posterior estimates are now conditioned on the input sequence. Hence we need to implement conditional distributions, either for transitions or emissions, which can be represented, for instance, by ANNs. The resulting training algorithm is thus a generalized EM, which is also guaranteed to converge. For transition probabilities, the output of the ANN would represent the posterior probability of each transition $p(q_t | q_{t-1}, y_t)$, with the constraint that all such probabilities from a given state sum to 1. For emission probabilities, the output of the ANN would represent the parameters of an unconditional probability distribution, such as a classical mixture of

Gaussians. Another implementation option would be to use an ANN to represent only the expectation $E[z_i|q_n, y_i]$ instead of the probability itself. For some applications such as prediction, this is often sufficient and more efficient.

An interesting extension of IOHMMs has been proposed in Bengio and Bengio (1996) in order to handle asynchronous input-output sequences, and thus to match input sequences that might be shorter or longer than output sequences. An obvious application of asynchronous IOHMMs is in speech recognition (or handwritten cursive recognition) where the input sequence represents the acoustic signal and the output sequence represents the corresponding phoneme transcription.

Discussion

We have discussed the use of deterministic and stochastic finite-state automata for sequence processing and have sought to present this information in a unified framework. As a particularly powerful instantiation of SFSAs and one of the most popular tools for the processing of complex piecewise stationary sequences, we also discussed HMMs in more detail. Finally, we described a few contributions of the ANN community to improving those SFSAs, mainly the hybrid HMM/ANN systems and IOHMMs. A more extensive discussion would have taken up related areas, such as transducers, linear dynamical systems, Kalman filters, and so on—all outside the scope of this brief introduction.

Road Maps: Learning in Artificial Networks; Linguistics and Speech Processing

Related Reading: Speech Recognition Technology; Temporal Pattern Processing

Hippocampal Rhythm Generation

Péter Érdi and Krisztina Szalisznyó

Introduction

Global brain states in both normal and pathological conditions may be associated with spontaneous rhythmic activities of large populations of neurons. Experimentally, these activities can be detected by recording both from large neural assemblies (as in the EEG) or from a single neuron of the cell population. Two main global hippocampal states that occur normally are known: rhythmic slow activity, called *theta rhythm*, with the associated *gamma oscillation*, and *irregular sharp waves*, with the associated *high-frequency (ripple) oscillation*. A pathological brain state associated with *epileptic seizures* and producing epileptiform patterns also frequently occurs in the hippocampus. This article reviews the basic phenomena and their models. The main hippocampal rhythms are shown in Figure 1.

The general belief is that the different hippocampal rhythms are strongly involved in cognitive functioning. However, the functional significance of these rhythms is mentioned here only briefly (for a fuller discussion, see, e.g., Buzsáki, 1996). The first part of this article describes the most important normal and pathological hippocampal rhythms and the possible underlying mechanisms. In the second part, several different modeling strategies for studying rhythmicity in the hippocampal CA3 region are compared.

References

- Baum, L., and Petrie, T., 1966, Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Statist.*, 37:1554–1563.
- Bengio, S., and Bengio, Y., 1996, An EM algorithm for asynchronous input/output hidden Markov models, in *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, Hong Kong, pp. 328–334.
- Bengio, Y., and Frasconi, P., 1995, An input/output HMM architecture, in *Advances in Neural Information Processing Systems 7*, Cambridge, MA: MIT Press, pp. 427–434.
- Bourlard, H., Konig, Y., and Morgan, N., 1994, *REMAP: Recursive Estimation and Maximization of a Posteriori Probabilities*, Technical Report TR-94-064, Berkeley, CA: International Computer Science Institute.
- Bourlard, H., Konig, Y., and Morgan, N., 1996, A training algorithm for statistical sequence recognition with applications to transition-based speech recognition, *IEEE Signal Process. Lett.*, 3:203–205.
- Bourlard, H., and Morgan, N., 1993, *Connectionist Speech Recognition: A Hybrid Approach*, Boston: Kluwer Academic. ♦
- Deller, J., Proakis, J., and Hansen, J., 1993, *Discrete-Time Processing of Speech Signals*, New York: Macmillan. ♦
- Gold, B., and Morgan, N., 2000, *Speech and Audio Signal Processing*, New York: Wiley. ♦
- Hennebert, J., Ris, C., Bourlard, H., Renals, S., and Morgan, N., 1997, Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems, in *Proceedings of Eurospeech'97*, pp. 1951–1954.
- Hopcroft, J., Motwani, R., and Ullman, J., 2000, *Introduction to Automata Theory, Language and Computations*, 2nd ed., Reading, MA: Addison-Wesley. ♦
- Jelinek, F., 1998, *Statistical Methods for Speech Recognition*, Cambridge, MA: MIT Press. ♦
- Liporace, L., 1982, Maximum likelihood estimation for multivariate observations of markov sources, *IEEE Trans. Inf. Theory*, IT-28:729–734.
- Richard, M., and Lippmann, R., 1991, Neural network classifiers estimate bayesian a posteriori probabilities, *Neural Computat.*, 3:461–483.
- Robinson, T., Hochberg, M., and Renals, S., 1996, The use of recurrent neural networks in continuous speech recognition, in *Automatic Speech and Speaker Recognition*, Boston: Kluwer Academic, pp. 233–258.

Normal Electrical Activity Patterns

Theta Rhythms

The phenomenon. The theta rhythm is a population oscillation with large (1 mV) amplitude and 4–12 Hz frequency. Theta rhythm occurs whenever an animal engages in behaviors such as walking, exploration, or sensory scanning, as well as during REM sleep (Buzsáki, 1996). Single-cell physiological studies showed different relations between the behavior of individual cells and the theta rhythm. Pyramidal cells in the hippocampus proper generally discharge with a very low frequency (0.01–0.5 Hz), although spatially sensitive “place cells” fire at 4–8 Hz when the rat is in its place field, and the position of the animal within a cell’s place field may be correlated with the phase of its firing relative to the theta rhythm.

Origin. Discharging neurons phase locked to hippocampal theta waves have been observed in the dorsal raphe nucleus, the nucleus reticularis pontis oralis and caudalis, the supramammillary region, the septum, and the entorhinal cortex. Some of these areas are reciprocally interconnected with the hippocampal formation.

In vitro experiments. Population oscillation at the theta frequency can be induced pharmacologically by instillation of carbachol into

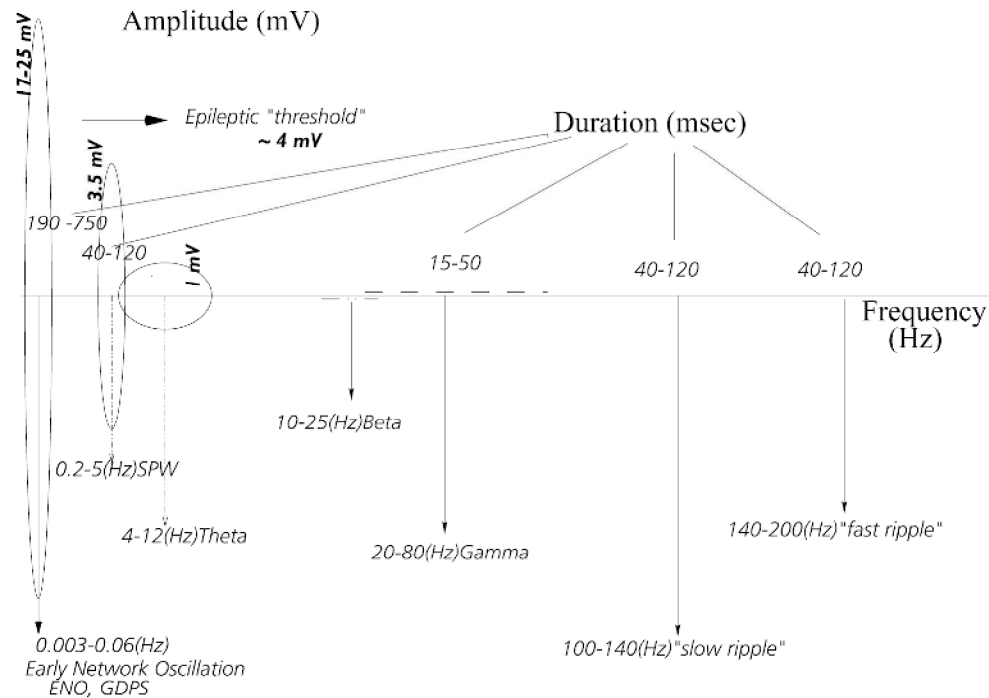


Figure 1. The main hippocampal rhythms occurring during physiological and pathological conditions: their frequency, duration, and amplitude.

hippocampal slices. Based on these and similar findings, it was suggested that theta rhythms may be generated not only extrahippocampally but also by the intrinsic membrane properties of the neurons of the CA3 region. The underlying single-cell firing patterns, however, may be different for in vivo and in vitro carbachol-induced oscillation: the pyramidal cells fire at a much higher frequency than in vivo. Some observations suggest a presence of two, relatively independent theta generators in the hippocampus that are mediated by the entorhinal cortex and the CA3-mossy cell recurrent circuitry, respectively. The CA3-mossy cell theta generator is partially suppressed by the dentate gyrus interneuronal output in the intact brain. Resonant properties of the CA1 neurons were found in the theta frequency range in vitro.

Septohippocampal pathway. The septohippocampal pathway has a cholinergic component, but it is not the only one to contribute to the generation of theta rhythm: atropine, a muscarinic antagonist of acetylcholine, does not entirely abolish the rhythmic slow activity. The GABAergic component of the septal afferents modifies the activity of the principal cells by disinhibition and is also involved in the generation of theta rhythm. These GABAergic cells are located mostly at the border of the medial and lateral septum, and terminate on the GABAergic interneurons of the hippocampus and on the non-GABAergic supramammillary cells, which are known to project to the septal complex and the hippocampus (Figure 2). The cholinergic input into the septum does not show target-selective innervation. There are some backpropagations from the hippocampus to the septum, some of them topographic.

Entorhinal cortex. The timing of the action potentials of pyramidal cells during the theta cycle might be determined by cooperation between the active CA3 neurons and the entorhinal input. Entorhinal excitatory transmitter-containing neurons can also depress the activity of supramammillary theta-generating/regulating cells via septal inhibitory neurons (Figure 2).

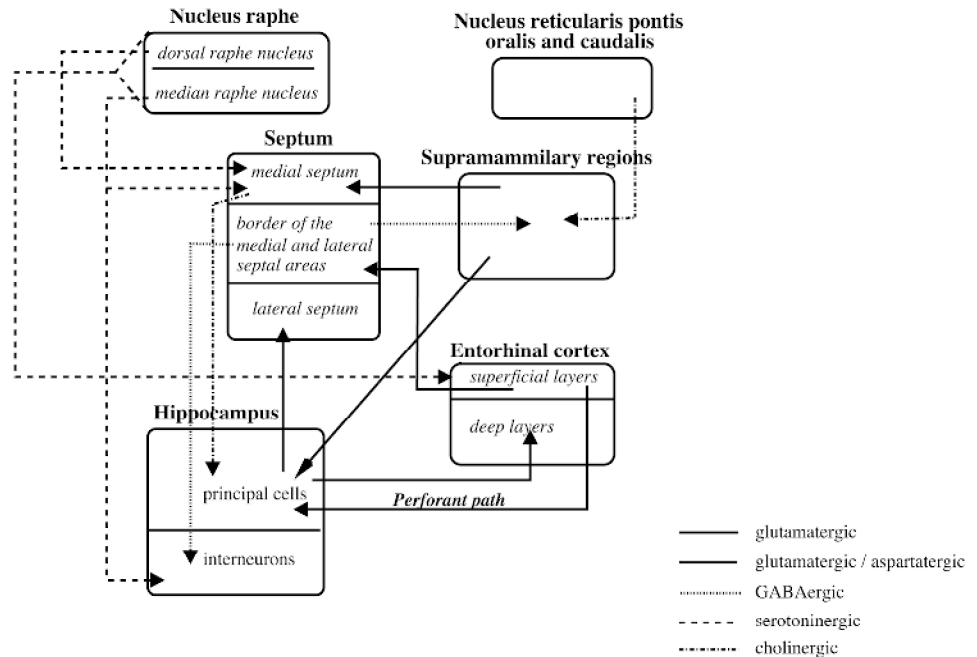
Dorsal and medial raphe nuclei. The main neurotransmitter of the raphe projections selectively innervates a subclass of the interneurons in the CA regions. A spatially segregated population of serotonergic neurons located caudally in the dorsal raphe nucleus was found to project only to the medial septum, and not to the hippocampus. In addition to the well-known serotonergic effect of the median raphe on hippocampal electrical activity, theta rhythm in the hippocampus may also be modulated by the dorsal raphe nucleus via the medial septum (Figure 2).

Nucleus reticularis pontis oralis, caudalis. A high degree of correlation was observed between theta waves in the nucleus reticularis pontis oralis, hippocampal fields, and nucleus reticularis pontis caudalis. The peak frequency of the theta rhythm, induced by stimulation of the reticularis pontis oralis nucleus in urethane-anesthetized rats, was decreased in aged rats compared with young and mature animals. The cholinergic system showed age-related deterioration in rats, including in the hippocampus (Figure 2).

Supramammillary nucleus. Neurons of the supramammillary nucleus fire phase locked to hippocampal theta rhythm. Stimulation of this area induces theta activity in the hippocampus via the medial septum and facilitates perforant pathway stimulation-evoked population spikes in the dentate gyrus even if the medial septum is inactivated. Most if not all postsynaptic targets of the supramammillary projection are principal cells in both the dentate gyrus and the CA2-CA3a subfields (Figure 2).

The functional significance of the theta rhythm, including its possible role in evolution, was recently reviewed by Kahana, Seelig, and Madsen (2001). Studies in rodents have clarified the involvement of theta rhythm in synaptic plasticity and neural coding. Specifically, the phase relationship between theta field activity and single-cell patterns codes the location of the exploring animal. Recently, enhanced theta activity was demonstrated during both verbal and spatial memory tasks.

Figure 2. The interaction of the hippocampus with cortical and subcortical structures. The neurochemical character of the afferent and efferent fibers are visualized.



Gamma Oscillations

The phenomenon. Gamma frequency field oscillations reflect synchronized synaptic potentials in neuronal populations within a range of approximately 10–40 ms. The frequency of this oscillation is in the 20–80 Hz range.

Intrahippocampal origin. The power of gamma oscillation in the hilus decreased significantly after bilateral removal of the entorhinal cortex but survived after surgical removal of the subcortical inputs of the hippocampus. The GABAergic perforant path input terminates exclusively on the interneurons. This pathway provides a rhythmic hyperpolarization of the interneurons, at theta frequency range, so the hypothesized voltage-dependent gamma oscillation of hippocampal interneurons will be periodically interrupted. Gamma oscillation can emerge independently in each subregion of the hippocampus, including the dentate gyrus, CA3, and CA1 regions, but it occurs with greatest power in the dentate gyrus.

Extrahippocampal origin. GABAergic neurons in the basal forebrain display a gamma oscillation. These GABAergic cells also project to the GABAergic reticular nucleus of the thalamus and the neocortex, where they preferentially terminate on the GABAergic interneurons. The gamma patterns are modulated by theta activity, and the dominant source of gamma activity after removal of the entorhinal input is the CA3 region.

Underlying mechanisms. The mechanisms underlying gamma oscillations are not fully understood. Studies of hippocampal formation have suggested that field oscillations in the gamma frequency band reflect synchronous IPSPs on the somata of principal cells. Population oscillation may emerge in interneuronal networks even when individual cells fire at remarkably higher frequencies. At least part of the gamma rhythm recorded in the extracellular space reflects synchronous membrane oscillation in the pyramidal cells brought about by the rhythmic IPSPs. Basket cells innervating the perisomatic region of pyramidal cells discharge at gamma frequency and are phase locked to the field oscillation. These findings indicate that the charges responsible for the rhythmic gamma waves

are mostly carried by chloride ions that enter through GABA_A receptors. Extracellularly observed gamma waves reflect summation of rhythmic EPSPs by the perforant path input at the dendrites (active inward currents) and IPSPs by the oscillating interneurons at the somata (active outward currents) of the granule cell population. Given the spatial segregation of the inward and outward currents (dendrites versus somata), these active currents would summate in the extracellular space. Previous hypotheses suggested that the origin of the synchronized gamma activity may be the mutual excitation among principal cells, or at the single-cell level (“chattering cells”; Traub, Jeffreys, and Whittington, 1999). Simulation studies later showed that interneuronal networks can be synchronized by GABA_A synapses preferentially within gamma frequency range. Resonant properties of pyramidal cells might facilitate network synchrony in the gamma frequency range (Orbán et al., 2001).

The finding that gamma oscillation occurred in different brain regions generated considerable excitement, since gamma oscillation is supposed to be responsible for the binding of perceived and recalled attributes of aspects and events, and for forming memory traces.

Irregular Sharp Waves

The phenomenon. Sharp waves (SPWs) have a very large amplitude (up to 3 mV), their duration is 40–120 ms, and their frequency can be between 0.2 and 5 Hz. Although maximal SPW frequencies do overlap theta frequencies, theta waves are much more regular than SPWs. SPWs also have behavioral correlates: they occur during awake immobility, drinking, eating, face washing, grooming, and slow wave sleep (Buzsáki, 1996).

Origin. During SPWs, pyramidal and inhibitory cells fire with increased frequency. Furthermore, there is a partial synchronous cellular activity of both pyramidal and inhibitory neurons. However, the degree of synchrony is below the threshold for induction of epileptic seizures. The amplitude and frequency of SPWs can be increased by high-frequency stimulation of the commissural system and the Schaffer collaterals, suggesting that such stimulation

enhances the efficacy of the excitatory synapses. The activity of neurons in the deep layers of entorhinal cortex is also correlated with SPWs.

SPWs are thought to be formed by internal intrahippocampal processes. One important precondition for SPW generation is the occurrence of a population burst in a small set of CA3 pyramidal cells. Their synchronization is mediated by excitatory synaptic connections.

The physiological mechanism of synchronization. The largest degree of synchronization occurs in the adult hippocampus during an irregular sharp wave under physiological conditions. CA3 pyramidal cells have recurrent excitatory connections that terminate within the CA3 region. The autoexcitation due to these connections produces large excitatory postsynaptic potentials (EPSP), which propagate to the CA1 region through the Schaffer collaterals. Inhibitory connections control the population activity in both regions. Although sharp waves were found in rat hippocampus during summatory behaviors and slow wave sleep, there is a normal human EEG phenomenon, called small sharp spikes (SSS), that is thought to be analogous to SPW, since it also results from partial synchronous cell firing. Not only “normal” but also epileptiform SPWs can occur. The latter are characterized an amplitude greater than 4 mV or by a less irregular pattern. Their duration is shorter than that of normal SPWs.

During irregular sharp waves, memory traces are supposed to be consolidated and transferred to neocortex.

High-Frequency (“Ripple”) Oscillations

The phenomenon. In conjunction with sharp wave bursts, CA1 pyramidal cells display a high-frequency (200 Hz) network oscillation (ripple) (Ylinen et al., 1995). Similar types of high-frequency oscillation were recorded from the entorhinal cortex and hippocampus of patients with medial temporal lobe epilepsy. The lower-frequency oscillation (80–160 Hz) was regarded as the human equivalent of normal ripples in the rat. The higher-frequency oscillation (250–500 Hz) was found in the epileptogenic regions and may reflect pathological hypersynchronous population spikes of bursting pyramidal neurons. Sleep is characterized by a structured combination of neuronal oscillations. In the hippocampus, slow-wave sleep (SWS) is marked by high-frequency network oscillations, and neocortical SWS activity is organized into low-frequency delta (1–4 Hz) and spindle (7–14 Hz) oscillations. The existence of temporal correlations between hippocampal ripples and cortical spindles is also reflected in the correlated activity of single neurons within these brain structures. This co-activation of hippocampal and thalamocortical pathways may be important for the process of memory consolidation, during which memories are gradually translated from short-term hippocampal to longer-term neocortical stores (Bragin et al., 1999).

Origin and underlying mechanisms. Single pyramidal cells discharge at a low frequency and are phase locked to the negative peak of the locally derived field oscillation. CA1 basket cells increase their firing rate during the network oscillation and discharged at the frequency of the extracellular ripple. These findings indicate that the intracellularly recorded fast oscillatory rhythm is not solely dependent on membrane currents intrinsic to the CA1 pyramidal cells but is a network-driven phenomenon dependent on the participation of inhibitory interneurons. One of the hypotheses was that fast field oscillation (200 Hz) in the CA1 region reflects summed IPSPs in pyramidal cells as a result of a high-frequency barrage of interneurons (Ylinen et al., 1995). The specific currents responsible for the ripple are believed to be synchronized somatic IPSPs interrupted by synchronous discharges of CA1 pyramidal

neurons every 5–6 ms. Concurrent with the hippocampal sharp wave, ripples are present also in the subiculum, parasubiculum, and deep layers of the entorhinal cortex, but the ripple frequency is fastest in the CA1 region. Recent experimental and computational simulation results suggest that ripple oscillation may be mediated by direct electrotonic coupling of neurons, most likely through gap-junctional connections (Traub et al., 1999).

Oscillation in Developing Hippocampus

Synchronous population activity is present both normally and in pathological conditions such as epilepsy. Low-frequency early network population oscillation (0.006–0.03 Hz), or ENO, was found in the hippocampus proper and in the gyrus dentatus region during the first few weeks of postnatal life. The underlying single-cell activity is the synchronous bursting activity, generated by pyramidal cells and interneurons, via GABA_A, NMDA, and AMPA receptors. The oscillation is phase locked to the intracellular Ca²⁺ increase, and the interneuronal network exhibits a Ca²⁺ burst in synchrony with the ENO-associated early pyramidal Ca²⁺ bursts. The ENO is totally blocked by the GABA_A receptor antagonist bicucullin and is reduced by the glutamatergic antagonist. The developmental change of the Cl[−] equilibrium potential is responsible for the change of the GABA_A Cl[−] ion current direction. The ENO-associated GABA_A LTP and LTD are supposed to be partly responsible for the formation of the interneuronal network, and it is consistent with the theory that the excitatory effect-related synaptic plasticity might play a role even in inhibitory synaptic formation.

Epileptic Seizures

The phenomenon. Epileptic activity occurs in a population of neurons when the membrane potentials of the neurons are “abnormally” synchronized. A certain degree of synchrony is necessary for normal theta and SPW behavior. However, there are some fundamental questions to be answered. Under what conditions does population firing become synchronized? What factors regulate the extent of synchronization? What are the critical factors that can influence the change from the physiological to the pathological level of the synchronization? Answers obtained in model studies are briefly summarized in the following sections.

In vitro models. Several in vitro models of seizures have been developed, including models invoking electrical stimulation or low calcium, low magnesium, or elevated potassium levels. (Of course, in vitro models have nothing to do with mathematical models.) The functional removal of some inhibitory interneurons from network activity might be another factor contributing to epileptic phenomena, either because of their inadequate excitatory drive or because of their depolarization blockade. Experiments with ion-selective microelectrodes revealed that a considerable activity of K⁺ ions appears temporarily in the extracellular space during enhanced neuronal activity and is removed from the extracellular space by diffusion. The elevated-potassium model of epilepsy suggests that a modest elevation in extracellular potassium ion concentration produces hypersynchronous epileptiform activity. One important element in epileptogenesis may be attenuation of the inhibitory synaptic inputs to pyramidal cells during high-K⁺ seizures. A few different epileptic phenomena are found in vitro, such as synchronized bursts, synchronized multiple bursts, and seizure-like events. Synchronized bursts last 50–100 ms, and interburst intervals are generally longer than 1 s. They are analogous to the so-called interictal events found in vivo, and can be elicited by applying a localized GABA_A-blocking agent (e.g., picrotoxin, bicucullin, penicillin). Synchronized multiple bursts are characterized by a series of up to about ten synchronized bursts occurring at 65–75 ms in-

tervals. The intervals between the complex events are longer than 10 s. This phenomenon can also be generated by applying GABA_A blocker. Seizure-like events can be generated both with and without the aid of chemical synapses. The latter was demonstrated in slices with low Ca²⁺ solutions. Low Ca²⁺ blocks spike-dependent synaptic transmission, yet spontaneous bursts of population spikes—"field bursts"—with underlying high-frequency firing of pyramidal cells are still evoked.

Computational Models of the CA3 Region: Comparative Studies

1. Compartmental Models

Some detailed single-cell multicompartmental hippocampal pyramidal cell models have been established (Traub et al., 1999, p. 17). Specifically, a 19-compartment pyramidal cable model of a guinea pig CA3 pyramidal neuron was developed and incorporated in several network models. Each compartment contains six active ionic conductances: gNa, gCa, gK(DR) (where DR stands for delayed rectifier), gK(A), gK(AHP), and gK(C). The conductance gCa is of the high-voltage activated type. The model kinetics incorporate voltage-clamp data obtained from isolated hippocampal pyramidal neurons. The model predicts that CA3 pyramidal neurons in media blocking synaptic transmission should fire a burst of action potentials following antidromic stimulation. This was confirmed experimentally in hippocampal slices.

Model reduction. Multicompartment models could be reduced to allow questions about larger-scale brain regions to be answered while still revealing something about activity at the single-cell level. The 19-compartment Traub model has been reduced to a two-compartment model (Pinsky and Rinzel, 1994). This two-compartment model (soma, dendrite) is able to qualitatively reproduce the salient stimulus-response characteristics of the Traub-Miles (1991) single-cell model, was useful for studying the effect of intercompartmental coupling on the responses generated, and proved to be a computationally effective unit for network simulations. The "slow" and "fast" currents are segregated in the dendrite-like and the soma-like compartments, respectively.

2. Network Studies: Model of an in Vitro CA3 Slice

Physiological synchronized population activities. Traub and Miles simulated hippocampal (mostly CA3) population activity by building "bottom-up" models from data on anatomic connectivities, ionic conductances, and synaptic properties (Traub and Miles, 1991).

Three basic cell types, pyramidal cells and two types of inhibitory cells, inh(1) and inh(2) cells, were assumed. The postsynaptic effect of the inh(1) cells is mediated by perisomatic GABA_A receptors, while the inhibition of the inh(2) cells is mediated by the dendritic GABA_B receptors.

Instead of giving detailed wiring, the strategy for specifying synapses was to define the statistical properties of the topology of the neural structures. Traub and Miles defined both globally and locally random networks. The probabilities of synaptic connections for a given type of cell pair are constant in the former case and decrease with distance in the latter case.

Fully and partially synchronized bursts, multiple bursts, synchronized population oscillations, and interictal epileptic seizures were physiologically measured and reproduced by this model (Traub and Miles, 1991; Whittington, Traub, and Jeffreys, 1995).

Epileptic interictal spikes. A computer model was constructed of the guinea pig hippocampal region in vitro that contained 100 py-

ramidal neurons modeled by the 19-compartment model detailed above. This approach has contributed to the understanding of brief (usually less than 100 ms) epileptic events known as interictal spikes. The neurons were randomly interconnected with excitatory synapses, each synapse exerting a fast voltage-independent (AMPA) component and a slower voltage- and ligand-dependent (NMDA) component.

Synchronized gamma oscillations in a hippocampus. 1. *Gamma frequency in the interneuronal network.* Using computer simulations the hypothesis that gamma rhythm can emerge in a random network of interconnected GABAergic fast-spiking interneurons was investigated. The amplitude of spike afterhyperpolarization was above the GABA_A synaptic reversal potential. The ratio between the synaptic decay time constant and the oscillation period was sufficiently large; the effects of heterogeneities were modest because of a steep frequency-current relationship of fast-spiking neurons. It has been demonstrated that large-scale network synchronization requires a critical (minimal) average number of synaptic contacts per cell, and this number is not sensitive to network size. The neuronal firing frequencies could be gradually and monotonically varied by changing the GABA_A synaptic maximal conductance, the synaptic decay time constant, and the mean external excitatory drive to the network, but the network synchronization was found to be high only within a frequency band coinciding with the gamma (20–80 Hz) range. The model predicts that the GABA_A synaptic transmission provides a suitable mechanism for synchronized gamma oscillations in a sparsely connected network of fast-spiking interneurons (Wang and Buzsáki, 1996).

2. *Gamma frequency in the network of interneurons and pyramidal cells.* A network simulation was used to investigate how pyramidal cells, connected to the interneurons and to each other through AMPA-type and/or NMDA-type glutamate receptors, might modify the interneuron network oscillation. With or without AMPA receptor-mediated excitation of the interneurons, the pyramidal cells and interneurons fired in phase during the gamma oscillation. Pyramidal cells caused the interneurons to fire spike doublets or short bursts at gamma frequencies, thereby slowing the population rhythm. Rhythmic synchronized IPSPs allowed the pyramidal cells to encode their mean excitation by their phase of firing relative to the population waves. Recurrent excitation between the pyramidal cells could modify the phase of firing relative to the population waves. This model suggested that pools of synaptically interconnected inhibitory cells are sufficient to produce gamma-frequency rhythms, but the network behavior can be modified by participation of pyramidal cells (Traub, Jeffreys, and Whittington, 1997).

High-frequency oscillation in the hippocampus. To explore the hypothesis that gap junctions occurring between axons could explain high-frequency oscillations, a network was constructed. It has been shown that in randomly connected networks with an average of two gap junctions per cell or less, synchronized network bursts can arise without chemical synapses, with frequencies in the experimentally observed range (spectral peaks 125–182 Hz). The critical assumptions were that (1) there is a background of ectopic axonal spikes that can occur at low frequency (one event per 25 s per axon), and (2) the gap junction resistance is small enough that a spike in one axon can induce a spike in the coupled axon at short latency. The result of the simulation was that axo-axonal gap junctions, in combination with recurrent excitatory synapses, induced the occurrence of high-frequency population spikes superimposed on epileptiform field potentials.

3. Population Models

There is a long tradition of trying to connect the "microscopic" single-cell behavior to the global "macrostate" of the nervous sys-

tem, analogously to the procedures applied in statistical physics. Global brain dynamics is handled by using continuous (neural field) description instead of the networks of discrete nerve cells. Both deterministic and statistical approaches have been developed.

Model framework. In a long series of papers, Ventriglia constructed a neural kinetic theory, i.e., a statistical field theory of large-scale brain activities. He assumed two types of entities, spatially fixed neurons and spatially propagating spikes (“impulses”) (Ventriglia, 1994). In the deterministic field-theoretical description, each neuron is represented as a point in the neural layer or field and a neuronal density is defined, while the statistical model describes neural population activity in terms of the probability density functions (p.d.f.’s) of (1) neurons and (2) spikes traveling between the neurons.

Synchronized population oscillation. A different version of neural kinetic theory has been invoked to simulate a range of epileptiform and nonepileptic rhythms (Barna, Gröbler, and Érdi, 1998; Gröbler, Barna, and Érdi, 1998). The synaptic strengths of excitatory and inhibitory synapses have been varied. The degree of pyramidal cell synchronization has been studied, even as single-cell activity (underlying population behavior) was monitored. Phenomena that were reproduced included fully synchronized population bursts, synchronized synaptic potentials, and low-amplitude population oscillation.

Wave propagation. The spatial pattern of propagation is shown in Figure 3. The model slice is shown with increasing time from top to bottom. High activity appears first in the stimulated subregion, then builds up in the neighboring regions, then propagates through the full length of the slice. The velocity of the simulated wave propagation exhibits a linear increase on the maximal synaptic conductance and is in the interval of 5–10 cm/s.

Discussion

This article reviewed the basic physiological mechanisms and models of normal and pathological hippocampal rhythm generation. The interplay of intrinsic cell properties and synaptic interactions contributes to the generation of rhythms at both single-cell and population levels. Hippocampal rhythms are strongly involved in many areas of cognitive functioning, including navigation (see HIPPOCAMPUS: SPATIAL MODELS), various memory phenomena, such as memory formation, consolidation, and amnesic syndromes.

The hippocampus has an important role in neurological diseases. Alzheimer’s disease, epilepsy, and ischemia are characterized by learning and memory impairment and accompanied by selective neuronal death or characteristic changes in the hippocampal circuitry. How the various hippocampal rhythms are involved in these disorders is an open question. Understanding the electrophysiology of the hippocampal area should contribute greatly to the development of pharmacological strategies to overcome the hippocampus-dependent disorders (see NEUROLOGICAL AND PSYCHIATRIC DISORDERS).

Complementary neural models are used to study the generation and control of hippocampal and other cortical rhythms. Multicompartment modeling techniques proved to be an efficient way to simulate the dynamics of even relatively large networks. Statistical population models are proper tools for describing large-scale population phenomena and wave propagation.

Road Maps: Biological Networks; Mammalian Brain Regions

Related Reading: EEG and MEG Analysis; Event-Related Potentials; Hippocampus: Spatial Models; Oscillatory and Bursting Properties of Neurons; Sleep Oscillations

References

- Barna, G., Gröbler, T., and Érdi, P., 1998, Statistical model of the hippocampal CA3 region: II. The population framework: Model of rhythmic activity in the CA3 slice, *Biol. Cybern.*, 79:309–321. ♦
- Bragin, A., Engel, J., Jr., Wilson, C. L., Fried, I., and Buzsáki, G., 1999, High-frequency oscillations in human brain, *Hippocampus*, 9:137–142.
- Buzsáki, G., 1996, The hippocampo-neocortical dialogue, *Cereb Cortex*, 6:81–92.
- Buzsáki, G., and Chrobak, J., 1995, Temporal structure in spatially organized neuronal ensembles: A role for interneuronal networks, *Curr. Biol.*, 5:504–510. ♦
- Gröbler, T., Barna, G., and Érdi, P., 1998, Statistical model of the hippocampal CA3 region: I. The single-cell module. Bursting model of the pyramidal cell, *Biol. Cybern.*, 79:301–308.
- Kahana, M. J., Seelig, D., and Madsen, R., 2001, Theta returns, *Curr. Opin. Neurobiol.*, 11:739–744. ♦
- Orbán, G., Kiss, T., Lengyel, M., and Érdi, P., 2001, Hippocampal rhythm generation: Gamma-related theta-frequency resonance in CA3 interneurons, *Biol. Cybern.*, 84:123–132.
- Pinsky, P. F., and Rinzel, J., 1994, Intrinsic and network rhythmogenesis in a reduced Traub model for CA3 neurons, *J. Computat. Neurosci.*, 1:39–60.

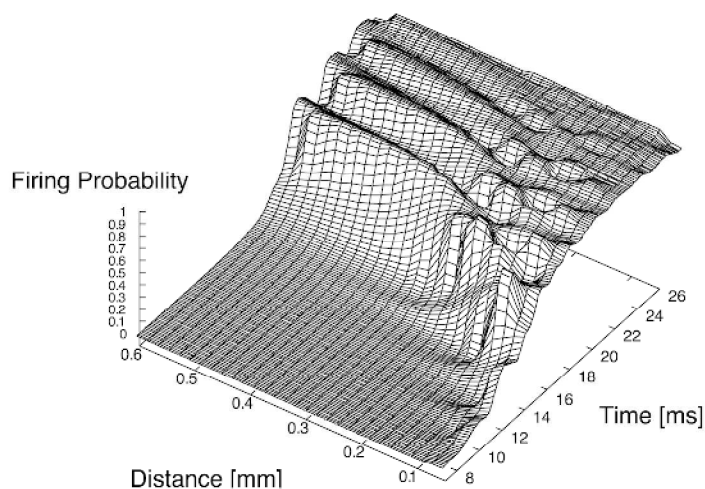


Figure 3. Propagation of activity wave in hippocampal slices. Activity is identified with firing probability.

- Traub, R. D., and Miles, R., 1991, *Neuronal Networks of the Hippocampus*, New York: Cambridge University Press. ♦
- Traub, R. D., Jeffreys, G. R., and Whittington, M. A., 1997, Simulation of gamma rhythms in networks of interneurons and pyramidal cells, *J. Comput. Neurosci.*, 4:141–150
- Traub, R. D., Jeffreys, G. R., and Whittington, M. A., 1999, *Fast Oscillations in Cortical Circuits*, Cambridge, MA: MIT Press. ♦
- Ventriglia, F., 1994, Towards a kinetic theory of cortical-like neural fields, in *Neural Modeling and Neural Networks*, New York: Pergamon Press, pp. 217–249.

- Wang, X. J., and Buzsáki, G., 1996, Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model, *J. Neurosci.*, 16:6402–6413.
- Whittington, M. A., Traub, R. B., and Jeffreys, J. J., 1995, Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation, *Nature*, 370:612–615.
- Ylinen, A., Soltesz, I., Bragin, A., Penttonen, M., Sik, A., and Buzsáki, G., 1995, Sharp wave-associated high-frequency oscillation (200Hz) in the intact hippocampus: Network and intracellular mechanisms, *J. Neurosci.*, 15:30–46.

Hippocampus: Spatial Models

Neil Burgess and John O'Keefe

Introduction

The hippocampus is the most-studied part of the brain, attracting interest because of its position (many synapses removed from sensory transducers or motor effectors), its role in human amnesia and Alzheimer's disease, and the discovery of long-term potentiation (LTP, see HEBBIAN SYNAPTIC PLASTICITY) and of spatially coded cell firing. Bilateral damage to the hippocampus and nearby structures in patient H.M., as treatment for epilepsy, produced a profound retrograde and anterograde amnesia, prompting extensive cross-species research to uncover the specific memory deficits that result from hippocampal damage (the most prominent of which, in the rat, appears to be a deficit in spatial navigation). In short, the hippocampus has become the primary region in the mammalian brain for the study of the synaptic basis of memory and learning. Structurally, it is the simplest form of cortex. It contains one projection cell type whose cell bodies are confined to a single layer and which receives inputs from all sensory systems and association areas (Figure 1).

Attempts to model the hippocampus differ both in the level of anatomical detail and in the functionality that they seek to reproduce. Marr (1971) proposed a theory for how the hippocampus could function as an associative memory, from which have grown many extensions, usually focusing on the role of the CA3 recurrent collaterals (see, e.g., McNaughton & Nadel, 1990; for collected works, see Gluck, 1996, and Burgess, Jeffery, and O'Keefe, 1999; for a review, see Burgess et al., 2001; see also ASSOCIATIVE NETWORKS). However, the precise contribution of the hippocampus to memory, as opposed to the contribution of nearby structures, remains controversial. In this article we specifically consider neuronal models of spatial processing in the rat hippocampus, the domain in which the least controversial experimental data are available. We introduce data on the spatial correlates of hippocampal cell firing and the idea of the hippocampus as a spatial map, and describe some models of the firing of hippocampal place cells and their role in navigation.

Basic Data and Issues

Electrophysiology

Single-unit recordings in freely moving rats have revealed "place cells" (PCs) in fields CA3 and CA1 of the hippocampus, so called because their firing is restricted to small portions of the rat's environment (the corresponding "place fields"). There is little topographic organization of PCs relating their positions in CA3 or CA1 to the positions of their firing fields. The firing properties of PCs

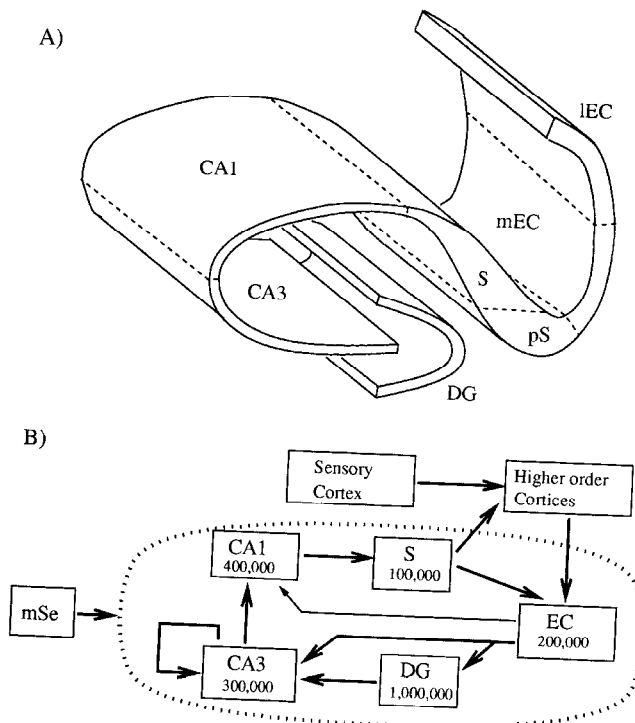


Figure 1. The hippocampus is formed from sheets of cells. *A*, A schematic section cut perpendicular to the longitudinal axis of the hippocampus. EC, entorhinal cortex (mEC, medial EC; IEC, lateral EC); S, subiculum; pS, pre- and parasubiculum; DG, dentate gyrus. *B*, The major projections between subfields (mSe, medial septum), and approximate numbers for the major cell type in each subfield (i.e., pyramidal cells, except for the DG, in which it is granule cells) in the rat. The human hippocampus contains one order of magnitude more cells. In the DG-CA3 projection a single mossy fiber projects from each granule cell, making very large synapses onto only 14 or so pyramidal cells. All the other projections have large divergence and convergence (many thousands to 1) and involve the type of synapse in which "Hebbian" LTP has been observed. A variety of interneurons provide feedforward and feedback inhibition. Cells in the mSe project into DG, CA3, and (less strongly) CA1, playing a role in producing the θ rhythm of the hippocampal EEG. Cells in CA3 and CA1 also project out to the lateral septum via the fornix. (Adapted with permission from B. L. McNaughton, 1989, *Neuronal mechanisms for spatial computation and information storage*, in *Neural Connections, Mental Computation* [L. Nadel, L. A. Cooper, P. Culicover, and R. M. Harnish, Eds.], Cambridge, MA: MIT Press.)

can be manipulated by changing the rat's environment: for example, rotating the major cues in an environment can cause the place fields to rotate. In environments in which direction of movement is restricted (e.g., mazes with narrow arms), PC firing rates appear to depend on the rat's direction of travel as well as its location.

Cells in the entorhinal cortex (the main cortical input to the hippocampus; see Figure 1) also have spatially correlated firing but tend to have larger, less well-defined place fields than those in CA3 or CA1. Cells whose primary behavioral correlate is "head direction" have also been found, in the (dorsal) presubiculum (see Figure 1), anterior thalamus, and mammillary bodies. They fire when the rat points its head in a specific direction relative to the cues in the environment, and independently of its location (see RODENT HEAD DIRECTION SYSTEM and Zhang, 1996, and Sharp, Blair, and Brown, in Gluck, 1996, for models).

The electroencephalogram (EEG) recorded in the hippocampus is the largest electrical signal in the rat brain. One form of the EEG, called the theta (θ) rhythm, is an oscillation of 7–12 Hz. O'Keefe and Nadel (1978) have suggested that in the rat, the θ rhythm coincides with displacement movements. PC firing has been found to have a systematic phase relationship to θ , discovered by O'Keefe and Recce in 1993 (see Burgess and O'Keefe in Gluck, 1996): when a rat on a linear track runs through a place field, the PC tends to fire groups of spikes, with each successive group occurring at an earlier phase of the θ cycle. Consistent with these data, PCs firing at a late phase tend to have place fields centered ahead of the rat, whereas those firing at an early phase tend to have place fields centered behind the rat in open field environments (see Burgess and O'Keefe in Gluck, 1996).

There are two features of PC firing that raise immediate problems for their use in navigation: (1) information about a place in an environment (i.e., the firing of the corresponding PCs) can only be accessed locally (by actually visiting that place), and (2) place fields appear to be no more affected by the location of the goal (which is obviously essential for navigation) than by the location of any other cue. Unfortunately, there are no reports to date of cells that code for the destination of a rat's current trajectory.

Path Integration

An animal may estimate its current position relative to some starting position purely on the basis of internal signals (e.g., vestibular or proprioceptive) relating to its movements in the intervening period. Such a process is often referred to as "path integration" (PI). Many animals appear to be able to use PI to return to a home location in the absence of external stimuli. Interestingly, once a rat has got its orientation from the array of cues, PCs can continue to fire in the correct places after all of the salient cues in an environment have been removed, or after the lights have been switched off (see also the role of PI in the RODENT HEAD DIRECTION SYSTEM). Experiments indicate that PC firing can be supported by visual, auditory, olfactory, tactile, or internal information, as available.

Cognitive Maps

Cognitive maps, or mental representations of the spatial layout of an environment, were first introduced by Tolman to explain place learning in rats, including, for example, their ability to take shortcuts (see COGNITIVE MAPS). An alternative view, suggested by Hull, is that navigation is achieved by following a list of stimulus-response-stimulus steps. O'Keefe and Nadel (1978) proposed that independent neural systems exist in the brain to support a "taxon" system for route navigation and a "locale" system for map-based navigation (for a synopsis, see O'Keefe, 1991). The "map" was taken to be a Euclidean description of the environment in an "allocentric" coordinate system (based on the world and not on some

part of the animal's body). They proposed that the locale navigation system resides in the hippocampus, based on (1) the firing properties of hippocampal PCs, (2) the presence of θ rhythm during displacement movement, (3) deficits in performance of spatial tasks, including the Morris water maze and the Olton eight-arm maze, following hippocampal lesions, and (4) the interpretation of the amnesic syndrome as the loss of episodic memory (memory for specific events set in a spatiotemporal context). Note, however, that the goal independence of PC firing indicates that such cells form only part of a cognitive map, some read-out mechanism being required to guide behavior.

O'Keefe and Nadel (1978) proposed that, while the hippocampal cognitive map was clearly tied to external cues, some form of PI might support its intrinsic distance and direction metric. In their original formulation, a PC could be activated by two independent means. First, PCs could be directly activated by the sensory inputs available to an animal in a particular location. Second, activation of the set of PCs corresponding to one location, coupled with inputs indicating the rat's performance of a movement translating and rotating it by a certain amount, would lead to the activation of the set of PCs corresponding to the new location. Mismatches between the two would provide the signal for exploration, which would bring the two sets of information into correspondence by strengthening some sensory inputs and weakening others (O'Keefe and Nadel, 1978, pp. 220–230).

Models of Place Cell Firing

Sensory Inputs

In this section we describe models of how the spatial firing of PCs develops and is maintained as the rat moves around an environment. Following an earlier mathematical model of PC firing (by Zipser in 1985), Sharp used a simple network with an input layer and two layers of cells governed by "competitive learning" dynamics (see COMPETITIVE LEARNING and Sharp, Blair, and Brown in Gluck, 1996). In this model two types of input (or "neocortical") cells respond to cues placed around the environment: a type 1 cell that "fires" whenever a particular cue is at a given distance from the rat, and a type 2 cell that likewise responds to a particular cue at a given distance, but only if the cue is within a certain range of angles relative to the rat's head. During exploration, competitive learning leads to unsupervised clustering of the input vectors: a PC learns to fire in a portion of the environment in which the inputs (i.e., the distances and angles of cues) are similar. Interestingly, if the simulated rat's exploration is restricted to movements consistent with being in an eight-arm maze, then PC firing tends to be much more strongly correlated with the orientation of the rat (as well as its location) than in the case of unrestricted exploration, which fits well with the experimental data. The place fields in this model are robust to the removal of a subset of the environmental cues. However, the model takes some time to develop realistic place fields, whereas experiments indicate that they are present, and orientation independent, as soon as they can be measured.

Attractors and Path Integration

Touretzky and Redish (in Burgess et al., 1999) proposed a model in which PCs form a coherent representation of location on the basis of estimates of the rat's location from PI and local-view information. The model investigates the interaction of frames of reference supported by different mechanisms and in different brain regions, assuming that resetting the PI system depends on the hippocampus. In a similar approach, Guazzelli, Bota, and Arbib (2001) proposed a feedforward competitive learning model that develops a PC representation by combining PI and sensory inputs, simulating

in some detail the effects of darkness, and the deletion or movement of extramaze cues (see also Arbib in Burgess et al., 1999).

Samsonovich and McNaughton (1997) went much farther in placing the hippocampus at the heart of a PI system. They proposed that region CA3 of the hippocampus forms “continuous attractors” (Zhang, 1996) such that sets of place fields form preconfigured “charts,” as follows. The recurrent connections between PCs fix the relative locations of place fields within a chart, and also ensure stable and coherent patterns of PC activity (i.e., activity consistent with the rat being in a single location). This system serves as the neural basis of a PI system driven by hardwired motion-related signals that shift PC activity so as to reflect the change in location of the rat corresponding to its movements. A particular chart becomes associated to the sensory stimuli in a particular environment so as to make a correspondence between locations on the chart and locations in reality. This model predicts that hippocampal lesions will impair PI, although the experimental evidence for this is controversial.

Kali and Dayan (2000) showed that continuous attractors could be formed by Hebbian learning in the recurrent connections during exploration. However, to create an attractor of equal depth across an unevenly sampled environment required that learning be mediated by novelty. Hasselmo, Wyble, and Wallenstein (in Gluck, 1996) suggest a mechanism for this. The CA3 representation reflects the expected state of the world, being influenced by the associations learned by the recurrent collaterals, while the CA1 representation reflects direct cortical input. Novelty is detected as a mismatch between the two representations, and mediates learning in CA3 by triggering the release of acetylcholine from the medial septum.

What Inputs Support a Place Cell’s Spatially Tuned Firing?

Recording the same PCs in boxes of varying shape and size provides insight into the environment determinants of place fields. In these experiments the location of peak firing tends to maintain a fixed distance from the nearest walls, and a symmetrical, unimodal, place field in a small box may be elongated or multimodal in a larger box. These results are qualitatively fitted by a simple model in which PC firing is a thresholded sum of up to four “boundary vector cell” (BVC) inputs, each tuned to respond maximally whenever there is a wall a given distance away along a given allocentric direction. The tuning to distance is Gaussian, with a width that increases with the distance of the peak response from the wall. (We note that the attractor models discussed earlier can show this behavior only to the extent that the behavior is determined by feed-forward BVC inputs.) A random selection of hardwired BVC inputs is sufficient to model the characteristics of populations of place fields, and choosing BVCs to fit a given cell’s firing in one set of environments enables prediction of its firing in a novel environment (Hartley et al., 2000).

While the model works well for the initial firing of PCs in an environment, a slow experience-dependent divergence (or “remapping”) of the representations of environments of different shape (Lever et al., 2002) indicates an additional role for synaptic plasticity. Because the PC representations in the models of Kali and Dayan (2000) and Gauzzelli et al. (2001) can be incrementally modulated by learning, these models can begin to be used to address the data showing varying degrees of remapping in different experiments, although it is not yet clear what changes will be necessary to provide an accurate model of the dynamics of these data.

Navigation

How could the hippocampus be used to enable navigation? The simplest map-based strategies are based on defining a surface, over

the whole environment, on which gradient ascent leads to the goal (see REINFORCEMENT LEARNING). These strategies tend to have a problem, namely, that to build up the surface, the goal must be reached many times, from different points in the environment. A new surface must be computed if the goal is moved, and multiple goals, as in the eight-arm maze, cannot be handled. Learning in these models seems slower and more goal dependent than in rats, and they are unable to perform “latent learning” (e.g., in rats, exploration in the absence of goals improves subsequent navigation). Interestingly, the performance of these models improves somewhat when a spatially diffuse representation (like place fields) rather than a punctate representation is built up during exploration (see Foster, Morris, and Dayan, 2000). Some recent models that have related navigation to the action of individual cells in the hippocampus are described in the following section.

Using the CA3 Recurrent Collaterals

A role in navigation was proposed for the CA3 recurrent collaterals (the axonal projections by which each CA3 PC contacts approximately 5% of the other CA3 PCs) by Muller and Stead (in Gluck, 1996). Given a model of LTP in which pre- and postsynaptic firing within a short time interval leads to a small increase in synaptic “strength,” the synaptic strength of a connection between two CA3 PCs can come to depend on the proximity of their place fields. This is also the condition for the formation of a continuous attractor representation, discussed earlier. After brief exploration, the synaptic strengths represent distances along the paths taken by the rat. The model proposes that the rat navigates by moving through the place fields of the cells most strongly connected to the cells with fields at the current and destination positions.

Blum and Abbott (1996) propose a related model in which the temporal asymmetry of LTP (synaptic strengthening can occur when presynaptic activity precedes postsynaptic activity; see TEMPORAL DYNAMICS OF BIOLOGICAL SYNAPSES) is invoked to strengthen recurrent connections from one PC to another if they fire in sequence along the rat’s trajectory during exploration. If synaptic modification is also weighted by how soon after the pre- and postsynaptic activity the goal was reached, then the effect of the current collaterals is to shift the location represented by PC firing from the rat’s current location toward the goal. This model proposes that the rat navigates by moving from the current location (e.g., read from CA1) to the shifted location in CA3.

Neither of these models makes clear how a direction of motion is actually generated, or if it would be able to generate a shortcut or detour. To build up a true distance metric in complex environments would take a long time and might best be achieved by reinforcement learning (see Foster et al., 2000). Clear experimental support for learned asymmetric connections between PCs comes from the observation that place fields tend to become elongated backward along a path during the first few times that a rat runs along it. Asymmetric recurrent connections have also been invoked to explain the phase precession effect. In these models the spread of activation from cells with fields early in a learned trajectory to cells with fields farther along the trajectory will, later in the θ cycle, cause a PC to fire before reaching the location on the path where its original (externally driven) place field was located. However, recent evidence indicates that blocking LTP blocks the development of asymmetric place fields but does not prevent the phase precession effect.

Local View Model

In 1989, McNaughton proposed that the hippocampus functions as an associative memory, as follows. As the rat explores, it learns to associate each local view and movement made with the local view

from the place visited as a result of the movement. Thus, routes through an environment are stored as a chain of local view/movement associations (see McNaughton and Nadel, 1990). The model is supported by the fact that, in some situations, PC firing depends on the rat's direction (and therefore its "local view"). Some major problems with this theory are the following: (1) simple route-following strategies appear to be the kind of navigation of which hippocampectomized rats are capable (see O'Keefe and Nadel, 1978); (2) it is difficult to know which particular route will lead to the desired goal: solving this problem leads one back to the reinforcement learning approach; (3) the model is not capable of more sophisticated navigation, such as taking shortcuts; and (4) in open fields, PC firing does not seem to depend on direction. It is also not clear whether this scheme is computationally feasible; see Sharp et al. (in Gluck, 1996) for a discussion of the limitations of simple associative nets in a related situation.

Centroid Model

O'Keefe (1991) proposed a navigational mechanism in which environments are characterized by two parameters: the centroid and the slope of the positions of environmental cues, which can be used as the origin and 0° direction of a polar coordinate framework. The firing of a PC could represent the average position of a small number of cues (their minicentroid), while head direction cells could represent the translation vector between pairs of cues. The environmental centroid and slope are then found by averaging the minicentroids and slopes. It was proposed that single cells could represent two-dimensional vectors as phasors: taking the θ rhythm of the EEG as a clock cycle, the amplitude of firing would code for proximity, and the phase of firing within a clock cycle would code for angle. Thus, summing the output of several neurons results in vector addition, and subtraction is equivalent to a phase inversion followed by addition. Thus, the summed PC activity could provide a vector $\vec{v}(t)$ continually pointing to the centroid of the environment, so that, if the "goal" was encountered at time t_g , storing $\vec{v}(t_g)$ (outside the hippocampus) would enable the translation vector $\vec{v}(t) - \vec{v}(t_g)$ from the rat to the goal to be calculated whenever the rat wanted to go to the goal. The advantages of this system include the ability to perform shortcuts, while the disadvantages include the sensitivity of the slope (i.e., small movements of cues could lead to reversal of the coordinate system).

Place Cell Firing and Navigation

The population vector model (see Burgess and O'Keefe in Gluck, 1996) is implemented at the neuronal level but also generates actual movement trajectories for the simulated rat, and aims to surmount some of the difficulties discussed in the previous section. It assumes that the output stage of the hippocampal system is groups of cells that represent the distance and direction from the rat of previously encountered goal locations as it moves around the environment, and that the input to the hippocampus is a set of "sensory cells" with tuning curve responses to the distance of cues from the rat. Each goal location is represented by the firing rates of a group of "goal cells" as a population vector (i.e., the vector sum of cells' preferred directions weighted by their firing rates; see MOTOR CORTX: CODING AND DECODING OF DIRECTIONAL OPERATIONS), and is used to guide the rat back to a goal location (Figure 2).

The network operates in a feedforward manner. During exploration, a representation of current location is learned in the intermediate layers, entorhinal cells (ECs), PCs, and subicular cells (SCs), which map the sensory input to the population vector output. A type of competitive learning governs the dynamics of the PCs and SCs, similar to that used by Sharp (see Sharp et al. in Gluck, 1996). Latent learning during exploration is expressed as the de-

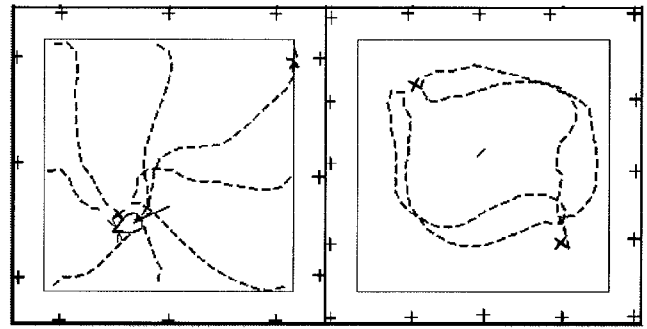


Figure 2. *Left*, Simulated trajectories from eight novel starting positions to a goal encountered after 30 seconds of exploration (at 60 cm/s) in a 135×135 cm² environment; the rat is shown to scale. *Right*, Simulated navigation between two goals with an "obstacle" in between. Cues are marked by +, goals by x, and obstacles by I. (Adapted with permission from Burgess, N., Recce, M., and O'Keefe, J., 1994, *Neural Netw.*, 7:1065–1081.)

velopment of large firing fields in SCs that avoids the locality of information access problem (so that goal cell firing fields cover the whole environment; see later discussion). Upon encountering a goal location, learning by modification of connections to goal cells results in each goal cell having a conical firing rate map whose peak is displaced from the goal position in a particular absolute direction (the "preferred direction" for that cell), creating the appropriate population vector.

The model relies on the phase of firing of SCs (relative to θ) being such that those firing late in a cycle have place fields centered ahead of the rat. This is achieved by making the phase of firing of ECs depend on the angle between the rat's heading direction and the direction of the centroid of the corresponding pair of cues (if the centroid is ahead, the cell fires at a late phase; if behind, it fires early). This property propagates throughout the PC and SC layers. When the rat is at the goal, the goal cell with preferred direction closest to the rat's heading direction receives a strong input, allowing connections to it to be switched on. This signal arrives at a late phase of the θ rhythm, and connections are switched on from those SCs active at that time (which tend to have firing rate maps that peak ahead of the rat). When a goal is encountered, the rat looks around in all directions, so that connections are switched on to goal cells representing each direction.

The direction and proximity (represented by the net firing of the group of goal cells) of interesting objects is the output of the hippocampal "map" and allows the simulated rat to navigate. A small number of obstacles can be avoided during navigation by subtracting the population vector of obstacle cells from the population vector of goal cells. The advantages of this model are that reasonable trajectories, including shortcuts, are performed after one visit to the goal following brief exploration, and its latent learning. The representation of directions is allocentric (e.g., north, south, etc.), and the necessary translation into left-right body turns is assumed to occur in parietal cortex (taking into account the current heading direction and the locations of obstacles) (see also Recce and Harris in Gluck, 1996; Arbib in Burgess et al., 1999; and Burgess et al., 2001).

Brown and Sharp (see Sharp et al. in Gluck, 1996) proposed a similar feedforward model of PC firing and navigation. In their model the output representation (in the nucleus accumbens) is of the egocentric directions (body turns) that lead to the goal. The association from PCs to turn cells is built up over many runs to a goal. As with Blum and Abbott's model, synaptic modification must be weighted by how soon the goal was reached after pre- and

postsynaptic activity. This model would not show latent learning, and navigation would be strongly affected if stereotyped routes were used during learning. Foster et al. (2000) suggest that, in addition to a fast learning mechanism related to that used by Burgess and O'Keefe (in Gluck, 1996), a slower process of reinforcement learning might build up an explicit metric representation of the environment over the course of several trials.

Discussion

We have reviewed several simple neuronal models of how the hippocampus takes in sensory information from environmental cues, turns it into a place cell representation of space, and uses this representation to support a spatial memory for where interesting things are located. Several of these models also consider the role of internal signals and recurrent connections in these processes. Together, these models represent some of the clearest examples of neuronal-level explanations of cognitive behavior. As noted earlier, how different environments are distinguished so that where things are in each environment can be remembered separately remains to be well understood, with data on remapping and the roles of the DG and the CA3 recurrent collaterals in providing an associative memory likely to play a part (see Marr, 1971; McNaughton and Nadel, 1990; Gluck, 1996; Burgess et al., 2001; Lever et al., 2002; and ASSOCIATIVE NETWORKS). A related issue is the problem of navigation in complex environments: how can local maps be patched together to guide behavior over long distances (see the "world graph" model of Arbib in Burgess et al., 1999).

With some progress made in understanding the role of the hippocampus in rat navigation, extending these models to include the role of the hippocampus in monkey and human behavior poses an exciting challenge. Recordings in monkey hippocampus that show neurons responding to the performance of actions in places (T. Ono and colleagues) or to the monkey looking in a particular place (see Rolls in Burgess et al., 1999) give an indication of how hippocampal function might generalize from rats to monkeys. In humans, there is evidence that the right hippocampus plays a role in navigation similar to the role it plays in the rat (see, e.g., Burgess et al., 1999, 2001). However, bilateral hippocampal damage in humans causes a general impairment in episodic memory. One idea relating the spatial and mnemonic roles is that a (nonverbal) episodic memory system could be formed in the right hippocampus by the addition of the human sense of linear time to the rat's spatial system; in the left hippocampus, the inputs have been supplemented by verbal information, producing memory for narratives (O'Keefe and Nadel, 1978). These issues, and the relationship between the hippocampus and the parietal cortex (generally considered the primary locus of spatial processing in primates) are explored further in Burgess et al. (1999). Finally, the need to impose

an egocentric point of view on the allocentric representations in long-term memory provides a starting point for modeling the role of the hippocampus and the head-direction system in episodic retrieval (Burgess et al., 2001; see also Recce and Harris in Gluck, 1996).

Road Maps: Mammalian Brain Regions; Mammalian Motor Control

Related Reading: Cognitive Maps; Hippocampal Rhythm Generation; Rodent Head Direction System; Short-Term Memory

References

- Blum, K. I., and Abbott, L. F., 1996, A model of spatial map formation in the hippocampus of the rat, *Neural Computat.*, 8:85–93.
- Burgess, N., Becker, S., King, J. A., and O'Keefe, J., 2001, Memory for events and their spatial context: Models and experiments, *Philos. Trans. R. Soc. Lond. B*, 356:1493–1503. ♦
- Burgess, N., Jeffery, K. J., and O'Keefe, J., Eds., 1999, *The Hippocampal and Parietal Foundations of Spatial Cognition*, Oxford, Engl.: Oxford University Press. ♦
- Foster, D. J., Morris, R. G., and Dayan, P., 2000, A model of hippocampally dependent navigation, using the temporal difference learning rule, *Hippocampus*, 10:1–16.
- Gluck, M. A., Ed., 1996, *Hippocampus* (Special issue on computational models of the hippocampus), 6:565–762. ♦
- Guazzelli, A., Bota, M., and Arbib, M. A., 2001, Competitive Hebbian learning and the hippocampal place cell system: Modeling the interaction of visual and path integration cues, *Hippocampus*, 11:216–239.
- Hartley, T., Burgess, N., Lever, C., Cacucci, F., and O'Keefe, J., 2000, Modeling place fields in terms of the cortical inputs to the hippocampus, *Hippocampus*, 10:369–379.
- Kali, S., and Dayan, P., 2000, The involvement of recurrent connections in area CA3 in establishing the properties of place fields: A model, *J. Neurosci.*, 20:7463–7477.
- Lever, C., Wills, T., Cacucci F., Burgess, N., and O'Keefe, V., 2002, Long-term plasticity in the hippocampal place cell representation of environmental geometry, *Nature*, 416:90–94
- Marr, D., 1971, Simple memory: A theory for archicortex, *Philos. Trans. R. Soc. Lond. B*, 262:23–81.
- McNaughton, B. L., and Nadel, L., 1990, Hebb-Marr networks and the neurobiological representation of action in space, in *Neuroscience and Connectionist Theory* (M. A. Gluck and D. E. Rumelhart, Eds.), Hillsdale, NJ: Erlbaum, pp. 1–63. ♦
- O'Keefe, J., 1991, The hippocampal cognitive map and navigational strategies, in *Brain and Space* (J. Paillard, Ed.), Oxford, Engl.: Oxford University Press, pp. 273–295.
- O'Keefe, J., and Nadel, L., 1978, *The Hippocampus as a Cognitive Map*, Oxford, Engl.: Clarendon Press. ♦
- Samsonovich, A., and McNaughton, B. L., 1997, Path integration and cognitive mapping in a continuous attractor neural network model, *J. Neurosci.*, 17:5900–5920.
- Zhang, K., 1996, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory, *J. Neurosci.*, 16:2112–2126.

Hybrid Connectionist/Symbolic Systems

Ron Sun

Introduction

There has been a great deal of research in integrating neural and symbolic processes from either cognitive or engineering standpoints. This research has led to the so-called hybrid systems. Hybrid connectionist-symbolic systems constitute a promising approach for developing more robust, more versatile, and more

powerful systems for modeling cognitive processes as well as for engineering practical intelligent systems. The need for such models has been growing steadily. Events such as the 1992 AAAI Workshop on Integrating Neural and Symbolic Processes, the 1995 IJCAI Workshop on Connectionist-Symbolic Integration, and the 1998 NIPS Workshop on Hybrid Neural Symbolic Integration have brought to light many ideas, issues, controversies, and syntheses

in this area. This article provides an overview and a categorization of hybrid systems.

The basic motivation for research on hybrid connectionist-symbolic systems can be summarized as follows:

- Cognitive processes are not homogeneous. A wide variety of representations and processes are likely employed. Some cognitive processes are best captured by symbolic models and others by connectionist models, just as both quantum mechanics and fluid dynamics are needed in order to model physical processes (Dreyfus and Dreyfus, 1987; Smolensky, 1988; Sun, 1994). The need for “pluralism” in cognitive modeling has led to the development of hybrid systems.
- The development of intelligent systems for practical applications would benefit greatly from a proper combination of different techniques, since currently no single technique can do everything successfully. Application domains range from loan approval to process control (Medsker, 1994).

The relative advantages of connectionist and symbolic models have been amply argued for (see, e.g., Feldman and Ballard in Waltz and Feldman, 1986; Smolensky, 1988; Sun, 1994; see also ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS). Once the relative advantages of each model are understood, the computational benefit of the combination of models is relatively easy to justify. Naturally, one would want to take advantage of both types of models, and especially their synergy. Dreyfus and Dreyfus (1987), Sun and Bookman (1994), and Sun and Alexandre (1997) include detailed justifications for using hybrid systems.

For example, Sun and Peterson (1998) presented a model, CLARION, for capturing human skill learning that went from implicit knowledge to explicit knowledge. By integrating symbolic models (for capturing explicit knowledge) and subsymbolic models (for capturing implicit knowledge), CLARION more accurately captured human data and provided a new perspective on human skill learning.

Issues

In developing hybrid connectionist-symbolic systems, we need to ask the following questions:

1. What types of problems are hybrid systems suitable for?
2. What are the relative advantages and disadvantages of each approach to hybridization?
3. How cognitively plausible is each approach?

Other important issues concern the architecture of hybrid systems and learning in these systems. Hybrid models likely involve a variety of different types of processes and representations, in both learning and performance. Therefore, multiple heterogeneous mechanisms interact in complex ways. Architectures, or ways of structuring these different components, thus occupy a prominent place in this area of research. Some architecture-related issues include the following:

- What type of architecture facilitates what type of process?
- Should hybrid architectures be modular or monolithic?
- For modular architectures, should different representations be used in different modules or should the same representation be used throughout?
- How does an investigator decide whether the representation of a particular part of an architecture should be symbolic, localist, or distributed?
- What are the appropriate representational techniques that bridge the heterogeneity of hybrid systems?

- How do representation and learning interact in hybrid systems? (In such systems, both aspects are likely to be more complex.)
- How should the different parts of a hybrid system be structured to achieve optimal results?

A second matter of concern is the increased difficulty of learning. Although purely connectionist models, which are part of any hybrid system, excel in their learning abilities, hybridization makes it more difficult for a system to perform learning. Very generally, the difficulty hybrid systems have with learning comes from the symbolic side, and that difficulty dilutes the advantage that the connectionist parent brings to learning. Some of the learning-related issues include the following:

- What kind of learning can be carried out in each type of architecture?
- How can complex symbolic structures, such as rules, frames, and semantic networks, be learned in hybrid models?
- How can learning algorithms be developed for (usually knowledge-based) structured connectionist networks?
- What is the relationship among symbolic learning methods, knowledge acquisition methods, and connectionist (neural network) learning algorithms in hybrid systems?
- Can each type of architecture itself be developed with various combinations of the methods listed above?

Although many interesting hybrid models have been developed, a broader understanding of these hybrids awaits future work. Any model proposed should be examined for its cognitive plausibility, application potentials, and its strengths and weaknesses. Such an examination can lead to a synthesis of existing divergent approaches and can provide useful insight for further advances in this area. In the following sections we provide a brief categorization of different existing systems.

Architectures and Representations

Various classification schemes of hybrid systems have been proposed (see, e.g., Sun and Alexandre, 1997). For now, we can divide hybrid systems into two broad categories: *single-module* architectures and *multimodule* architectures (Figure 1).

In single-module systems, the *representation* dimension can be of the following forms (Sun and Bookman, 1994): symbolic (as in conventional symbolic models), localist (with one distinct node for representing each concept; see, e.g., Lange and Dyer, 1989; Shastri and Aijanaagadde, 1993; Barnden in Sun and Bookman, 1994), or distributed (with a set of nonexclusive, overlapping nodes for representing each concept; see, e.g., Pollack, 1990; Miikkulainen in Sun and Bookman, 1994; Plate in Sun and Alexandre, 1997; Sperduti et al. in Sun and Alexandre, 1997; see also CONNECTIONIST AND SYMBOLIC REPRESENTATIONS). Usually it is easier to incorporate prior knowledge into localist models, since their structures can be made to directly correspond to the structure of symbolic

- | |
|---|
| <ol style="list-style-type: none"> 1. Single-module <ul style="list-style-type: none"> • Representation: Symbolic, localist, distributed • Mapping: Direct translational, transformational 2. Heterogeneous multimodule <ul style="list-style-type: none"> • Components: Localist + distributed, symbolic + connectionist • Coupling: Loosely coupled, tightly coupled • Granularity: Coarse-grained, fine-grained 3. Homogeneous multimodule <ul style="list-style-type: none"> • Granularity: Coarse-grained, fine-grained |
|---|

Figure 1. Classification of hybrid systems.

knowledge. (For more details on localist models, see STRUCTURED CONNECTIONIST MODELS). On the other hand, connectionist learning usually leads to distributed representation, as in the case of backpropagation. Distributed representation has some unique and useful properties. (For more details on distributed models, see PERCEPTRONS, ADALINES, AND BACKPROPAGATION and COMPOSITIONALITY IN NEURAL SYSTEMS.)

A question that may naturally arise is: Why should we use connectionist models (especially localist ones) for symbol processing, instead of symbolic models? Possible reasons for using connectionist models may include the following. (1) Connectionist models are believed to be a more apt framework for capturing a wide variety of cognitive processes (Waltz and Feldman, 1986; Sun, 1994). (2) Some inherent processing characteristics of connectionist models (such as similarity-based processing) make them more suitable for certain tasks (especially in cognitive modeling of human reasoning and learning). (3) Learning processes may be more easily developed in connectionist models, using, e.g., gradient descent and its various approximations, the Expectation-Maximization algorithm, the Baum-Welch algorithm and so on.

For multimodule systems, we can distinguish between *homogeneous* and *heterogeneous* systems. *Homogeneous* systems are similar to single-module systems except that they contain several replicated copies of the same structure, each of which can be used for processing the same set of inputs, to provide redundancy for various reasons. For example, there may be a set of competing experts for the same domain, each of which may vote for a particular solution. Or, each module can be specialized (with regard to content) for processing a particular type of input. For example, there may be different experts with the same structure but with different content knowledge for dealing with different situations.

Heterogeneous multimodule systems are more interesting. This category is the most hybrid of hybrid systems; CONSYDERR (Sun, 1994), SOAR/ECHO (Johnson et al. in Sun and Alexandre, 1997), and SCREEN (Weber and Wermter in Wermter, Riloff, and Scheler, 1996) belong to this category. A variety of distinctions can be made here. The first distinction has to do with *representations* of constituent modules. In heterogeneous multimodule systems, there can be different combinations of different types of constituent modules. For example, a system can be a combination of localist modules and distributed modules (e.g., CONSYDERR: Sun, 1994), or it can be a combination of symbolic modules and connectionist modules (either localist or distributed; e.g., SCRUFFY: Hendler in Barnden and Pollack, 1991). Some of these combinations can be traced to the ideas of Smolensky (1988), who argued for the dichotomy of conceptual and subconceptual processing, and Dreyfus and Dreyfus (1987), who put forward the distinction between analytical thinking and intuitive thinking.

A second distinction that can be made among heterogeneous multimodule systems has to do with the *coupling* of modules. A set of modules can be either loosely coupled or tightly coupled (Medsker, 1994). In loosely coupled situations, modules communicate with each other, primarily through message passing, shared memory locations, or shared files; an example is SCRUFFY (see Hendler in Barnden and Pollack, 1991). Loose coupling enables some loose forms of cooperation among modules. An example of cooperation is pre- and postprocessing versus main processing: while one or more modules take care of pre- and postprocessing, such as transforming input data or rectifying output data, a main module focuses on the main part of the task. Commonly, pre- and postprocessing are done using a neural network, while the main task is accomplished through the use of symbolic methods. Another form of cooperation is a master-slave relationship: while one module maintains control of the task at hand, it can signal other modules to handle some specific aspects of the task. For example, a symbolic expert system, as part of a rule, may invoke a neural network to

perform a specific classification or decision making. Yet another form of cooperation is the equal partnership of multiple modules. In this form, the modules—the equal partners—may consist of (1) complementary processes, such as in SOAR/ECHO (Johnson et al. in Sun and Alexandre, 1997) or (2) multiple functionally equivalent but structurally and representationally different processes, such as in CLARION (Sun and Peterson, 1998); or (3) they may consist of multiple differentially specialized and heterogeneously represented experts, each of which constitutes an equal partner in accomplishing a task.

In tightly coupled systems, on the other hand, the constituent modules interact through multiple channels or may even have node-to-node connections across two modules, as in CONSYDERR (Sun, 1994), in which each node in one module is connected to a corresponding node in the other module. Various forms of cooperation among modules exist, in ways similar to loosely coupled systems.

Yet another distinction that can be made in multimodule systems has to do with the *granularity* of modules in such systems: they can be either coarse-grained or fine-grained. At one end of the spectrum, a multimodule system can be very coarse-grained, so that it contains only two or three modules (such as the examples cited above). At the other end, a system can be so fine-grained that it can contain numerous modules. In an extremely fine-grained system, each tiny module may contain both a (simple and tiny) symbolic component and a (simple and tiny) connectionist component, as exemplified by Kokinov's model or Stevenson's model (see their chapters in Sun and Alexandre, 1997). The computational advantage of such a microlevel integration is that a vast number of simple "processors" can exist that make up an efficient, massively parallel system.

Learning

Learning, which can include learning content (i.e., knowledge) in a certain architecture or learning and developing an architecture itself, is a fundamental issue, and one that is clearly difficult. Learning is necessary not just because it is a fundamental aspect of cognition, but also because it is indispensable if hybrid systems are ever to be scaled up. Earlier work on hybrid models focused mostly on representational issues (see, e.g., Sun and Bookman, 1994). Such a focus might be justified at an early stage of research, since before we learn complex symbolic representations in hybrid models, we need to understand ways of representing complex symbolic structures in the first place. Over the years, some progress on learning has been made. While some have tried to extend neural learning algorithms (such as backpropagation) to learning complex symbolic representations (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION), others have instead incorporated symbolic learning methods into hybrid models.

Let us look at a few models that incorporated symbolic learning methods. Sun and Peterson (1998) presented the two-module model CLARION for sequential decision tasks. In this model, symbolic knowledge is extracted on-line from a reinforcement learning neural network and in turn is used to speed up neural learning and to facilitate transfer of learned knowledge. The work showed not only the synergy between neural and symbolic learning, but also that symbolic knowledge can be learned autonomously on-line, from subsymbolic knowledge, which is essential for developing autonomous agents.

In a similar vein, Johnson et al. (in Sun and Alexandre, 1997) developed a two-module model SOAR/ECHO for abductive reasoning through a combination of symbolic and connectionist learning (symbolic explanation-based learning in SOAR and neural learning in ECHO).

Furthermore, Thrun (1996) developed a method that integrates explanation-based learning and neural learning in a connectionist framework. In his model there is no separate symbolic and connectionist module (thus the model is a single-module one). The model uses initial domain knowledge (in the form of a trained neural network) and an explanation-based learning process to learn a complete domain theory from the initial knowledge (utilizing “explanations” based on the slopes of activation functions of the initial neural network).

There are a variety of other proposals as well (including rule extraction or insertion algorithms). Future advances in hybrid systems are likely to depend heavily on the development of new learning methods for hybrid systems and on the integration of learning and complex symbolic representations. As mentioned earlier, symbolic representation and reasoning may well emerge from subsymbolic processes through learning, and thus an intimate, synergistic combination of symbolic and subsymbolic learning is desirable and should be further pursued.

Application Domains

To see the breadth of the hybrid systems research, let us look at a brief summary of applications.

Cognitive Science

The development of hybrid systems covers most of the traditional topics in cognitive science, among them the following:

- *Reasoning*, which includes work on hybrid models for logical reasoning, case-based reasoning, and schema-based reasoning. Among existing work, Shastri and Ajjanagadde (1993) took a logic-based approach, while Barnden (in Sun and Bookman, 1994) took a case-based approach. Lange and Dyer (1989) performed reasoning based on schemas (frames). Sun (1994) combined logic-based, case-based, and schema-based approaches. These models went beyond existing nonhybrid models in capturing complex cognitive processes involved.
- *Memory*, which is an area where many models, including hybrid models, are being developed by the psychology community.
- *Classification and categorization*, which involve a variety of hybrid models, conceptual or computational (see CONCEPT LEARNING). These hybrid models often more accurately capture human data than their nonhybrid counterparts.
- *Skill learning*, including, e.g., Sun and Peterson (1998) and Johnson et al. (in Sun and Alexandre, 1997). These models have some important advantages. For example, Sun and Peterson (1998) better explained a type of skill learning than any existing nonhybrid models.
- *Word sense disambiguation*, including work by Lange and Dyer (1989), Bookman (in Sun and Bookman, 1994), and Wermter et al. (1996).
- *Natural language processing* in general, including both syntactic and semantic processing. There are many existing models. (See, e.g., Mikkulainen’s and Dyer’s chapters in Sun and Bookman, 1994, and all relevant chapters in Wermter et al., 1996.)

These hybrid systems use the synergy between connectionist and symbolic processes to more accurately capture human cognitive processes, either quantitatively or qualitatively.

Industrial Applications

Hybrid systems of various sorts have a large variety of practical applications. Because they combine characteristics of connectionist and symbolic models, hybrid systems are often able to perform

better at various tasks. For a summary of early work in this area, see Medsker (1994). See also the application-oriented chapters in Sun and Alexandre (1997), and recent issues of various *IEEE Transactions*.

Discussion

Progress in hybrid systems is occurring steadily, if slowly. Because of the many advantages of hybrid systems, there is reason to expect further significant progress in this area. A number of possibilities exist for architectures, representations, and learning. This abundance suggests exciting possibilities in theoretical advances and in applications.

A distillation of the various extant proposals suggests two different approaches, which can be characterized as (1) incorporating symbolic mechanisms and connectionist models, and (2) stretching connectionist models all the way. In the first approach (see, e.g., Hendler in Barnden and Pollack, 1991; Sun, 1994; Wermter et al., 1996), the representation and learning techniques from both symbolic processing models and neural network models are used to tackle problems that neither type of model handles very well alone. Such problems include modeling cognition, which requires dealing with a variety of cognitive capacities. Several researchers (e.g., Smolensky, 1988; Dreyfus and Dreyfus, 1987) have argued that cognition is better captured with a combination of symbolic and neural components. The second approach, stretching connectionist models to their limit (e.g., Pollack, 1990; Giles and Gori, 1998), is predicated on the assumption that one can perform complex symbolic processing using neural networks alone, with, for example, tensor products, RAAM, or holographic models (see relevant chapters in Sun and Alexandre, 1997). Thus far, both approaches have flourished.

Despite the diversity of approaches, there is clearly an underlying theme, namely, the bringing together of symbolic and connectionist models to achieve the synthesis and synergy of two seemingly different paradigms. The various proposed methods, models, and architectures reflect the common belief that connectionist and symbolic methods can be usefully combined and integrated, and that such integration may lead to significant advances in our understanding of cognition.

Road Map: Artificial Intelligence

Related Reading: Artificial Intelligence and Neural Networks; Compositionality in Neural Systems; Connectionist and Symbolic Representations; Decision Support Systems and Expert Systems; Modular and Hierarchical Learning Systems; Schema Theory; Structured Connectionist Models

References

- Barnden, J. A., and Pollack, J. B., Eds., 1991, *Advances in Connectionist and Neural Computation Theory*, Hillsdale, NJ: Erlbaum.
- Dreyfus, H., and Dreyfus, S., 1987, *Mind Over Machine*, New York: Free Press.
- Giles, L., and Gori, M., 1998, *Adaptive Processing of Sequences and Data Structures*, New York: Springer-Verlag. ♦
- Lange, T., and Dyer, M., 1989, High-level inferencing in a connectionist network, *Connect. Sci.*, 1:181–217.
- Medsker, L., 1994, *Hybrid Neural Networks and Expert Systems*, Boston: Kluwer.
- Pollack, J., 1990, Recursive distributed representation, *Artif. Intell.*, 46:77–106.
- Shastri, L., and Ajjanagadde, V., 1993, From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings, *Behav. Brain Sci.*, 16:417–494.
- Smolensky, P., 1988, On the proper treatment of connectionism, *Behav. Brain Sci.*, 11:1–74.
- Sun, R., 1994, *Integrating Rules and Connectionism for Robust Commonsense Reasoning*, New York: Wiley. ♦

- Sun, R., and Alexandre, F., Eds., 1997, *Connectionist Symbolic Integration*, Hillsdale, NJ: Erlbaum.
- Sun, R., and Bookman, L., Eds., 1994, *Architectures Incorporating Neural and Symbolic Processes*, Boston: Kluwer. ♦
- Sun, R., and Peterson, T., 1998, Autonomous learning of sequential tasks: Experiments and analyses, *IEEE Trans. Neural Netw.*, 9:1217–1234.

- Thrun, S., 1996, *Explanation-Based Neural Network Learning*, Boston: Kluwer.
- Waltz, D., and Feldman, J., Eds., 1986, *Connectionist Models and Their Implications*, Norwood, NJ: Ablex.
- Wermter, S., Rilloff, E., and Scheler, E., Eds., 1996, *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, Berlin: Springer-Verlag.

Identification and Control

Kumpati S. Narendra

Introduction

System characterization and system identification are fundamental problems in systems theory. The problem of *characterization* is concerned with the mathematical representation of a system as an operator S which maps input signals into output signals. The problem of *identification* is to approximate S using an identification model, along with the measured inputs and outputs to the system. In the past five decades, systems theory has made major advances through a combination of mathematics, modeling, computation, and experimentation. At the same time, there has also been an explosive growth in pure and applied research related to neural networks. This article briefly explores how the concepts and methods developed in the two areas are being combined to generate general principles for the identification and control of complex nonlinear dynamical systems.

Systems can be classified as either continuous time (in which all variables of the systems are defined for all values of time $t \in R$) or discrete time (in which they are defined at integer values, i.e., $t = 0, 1, 2, \dots$). A very general method of representing multi-input, multi-output continuous-time and discrete-time dynamical systems is by using vector differential and difference equations, as follows:

$$\begin{array}{ll} \text{Continuous-Time Systems} & \text{Discrete-Time Systems} \\ \dot{x}(t) = f[x(t), u(t)] & x(k+1) = f[x(k), u(k)] \\ y(t) = h[x(t)] & y(k) = h[x(k)] \end{array} \quad (1)$$

where $u(t)(u(k)) \in R^r$ is an input vector, $x(t)(x(k)) \in R^n$ is the state vector, and $y(t)(y(k)) \in R^m$ is an output vector. From Equation 1 it follows that given the state at time $t_0(k_0)$ and the input (input sequence) for $t(k) \geq t_0(k_0)$, the corresponding state and output can be determined. Equation 1 emphasizes the central role played by the state of the system, since, if the state at time t_0 is known, the past history of the system is not relevant, and the output is determined uniquely by the input from time t_0 .

From the very beginning it has been realized by systems theorists that most real dynamical systems are nonlinear. However, linearizations of such systems around equilibrium states yield linear models that are mathematically tractable. For example, the linearization of Equation 1 around an equilibrium state can be described by the equations:

$$\begin{array}{ll} \dot{x}(t) = Ax(t) + Bu(t) & x(k+1) = Ax(k) + Bu(k) \\ y(t) = Cx(t) & y(k) = Cx(k) \end{array} \quad (2)$$

where A , B , and C are constant $n \times n$, $n \times r$, and $m \times n$ matrices, respectively. (In this article, we confine our attention to nonlinear discrete-time dynamical systems of the form given in Equation 1, and to their linearization, given in Equation 2.) The systems given by Equation 2 are said to be time invariant and are completely parameterized by the triple A, B, C . The analytical tractability of the above linear models can be attributed to the fact that the su-

perposition principle applies to them. If any two input sequences $\{u_1(k)\}$ and $\{u_2(k)\}$, when applied to the system at rest, result in output sequences $\{y_1(k)\}$ and $\{y_2(k)\}$, respectively, an input sequence $\{\alpha u_1(k) + \beta u_2(k)\}$ for $\alpha, \beta \in R$ will result in an output sequence $\{\alpha y_1(k) + \beta y_2(k)\}$. Further, in many engineering problems it has been found that most nonlinear systems can be approximated satisfactorily by such linear models in their normal ranges of operation, and this has made the latter attractive in practical contexts as well. It is this combined effect of ease of analysis and practical applicability that accounts for the great success of linear models and has made them the subject of intense study for more than four decades.

The objective of control is to influence the behavior of dynamical system in some desired fashion. The latter includes maintaining the outputs of systems at constant values (regulation) or forcing them to follow prescribed time functions (tracking). Maintaining the altitude of an aircraft or the glucose level in the blood at a constant value are simple examples of regulation; controlling a rocket to follow a given trajectory is a simple example of tracking. For the same reasons given earlier in the context of system representation, the best-developed part of control theory deals with linear time-invariant systems, for which design methods are currently well established.

The demands of a rapidly advancing technology for faster and more accurate controllers have always had a strong influence on the progress of control theory. Applications in new technologies such as robotics, manufacturing, space technology, and medical automation, as well as those in older technologies such as process control and aircraft control, are providing a wealth of new problems in which nonlinearities and uncertainties play a major role and linear approximations are no longer satisfactory. To cope with such problems, research on both identification and control using neural networks has been under way. This article describes why neural networks are attractive in this context, as well as the theoretical assumptions that have to be made when such networks are used as identifiers and controllers.

Artificial Neural Networks

The term artificial neural network (ANN) has come to mean any computing architecture that consists of massively parallel interconnections of simple computing elements. From a systems-theoretic point of view, it can be considered as a conveniently parameterized and easily implementable class of nonlinear maps. In the early 1980s, elaborate ANNs were constructed and empirically demonstrated (using simulation studies) to approximate quite well nearly all functions encountered in practical applications. However, only after numerous authors had demonstrated that such networks are capable of universal approximation in a very precise and satisfac-

tory sense did their study leave its empirical origins to become a mathematical discipline.

Even as these theoretical developments were in progress, empirical investigations continued, and neural networks were used not only for function approximation but also in pattern recognition and optimization problems. In 1990, Narendra and Parthasarathy suggested that feedforward neural networks could also be used as components in feedback systems. The approximation capabilities of such networks could be used in the design of both identifiers and controllers, and their analysis and synthesis could be carried out within a systems-theoretic framework. Following the publication of Narendra and Parthasarathy's paper, there was a frenzy of activity in the area of neural network-based identification and control, and a profusion of methods was suggested for controlling nonlinear systems. Although much of the research was heuristic, it provided empirical evidence that neural networks could outperform traditional methods. However, it soon became evident that more formal methods would be needed to quantitatively assess the scope and limitations of neural network-based control.

MLP and RBFN Networks

The most commonly used network structures for approximating nonlinear maps are the multilayer feedforward networks, also known as the multilayer perceptron (MLP) and the radial basis function network (RBFN). These two classes of neural networks (which are described in greater detail in other articles in this *Handbook*) form the principal building blocks of the dynamical systems considered in the following sections. The n -layer MLP with input u and output y is described by the equation

$$\Gamma[w_n \Gamma[w_{n-1} \cdots \Gamma[w_1 u + b_1] + \cdots + b_{n-1}] + b_n] = y \quad (3)$$

where w_i is a weight matrix in the i th layer and the vectors b_i represent the threshold value for each node in the i th layer. $\Gamma[\cdot]$ is a nonlinear operator with $\Gamma(x) = [\gamma_1(x), \gamma_2(x), \dots, \gamma_n(x)]$, where $\gamma_i(\cdot)$ is a smooth activation function (generally a sigmoid). An alternative to the MLP is the RBFN, which can be considered a two-layer network in which the hidden layer performs a nonlinear transformation on the inputs (see RADIAL BASIS FUNCTION NETWORKS). The output layer then combines the outputs of the first layer linearly, so that the output is described by the equation

$$y = f(u) = \sum_{i=1}^N w_i R_i(u) + w_0 \quad (4)$$

The functions R_i are termed radial basis functions, and typically these are Gaussian functions.

In both networks, the parameters (weights) are adjusted to decrease the error between the output of the function to be approximated and that of the network, along the negative gradient. Numerous algorithms have been proposed to improve the convergence properties of MLP networks, but invariably all of them are modifications of the gradient approach. When RBFNs are used to approximate an unknown function based only on inputs and outputs, the convergence is substantially faster, since the output error is linear in the parameters.

A question that naturally arises is why neural networks should be preferred to other methods for approximating nonlinear functions in identifiers and controllers. The principal reason, already referred to, is the universal approximating capability of such networks. Empirical evidence is also available that they are fault tolerant and robust in the presence of noise. The fact that they are implementable in hardware makes them attractive in practical applications. However, a convincing theoretical argument was provided by Barron (1993), who showed that the effectiveness of

MLPs (in which the output depends nonlinearly on the weights) for approximating a general class of nonlinear functions increases with the dimension of the input space, as compared to networks in which the output depends linearly on the weights. Since the dimension of the input vector used for identification and control in complex dynamical systems is generally quite high, the advantages assured by Barron are very attractive. It must be added here that other authors have taken issue with Barron on this matter by suggesting that RBFNs can enjoy the same advantages if the characteristics of the nonlinear activation functions used in them are also varied.

Identification

If an ANN is to be used for identifying a nonlinear dynamical system, the existence of a dynamic nonlinear map relating the input and output spaces must first be established. In many practical applications, identification has to be carried out using only observed input-output data. In such cases, it is tacitly assumed that there is an underlying nonlinear dynamic map relating the two. An example of such a map is given by the state equation given in Equation 1. The objective would then be to identify the maps f and h . If measurements on the state vector $x(k)$ of the unknown system (generally referred to as the *plant*) are available, a model of the nonlinear system can be set up as

$$\begin{aligned} \hat{x}(k+1) &= N_f[x(k), u(k)] \\ \hat{y}(k) &= N_h[x(k)] \end{aligned} \quad (5)$$

where N_f and N_h are neural networks approximating f and h , respectively. Since $x(k)$ and $u(k)$ are accessible, standard approximation methods may be used to adjust the parameters of N_f and N_h based on the errors $\tilde{x}(k) = \hat{x}(k) - x(k)$ and $\tilde{y}(k) = \hat{y}(k) - y(k)$, respectively. It must be noted that the state/output of the system $x(k)/y(k)$, rather than $\hat{x}(k)/\hat{y}(k)$ of the network, is used in the right-hand side of the models given in Equation 5. This is generally referred to as a series-parallel model. This makes the identification procedure substantially simpler and hence practically attractive. However, a more realistic model would have the form:

$$\begin{aligned} \hat{x}(k+1) &= N_f[\hat{x}(k), u(k)] \\ \hat{y}(k) &= N_h[\hat{x}(k)] \end{aligned} \quad (6)$$

Since the estimated state of the system is used to determine $\hat{y}(k)$, we have a feedback system in this case. This implies that the model can be unstable. This model is also referred to as a recurrent network. Determining the gradient of the performance index with respect to the adjustable parameters of the network is no longer simple. Static backpropagation, used in Equation 5, has to be replaced by dynamic backpropagation, which is computationally intensive. A network described by Equation 6 is referred to as a recurrent network, and the stability properties of such networks are not well understood. Hence, the series-parallel networks described by Equation 5 are preferred in practice.

The Nonlinear Autoregressive Moving Average (NARMA) Model

The identification problem using a state representation becomes substantially more complex when the state variables of the plant are unknown, and identification has to be carried out using only input-output data. It is well known that even in the linear case, a unique parameterization of the plant no longer exists. The success of nonlinear identification techniques therefore strongly depends on the specific parameterizations used.

For a single-input single-output (SISO) linear time-invariant system described by the state Equations 2, it can be shown that, if the

system is observable, the input and output are related by the equation

$$y(k+1) = \sum_{i=0}^{n-1} \alpha_i y(k-i) + \sum_{j=0}^{n-1} \beta_j u(k-j) \quad (7)$$

where α_i and β_j ($i, j = 0, 1, \dots, n-1$) are the parameters of the system. Equation 7 is known as the autoregressive moving average (ARMA) representation of the given plant and expresses the output at any instant as a linear combination of the past n values of the input and the n values of the output. The ARMA representation has found extensive application in linear systems theory.

Motivated by the ARMA model in the linear case, efforts were made in the early 1990s to suggest models for nonlinear system identification. If the linearized system around the equilibrium state of the system in Equation 1 is observable, it was shown (Levin and Narendra, 1993), using the implicit function theorem, that the state $x(k)$ and hence the output $y(k)$ can be determined using n past values of the input and output, as follows:

$$x(k) = G[y(k), y(k-1), \dots, y(k-n+1), u(k), \dots, u(k-n+1)] \quad (8)$$

and

$$y(k) = F[y(k), y(k-1), \dots, y(k-n+1), u(k), \dots, u(k-n+1)] \quad (9)$$

where G and F are functions mapping $R^m \times R^{2n}$ into R^n and R , respectively. Equation 9 is an exact mathematical representation of the given system (from Equation 1) in a neighborhood of the equilibrium state. It is referred to as the NARMA model and provides a rigorous mathematical basis for the synthesis of identification models using input-output data.

As mentioned in the Introduction to this article, once the existence of a nonlinear dynamic map from the input space to the output space is established, a neural network can be used to approximate the map using available data. In the present case, identification reduces to the approximation of the function F in Equation 9. A neural network model of the system is then given by

$$\hat{y}(k+1) = N_F[y(k), y(k-1), \dots, y(k-n+1), u(k), \dots, u(k-n+1)] \quad (10)$$

where N_F is a neural network whose $2n$ inputs are the past n values of the input and output, respectively. The parameters of the network are adjusted to minimize the error $e_i(k) = \hat{y}(k) - y(k)$, between the output of the network and the output of the given plant. Equation 10, approximating the NARMA model, has been used extensively for the practical identification of nonlinear systems using neural networks.

Comments

The NARMA representation given in Equation 9 and the neural network approximation given in Equation 10 raise numerous theoretical and practical questions.

From a theoretical point of view, one is interested in relating the number of past values of the output and the input needed with the dimension of the state space n of the system (in Equation 9 this is chosen to be n). Also, Equations 9 and 10 describe the evolution of two dynamical systems, and one must establish in what sense the proximity of F and N implies the proximity of the trajectories of the two systems.

From a practical point of view, the choice of the number of nodes and the number of layers in the network to approximate F is important. However, at the present time, this is very much of an art, and the values are chosen by trial and error. There are also numerous gradient-based methods for the adjustment of parameters, but

the performance criterion varies from design to design. A criterion function that has proved effective in applications is one in which the output errors as well as the incremental inputs are weighted over a finite interval of time (i.e., $\sum_k [e_i^2(k) + (u(k) - u(k-1))^2]$).

Recently, efforts have been made to approximate models of Equation 10 in such a manner that the input $u(k)$ appears linearly. They have the following form:

$$\begin{aligned} \hat{y}(k+1) = & F_0[y(k), y(k-1), \dots, y(k-n+1)] \\ & + \sum_{j=0}^{n-1} F_j[y(k), \dots, y(k-n+1)]u(k-j) \end{aligned} \quad (11)$$

The need for such models is discussed in the following sections in the context of control.

Control

Two distinct classes of problems that are encountered with increasing frequency in industry deserve special attention from the point of view of both the control theorist and the practicing engineer. The first is encountered in the context of systems that are already in existence and that were designed satisfactorily in a small neighborhood of the equilibrium in the state space. In this domain, linear models are used to describe the systems, and linear controllers based on them are found to perform satisfactorily. However, because of the demands of technology, the systems are required to operate in larger regions in the state space, where their characteristics are distinctly nonlinear. Both the identification model and the linear controller are then found to be inadequate to achieve the desired level of performance.

In the second class of problems, mathematical models of the plant (or process) to be controlled cannot be developed from first principles, but the process is known to be distinctly nonlinear. A finite amount of stored input-output data of the plant is available from which an adequate model of the process and a suitable controller are to be determined to satisfy stringent performance criteria.

For both classes of problems stated above, numerous decisions have to be made concerning the representation to be used, the prior information to be obtained to ensure the existence of a controller, the architecture of the identifier and controller, and the algorithms to be used in training their parameters. In the next sections we address some of the theoretical and practical questions that arise in neurocontrol.

Regulation and Tracking

In the Introduction we noted that regulation and tracking are two control problems of general interest. Regulation involves the generalizations of a control input that stabilize the system around an equilibrium state. In the tracking problem, a reference output $y^*(k)$ is specified and the output $y(k)$ of the plant is to approximate it in some sense, e.g., $\limsup_k \|y(k) - y^*(k)\| \leq \varepsilon$. For theoretical analysis, ε is assumed to be zero, so that asymptotic tracking is achieved.

In the following, we assume that linearization of the system around the equilibrium state is both controllable and observable. Controllability implies that the state of the system can be transferred to any other state by the application of a suitable control input. Observability implies that the state of the system can be determined by observing the output over a finite interval of time. If the linearized system is controllable and observable, it follows that the nonlinear system is also locally controllable and observable. Almost all control systems in operation that were designed using linear control principles satisfy the above conditions. This conclusion,

in turn, implies that such systems can be controlled as described in the next section, even in regions where their dynamics are nonlinear.

The Identification and Control Procedure

The process of identifying and controlling a plant using neural networks can be summarized as a three-step procedure. In the first step, a neural network is used to approximate the behavior of the given system. In the second step, after a sufficiently accurate model has been obtained, a neural network is designed to control the model (rather than the plant) to achieve the desired performance. In the third step, the same controller is used to control the system. If the performance does not meet specifications, the common procedure is to increase the size(s) of the neural networks used for identification and control, and repeat the entire procedure. If both identification and control are carried out concurrently, we have an adaptive system.

We first consider the dynamical system described by the state Equation 1, in which the entire state vector $x(k)$ is accessible, and attempt to regulate it around the origin. If the linearized system around the origin is controllable, it has been shown (Levin and Narendra, 1993) that $u = g(x)$ (i.e., nonlinear state feedback exists) such that the equilibrium state is stable. Hence, in principle, a neural network can be used to approximate $g(x)$. If $u(k) = N_c[x(k)]$, the overall feedback system is described by the equation

$$x(k+1) = f[x(k), N_c[x(k)]] \quad (12)$$

The parameters of the neural network are then adjusted using dynamic back propagation. In Levin and Narendra (1993), different methods are described for accomplishing this, based on the measured state $x(k)$. Simulation results indicate that this method is far superior to that obtained using linear theory.

Tracking

The problem of tracking a reference signal when the dynamics of the plant are unknown poses a real challenge to the control engineer. In this case, the principal question that arises is whether an input $u(\cdot)$ exists so that the output of the plant can asymptotically follow the reference input. Assuming that such an input exists, the next problem is to determine a controller structure as well as the inputs to the controller so that the output of the controller is the desired input to the plant.

Various authors have investigated the application of neural networks for tracking. In the pioneering work of Widrow (e.g., Widrow and Wallach, 1996), the nonlinear plant itself is used in place of the identification model while training the controller. In the work of Jordan (1990) and Kawato (1990), inverse modeling using neural networks is used for solving inverse kinematics and inverse dynamics problems. Qin, Su, and McAvoy (1992) have used a similar methodology for process control. More recently, Cabrera and Narendra (1999) studied in detail the conditions that a plant must satisfy for a bounded control to exist in order to achieve asymptotic tracking of any specified bounded signal in the region of interest. In the following paragraphs, we summarize the results contained in the latter that are relevant to the present discussion.

The relative degree and zero dynamics of a nonlinear discrete-time system are important concepts in attempting to determine the control input to a dynamical system to track a desired output. The definitions of relative degree and zero dynamics have been discussed by numerous authors, but the analytical issues involved are for the expert. In qualitative terms, a relative degree d implies that for any arbitrary initial condition in a neighborhood of the origin, the effect of a control input $u(k)$ is felt only at time $k+d$. Hence, for the purposes of our discussion here, the *relative degree* of the

system can be defined as the delay of the system. The *zero dynamics* of the system describe the behavior of the system when the input and the initial conditions are jointly chosen in such a way that the output is identically zero. One of the principal results given by Cabrera and Narendra (1999) is that if the relative degree of a dynamic system is well defined and its zero dynamics are asymptotically stable, the asymptotic tracking problem can be solved using an analytic controller of the form

$$u(k) = \gamma[y(k), y(k-1), \dots, y(k-n+1), u(k-1), \dots, u(k-n+1), y^*(k), y^*(k+1) \dots y^*(k+d)] \quad (13)$$

where $y^*(k)$ is the desired output at time k , and d is the relative degree of the system. In some cases only the reference input $y^*(k+d)$ may be needed at time k .

The above results have great practical significance, since the existence of a map γ implies that a neural network N_γ can be used to approximate it.

Practical Considerations

Numerous difficulties are encountered when a neural network is used to control a dynamical system. These can be briefly listed as follows:

1. Since the neural network is in a feedback loop with the controlled plant, dynamic rather than static backpropagation is needed to adjust the parameters along the negative gradient. However, in practice, only static backpropagation is used.
2. From both theoretical and practical viewpoints, the best approach in the author's experience is to use the approximate model of Equation 11, in which the control input $u(k)$ can be computed algebraically. To improve the performance further, the parameters of an additional neural network may be adjusted, as in point 1.
3. Because of the complexity of the structure of an MLP neural network and the nonlinear dependence of its map on its parameter values, stability analysis of the resulting system is always very difficult and quite often intractable. However, many interesting results have been obtained by Sanner and Slotine (1992), Jagannathan and Lewis (1996), Polycarpou (1996), and Chen and Khalil (1995), most of which are applicable in specific contexts.
4. In practice, the three-step procedure described earlier is used in industrial problems, to avoid adaptation on-line and the ensuing stability problems. However, if the feedback system is stable, on-line adaptive adjustments can improve performance significantly, provided such adjustments are small.

Conclusions

Numerous control methods have been suggested by different authors over the past decade for the practical control of dynamical systems using neural networks. These methods include supervised control, inverse control, internal model control, and model reference control. In many cases, the authors have attempted to realize identifiers and controllers as static maps using neural networks both to improve convergence and to simplify analysis. In many practical applications, such methods have performed satisfactorily. However, to our knowledge, the identification and control models described in this article are the only ones that have been rigorously derived using theoretical results from nonlinear and adaptive control.

Most of the current interest in the application of neural networks is in static systems, particularly in pattern recognition, where they have been very successful. The need for the control of nonlinear dynamical systems is, however, increasing in new technologies in

which nonlinearities, uncertainties, and complexity play a major role. Neural networks are particularly attractive in such situations. However, the results obtained in nonlinear control theory, the concepts and structures provided by linear adaptive control, and the approximating capabilities of neural networks have to be judiciously combined to deal with the control problems that arise in complex dynamical systems (Chen and Narendra, 2001; Jagannathan, 2001).

The presence of a feedback loop in a system implies that stability issues have to be addressed in their design. At present, the neuro-control problems being attempted in industry are those in which improving performance rather than ensuring stability is the main consideration. However, as faster response is required in industrial applications, stability questions are bound to become more important.

Practical systems design is driven by both cost and operational requirements. Anyone familiar with industrial problems is only too aware that bridging the gap between theoretical principles, on the one hand, and development, testing, and implementation on the other is a slow process. On the basis of theoretical advances made thus far, and the great successes that have already been reported in some cases, there is every reason to believe that neural network-based control systems will be developed in many industries in the next decade.

Road Map: Robotics and Control Theory

Background: Perceptrons, Adalines, and Backpropagation

Related Reading: Motor Control, Biological and Theoretical; Sensorimotor Learning

References

- Barron, A. R., 1993, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory*, 39:930–945.
- Cabrera, J. B. D., and Narendra, K. S., 1999, Issues in the application of neural networks for tracking based on inverse control, *IEEE Trans. Autom. Control*, 44:2007–2027. ♦
- Chen, F.-C., and Khalil, H. K., 1995, Adaptive control of a class of nonlinear discrete-time systems using neural networks, *IEEE Trans. Autom. Control*, 40:791–801.
- Chen, L. J., and Narendra, K. S., 2001, Nonlinear adaptive control using neural networks and multiple models, *Automatica*, 37:1245–1255.
- Jagannathan, S., 2001, Control of a class of nonlinear discrete-time systems using multilayer neural networks, *IEEE Trans. Neural Netw.*, 12:1113–1120.
- Jagannathan, S., and Lewis, F. L., 1996, Multilayer discrete-time neural-net controller with guaranteed performance, *IEEE Trans. Neural Netw.*, 7:107–130.
- Jordan, M. I., 1990, Learning inverse mappings using forward models, in *Proceedings of the 6th Yale Workshop on Adaptive Learning Systems*, pp. 146–151.
- Kawato, M., 1990, Computational schemes and neural networks models for formation and control of multijoint arm trajectory, in *Neural Networks for Control* (W. T. Miller III, R. S. Sutton, and P. J. Werbos, Eds.), Cambridge, MA: MIT Press, pp. 197–228.
- Levin, A. U., and Narendra, K. S., 1993, Control of nonlinear dynamical systems using neural networks: Controllability and stabilization, *IEEE Trans. Neural Netw.*, 4:192–206. ♦
- Levin, A. U., and Narendra, K. S., 1996, Control of nonlinear dynamical systems using neural networks: Part II. Observability, identification and control, *IEEE Trans. Neural Netw.*, 7:30–42.
- Narendra, K. S., and Parthasarathy, K., 1990, Identification and control of dynamical systems using neural networks, *IEEE Trans. Neural Netw.*, 1:4–27.
- Polycarpou, M. M., 1996, Stable adaptive neural control scheme for nonlinear systems, *IEEE Trans. Autom. Control*, 41:447–451.
- Qin, S., Su, H., and McAvoy, T. J., 1992, Comparison of four neural net learning methods for dynamic system identification, *IEEE Trans. Neural Netw.*, 3:122–130.
- Sanner, R. M., and Slotine, J.-J. E., 1992, Gaussian networks for direct adaptive control, *IEEE Trans. Neural Netw.*, 3:837–863.
- Widrow, B., and Walach, E., 1996, *Adaptive Inverse Control*, Englewood Cliffs, NJ: Prentice-Hall. ♦

Imaging the Grammatical Brain

Yosef Grodzinsky

Introduction

What do students of language do? Linguists characterize linguistic knowledge; psycholinguists model the algorithms that implement this knowledge in speaking and understanding; and neurolinguists are interested in the neural mechanisms that realize these algorithms. One can imagine a research program in which these perspectives cohere, attempting to understand knowledge of language, its acquisition, processing mechanisms, and neural computation. This is the neurobiological project that attempts to characterize human language. This chapter describes attempts to reconstruct an image of language mechanisms through the analysis of lesion data and functional neuroimaging. I argue that a correct choice of the unit for functional analysis of behavior leads to a clearer image of the linguistic brain.

Innovative technologies have recently made the goal of a coherent, focused picture of the neural basis of language closer than ever before. Linguistic theory provides a sophisticated technology for the analysis of the linguistic signal; instruments that measure neural activity have become less invasive, with high resolution in both time and space; experimental ingenuity may lead to new solutions to old (and new) problems. This research enterprise must thus define brain/language relations in the form of an equation, both sides of which contain complex terms: On the one side there is linguistic behavior, described in the best theoretical vocabulary one can find,

and on the other side there are brain mechanisms, interpreted by whatever techniques neuroscience can offer. The relation between the two sides is also extremely complex, and it is here that disagreements arise. Some neurolinguists study words, some study sentences, and others investigate not linguistic units, but activities, such as speaking, listening, reading, and writing. It is quite difficult to find a unit of analysis on which a consensus (one that would hopefully reflect understanding) exists. In this respect, the study of language is unique. Compare it to the study of the visual system—an uncontroversial success story. Debates in vision exist, yet none regarding the basic unit of analysis. In low-level vision, lines are lines, angles are angles, and edges are edges—elemental parts that quite clearly play a constitutive role in forming our visual experience. Likewise, in visual object recognition, some basic units of analysis—objects organized in hierarchically structured categories—also seem consensual. In the study of language, by contrast, little is agreed upon. This weakness threatens to hinder the effort to image the neural basis of language (NEUROLINGUISTICS).

The present perspective sees linguistic capacity as critically involving the pairing of sound sequences and meanings, aided by inventories of combinatorial rules, and stores of complex objects of several types, over which these rules operate as language is practiced through its various modalities. The language faculty, in this view, inheres in a cerebrally represented knowledge base (rule system), and in algorithms that instantiate it in use. It is divided

into levels of representation, reflected in language processing: a level for the identification and segmentation of speech sounds (universal phonetics), and a system that enables the concatenation of phonetic units into sequences (phonology), then into words (morphology, where word structure is computed), sentences (syntax), and meaning (lexical and compositional semantics). This rich system of knowledge is cerebrally represented, and has several important properties: At every level beyond phonetics, linguistic units are taken to be discrete; the algorithms that concatenate them do so in keeping with formal rules, some of which are recursive, hence capable of handling strings of arbitrary length; these systems are universal—shared by speakers of all the world's languages. Language-particular rules are encoded in parameters that are embedded within the universal rule system (SPEECH PROCESSING; PSYCHOLINGUISTICS).

Although not all students of language share this view, as should be evident from this handbook (see LANGUAGE PROCESSING; PAST TENSE LEARNING), there is considerable empirical evidence that supports it. This chapter thus briefly reviews some central results emanating from investigations into the brain/language juncture, which support the neural reality of linguistic rules as a constitutive element of the human language faculty. The focus here is on linguistic combinations at the sentence level; but first, some key results in two other successful areas of research are reviewed: the cerebral representation of phonological units, and of word meaning in its isolated and compositional aspects.

Combinatorial Linguistic Systems: The Neural Representation of Sound and Meaning

A central concern of linguists who investigate the phonetic/phonology interface is the nature of the basic building blocks of speech. Rules of concatenation in language typically operate on discrete units, hence the issue of discreteness is critical. Indeed, some recent results suggest that abstract representations of discrete linguistic objects have neural reality. Experiments in magnetoencephalography (MEG) show that auditory cortex can very rapidly construct abstract representations of discrete linguistic objects that go beyond phonetics. Specifically, representations of discrete phonological categories are available already at the earliest stages of processing. Phillips et al. (2000), for example, have shown this through an “oddball paradigm.” They recorded brain activity by MEG, while subjects listened passively to synthetic speech sounds, presented in a phonological and an acoustic condition. The former contrasted stimuli from an acoustic /dæ/-/tæ/ continuum, and elicited a magnetic mismatch field in a sequence of stimuli in which phonological categories occurred in a many-to-one ratio, but no acoustic many-to-one ratio was present. Phillips et al. compared these results to the acoustic condition, where the many-to-one distribution of phonological categories was removed. No such response was elicited, although the stimuli came from an acoustic continuum identical to the phonological condition. Thus, the all-or-nothing property of phonological category membership, as opposed to phonetic stimuli, was demonstrated through MEG, supporting the existence of a neural code for a phonology with discrete categories (see also NEUROLINGUISTICS).

Another issue that neurolinguists are interested in concerns the combination of meaning-bearing units. Some hints regarding this problem may come from a comparison of neural processes involved in the recognition of words in isolation versus sentential context. Posner and Pavese (1998) tried to investigate this question through a paradigm that utilized Evoked Response Potentials (ERPs). They presented sentences like “he ate his food with a _____,” in which the final word was either an appropriate artifact (fork), or an appropriate natural object (fingers), or an inappropriate artifact/natural object (tub/bush). In one condition, subjects were asked to decide

whether the last word referred to an artifact or a natural object (word in isolation); another condition asked whether or not the last word fit the sentential context (sentence task).

Importantly, both conditions focused more on meaning than form. When ERPs obtained in the two tasks were compared, it turned out that left frontal regions were more activated in the lexical task around 120–500 msec into the task, whereas left posterior regions were more activated in the sentence task around half a second into the task. This result suggests that lexical elements are interpreted at different times and locations than sentences, and adds to the growing literature that points to the left frontal and temporal cortices as loci of lexical semantics. It refines the picture in making an initial step toward an identification of regions in which compositional semantic processes take place.

Having discussed sound and meaning in brief, we can now move on to our main topic: how and where in the brain sentences are processed.

Reconstructing the Image of Sentence Grammar: Lesion Data

The nineteenth-century “Connectionist” school, founded by Broca, Wernicke, and Lichtheim, and revived in our time by Geschwind (1979), began modern neuropsychology. This approach emphasized connections between brain regions (rather than the synaptic connections of present-day connectionism), and fortified belief in the existence of cerebral language centers. A clinically oriented approach, it emphasized patients' communicative skills, viewing language as a collection of *activities*, practiced in the service of communication: speaking, listening, reading, writing, naming, repetition, etc. The characterization of the language centers derived from this intuitive theory—each activity was associated with a cerebral locus. Activities are building blocks of the resulting theory of localization, and they are taken to be the essence of human linguistic capacity (NEUROLINGUISTICS).

Since the 1960s, psycholinguists have challenged this view, using theoretical and experimental tools borrowed from linguistics and psycholinguistics (e.g., Goodglass and Berko, 1960; Zurif and Caramazza, 1976). Not denying the relevance of activities, they focused on linguistic distinctions. Language became a structure-dependent piece of knowledge, divided into *levels of representation*. A variety of experiments in the 1970s proved this approach worthwhile, in that results indicated that the brain makes distinctions between types of linguistic information. Such results could not be couched in the standard view, and thus the centers were “redefined” (Zurif, 1980): each anatomical center was now said to contain devices used for the analysis and synthesis of linguistic objects. Roughly, Broca's region (Brodmann's Area BA 44, 45, 47, see Figure 1) was said to house syntax (for both receptive and productive language), while semantics was to reside in Wernicke's area (BA 22, 42, 39). Neuroanatomy also witnessed parallel advances. As large samples of patients became available, it became increasingly clear that language occupies larger areas than previously supposed. As analytic and experimental tools improved, the involvement of both hemispheres in aspects of linguistic activity was documented (cf. Ojemann, 1991).

Yet as findings accumulated—from different tasks, languages, stimulus types, and laboratories—contradictions within the behavioral data began to surface: In some cases, Wernicke's aphasics showed syntactic disturbances; Broca's patients, on the other hand, while failing certain tasks that probe syntactic abilities, succeeded in others. Serious doubts were cast on the new model, in which Broca's (but not Wernicke's) area supports receptive syntactic mechanisms. Attempts to reconcile the findings with the prevailing view argued that regions are organized not just by activities and

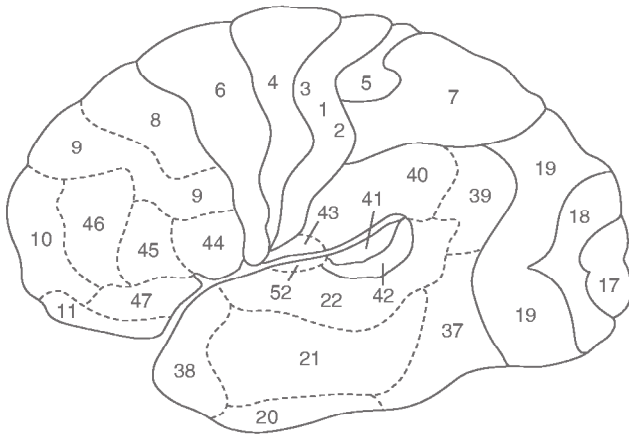


Figure 1. Brodmann's division of the left cortical surface into the areas referred to in the text as BA 22, BA 44, etc.

linguistic levels, but also by *tasks*, saying that “syntactic comprehension is compromised,” and “grammaticality judgment is intact.”

Upon examination, these analyses share a common thread: Although they were detailed in the description of tasks and activities, they were all rather holistic in their approach to the linguistic signal. Gross distinctions between form and meaning seem sufficient, and hence, less attention is paid to detailed structural properties of linguistic stimuli. Still, the amended neurological model of language could continue and prevail.

An exception to this description is the study of phonology and phonetics (cf. Blumstein (1994), NEUROLINGUISTICS). In these areas fine, theoretically motivated distinctions have long been used, and landmark discoveries of subtle distinctions that the brain makes among informational types have been made. Students working in other domains of language, however, were slower in making connections between matters neuropsychological and linguistic. Yet, when it turned out that the task-oriented approach was incorrect, the next move was to try and argue that the inconsistencies in results discussed previously were just apparent, due to our failure to make distinctions among linguistic types. It was argued that systems of grammatical knowledge are complex, and as such, can experience partial breakdown subsequent to focal brain damage. The next step, then, was to seek linguistic frameworks within which patterns of impairment and sparing in aphasia could be couched, and which in turn would give rise to a more finely grained theory of brain/language relations. This resulted in an investigation into the cerebral localization of *grammatical rule systems*. It was shown that despite the importance of channels through which language is practiced, the correct (in fact most telling) unit of analysis for the interpretation of lesion data is the particular rule type.

It is here that considerations pertaining to the structure of language began to matter heavily. As the most important aspect of our “mental organ for language” is its combinatorial nature, that is, the knowledge base and algorithms for the concatenation of linguistic sequences at all levels. Activities and tasks no doubt play a mediating role in linguistic communication, yet the defining characteristic of the language faculty is its being composed to rule systems. How these rule systems are instantiated in neural tissue thus seems the central question in neurolinguistics.

Thus it was shown that in the domain of language production, the brain makes fine distinctions among rule types: Broca's (but not Wernicke's) aphasics are deficient in producing Tense inflection, but intact in Agreement inflection (as shown in a wide variety of languages). Cross-linguistic studies further indicated that not

only inflection type, but also, the position of the verb in the sentence determines its appearance subsequent to a lesion in Broca's region (Friedmann, 1998). In receptive language, the distinction between transformational and nontransformational sentences, yields a big performance contrast: aphasics with lesions in Broca's region understand active sentences, subject relatives, subject questions and the like normally, yet fail on their transformational counterparts: passives, object relatives and questions, etc. This led to the claim that in receptive language, Broca's aphasics are unable to compute transformational relations. This generalization helps localize this grammatical operation in the brain (Trace-Deletion Hypothesis, TDH, cf. Grodzinsky, 2000). Furthermore, the highly selective character of this deficit has major theoretical ramifications to linguistic theory and the theory of sentence processing. A particularly compelling argument that supports the localization of transformations in Broca's region comes from cross-linguistic comparisons: Chinese, Japanese, German, Dutch, Spanish, and Hebrew have different properties, and the performance of Broca's aphasics is determined by the TDH as it interacts with the particular grammar of each language. In English, aphasics comprehend active sentences properly. Yet the results for Japanese, which has two types of actives, are different. *Taro-ga Hanako-o nagutta* (Taro hit Hanako)—Subject Object Verb, and *Hanako-o Taro-ga nagutta*—Object Subject Verb. These constructions are simple, they mean the same, and they are identical on every dimension, except in that the latter is derived transformationally, with the bolded element fronted to the left edge of the sentence. Remarkably, Broca's aphasics' comprehension splits: they handle the SOV type properly, and are at chance level on the OSV.

In Chinese, an otherwise SVO language like English, (bolded) heads of relative clauses (1a, 2a) follow the (parenthesized) relative, unlike English (1b, 2b) in which they precede it. Remarkably, this reversed order correlates perfectly with the cross-linguistic results in aphasia: subject relatives (1) are comprehended at chance in Chinese and above chance in English, whereas object relatives (2) yield the opposite pattern:

- | | |
|---|---------------------|
| (1) a. [_ zhuei gou] de mau hen da | <i>chance</i> |
| chase dog that cat very big | |
| b. The cat that [_ chased the dog] was very big | <i>above chance</i> |
| (2) a. [mau zhuei _] de gou hen xiao | <i>above chance</i> |
| cat chased that dog very small | |
| b. The dog that [the cat chased _] was very small | <i>chance</i> |

English and Chinese thus yield mirror-image results, which correlate with a relevant syntactic contrast between the two languages. Other intriguing cross-linguistic contrasts also exist, providing further evidence that Broca's region is critically involved in transformational analysis. Moreover, reflections of the same disruption are also found in the domain of real-time processing (Zurif, 1995). This rich database is further augmented by results regarding grammatical aspects of the mental lexicon. These are also localizable, as they appear retained in Broca's aphasia, but severely disrupted after a lesion in Wernicke's area.

In sum, over the past decade or so, a new, intriguingly complex model of grammar/brain relations has emerged: Aspects of receptive syntax—those dedicated to the computation of transformational relations—are represented in Broca's region, and to some extent in Wernicke's region; the linguistic lexicon is in the latter region, whereas other parts of receptive syntax, while clearly residing in the left hemisphere, are not localizable as of yet; in productive language, Broca's region is dedicated to extremely limited aspects of structure that pertain to the upper, leftmost end of the syntactic tree (Friedmann and Grodzinsky, 2000). Most importantly, linguistic tools appear critical for the analysis of brain/

language relations (for alternative views, see NEUROLINGUISTICS; LESIONED NETWORKS AS MODELS OF NEUROPSYCHOLOGICAL DEFICITS).

Reconstructing the Grammatical Brain: Neuroimaging

The lesion studies story may have an important lesson regarding functional neuroimaging. Early studies that used this experimental methodology grappled with many hard questions, one of which had to do with the choice of experimental materials, determined, to a large extent, by the experimenter's theoretical tastes. One would have expected neuropsychological data to play a central role in this new effort; in practice, functional imaging of language witnessed an attempt to start almost from scratch. Caught by the excitement that swept the field when neuroimaging techniques were introduced, many investigators have largely tended to dismiss aphasia data, rather than seek cross-methodological convergence. Some important mistakes were repeated as a result. Preliminary studies conducted contrastive investigations of *activities and modalities*. The first ones (Petersen et al., 1990) investigated the production versus comprehension of various linguistic stimuli in PET and then fMRI; and although they made a distinction between overt and covert sentence production, the nature of stimuli—their structure—remained unanalyzed and unspecified. No wonder, then, that anatomical overlap among studies was very limited: verb production versus comprehension, for instance, activated the cerebellum and culliculi for Petersen et al., whereas in more recent studies it was localized in the left posterior temporal lobe and the anterior insula bilaterally.

Early studies were also concerned with *cross-language comparisons*, with language once again taken as one unanalyzed whole, leading to great variation in stimuli (and subsequent anatomical variation). Thus, Mazoyer et al. (1993) conducted a PET investigation of the functional anatomy of sentence comprehension in a known (in fact, native) versus unknown language (French vs. Tamil); other authors looked at PET activations during the comprehension of active declarative sentence in spoken language, as compared to similar stimuli in sign language, finding multiple activations in the left frontal lobe, as well as in the temporal lobe bilaterally. Still others compared the fMRI activation during the comprehension of native (Japanese) versus second (English) and an unknown (Hungarian) language in the same speakers. Here, some aspects of the frontal cortex was activated bilaterally, whereas Broca's and Wernicke's regions, as well as some neighboring ones, were activated only on the left side. Similarly, the BOLD response in a comparative fMRI study of English sentences versus sentences in Mandarin Chinese resulted in bilateral activations in the inferior prefrontal cortex (BA 44, 45, 47, Figure 1), bilateral middle prefrontal cortex (BA 6, 8, 9) and secondarily in the left temporal region (BA 22, 21, 38), the left angular gyrus (BA 39), and bilateral activations in the anterior supplementary motor area (BA 8), the superior parietal region (BA 7), and in some occipital regions. So, while most of these studies demonstrated activation in the left Broca's area and around left Wernicke's area, scattered activations in many more regions—in both the left and the right hemispheres—were also recorded, thus blurring the picture, and making it much less stable than we would like it to be.

Yet, when previous neuropsychological data and linguistic considerations are taken into account, it is quite possible that activities or languages may not be the correct units of analysis for a precise characterization of brain/language relations. One possible reason for the lack of anatomical congruence among past studies, then, is that they made incorrect choices of analytic units, and as a consequence, they simply did not use appropriately minimal contrasts in their comparisons. From this perspective, a sentence in sign language is an incorrect control for a condition that contains English

sentences—it may be as inappropriate a control as a Mozart symphony would be for a test for visual object recognition. Psychologists have realized this, and as a next step, set themselves to study more finely grained distinctions. Again, following the neuropsychological tradition, some have attempted to test distinctions among *levels of linguistic description*. Yet here, too, localization has been somewhat disappointing.

Friederici and her colleagues conducted a series of studies that also contrasted syntactic with semantic variables, and sought neural correlates for it, as monitored through MEG. In one study they tried to localize syntactic processes through the measurement of magnetic response during auditory exposure to “syntactically correct” and “syntactically incorrect” sentences. They found that “early syntactic parsing processes” activated temporal regions, possibly the planum polare, as well as fronto-lateral regions. They further comment that “the contribution of the left temporal regions to the early syntactic processes seems to be larger than that of the left fronto-lateral regions.” Friederici (2000) summarizes the results from PET and fMRI studies: “The posterior region of the left superior temporal gyrus and the adjacent planum temporale is specifically involved in auditory language comprehension.” There is also “an involvement of left inferior frontal regions in phonetic processing,” and for syntax there is “maximal activation in the left third frontal convolution . . . but additional activation in the left Wernicke's area as well as some activation in the homotopic areas in the right hemisphere.” Another experiment this group conducted sought to dissociate the phonological, semantic, and syntactic subsystems. They presented active declarative sentences along with sentences with the same syntactic “frame” but with nonsense words, and with unstructured word lists and non-word lists. When sentences with real and pseudo words were compared to word and non-word lists (as a reflection of syntax), certain bilateral temporal, parietal, frontal, and subcortical areas were activated.

In a similar vein, Dapretto and Bookheimer (1999) tried to dissociate syntax from semantics through fMRI. They asked subjects to make same/different judgments on sentence pairs of two types: one involving the same sentence structure but with one different word; another involving same meaning but different sentence structure (active versus passive). For both the semantic and syntactic comparisons, they report Broca's region and its vicinity (BA 44, 45, 47) and the superior and middle temporal gyri (BA 42, 22, 21) bilaterally as the main activated area in the comparison, with some more activation on the left for the syntactic comparison (BA 44).

The reader may have noticed that here, too, the anatomical overlap between studies is not very promising. Again, Broca's and Wernicke's regions are activated, providing support to the view—originating in Broca's and Wernicke's writings—that these regions are crucial parts of the language faculty. Yet this is not enough: other regions are activated in a nonoverlapping manner, and we must try and understand what this may mean. Three interpretations are imaginable: either language is widely distributed in the brain, and moreover, linguistic representations are unstable, varying from one individual to the next in a manner that affects findings; or the available imaging technology is unreliable; or experiments do not test what they purport to test.

My own tendency is optimistic, leaning toward the third possibility: while there is clearly individual variation in the precise size, location, and structure of the language areas (cf. Amunts et al., 1999, for compelling cytoarchitectonic evidence), brains appear to be relatively stable in what they represent. A large amount of functional variation and spreading, I would argue, is a consequence of the great variation among experiments at this point, caused mainly by an insufficiently refined view of linguistic structure. The fact that “syntax,” “phonology,” and the like are undifferentiated is likely an important reason for the wide range of anatomical loci

imputed to sentence processing. A linguistic perspective—especially one that seeks to account not just for the functional imaging data, but also the rich body of knowledge that comes from lesion studies—might make matters more uniform.

An attempt to be more detailed psycholinguistically has been made by Caplan and by Just and Carpenter. These groups have attempted to view language processing in the brain from the point of view of the putative processing difficulty of different sentence types. Just et al. (1996) looked at the comprehension of three sentence types in fMRI, and found that they all activated left and right Broca's and Wernicke's regions, yet the magnitude of the effect grew with processing difficulty. Using PET, Stromswold et al. (1996) showed differential activation in left Broca's region for differentially difficult relative clauses; and study by the same group conducted a PET study with similar materials, yet with a slightly different task, and found activations in the centromedian nucleus of the left thalamus, the medial frontal gyrus, Broca's area, and the posterior cingulate gyrus. Differential processing difficulty, used as a marker that delineates the language faculty, again results in poor anatomical overlap. The similarity in the questions posed by most of these studies suggests that discrepancies in anatomical findings may either be due to different imaging devices, or choice of tasks and materials. Still, experience with the linguistic interpretation of lesion data leaves one with a gnawing sense that systematic linguistic description of functional imaging results—and subsequent planning of linguistically motivated experiments—is somewhat lacking. In the case of complexity, the blurred picture may well be due to the fact that the linguistic complexity is not a well-defined notion, and its varying construal affects the nature of experimental materials and the analysis and interpretation of results. It is perhaps advisable to go back to studying aphasia, to try and find some hints there. The strong link between grammatical transformations and the language areas may be a good place to start, if we seek to tease particular components *within* the grammar apart from others.

When transformations are separated from complexity, and tested in fMRI, a fairly clear picture emerges: left Broca's region (BA 44, 45) and to a lesser extent, both Heschl's gyri, are most strongly involved in transformational analysis. Ben-Shachar et al. (2001) have conducted this experiment in Hebrew, searching for a T(transformational)-effect. They used minimal pairs of equally complex sentences, except that one set contained a transformation (3a) and another did not (3b):

- (3) a. I helped the nurse [that John saw ___ in the living room]
 b. I told John [that the nurse slept in the living room]

A T-effect was found in left Broca's area (BA 44, 45): A higher BOLD signal was detected for + Transformation sentences relative to – Transformation sentences (Figure 2). These results suggest a critical role for Broca's region in the analysis of transformations in the healthy brain, and converge on the available lesion data. An ROI approach detected activations in the posterior inferior frontal gyrus and the anterior insula, the posterior superior temporal sulcus and Heschl's complex. Thus, the core computational resource for Movement structures is in areas 44, 45. Auxiliary computations occur at temporal areas bilaterally.

Discussion

We have reviewed results that point to the neurological distinctness and locus of the transformational component of syntax. They also suggest that at least some of the results obtained in the fMRI and PET syntactic complexity experiments could be recast in transformational terms, which may lead to a radical reduction in the amount of variation, and to convergence of cross-linguistic and cross-

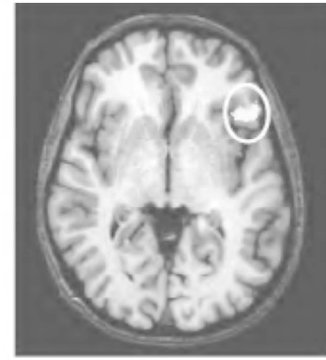


Figure 2. A statistical map associated with +T sentences. Left IFG is the most activated region (Ben-Shachar et al., 2001)

methodological data from lesions studies, as well as from PET and fMRI.

So what is the image of the linguistic brain? We are just beginning to reconstruct it. Whether the somewhat variable anatomy will ever permit precise localization is still an open question; and gross localization, after all, is just a small step toward understanding. Yet our best bet, it seems, is to take linguistic rules as the basic unit of functional analysis of the intricate relationship between language and the brain.

Road Maps: Cognitive Neuroscience; Linguistics and Speech Processing
Related Reading: Imaging the Motor Brain; Imaging the Visual Brain; Lesioned Networks as Models of Neuropsychological Deficits; Neurolinguistics

References

- Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B. M., and Zilles, K., 1999, Broca's region revisited: Cytoarchitecture and intersubject variability, *J. Comp. Neurol.*, 412:319–341. ♦
- Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., and Grodzinsky Y., 2001, Grammatical transformations activate Broca's region—An fMRI study. Presented at the Cognitive Neuroscience Society, New York.
- Blumstein, S. E., 1994, The neurobiology of the sound structure of language, in *Handbook of Cognitive Neuroscience* (M. Gazzaniga, Ed.), Cambridge, MA: MIT Press.
- Dapretto, M., and Bookheimer, S. Y., 1999, Form and content: dissociating syntax and semantics in sentence comprehension, *Neuron*, 24(2):427–432.
- Friederici, A., 2000, The neural dynamics of language comprehension, in *Image, Language, Brain* (A. Marantz, Y. Miyashita, and W. O'Neil, Eds.), Cambridge, MA: MIT Press.
- Friedmann, N., 1998, *Functional Categories in Agrammatism*, Doctoral dissertation, Tel Aviv University.
- Friedmann, N., and Grodzinsky, Y., 2000, Neurolinguistic evidence for split inflection, in *The Acquisition of Syntax* (M. A. Friedemann and L. Rizzi, Eds.), London: Blackwell.
- Geschwind, N., 1979, Specializations of the human brain, *Scientific American*, September, 241(3):180–199. ♦
- Goodglass, H., and Berko, J., 1960, Agrammatism and inflectional morphology in English, *J. Speech Hear. Res.*, 7:257–267.
- Grodzinsky, Y., 2000, The neurology of syntax: language use without Broca's area, *Behav. Brain Sci.*, 23:1–71. ♦
- Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R., 1996, Brain activation modulated by sentence comprehension, *Science*, 274:114–116.
- Mazoyer, B. M., Dehaene, S., Tzourio, N., Frak, V., Murayama, N., Cohen, L., Levrier, O., Salamon, G., Syrota, A., and Mehler, J., 1993, The cortical representation of speech, *J. Cognit. Neurosci.*, 5:467–497.

- Ojemann, G. A., 1991, Cortical organization of language, *J. Neurosci.*, 11:2281–2287. ♦
- Petersen, S. E., Fox, P. T., Snyder, A. Z., and Raichle, M. E., 1990, Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli, *Science*, 249:1041–1044.
- Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K. McGinnis, M., Poeppel, D., Roberts, T., 2000, Auditory cortex accesses phonological categories: An MEG mismatch study, *J. Cognit. Neurosci.*, 12:1038–1055.
- Posner, M. I., and Pavese, A., 1998, Anatomy of word and sentence meaning, *Proc. Natl. Acad. Sci. USA*, 95:899–905.
- Stromswold, K., Caplan, D., Alpert, N. and Rauch, S., 1996, Localization of syntactic comprehension by positron emission tomography, *Brain Lang.*, 52:452–473.
- Zurif, E. B., 1980, Language mechanisms: A neuropsychological perspective, *Am. Sci.*, 68:305–311. ♦
- Zurif, E. B., 1995, Brain regions of relevance to syntactic processing, in *An Invitation to Cognitive Science*, Vol. I (L. Gleitman and M. Liberman, Eds.), 2nd ed., Cambridge, MA: MIT Press. ♦
- Zurif, E. B., and Caramazza, A., 1976, Linguistic structures in aphasia: Studies in syntax and semantics, in *Studies in Neurolinguistics*, Vol. 2 (H. Whitaker and H. H. Whitaker, Eds.), New York: Academic Press.

Imaging the Motor Brain

John Darrell Van Horn

Introduction

Functional imaging of the human brain during the performance of motor tasks examines one of the most basic, as well as one of the most complex, facets of brain function. Early studies assessing regional cerebral blood flow (rCBF), using positron emission tomography (PET), and investigations of blood oxygenation effects, with functional magnetic resonance imaging (fMRI), relied on simple finger opposition tasks to produce large and robust activation patterns in the motor cortex (e.g., Figure 1). However, the results of recent motor neuroimaging studies suggest that the behavioral form and context of a movement are important determinants of functional activity within cortical motor areas and the cerebellum. Unlike the consideration of higher cognitive processes (e.g., memory, learning, vision, etc.), functional imaging of the human motor system is based on the need to understand the interaction of neurological and cognitive processes with the biomechanical characteristics of the limb (e.g., reaching, grasp, rotation, flexion, etc.). This implies a dependency on other neural systems, such as vision (see EYE-HAND COORDINATION IN REACHING MOVEMENTS), to help supply information for the purposes of learning the parameters of motor tasks. Neuroimaging evidence from such studies has accumulated, indicating that multiple neural systems and their functional interactions are needed to successfully perform motor tasks, encode relevant information for motor learning, and update behavioral performance in real time. Thus, more than in any other domain, motor neuroimaging is moving beyond the concept of localization-based modularity and into that of the functional interaction of brain systems for the construction of theories of motor function.

Accompanying the rise in functional neuroimaging as a research tool has been the increase in sophistication of task paradigms employed for probing the motor system. Investigations utilizing both PET and fMRI are now examining complex sequences of finger movements and continuously generated motor behaviors to investigate in finer detail the formation of motor programs, the role of proprioceptive mechanisms, anticipation of movement, and the generation of internal models. For fMRI, in particular, this has necessitated novel approaches to how the behavioral and functional data are collected simultaneously. Also, the mathematical and statistical modeling of these data has required the careful consideration of the temporal characteristics of both the measured blood oxygenation level dependent (BOLD) signal as well as the associated motor behavioral output. These emerging experimental frameworks represent a considerable departure from the traditional “task minus control”-style experimental designs traditionally employed for fMRI. Using such methods, researchers are rapidly gain-

ing insight into motor processes that are hugely dynamic, shaped not only by the demands of the task paradigm itself, but also on the basis of error feedback, behavioral skill acquisition, and automaticity.

In this chapter, the following aspects of the examination of the human motor system with neuroimaging will be discussed: (1) how evidence from functional imaging studies is lending support to current constructs of motor theory concerning the development of internal models of movement, (2) how this information is being functionally integrated by the brain, (3) motor automaticity, and (4) experimental design and data modeling considerations for func-

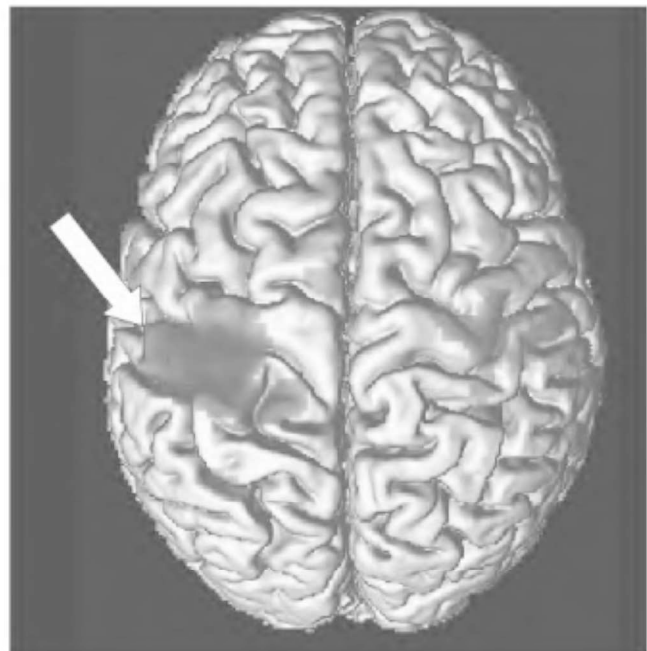


Figure 1. Block-design motor activation. A single male subject alternately rotates a small object in his dominant (right) hand or is at rest in 15-s intervals, over a 4.5-min functional (EPI) scanning session (General Electric Horizon 1.5 Tesla, TR = 2000 ms, TE = 500 ms, FOV = 24 cm, 27 slices). Significant Student's *t*-test ($p < 0.001$) activation, overlaid on a rendering of the subject's cortex as measured via a high resolution spoiled gradient echo (SPGR) structural scan, is evident in motor and premotor regions, being most extensive contralateral (left; arrow) to the side of movement.