Third edition

DANIEL S. LEVINE

Introduction to Neural and Cognitive Modeling

"Levine's book achieves an impressive synthesis of historical trends and current research results in both biological and artificial neural network research. This synthesis clarifies that the currently popular Deep Learning is just one contribution to this burgeoning field, and one that does not incorporate many of the most powerful properties of biological learning. Levine's book provides an accessible introduction to many of these properties, while also reviewing important properties of neural models of vision and visual attention, sequence learning and performance, executive function, and decision-making, among its other expository accomplishments."

—**Stephen Grossberg**, *Boston University, USA*

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

INTRODUCTION TO NEURAL AND COGNITIVE MODELING

Providing a thorough introduction to the field of neural networks, this edition concentrates on networks for modeling brain processes involved in cognitive and behavioral functions. Part I explores the philosophy of modeling and the field's history, starting from the mid-1940s, and then discusses past models of associative learning and of short-term memory that provide building blocks for more complex recent models. Part II of the book reviews recent experimental findings in cognitive neuroscience and discusses models of conditioning, categorization, category learning, vision, visual attention, sequence learning, behavioral control, decision-making, reasoning, and creativity. The book presents these models both as abstract ideas and through examples and concrete data for specific brain regions.

The book includes two appendices to help ground the reader: one reviewing the mathematics used in network modeling, and a second reviewing basic neuroscience at both the neuron and brain region level. The book also includes equations, practice exercises, and thought experiments.

Daniel S. Levine is Professor of Psychology at the University of Texas at Arlington. He is a fellow and former president of the International Neural Network Society. His research involves computational modeling of brain processes in decision-making and cognitive-emotional interactions.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

INTRODUCTION TO NEURAL AND COGNITIVE MODELING

Third edition

Daniel S. Levine

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Third edition published 2019
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Psychology Press 1991

Second edition published by Routledge 2000

Library of Congress Cataloging-in-Publication Data
Names: Levine, Daniel S., author.
Title: Introduction to neural and cognitive modeling / Daniel S. Levine.
Description: Third edition. | New York, NY : Routledge, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2018018582| ISBN 9781848726475 (hbk) | ISBN 9781848726482 (pbk) | ISBN 9780429448805 (ebk)
Subjects: LCSH: Neural networks (Neurobiology) | Neural networks (Computer science)
Classification: LCC QP363.3 .L48 2019 | DDC 612.8/2–dc23
LC record available at https://lccn.loc.gov/2018018582

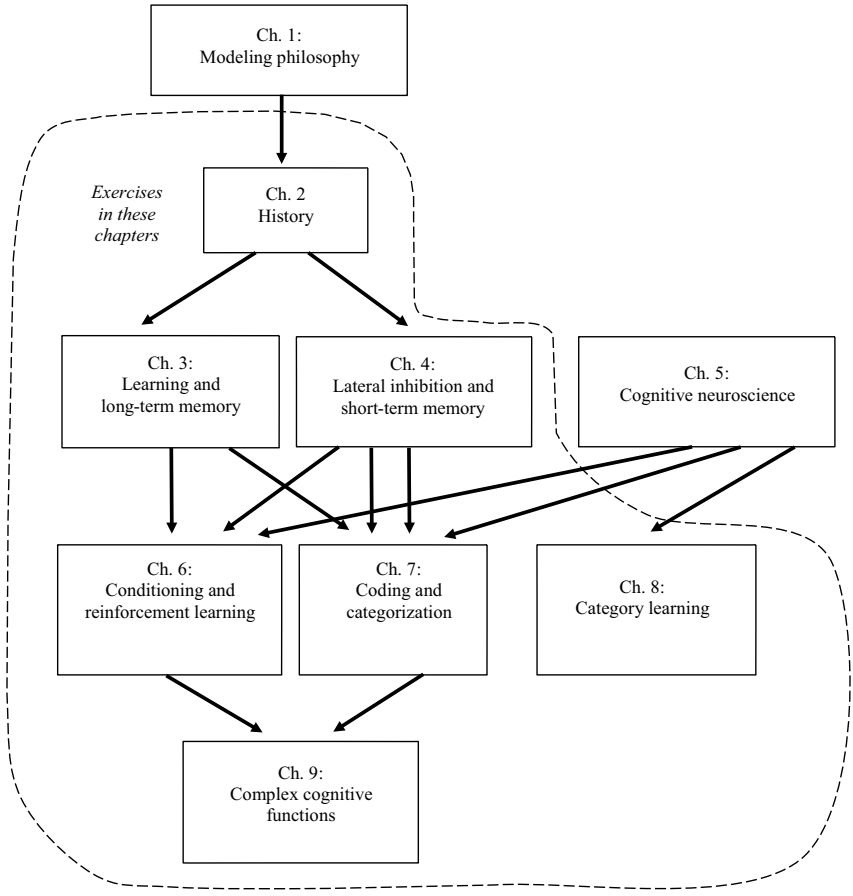
ISBN: 978-1-84872-647-5 (hbk)
ISBN: 978-1-84872-648-2 (pbk)
ISBN: 978-0-429-44880-5 (ebk)

Typeset in Times New Roman and Adobe Garamond
by Florence Production Ltd, Stoodleigh, Devon, UK

CONTENTS

FLOW CHART OF THE BOOK

Here is the structure of the book. Chapters depend in part on previous chapters that have arrows pointing to them.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PREFACE TO THE THIRD EDITION

How do you eat an elephant? One bite at a time.
Beverly Johnson, after climbing the rock face of
El Capitan, Yosemite National Park

Since the second edition of *Introduction to Neural and Cognitive Modeling* appeared in 2000, the growth of neuroscience, and of cognitive neuroscience in particular, has been explosive. For example, I did a search of the library database PsycInfo under the keywords "fMRI or functional magnetic resonance imaging," which is the most frequently used experimental method for studying the relationship between brain physiology and human cognitive function. This search yielded only 929 results between the years 1943, the date of the first seminal article in neural networks, by McCulloch and Pitts, and 2000. Yet, the same search yielded 31,107 results between 2001 and 2017.

The growth in computational models of neural and cognitive processes has not been quite as spectacular as the growth in experimental studies but is still impressive. One result has been the formation of distinct niches within the field of neural networks. The changes in the field over 17 years were sufficient to warrant a new edition of this book. The first and second editions were intended for the varied audience of the field, those primarily interested in using neural networks for intelligent computing and engineering applications as well as those primarily interested in understanding brain processes. By contrast, this edition primarily targets those interested in understanding brain processes. However, most of the more general foundational material from the earlier editions remains with updating in Chapters 2–4, so the neural engineer or computer scientist can still find the book a valuable source.

This book is intended both as a textbook for a graduate or advanced undergraduate course in neural networks and as a general introduction to the field. Its focus is on the relationships between neural structure and cognitive function. Details of neuronal biophysics are included only in so far as they illuminate the cognitive and behavioral implications of neural structures: other books (e.g., Dayan & Abbott, 2005) present more computational and mathematical models at the neuronal level.

The cognitive functions discussed herein include learning, perception, attention, memory, pattern recognition, categorization, executive function, decision-making, and inference. The neural networks modeling these functions sometimes incorporate organizing principles such as competition, association, opponent processing, and error correction, principles that can be suggested either by the exigencies of modeling psychological data or by the description of known neuroanatomical structures. These principles are developed in early chapters and appear throughout the book.

In keeping with the goal of accessibility to a varied audience, technical prerequisites in any one discipline are kept to a minimum. Recent advances in computing make the field accessible to many more people than before. For those needing additional background in mathematics or in neurobiology, appendices in those fields are included; the appendices also list sources for more detailed coverage.

A word should be said here about equations. The last section of each of Chapters 2–4 and 6–9 includes differential or difference equations for some of the networks discussed in that chapter, so that the reader can gain hands-on experience in computer simulation of the networks. On first reading, the student without mathematical background can skip these equations and follow the development of networks by means of the figures. On second reading, the same student can turn to Appendix 1 for explanations of how equations reflect the qualitative relationships in networks, and simple algorithms for simulating such equations. The first half of Appendix 1 is written so as not to require previous background; notions needed from calculus are redefined and motivated in the context of neural network applications.

The book is divided into two sections. The first section, Chapters 1–4, is more foundational. Chapter 1 gives the underlying philosophy of the book and Chapter 2 gives the historical background for its main ideas. The next two chapters describe models of common "building blocks" for network models of cognitive functions: associative learning in Chapter 3 and lateral inhibition in Chapter 4.

The second section, Chapters 5–9, discusses models of cognitive functions, building on the work of the previous chapters. Chapter 5 reviews recent advances in experimental cognitive neuroscience. The networks discussed in Chapter 6–9 simulate some of the data discussed in Chapter 5 and often incorporate simpler network structures developed in Chapters 3 and 4.

Chapter 6 on conditioning and reinforcement learning, Chapter 7 on unsupervised categorization, and Chapter 8 on supervised categorization include much of the material of Chapters 5 and 6 of the second edition but add a considerable amount on models developed since 2000. Chapter 9 is entirely new, covering models of a variety of complex cognitive functions: vision and visual attention, sequence learning and performance, executive function, decision-making, reasoning, and creativity.

Chapters 2–4 and 6–9 contain homework exercises. Some exercises are thought experiments and others are computer simulations of various neural network models. The exercises herein are a small sampling of the possible questions that can be asked about the material discussed in the book, and the instructor is encouraged to supplement them as he or she sees fit. Many of the thought questions asked here do not have right and wrong answers, only a variety of better and worse answers. The reader should approach the field with at least as much intellectual flexibility and curiosity as possessed by the systems we model. A solution manual has been provided for instructors, with links from the book's website. The manual can provide hints for exercises, particularly for the harder ones in Chapters 6 through 9.

The flowchart in the front of the book illustrates how the understanding of each chapter depends on previous chapters. After the introduction in Chapter 1 and the historical account in Chapter 2, Chapters 3 through 9 (excluding Chapter 5, which is a review of experimental data) reflect a hierarchy in models from simpler to more complex neural and cognitive processes. Chapters 2–4, which are foundational for the more complex models in Chapters 6–8, are the chapters least changed from the second edition, though they are updated based on recent scientific advances.

ACKNOWLEDGMENTS

The psychology editor at Taylor & Francis, Paul Dukes, encouraged this book project from the very beginning. When Taylor & Francis became the parent company of Lawrence Erlbaum Associates in 2007, Paul noted that previous editions had been among LEA's most successful books until about 2005, when the field had grown large enough to spawn competing books from other publishers. He enlisted me to update my book for a twenty-first-century audience. He has worked hard with other Taylor & Francis staff members to obtain helpful reviews both before and after my signing of the book contract, and to provide infrastructure both for publicizing the book and making it suitable for classroom use. Claudia Bona-Cohen, the editorial assistant at Taylor & Francis, saw the book through efficiently at the production stage.

Several leading neural network researchers gave feedback both on the book proposal and on earlier drafts of parts of the text that markedly improved its content. These researchers included Jeffrey Bowers, Joshua Brown, Daniel Bullock, Gail Carpenter, Colin Davis, Stephen Grossberg, Sebastian Hélie, Ali Minai, and Paul Werbos.

Bakur AlQaudi, who is a faculty member at Yanbu Industrial College (Saudi Arabia) with a PhD in electrical engineering from the University of Texas at Arlington (UTA), tested all the computational exercises in MATLAB on a Lenovo P40 Yoga Laptop (ThinkPad) Windows 10 (64-bit). He worked with me to "vet" the exercises and helped me edit their text to make them as suitable as possible for the level of the graduate students who are likely to run them in class work. In addition, the UTA Department of Psychology provided me with comfortable office, laboratory, and computing space that expedited the book's production. My student Amandeep Dhaliwal helped with the quality of the figures.

Finally, my wife, Lorraine Levine, lived patiently with the highs and the rebound lows associated with the book's composition. She combined an appreciation of the project's value with a warm sense of humor that kept me on course but helped me avoid the perils of overly grim determination.

NOTATION USED IN THE BOOK

In figures

An arrow (→) denotes an excitatory connection
A filled circle (●) denotes an inhibitory connection
A filled semicircle (◗) denotes a modifiable connection

In exercises

A single asterisk (*) denotes a computer simulation exercise
A double asterisk (**) denotes a mathematical problem
An open circle (○) denotes an open-ended conceptual exercise

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PART I

Foundations of Neural Network Theory

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

NEURAL NETWORKS FOR MODELING BEHAVIOR

My mind to me a kingdom is,
Such perfect joy therein I find
As far exceeds all earthly bliss
That God or nature hath assigned.
Edward Dyer

What is mind? No matter. What is matter? Never mind.
Thomas Hewitt Key (epigram in *Punch*)

1.1. What Are Neural Networks?

The birth of the current field of neural network modeling can be traced to the cybernetic revolution of the 1940s and 1950s. At that time, scientists across many disciplines became excited about the notion that neurons are digital on–off switches (either firing or not firing), and thus that brains and the newly emerging digital computers had similar structural organizations (Wiener, 1948). Before long, biologists discovered that the digital metaphor was an inadequate one for capturing what was known about neurobiology and psychology. It was found necessary to understand the graded (or continuous, or analog, or grayscale) as well as the all-or-none (or digital) components of neuron responses (see, e.g., Thompson, 1967, Ch. 1).

Yet, if one seeks to understand cognitive and behavioral functioning, it is not enough simply to understand neurons and their assemblies. It is just as important to understand architectural principles guiding network connection, principles that enable key psychological processes to occur and thereby allow the assembly of neurons to instantiate significant classes of cognitive and behavioral data.

How do we develop theories of brain–behavior relationships? Elsewhere (Levine, 1999) I have described the theory building process as analogous to the building of the tunnel across the channel between England and France. The theorist starts at both the behavioral and the neural ends, and then builds "tunnels" of connection from each end toward the other (Figure 1.1).

Neural network models have gradually evolved in the direction of greater correspondence with brain structure and function, from the 1980s to the present. Techniques such as taking magnetic resonance scans of whole brain regions and recording with electrodes from up to 50 neurons at once have made neurophysiology more amenable to quantification. At the same time, advances in computing (personal computers, supercomputers, and interfaces with recording devices) have made simulation of biological data easier and more practical. These technical developments are enabling theorists building on earlier, more abstract cognitive models to create theories with real explanatory and predictive power. The review article of Ashby and Hélie (2011) calls the newer models that incorporate sophisticated neuroscience data by the name of *computational cognitive neuroscience* models, to distinguish them from more loosely brain-related cognitive models that are called by the older name of *connectionist* models (e.g., Feldman & Ballard, 1982; Rumelhart & McClelland, 1986a).

The development of neural network or connectionist theories has been spurred not only by interest in how the brain works but equally by the interest in potential engineering applications of intelligent computing. The International Neural Network Society's several hundred members, and several annual meetings including the International Joint Conferences on Neural Networks, span research in both the neuroscientific and engineering applications of the field, and some studies that overlap neuroscience and engineering.

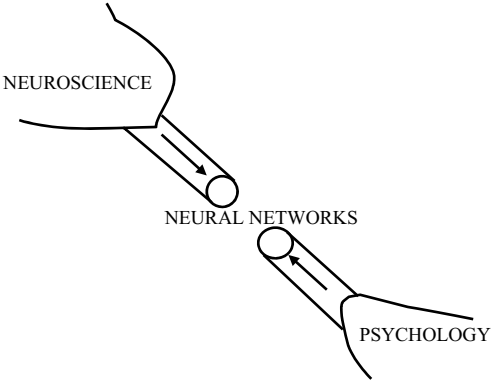


FIGURE 1.1 "Tunnel" metaphor for integrating neuroscience and psychology; see the text.

Designers of machines for performing cognitive functions generally are interested in learning, without slavishly imitating, how the brain performs those functions. Consequently, such machines have often been built like simulated brain regions, with nodes corresponding to neurons or neuron populations, and connections between the nodes; at times, their designers have borrowed ideas from recent experimental results on the brain's analog responses. The industrial applications of connectionist theory are often called *artificial neural networks* or *ANNs* (e.g., Hecht-Nielsen, 1986).

The much older term *neural networks* is usually considered to encompass both theoretical and applied models that provide mechanistic bases for cognitive functions. The functional units in neural networks have alternatively been called "nodes," "units," "neurons," and "populations." The first two terms are most commonly used, in this book and elsewhere, because they do not commit the user to an assumption that units correspond to either single or multiple neurons. Both the earlier connectionist networks and the networks in computational cognitive neuroscience models include nodes, connections, and equations describing the interactions of node activities and connection strengths.

1.2. What Are Some Principles of Neural Network Theory?

While the neural network field has experienced upward and downward surges in popularity, these surges conceal the field's maturity. The early history of neural network models, summarized in the review article of Levine (1983) and discussed further in Chapter 2, shows that most modern ideas in network design have much earlier antecedents. For example, the popular distinction between input, hidden, and output units (Rumelhart & McClelland, 1986a), which led to back propagation and then to deep learning, owes much to the early work of Rosenblatt (1962) on networks with sensory, associative, and response units. Rosenblatt, in turn, combined extensions of the linear threshold law of McCulloch and Pitts (1943) with various learning laws, some of them inspired by the work of Hebb (1949). A precise mathematical formulation for many examples of this type of network was given in the dissertation of Paul Werbos (1974; see Werbos, 1993, for a more accessible version). Moreover, several researchers who published seminal neural network articles in the late 1960s or early 1970s, including Michael Arbib, Jack Cowan, Walter Freeman, Stephen Grossberg, and Teuvo Kohonen, either remain active in the field today or did so through the early twenty-first century.

As the neural network literature grows, it is essential to find criteria for making distinctions among competing models. Meeter, Jehee, and Murre (2007) observed that fitting both behavioral and neural data is one of the important criteria but not the only one. Ideally, these authors note, in addition to fitting existing data and making nontrivial predictions, a model should be

based on assumptions that make sense from a biological and/or behavioral viewpoint. Also, it is desirable to have several interconnected and mutually consistent models for the same process at different levels of representation, and such hierarchies of models are particularly useful when some of the relevant data are unavailable. A similar point was made by the neural network pioneer Stephen Grossberg (Grossberg, 2006): "One works with large amounts of data because otherwise too many seemingly plausible hypotheses cannot be ruled out." The purpose of interconnected models, Grossberg goes on to say, is to capture the fact that "the brain . . . can be successfully understood as an organ that is designed to achieve successful autonomous adaptation to a changing world."

In order to understand the behaviors leading to autonomous adaptation, we need first to understand subsystems that perform parts of the tasks we want the system to perform. Subsystem identification is dramatized by the parable of the watchmakers (Simon, 1969, pp. 90–93). One watchmaker tries to fashion a whole watch simply by fitting parts together. Another watchmaker instead starts with the same parts but puts some of them together into subsystems. Not until the subsystems are working does he then join them into a watch. The second watchmaker prospers, while the first has to start all over again whenever he is interrupted.

Likewise, any major cognitive process needs to be analyzed into subprocesses. The understanding of the subprocesses then suggests principles that can also be used in models of a wide variety of other processes. Agreement on the principles of how to organize neural networks into subnetworks is far from universal. In part, this reflects the incredible variety of ways in which any given cognitive function is organized across many biological species (indeed, phyla) and individuals within these species. In part, it reflects the variety of theoretical perspectives brought to bear on these problems. Yet, it is possible to see through this diversity a few subnetwork organizing themes that are common to network models arising from many sources.

Levine (1989) developed a thought experiment that illustrates two generic examples of analyzing a mental process into subprocesses. Both examples relate to main topics of later chapters in this book.

The first example, discussed more fully in Chapter 7, is the categorization of sensory patterns. Categorization is necessary for an organism or network to make sense of its environment and make predictions about novel stimuli. For definiteness, say that the network is processing hand-printed characters and attempting to match each one to a known letter of the alphabet. The simplest (although not the only) way to understand this process is to include in the network some units ("feature nodes") that respond to presence or absence of writing at particular locations, and other units ("category nodes") that respond to patterns of feature node activation representing particular letters (see Figure 1.2).

To carry our thought experiment further, how are feature nodes and category nodes likely to be connected? If these connections were hard-wired, the network could not adjust to significant changes in its patterns of activation – for example, receiving inputs of Japanese or Russian instead of Roman characters. Hence, it is usually desired that the strength of the connection between any specific feature node and any specific category node be allowed to change over time, as a result of repeated activation of the connection. Such a change is often accomplished by a principle called *associative learning*; associative learning laws in neural networks are the main topic of Chapter 3.

Hence, associative learning has been one of the common subnetwork organizing principles in neural networks from the early 1960s to the present. There are other important subnetwork principles: many categorization models particularly rely on a principle called *competition*. To motivate the idea of neural competition, consider a sloppily written letter that is ambiguous (, say; it could either be an E or an F. The network needs a method for deciding which of the two letters is the more likely one. The feature nodes are activated, to varying degrees, by the incoming letter, and in turn activate their category nodes via the internode connections. As shown in Figure 1.2, there is mutual inhibition between the category nodes. Thus, if both the "E" and "F" nodes are activated but the activation of the "E" node is greater than that of the "F" node, the system makes a decision that the letter is an E. Inhibition between nodes at the same level of the network (in this case, the category level) is often

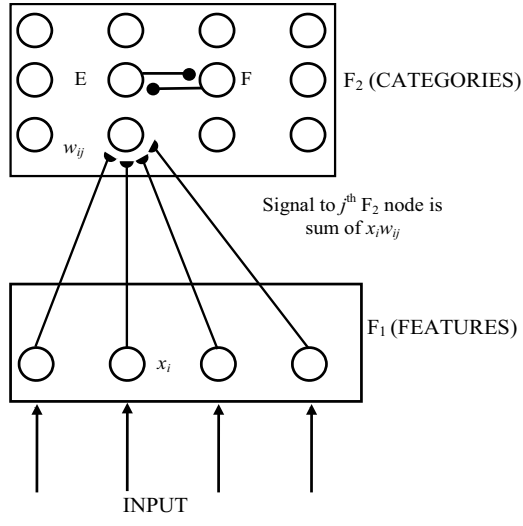


FIGURE 1.2 Generic categorization network combining associative learning and competition.

Source: Modified from Levine, 1989, with permission of Miller Freeman Publishers.

regarded as competition between the different cognitive entities coded by those nodes; competition laws in neural networks are the main topic of Chapter 4.

Associative and competitive principles are likely to combine in many other cognitive processes besides categorization. Our second example is attentional modulation of conditioning, which is discussed more fully in Chapter 6. Suppose that a neutral stimulus such as the sound of a bell has become associated with food. Then how does the animal learn to pay more attention to the bell than to other neutral stimuli in its environment?

Again, the psychological results suggest the neural network principle of competition. If different nodes develop representations of sensory stimuli, then the bell node should somehow "win" a competition with other sensory nodes for storage in short-term memory.

But how does the competition among sensory representations become biased in favor of the bell? One plausible answer suggests the neural network principle of associative learning. Other things being equal, the bell node tends to be activated because of prior association between the bell and food, or, more abstractly, between the bell and satisfaction of the hunger drive (see Figure 1.3). Hence, if there is another node representing the primary reinforcer of food or the hunger drive itself, repeated pairing of the bell and food tends to activate the bell representation if the animal is hungry. Strengthening pathways based on pairing of stimuli is an example of associative learning.

The thought experiment described here illustrates the potential of network theories to unify disparate areas of psychology. In the early to middle twentieth century, the behaviorist school was ascendant in academic psychology, with its focus on outwardly measurable performance and its distrust of any consideration of internal states, beliefs, or self-reports. But Behaviorism was challenged in the 1960s and 1970s with the rise of cognitive psychology, whereby internal states became of paramount importance (Reisberg, 2016, Chapter 1). The thought experiment tells us that classical conditioning, a major concern of the behaviorists, and categorization, which falls into the cognitive realm, can be understood using different combinations of the same theoretical "building blocks."

In the course of this exposition, a few other major subnetwork organizing principles emerge besides association and competition. One of these principles is *opponent processing*. This means that neural architectures are organized into pairs of pathways with opposite significance (e.g., light and dark, or reward and punishment) in such a manner that a sudden decrease in activity of one pathway transiently activates the opposing pathway. Like association and competition, opponent processing plays a role in some of the conditioning models discussed in Chapter 6 and some of the categorization models discussed in Chapters 7 and 8.

All these principles have been suggested by a heterogeneous database that is partly physiological and partly psychological. In some cases, the theory was

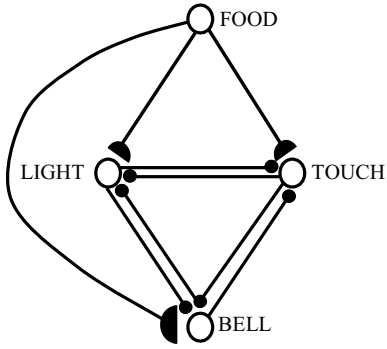


FIGURE 1.3 Generic selective attention network combining associative learning and competition. Filled semicircles represent learnable connections. The larger semicircle on the path from food to bell represents a connection strengthened by learning, based on pairing of the bell with food. This biases competition among light, touch, and bell nodes in the bell's favor.

Source: Modified from Levine, 1989, with permission of Miller Freeman Publishers.

first suggested by psychological results and later, at least qualitatively, supported neurophysiologically. An example is associative learning theory. Hebb (1949), inspired by Pavlovian conditioning data, proposed that if one neuron connects to another via a synapse, and the firing of the first neuron is repeatedly followed by firing of the second, then the synapse should become strengthened. As discussed in the early part of Chapter 3, examples of neuronal behavior consistent with this general hypothesis (though not in the precise form that Hebb had proposed) were discovered first in invertebrates (e.g., Kandel & Tauc, 1965) and later in vertebrates (e.g., Bliss & Lømo, 1973). More recent experimental studies, moreover, have partially supported mathematical variations of Hebb's learning law proposed by modelers in the 1960s and 1970s.

This book talks rather freely of "representation" by network nodes of features, categories, or other concepts. There is substantial debate among philosophers of science as to what a true representation is, either in the brain or a model (see Bechtel, Mandik, Mundale, & Stufflebeam, 2001; Bechtel, 2008). Bechtel and his colleagues note that the different disciplinary perspectives of cognitive scientists and neuroscientists have led to different concepts of representation. They also note that a school of modelers has arisen that study large networks by means of the mathematical theory of dynamical systems and have concluded that considering representations of part of a system detracts from considerations of the whole system (see, e.g., Freeman & Skarda, 1990).

This book presents models originating from a variety of perspectives and disciplines, so does not take a position in the philosophical debates about

representation. I believe, though, that the "tunnel-building" approach described here helps to unify the varying disciplinary perspectives from neuroscience, psychology, and cognitive science. Also, the models described in this book make extensive use of dynamical systems, but in some cases the modelers still find it a useful simplification to talk about some part of their network being a representation of some concept.

1.3. Problems, not Trends

This book's organization is unlike that of most other introductory books on neural networks. In keeping with its emphasis on modeling the connections between neuroscience and psychology, the primary organization of the chapters is by the behavioral and cognitive problems that the neural models were designed to address. Organization by different trends, approaches, and schools of modeling is secondary and shows up mainly in the headings of sections and subsections.

For this reason, such recent trends as deep learning and Bayesian learning are not treated in separate chapters but as aspects of models that reproduce specific behavioral and neural capabilities or datasets. Also, because the book more than its previous editions emphasizes neuroscience and psychology applications, engineering and machine learning applications are discussed only insofar as they illuminate the modeling of behavioral and neural phenomena.

Modelers of psychological phenomena frequently make a distinction between *tokens* (specific instantiations of models) and *types* (generic instantiations of models) (see, e.g., Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). There has been considerable work on *model mimicry*, the ability of one model to account for data generated by a competing model, at the level of tokens, that is, different parametrizations of a model with a given set of assumptions. Much of that work is beyond the scope of this book, because the book concentrates more on types than on tokens. The primary focus is on discerning the right set of model interactions, before trying to optimize the numerical strength of those interactions.

1.4. Methodological Considerations

Subnetworks incorporating principles such as associative learning, competition, and others to emerge in later chapters can be thought of as part of the neural network modeler's "tool kit." So can larger networks that have been developed by major researchers in the field. Hence, when neural networks are used in brain modeling, new discoveries on the brain or on biological cognition may force modifications of a theory rather than abandonment of the entire structure.

Also, a general qualitative theory may be widely applicable but the detailed instantiation of that theory may vary enormously between individuals and species.

The ability to break down many complex networks into functionally significant subnetworks does *not* necessarily mean that these subnetworks should be treated as independent modules, as in early artificial intelligence models. The function and even structure of each of the smaller networks can be strongly affected by its interactions with the other smaller networks. For example, in many models of vision the systems for processing shape, color, and motion all depend in part on interactions with the others, as is seen in Chapters 4 and 9.

The models discussed in this book vary in how faithful they are to known neuroanatomy and neurophysiology. The models discussed in Chapters 6–9 tend to be more neurobiologically correct or detailed than those discussed in Chapters 3 and 4, yet many of the models in Chapter 6–9 include parts that are modifications of subprocess models from the earlier chapters. The units, or nodes, in neural networks are often regarded as populations of neurons that are unified in some functional sense (e.g., those cells responsive to light in a particular part of the visual field, or to the hunger drive). Analogies of these units to averaged cell types in particular brain areas are quite close at times, more remote at other times. Hence, the book title uses the broad term *neural and cognitive modeling* to encompass models with different degrees of fidelity to real brain structure. The boundary between more and less "brain-faithful" networks is a fluid one. In fact, as the above example of associative learning shows, models that are more "cognitive" than "neural" sometimes lead to "neural" predictions that are later supported by data.

As neural networks have become popular, it is often asked how much neural network theory has really accomplished in its efforts to explain neurobiological data. The problems that neural network theory addresses are complex, and no single model has yet "cracked" the problem of categorization, or memory, or decision-making, or emotion. What has happened, instead, is that neural network theory has been part of a slow, steady increase in overall understanding of brain and cognitive functions. Hence, I believe that researchers in the field are justified in saying, in the words of the neural network pioneer Warren McCulloch, "Don't bite my finger, look where I am pointing" (McCulloch, 1965, Page xx).

The technological applications of neural networks are somewhat ahead of the biological modeling applications. Artificial neural networks have achieved considerable success in diverse areas, such as robotics, speech recognition and synthesis; decisions on whether to grant mortgage insurance; classification of some types of medical information, and classification of radar patterns. Yet in the last ten years the neurobiological modeling has begun to catch up to the technology.

Neural network models have led to experimental predictions, most notably in vision, motor control, and classical conditioning but to a lesser extent in more opaque areas such as decision-making and the effects of specific mental illnesses. Since the early 1990s, an increasing number of neuroscience and psychology laboratories are including computational modeling as part of their operations and actively testing some of the predictions of their models.

The scientific approach to knowledge, broadly speaking, argues that mental phenomena should have *some* mechanistic basis that will eventually be understandable by human beings. As Wiener (1954, p. 263) said, the faith of scientists is that nature (including mind) is governed by ordered laws, not by the capricious decrees of a tyrant like Lewis Carroll's Red Queen. Neural network modeling provides the best methodology now available for building mechanistic theories of mental and psychological functions.

Since all current neural models are subject to modification, this book is written to give the student or other reader hands-on experience in thinking about, simulating, and ultimately designing neural networks. It begins in Chapter 2 with a historical overview of major trends and the roots of current key ideas. Chapters 3 and 4 review the most important established models of associative learning and competition, respectively. Chapter 5 reviews some of the most significant recent findings in cognitive neuroscience. Chapters 6–9, building in part on some of the network structures introduced in Chapters 3–4, develop recent models (many of which Ashby and Hélie, 2011, would call CCN models) that reproduce some of those cognitive neuroscience results.

2

HISTORICAL OUTLINE

> Faithfulness to the truth of history involves far more than a research . . . into special facts. . . . The narrator must seek to imbue himself [*sic*] with the life and spirit of the time.
>
> Francis Parkman, *Pioneers of France in the New World*

> The abuse of truth should be as much punished as the introduction of falsehood.
>
> Blaise Pascal, *Pensées*

2.1. Digital Approaches

Neural network modeling as we know it today is partly rooted in the computer–brain analogy that captured the imagination of 1940s scientists, based on the fact that neurons are all-or-none, either firing or not firing, just as binary switches in a digital computer are either on or off. Since that time, neurophysiological data have indicated that the all-or-none outlook is oversimplified. Also, in a large number of the neural network models developed since then the functional units are neuron populations rather than single neurons. In spite of these technical advances, current approaches still owe many of their formulations to pioneers from the 1940s, such as McCulloch, Pitts, Hebb, and Rashevsky.

2.1.1. The McCulloch–Pitts Network

This inquiry essentially began with the classical study of all-or-none neurons by McCulloch and Pitts (1943). In this article, hidden under some elaborate symbolic logic, is a demonstration that any logical function can be duplicated

by some network of all-or-none neurons. That is, a neuron can be embedded into a network in such a manner as to fire selectively in response to any given spatiotemporal array of firings of other neurons in the network.

The rules governing the excitatory and inhibitory pathways in McCulloch–Pitts networks are the following:

1. All computations are carried out in discrete time intervals.
2. Each neuron[1] obeys a simple form of a *linear threshold law*: it fires whenever at least a given (threshold) number of excitatory pathways, and no inhibitory pathways, impinging on it are active from the previous time period.
3. If a neuron receives a single inhibitory signal from an active neuron, it does not fire.
4. The connections do not change as a function of experience. Thus the network deals with performance but not learning.

More general linear threshold laws are considered later in this section, in reference to the work of Rosenblatt (1962).

An example of an all-or-none neural network is reproduced in Figure 2.1. This network was designed by McCulloch and Pitts (1943) as a minimal model of the sensation of heat obtained from holding a cold object to the skin and then removing it. The cells labeled "*1*" and "*2*" are, respectively, heat and cold receptors on the skin, whereas heat and cold are felt when cells "*3*" and "*4*" fire, respectively. Each cell has a threshold of 2, hence it fires whenever it receives two excitatory (arrow) and no inhibitory (filled circle) signals from other cells active at the previous time.

In the network of Figure 2.1, if a cold object is presented and then removed, this means that at time 1, the only cell firing is Cell *2*. At time 2, Cell *A* fires because it receives two excitatory signals from Cell *2*. Since the cold has been removed, Cell *2* does not fire again, nor do any of the other cells in the network. At time 3, Cell *B* fires because it receives two excitatory signals from Cell *A*. At time 4, the two excitatory signals from *B* to *3* cause *3* to fire, meaning that heat is felt. The time sequence of firing patterns is shown in Table 2.1(a). In contrast, consider the same network's response to the cold object being on the skin continuously, as shown in Table 2.1(b). At time 2, Cells *2* and *A* will both be firing. At time 3, Cell *B* will not fire because the inhibitory signal from Cell *2* prevents *B*'s firing in response to *A*. Cell *4*, however, will fire because it receives excitation from *both* Cells *2* and *A*; hence, cold will be felt.

The McCulloch–Pitts model, although it uses an oversimplified formulation of neural electrical activity patterns, presages some issues that are still important in current cognitive models. For example, some of the best known modern connectionist networks contain three types of units or nodes – *input units*, *output units*, and *hidden units*. The input units react to particular data

features from the environment (e.g., "cold object on skin," "black dot in upper left corner," "loud noise to the right"). The output units generate particular organismic responses (e.g., "I feel cold," "the pattern is a letter A," "walk to the right"). The hidden units (a term popularized by Rumelhart & McClelland, 1986a) are neither input nor output units themselves but, via network connections, influence output units to respond to prescribed patterns of input unit firings or activities. The input–output–hidden trilogy can at times be seen as analogous to the distinction between sensory neurons, motor neurons, and all other neurons (interneurons) in the brain. At other times, though, a model neural network is designed to represent a small part of a larger behavioral process. The output may therefore not be a motor output but a particular internal state, such as a categorization or an emotion, which could be preparatory to a present or future motor response.

Note that, in the McCulloch–Pitts network of Figure 2.1, there are already input units (Cells *1* and *2*), hidden units (Cells *A* and *B*), and output units (Cells *3* and *4*). This distinction becomes explicit in more sophisticated linear threshold networks that are discussed below. In particular, the *perceptrons* developed by Rosenblatt (1962) contained units classified as "sensory," "associative," or "response."

Another cognitive issue raised by the "feel hot when cold is removed" network of Figure 2.1 is how to create output unit responses to given inputs that depend on the context of previous inputs. Specifically, this network responds to difference of the present input from a previous one; this may be called *temporal contrast enhancement*, by analogy with the *spatial contrast enhancement* (particularly observed in visual responses), which is a main topic

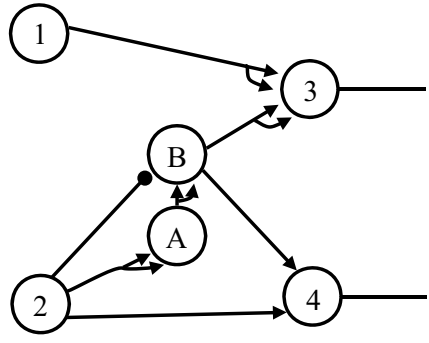


FIGURE 2.1 Example of an all-or-none network. Neurons labeled "*1*" and "*2*" are heat and cold receptors on skin. Heat and cold are felt when neurons "*3*" and "*4*" are active, respectively. Each neuron has threshold 2. A cold object held to the skin and then removed causes a sensation of heat.

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

of Chapter 4. Various forms of temporal contrast enhancement have been combined with learning in many neural network models (e.g., Barto, Sutton, & Watkins, 1990; Bear, Cooper, & Ebner, 1987; Brown, Bullock, & Grossberg, 1999; Grossberg, 1972b, 1972c; Grossberg & Schmajuk, 1987; Klopf, 1986; Seymour et al., 2004; Suri & Schultz, 2001; Sutton & Barto, 1981). Some of these networks model such psychological effects as a motor act becoming rewarding when it turns off an unpleasant stimulus (relief), the withholding of an expected reward being unpleasant (frustration), and the reward value of food being enhanced if the food is unexpected (partial reinforcement acquisition effect).

McCulloch and Pitts also confronted the issue of how memory is stored. Figure 2.2(a) shows a network of the McCulloch–Pitts type in which a neuron fires if a given input (say, a light) is on for three time units in a row. A similar network can easily be constructed to respond to any fixed number of consecutive occurrences of an input. Figure 2.2(b) shows a network in which a neuron is made to fire if the light has been on at *any time in the past*. Note that the mechanism for such memory storage is a reverberatory circuit. The concept of reverberation remains central to the understanding of memory today, and some advantages and limitations of the mechanism are discussed below.

McCulloch and Pitts noted the absence of a precise sense of timing in their model (1943): "the regenerative activity of constituent circles renders reference

Time	*Cell 1*	*Cell 2*	*Cell a*	*Cell b*	*Cell 3*	*Cell 4*
1	No	Yes	No	No	No	No
2	No	No	Yes	No	No	No
3	No	No	No	Yes	No	No
4	No	No	No	No	Yes	No
					FEEL HOT	

(a)

Time	*Cell 1*	*Cell 2*	*Cell a*	*Cell b*	*Cell 3*	*Cell 4*
1	No	Yes	No	No	No	No
2	No	Yes	Yes	No	No	No
3	No	Yes	Yes	No	No	Yes
						FEEL COLD

(b)

Table 2.1 Firings of neurons in the network of Figure 2.1 at successive time steps.

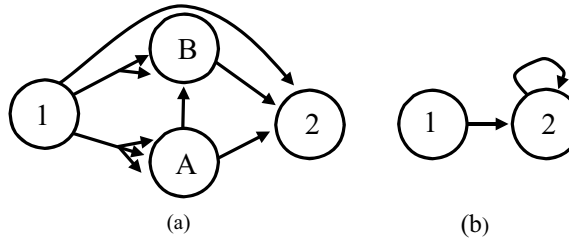


FIGURE 2.2 Two more all-or-none neural networks. In both networks, neuron "*1*" responds to a light being on. (a) Each neuron has threshold 3, and neuron "*2*" fires after the light has been on for three time units in a row. (b) Neuron "*2*," which has threshold 1, fires if the light has ever been on in the past.

indefinite as to time past" (p. 130). To them, this makes the model useful in certain ways: "This ignorance, implicit in all our brains, is the counterpart of the abstraction which renders our knowledge useful" (p. 131). Yet obviously a sense of timing is necessary for some other cognitive processes. For those processes, it is necessary to include, as later models do, the possibility of changing connection strengths over time.

2.1.2. Early Approaches to Modeling Learning: Hull and Hebb

At the same time that McCulloch and Pitts were developing a neural network formalism, psychologists were starting to consider mechanistic frameworks for studying learning and memory. This led to consideration of the issue of whether short-term memory (STM) can be distinguished from long-term memory (LTM). Hull (1943) proposed that the two memory processes involved the storage of two sets of traces. For example, consider the classic experiment of Pavlov (1927), where a bell is repeatedly paired with food until a dog salivates to the bell alone. After the experiment is stopped, conscious memory of the bell will be gone, since the dog is concentrating on other things. The memory of the *bell–food association*, however, will still be present, enabling the dog to salivate quickly on the next presentation of the bell. Hull thus distinguished between *stimulus traces* subject to rapid decay and *associative strengths* (or, in his terms, habit strengths) able to persist over a longer time period.

Hull's stimulus traces can be considered as the amounts of activity of particular nodes or functional units in a neural network. His associative strengths, then, are the strengths of connections between nodes. This suggests first that such connection strengths should change with experience, and second that they should correspond to some variable related to the *synapse*, or junction between neurons.

Hebb (1949) interpreted these memory issues with a theory that attempted to bridge psychology and neurophysiology. He declared that reverberatory

feedback loops, which had been suggested as a memory mechanism by McCulloch and Pitts (1943), could be a useful mechanism for STM but not for LTM. Concerning traces arising in such reverberatory loops, Hebb (1949) said: "Such a trace would be unstable. A reverberatory activity would be subject to the development of refractory states in the cells of the circuit in which it occurs, and external events could readily interrupt it" (p. 61). He was one of the first to recognize that a stable long-term memory depended on some structural change. But at the same time, he proposed (1949) that "A reverberatory trace might cooperate with the structural change and *carry the memory until the growth change is made*" (p. 62, author's italics).

Hebb went on (1949) to describe a hypothesis for the structural change involved in long-term memory:

> When the 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.
>
> (p. 62)

As for the nature of the structural change, Hebb proposed that if one cell repeatedly assists in firing another, the knobs of the synapse between the cells could grow so as to increase the area of contact (see Figure 2.3). The idea that learning is based on changes at neuronal connections goes back to Freud (1895/1953), who suggested it on intuitive grounds before enough neuroscience was known to provide a basis for it – in fact, before what we know now about synapses had been established. The notion of the synapse between two neurons as the nervous system's primary communication link was developed soon afterward by Sherrington (1906/1947), who coined the term "synapse" and established that neurons were physically separate from one another.

Neurophysiological data have suggested that actual growth of synaptic knobs can sometimes occur (e.g., Anderson et al., 1989; Bourne & Harris, 2008; Robinson & Kolb, 1999; Trommald, Hulleberg, & Anderson, 1996; Tsukahara & Oda, 1981). More frequently, as seen in Chapter 3, there has been experimental support for cellular and synaptic processes that do not involve gross structural changes but that alter the effective strength of connections in other ways. Such processes can embody an associative rule such as Hebb's for changes in connection strength between cells. This has led various neural network modelers, starting in the 1960s, to develop networks with rules whereby a connection weight (i.e., synaptic efficacy) increases with repeated pairing of presynaptic and postsynaptic activities; such rules are often called *Hebbian* rules in homage to Hebb's hypothesis.

There has also been extensive theoretical work on alternative rules for learning of connection weights and network modeling based on these rules.

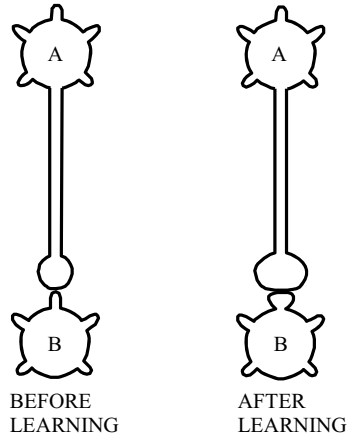


FIGURE 2.3 Diagram of Hebb's structural change hypothesis. The synaptic knob from *presynaptic* cell A to *postsynaptic* cell B gets larger after firing of A is repeatedly followed by firing of B.

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

For example, sometimes the connection weight changes as a function of *change* in either presynaptic or postsynaptic activity. Or sometimes the connection weight changes in a direction designed to make the network emit a desired response, typically a response determined at a different location in the network from the connection. More recently, learning rules have been developed whereby the weight changes are sensitive to the timing of presynaptic and/or postsynaptic spikes (see Section 3.5).

Various researchers commented that Hebb's rule proposed a way for connection strengths to increase, but the nervous system could become unstable if there was not a corresponding way for connection strengths to decrease. Though Hebb himself was aware of this issue, the first author to propose a rule of decrease complementary to Hebb's was probably Stent (1973), who also suggested detailed physiological mechanisms for implementing both Hebb's rule and his own. Stent's complementary rule was:

> When the presynaptic axon of cell A repeatedly and persistently fails to excite the postsynaptic cell B while cell B is firing under the influence of other presynaptic axons, metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is decreased.

Later work on physiological mechanisms for variations on both Hebb's and Stent's rules is discussed in Section 3.1.

In the early days of neural network modeling, considerable attention was paid to incorporating Hebb's rule and others for learning into a network of all-or-none neurons similar to that of McCulloch and Pitts. The modelers building adaptive networks of this variety included Rosenblatt (1962), Widrow (1962), and Selfridge (1959). In these networks, the McCulloch–Pitts form of the linear threshold law was generalized to laws whereby activities of all pathways impinging on a neuron are computed and the neuron fires whenever some weighted sum of those activities is above a given amount.

2.1.3. Rosenblatt's Perceptrons

The work of Rosenblatt was particularly influential and anticipated many of the themes of modern adaptive networks such as those of the PDP research group (cf. Rumelhart & McClelland, 1986a) and deep learning (Hinton, Osindero, & Teh, 2006; Schmidhuber, 2015). In fact, the latter type of network is often called *multilayer perceptrons*. The main function he proposed for his perceptrons was to make and learn choices between different patterns of sensory stimuli.

Rosenblatt set out to study the pattern classification capabilities of networks of sensory (*S*), associative (*A*), and response units (*R*) with various structures of active connections between units. Figure 2.4 shows examples of perceptrons with four possible connection structure types. These types are, in order, *three-layer series-coupled* (connections one-way from *S* to *A* to *R*); *multilayer series-coupled* (connections from *S* to one level of *A* to another level of *A* to *R*); *cross-coupled* (like three-layer series-coupled with the addition of cross links between *A*-units), and *back-coupled* (like series-coupled with the addition of feedback links from *R*- to *A*-units).

Rosenblatt first considered what he called elementary perceptrons (see the end of this chapter for the mathematical definition). An elementary perceptron is series-coupled, with connections only from *S*- to *A*-units and from *A*- to *R*-units, with only one R-unit.

Rosenblatt's book consisted of descriptions of a large number of mathematical and computer experiments on how well these different types of networks could either classify or generalize sensory patterns. The approach to modeling was described as *genotypic* rather than *monotypic*. These terms were defined as follows (Rosenblatt, 1962):

> Instead of beginning ("monotypic") with a detailed description of functional requirements and designing a specific physical system to satisfy them, this approach ("genotypic") begins with a set of rules for generating a set of physical conditions, and then attempts to analyze their common functional properties.
>
> (p. 22)

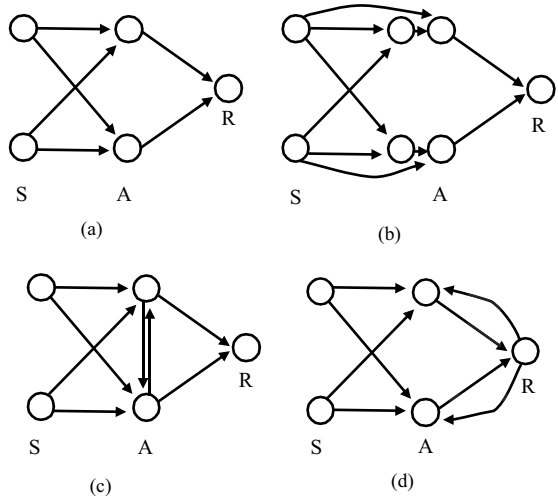


FIGURE 2.4 Examples of some classes of perceptrons: (a) three-layer series-coupled; (b) multilayer series-coupled; (c) cross-coupled; (d) back-coupled.

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

The learning rules for perceptrons, which Rosenblatt called the *reinforcement system*, were influenced by ideas of Hebb (1949). He distinguished two major types of reinforcement systems, *alpha* versus *gamma* systems. In the alpha system, all active connections terminating on a given active cell are changed by equal amounts, whereas inactive connections are not changed at all. In the gamma system, the total value of connection strengths is conserved, so that inactive connections are decreased while active ones are increased.

The amount of the connection change associated with reinforcement was a value δ determined by one of three *training procedures*. In a *response-controlled system*, the magnitude of δ is constant and its sign is determined by the response (that is, by the vector of R-element activities). In a *stimulus-controlled system*, the magnitude of δ is again constant but its sign is determined by the stimulus (that is, by the vector of S-element activities). In an *error-correcting* system, δ is 0 unless the response is determined elsewhere to be "incorrect." Also, reinforcement can be either *positive* or *negative*, that is, going in either the same direction as or the opposite direction to the current response.[2]

2.1.4. Some Experiments With Perceptrons

Rosenblatt (1962) ran simulation experiments in which these different types of perceptrons were taught to discriminate classes of stimuli. A number of distinctions were found between the capabilities of perceptrons with different

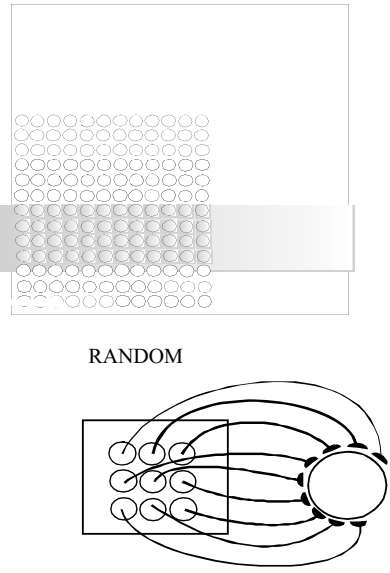
reinforcement rules and different training procedures, distinctions which are now mainly of historical interest. Not surprisingly, the perceptrons with error-correcting reinforcement converged faster than those with either stimulus-controlled or response-controlled reinforcement. Reinforcement rules of the error-correcting type were concurrently developed by Widrow and Hoff (1960) and are still used widely (e.g., Abdi, Valentin, Edelman, & O'Toole, 1996; Anderson & Murphy, 1986; Bullock & Grossberg, 1988, 1989; Cohen & Servan-Schreiber, 1992; Pineda, 1995; Stone, 1986).

As for the distinction between alpha and gamma reinforcement, the results of the simulation experiments were equivocal. A slight advantage was found for the gamma rule if the various stimuli presented were of unequal size or frequency, whereas the alpha rule seemed to carry some advantage if the system included an error correction mechanism. Conservation laws similar to the gamma rule have been used in more recent neural networks. Rosenblatt found that the conservation rule made the network's responses more likely to be stable. This same property was used in later neural network models by Malsburg (1973) and Wilson (1975), both of whom thought this "principle of constant synaptic strengths" could be explained in terms of conservation of some chemical substance at or near synapses. Synaptic conservation has continued to appear in more recent network models, such as the model by Choe and Miikkulainen (2004) of contour perception in the visual cortex.

In one of Rosenblatt's major experiments (see Figure 2.5), the *S*-units are arranged in a rectangular grid. Connections from *S*- to *A*-units are random, whereas all A-units connect to the single *R*-unit. The perceptron (elementary, series-coupled) was taught to discriminate vertical from horizontal bars; variants of this experiment are given in the exercises for this chapter.

Rosenblatt found that if all possible vertical and horizontal bars are presented to the elementary series-coupled perceptron, and the perceptron is reinforced positively for responding to the vertical bars and negatively for responding to the horizontal, then eventually the network gives the desired response reliably to each one. However, if only some of the vertical and horizontal bars are presented and positively or negatively reinforced, the series-coupled perceptron is unable to generalize its behavior to other vertical or horizontal bars that have not been presented. What generalization the network can do is based on location rather than on any more fundamental properties of the input patterns. In models of visual pattern discrimination, issues like translation invariance (ability to recognize a given pattern regardless of where it is in the visual field) remain difficult ones today. This property is exhibited by the Neocognitron of Fukushima (1980) and the What-and-Where filter inspired by architecture of the visual part of the cerebral cortex (Carpenter, Grossberg, & Lesher, 1998).

Inability to generalize is related to another weakness of series-coupled perceptrons: their inability to separate out parts (features) of a complex pattern. This means that, for a perceptron to perform categorizations, it needs an



excessively large number of nodes. Minsky and Papert (1969) remarked about a similar network that "along with its never-forgetting, it brings other elephantine characteristics" (p. 161). A third weakness is that these systems rely on a reinforcement signal external to the perceptron.

Further experiments and some theorems showed that generalization can be markedly improved by adding more connections to the perceptron. This can be accomplished either by interposing extra layers of associative units or by cross-coupling existing associative units. These additional connections also remove much of the perceptron's dependence on an external reinforcer. The separation of features from an overall pattern – for example, the classification of patterns on a rectangular grid ("retina") by whether there is a square in the center – proved more difficult for perceptrons. Preliminary simulations indicated that feature detection might be improved by back-coupling, that is, adding feedback from *R*-units to *A*-units.

Rosenblatt developed another idea that is a variant of back-coupling – namely, back propagation of errors. This procedure was later mathematically formalized by Werbos (1974; see also Werbos, 1993), modified in various ways by LeCun (1985) and Parker (1985), and then popularized by Rumelhart, Hinton, and Williams (1986). The conceptual principle of back propagation, as stated by Rosenblatt (1962), is:

> The procedure to be described here is called the "back-propagating error correction procedure" since it takes its cue from the error of the *R*-units, propagating corrections back toward the sensory end of the network if it fails to make a satisfactory correction quickly at the response end.
>
> (p. 292)

In other words, if some *A*-to-*R* connection strengths need to be corrected for satisfactory response, inferences can be drawn regarding which *S*-to-*A* connections need to be changed as well.

The mathematics of back propagation, and the possible biological basis for it, are discussed in Chapters 3 and 8. It is one of a variety of neural network schemas related to solving the artificial intelligence problem that Minsky (1961) called *credit assignment*, that is, deciding which part of a system is most responsible for an overall outcome and making the best corrections to the system for changing the outcome in the desired direction. Some more sophisticated credit assignment networks are discussed in Chapter 6, but were presaged by many of the inquiries in Werbos's (1974, 1993) dissertation. The more recently developed deep learning networks (e.g., Hinton et al., 2006; LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015), discussed in Chapter 8, have expanded the capabilities of back propagation into larger networks, including refinements of the learning rules that are arguably more biologically plausible than the original ones (e.g., Lillicrap, Cownden, Tweed, & Akerman, 2016).

2.1.5. The Divergence of Artificial Intelligence and Neural Modeling

From the late 1960s to the early 1980s, researchers in artificial intelligence largely abandoned neural networks of the linear threshold variety in favor of heuristic computer programs; this history was discussed in Levine (1983, Section 3.2). During this period, other linear threshold models contemporary with Rosenblatt's had some, although relatively minor, impact on artificial intelligence and neural modeling. Widrow (1962) developed the ADALINE (for "adaptive linear neuron"). Contrary to its author's intentions, this work was more influential among electrical engineers doing signal processing than among any group directly studying intelligent systems (Widrow, 1987). Selfridge (1959) developed the PANDEMONIUM model, which got its name from the different modules called "demons," each of them feature detectors with access to partial information from the environment. Decisions of the entire network were based on a weighted average of the decisions of the different demons. The demon approach had some influence on some early computational models of specific brain areas such as the reticular formation (Kilmer, McCulloch, & Blum, 1969) and the hippocampus (Kilmer & Olinski, 1974).

However, at that time, the detailed physiology of these brain areas was not understood well enough for such models to be widely accepted. Selfridge's work also inspired some of the abstract computational geometry of Minsky and Papert (1969).

Minsky and Papert (1969) developed their outlook in a book titled *Perceptrons*. The title was inspired by Rosenblatt's previous work, but the devices that Minsky and Papert studied are not exactly a subclass of Rosenblatt's. These abstract machines do, however, have parts that correspond loosely to "sensory," "associative," and "response" areas. The Minsky–Papert perceptron starts with a *retina*, which is a grid consisting of small squares, each of which is at any time active or inactive ("light" or "dark"). Downstream from the retina are units that compute *partial predicates*. Each partial predicate outputs a value of 1 or 0 based on some rule depending on the activity or nonactivity of units in a given subset of the retina. The maximum size of that subset over all predicates is called the *order* of the perceptron. Finally, there is a decision-making unit that computes a linear function of those predicate outputs and responds when that linear function is above some threshold.

Minsky and Papert proved that their abstract form of the perceptron can learn any classification of patterns on its retina. However, many of the theorems stated that, for a perceptron to make some geometrically important classifications, the order of the perceptron has to get arbitrarily large as the size of the retina increases. Theorems of this sort were widely interpreted as discrediting the utility of perceptron-like devices as learning machines. But Minsky later said that, in retrospect, the discrediting of perceptrons seemed like an overreaction (Rumelhart & McClelland, 1986a, Vol. 1, pp. 158–159).

Moreover, some of the visual discriminations that are difficult for perceptrons are also difficult for humans. For example, consider the distinction between connected and disconnected figures, as shown in Figure 2.6. It is easy for the unaided eye–brain combination to tell that a filled-in circle is connected,

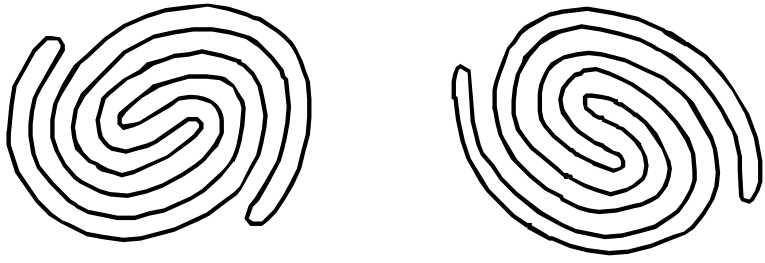


FIGURE 2.6 The finite-order perceptrons of Minsky and Papert (1969) cannot tell that the curve on the left is connected, whereas the curve on the right consists of two disjoint arcs. Can *you* tell that by visual inspection?

Source: Reprinted from Minsky and Papert, 1969, with permission of MIT Press.

whereas a pattern of two filled-in circles side by side is disconnected. But it is next to impossible for the eye and brain to tell which of the two convoluted patterns in Figure 2.6 is connected and which is disconnected without some help from finger tracing.

The models of the group of researchers that called themselves PDP, which originated about 1981 and most of which are summarized in Rumelhart and McClelland (1986a), recaptured some of the threads from Rosenblatt's work. They showed that some of the distinctions that are impossible for Minsky and Papert's kind of simple perceptrons (such as between inputs that activate an odd versus an even number of retinal units) can be made by perceptrons with additional "hidden unit" layers (cross-connections) *and* nonlinear activation functions. Some of this work is discussed in Chapters 3 and 8.

2.2. Continuous and Random Net Approaches

While the cybernetic revolution was stimulating discrete (digital) models of intelligent behavior, there was a concurrent proliferation of results from both experimental neurophysiology and psychology. Some of these experimental results stimulated the development of continuous (analog) neural models. This section reviews continuous approaches, random net approaches, and finally some partial syntheses of continuous and discrete approaches.

2.2.1. Rashevsky's Work

One of the pioneers in the development of continuous neural models was Rashevsky. The best exposition of his outlook was in his 1960 book *Mathematical Biophysics*. The first edition of this book had been written in 1938 – five years before the seminal article of McCulloch and Pitts (1943). Subsequently, the evolution of his thinking had been altered by the McCulloch–Pitts article (which was published in a journal that Rashevsky himself founded and edited).

In most applications of mathematics to physical phenomena, including the biophysics of electrical current flow in single neurons, there are variables that are not all-or-none but may take on any of a range of values. Hence, such processes are typically modeled using differential equations, which are equations describing continuous changes over time in an interacting collection of physical variables. (For those desiring a "primer" in differential equations and their utility in neural network modeling, please refer to Appendix 1.) Rashevsky (1960) described how the earlier edition of his book had used differential equations to model various data in the psychophysics of perception. These data included the relation of reaction times to stimulus intensities, and the just noticeable differences among intensities.

Rashevsky went on to describe how his thinking had been influenced by the article of McCulloch and Pitts (1943), which used all-or-none neurons. He stated (Rashevsky, 1960) that "the proper mathematical tool for representing the observed *discontinuous* interaction between neurons was not the differential equation but the Boolean Algebra or Logical Calculus" (p. 3). Yet it was difficult to model the observed psychophysical data using the McCulloch–Pitts postulates. This paradox was resolved with the observation that such behavioral data reflect the combined activity of very large numbers of neurons. Hence, "the discontinuous laws of interaction of individual neurons lead to a sort of average continuous effect which is described by the differential equations postulated originally" (also from p. 3 of Rashevsky, 1960).

The reconciliation effected by Rashevsky and others between continuous and discrete models is still in common use today. The description in terms of average activity is in line with the trend toward building models based on functional units or nodes that may represent large numbers of neurons (see Chapter 1). This is an idea that actually dates back to Hebb (1949), who proposed that significant percepts or concepts are coded not by neurons but by groups of neurons that he called *cell assemblies*.

The boundaries of "functional units" or "cell assemblies" in actual mammalian brains have yet to be defined precisely. Edelman (1987) speculated that units on the order of several thousand neurons in size encode stimulus categories of significance to the animal. Abeles (1991), Burnod (1988), and others have stressed the functional importance of cell assemblies in the mammalian cerebral cortex, or outermost brain layer, which are arranged roughly in columns. Other theorists (e.g., Crick & Koch, 1990; Koch & Crick, 1994; Milner, 1974) have speculated that significant concepts or percepts could be coded by the synchronized electrical activity of large distributed groups of neurons, an idea that has received some neurophysiological support (Eckhorn et al., 1988; Gray, König, Engel, & Singer, 1989; Gray & Singer, 1989). A more recent and somewhat related idea is the formation of "cognits" (Fuster & Bressler, 2012), which are interconnected networks that encode the contents of long-term memory.

Whatever its neurobiological mechanism turns out to be, averaging across many neurons also allows the use of deterministic equations for unit activity even if the behavior of single neurons includes a random component. A neuron fires (i.e., transmits an impulse or, more technically, an *action potential*) if its transmembrane voltage exceeds a value called the *threshold* (see Appendix 2 for details). This threshold is widely believed to vary according to some probability distribution, such as the Gaussian or normal distribution (see below). Neural models frequently average such random single-neuron effects across the functional groups of neurons that constitute network nodes; hence, the interactions between nodes become deterministic. In addition, many models

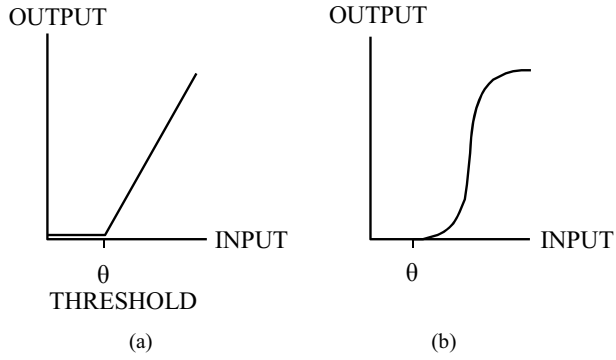


FIGURE 2.7 Schematic of linear (a) and sigmoid (b) functions of suprathreshold activity.

average random effects over short time intervals, so that the node activity variable is interpreted as representing a firing frequency rather than a voltage.

Rashevsky, however, made some simplifying assumptions about the neural averaging process. For example, he assumed that the frequency of impulses transmitted by a neuron is, on the average, a linear function of the cell's suprathreshold activity (see Figure 2.7a). That has proved to be a useful assumption for some neural models of sensory transduction, such as the model of the horseshoe crab retina developed by Hartline and Ratliff (1957). Yet averaging considerations can also lead one to consider input–output functions that are nonlinear, such as *sigmoid* functions (Figure 2.7b). As shown in Figure 2.8, if the firing threshold of an all-or-none neuron is described by a random variable with a Gaussian (normal) distribution, then the expected value of its output signal is a sigmoid function of activity. For this reason sigmoids have

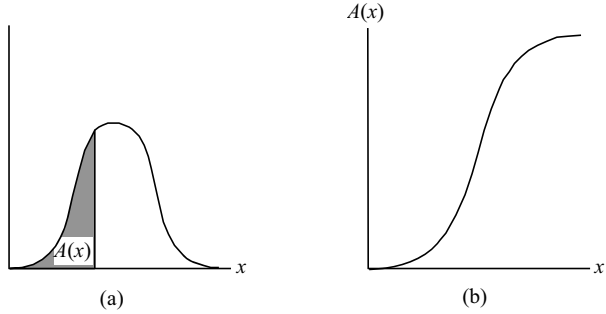


FIGURE 2.8 One possible biological basis for sigmoid functions: (a) Gaussian (normal) distribution of firing thresholds. If the activity (transmembrane voltage in the case of a single cell) is x, the node fires if the threshold is less than x. The probability of that happening is the area under the shaded part of the curve. (b) Schematic graph of the area $A(x)$ in (a) as a function of x.

become increasingly popular in recent neural models, such as the on-center off-surround models discussed Chapter 4 and the back propagation models discussed in Chapters 3 and 7. Also, there has been some physiological verification of sigmoid input–output functions at the neuron level (Brozović, Abbott, & Andersen, 2008; Kernell, 1965; Rall, 1955).

Despite his simplifications, Rashevsky inspired a generation of models that incorporate known neural phenomena into large networks of neurons connected more or less at random. One of these phenomena is the graded (not all-or-none) electrical potentials that occur at the dendrites of a neuron in response to all-or-none action potentials at other cells connected to it. Another is the *refractory period*, the short period of time in which a cell that has just fired (had an action potential) must remain inactive. (A historical outline of some relevant experimental findings appears in Katz, 1966.) Some of these models incorporated the averaging considerations described above, but others used units that were explicitly treated as single neurons.

2.2.2. Early Random Net Models

Many of the early random net models were discussed in the last sections of the review article by Harmon and Lewis (1968). The first attempts at random net modeling include only excitatory connections and no inhibitory ones. The absence of inhibition in model networks, which is unrealistic from the standpoint of known neuroanatomy, also led to unrealistic patterns of electrical activity. The excitatory nets developed by Beurle (1956) and Ashby, Foerster, and Walker (1962) tend, as time becomes large, to approach one of two extremes of activity: maximal activity leading to saturation of the entire net, or quiescence. The intermediate level of activity found in actual brains was not modeled by these nets. Griffith (1963a, 1963b, 1965) showed that, in this random net framework, stable submaximal activity is possible if inhibition is included.

In the years following Griffith's articles, other modelers tried to develop general theories for random neural networks with both excitatory and inhibitory connections. Most of these theories were based on differential or difference equations that include probabilistic terms. In addition, there were some neural net models inspired by specific formalisms from other scientific fields. Examples are models derived from statistical mechanics (Cowan, 1970) and from nonequilibrium thermodynamics (Freeman, 1975a; Katchalsky, Rowland, & Blumenthal, 1974; Prigogine, 1969). Application to neural network modeling of analogies with other fields remained popular for several decades. Some neural networks, for example, have been described as arrays of two-state units. This has led to analogies with the physics of *spin glasses*, which are structures with an array of magnetic spins that have one of two possible values (Amit, Gutfreund, & Sompolinsky, 1985; Chowdhury, 1986; Hopfield, 1982).

Yet analogies are limited by the fact that many nervous system properties are uniquely neural, brainlike, or cognitive. Hence, the further development of continuous and random models since the early 1970s has been influenced less by specific mathematical structures than by neuroanatomical, neurophysiological, and behavioral data. In particular, some data have indicated that brain connections may be random *within* certain neural populations and specific *between* these populations.

2.2.3. Reconciling Randomness and Specificity

The classic experiments of Lashley (1929) showed that many psychological functions, such as ability to remember specific events, are retained after extensive brain lesions. Lashley's experiments were among the first to inspire the idea, by now common, that representations of events are distributed throughout the brain rather than localized. Other experiments showed, however, that specific connections are important for other functions. Mountcastle (1957) found that the somatosensory (touch-sensitive) area of the cerebral cortex includes a well-organized topographic encoding of the body. Similarly, Hubel and Wiesel (1962, 1965) found that cells in the visual area of the cortex are organized into columns that code specific retinal positions or line orientations. (It is important to note, however, that visual and somatosensory maps are modifiable; the somatosensory maps, at least, can be altered even in adult life. If the connection to a given area of the cortex from the retinal or body area it would normally code is either cut or inactivated, the same area of cortex can learn to code a different, nearby area. Some of this evidence is summarized in Edelman, 1987).

The paradox between the Lashley data and the Hubel–Wiesel or Mountcastle data is resolved by means of a principle described in Anninos, Beek, Csermely, Harth, and Pertile (1970) as "randomness in the small and structure in the large" (p. 121). This section considers some models whose equations are explicitly based on this principle. The same principle is implicit in many models discussed in the next two chapters. The latter models use purely deterministic equations at the population level that reflect the averaging over large ensembles of probabilistic effects at the single-cell level. Some of the more biologically sophisticated models discussed in Section 3.5 and Chapters 6–8 bring back single-cell details such as spiking.

The article of Anninos et al. (1970) is one of a series of related articles (e.g., Anninos, 1972a, 1972b; Harth, Csermely, Beek, & Lindsay, 1970; Wong & Harth, 1973). In this series of models, neurons are organized with random connectivities into "netlets," and netlets in turn are organized deterministically into larger nets. Evidence for such netlets was found, for example, in the organization of the somatosensory and visual areas of the cortex into functional columns (Hubel & Wiesel, 1962, 1965; Mountcastle, 1957).

Using many cell properties such as refractory periods, Anninos et al. (1970) derived an expression for the expected activity (defined as fractional number of neurons firing) at (discrete) time $n + 1$ as a function of activity at time n. The crucial variable for determining long-term behavior is a parameter δ describing the number of excitatory postsynaptic potentials (here from within a netlet) needed to cause a cell to fire in the absence of inhibitory inputs. If δ is very small, netlet activity always tends to a unique positive stable steady state. If δ is very large, netlet activity always tends to 0. If δ is in a middle range, there are two stable steady states, one quiescent and one active, and a threshold exists for reaching the active state.

Anninos (1972a) pursued these principles of network organization further with simulations of multi-netlet nets. He found, for example, that the dependence of activity of a single netlet in such a network on some external input can exhibit *hysteresis cycles*. That is, the effect of an input can depend on the past history of stimulation. He hinted, without giving details, that such hysteresis could be a mechanism for short-term memory.

Amari (1971, 1972, 1974) described random networks by means of differential or difference equations with two variable parameters – averaged connection weight and averaged threshold. Depending on the values of these two parameters, the network can have either a single stable steady state, many, or none. If the system has excitatory and inhibitory subnetworks, there can be oscillations of very long period. Amari's systems also modeled association of ideas, by means of connection weights.

A confluence of random net modeling with experimental data occurred in the work of Freeman (1972a, 1972b, 1975a, 1975b), much of which led to models of the olfactory cortex. He laid out some general principles for forming waves from pulses in large neural masses, and showed how this neural mass theory could be used to model EEG (brain wave) patterns and predict their frequencies. He continued this general line of work to recent years, with some results indicating that EEG patterns in the olfactory cortex tend to be chaotic (in the mathematical sense) in the absence of an odorant stimulus but synchronized in the presence of an odor (Freeman, 1992; Kozma & Freeman, 2009; Skarda & Freeman, 1987).

Both random and chaotic elements are common in contemporary neural network models, but there are also models in which random elements at the neuron level are averaged into deterministic models at the neural population level. A mathematical justification for this averaging process, using stochastic differential equations, was given by Geman (1979, 1980). In the deterministic approach, the networks often have particular connection patterns suggested by the cognitive task involved (such as associative learning, pattern storage, conditioning, or categorization).

The next two chapters will discuss in turn some of the early models of associative learning (in Chapter 3) and competition (in Chapter 4), two of the

principles mentioned in Chapter 1. Most of these models use deterministic differential or difference equations. Many of them were developed in the 1960s, 1970s, and early 1980s by pioneers who are still active in the neural network field as of this writing. Because of the state of knowledge in neuroscience at the time, these models were less closely tied to neural data than typical current models, some of which include features such as calcium-dependent spiking, spike timing–dependent plasticity, and enzymes that mediate neurotransmitter function. Yet many of the more "abstract" models of Chapters 3 and 4 provided "building blocks" for the more biologically realistic models of complex processes discussed in Chapters 6–9.

A few of Rosenblatt's definitions

(Numbers are the ones from Rosenblatt's book)

Definition 6: A *sensory unit* (*S*-unit) is any transducer responding to physical energy (e.g., light, sound, pressure, heat, radio signals, etc.) by emitting a signal which is some function of the input energy. The input signal at time t to an *S*-unit s_i from the environment, W, is symbolized by $c^*_{wi}(t)$. The signal that is generated at time t is symbolized $s^*_i(t)$.

Definition 7: A *simple S-unit* is an *S*-unit which generates an output signal $s^*_i = +1$ if its input signal, c^*_{wi} exceeds a given threshold, θ_i, and 0 otherwise.

Definition 8: An *association unit* (*A*-unit) is a signal-generating unit (typically a logical decision element) having input and output connections. An *A*-unit a_j responds to the sequence of previous signals c^*_{ij} received by way of input connections c_{ij}, by emitting a signal $a^*_j(t)$.

Definition 9: A *simple A-unit* is a logical decision element, which generates an output signal if the algebraic sum of its input signals, α_i, is equal or greater than a threshold quantity, $\theta > 0$. The output signal a^*_i is equal to +1 if $\alpha_i \geq \theta$ and 0 otherwise. If $a^*_i = +1$, the unit is said to be *active*.

Definition 10: A *response unit* (*R*-unit) is a signal-generating unit having input connections, and emitting a signal that is transmitted outside the network (i.e., to the environment, or external system). The emitted signal from unit r_i will be symbolized by r^*_i.

Definition 11: A *simple R-unit* is an *R*-unit that emits the output $r^* = +1$ if the sum of its input signals is strictly positive, and $r^* = -1$ if the sum of its input signals

is strictly negative. If the sum of the inputs is zero, the output can be considered to be equal to zero or indeterminate.

Definition 12: *Transmission functions* of connections in a perceptron depend on two parameters: the *transmission time* of the connection, τ_{ij}, and the *coupling coefficient* or *value* of the connection, v_{ij}. The transmission function of a connection *cij* from *ui* to *uj* is of the form: $c^*_{ij}(t) = f[v_{ij}(t), u^*_i(t - \tau_{ij})]$. Values may be *fixed* or *variable* (depending on time). In the latter case, the value is a *memory function*.

Definition 22: A *simple perceptron* is any perceptron satisfying the following five conditions:

1. There is only one *R*-unit, with a connection from every *A*-unit.
2. The perceptron is series-coupled, with connections only from *S*-units to *A*-units, and from *A*-units to the *R*-unit.
3. The values of all sensory to *A*-unit connections are fixed (do not change with time).
4. The transmission time of every connection is either zero or equal to a fixed constant, τ.
5. All signal-generating functions of *S*-, *A*-, and *R*-units are of the form $u^*_i(t) = f(\alpha_i(t))$, where $\alpha_i(t)$ is the algebraic sum of all input signals arriving simultaneously at the unit u_i.

Detailed Description: Perceptron to Discriminate Vertical versus Horizontal

This is a description of the time course of Rosenblatt's simulation of teaching an elementary perceptron to distinguish between 20-by-4 vertical bars and 20-by-4 horizontal bars. As in Figure 2.5, S-units are arranged in a 20 × 20 grid ("retina"). There are nine A-units and one *R*-unit. Only the *A*-to-*R* connections are modifiable.

Each *A*-unit receives eight excitatory and two inhibitory connections from *S*-units, and one needs to program a random number generator to find out *which S*-units they come from! The connection strengths v_{ij} are +1 for excitatory pathways and – 1 for inhibitory pathways. For each of the nine *A*-units, a program similar to the one listed in Figure 2.9 must be used ten times, one for each connection from an *S*-unit, to generate a random number uniformly distributed between 0 and 1. Then the random number generated is multiplied by 400 and truncated to an integer, and 1 is added to get an integer between 1 and 400. The last integer determines which *S*-unit connects to the current *A*-unit, the *S*-units in the first row being numbered 1 through 20, the second row 21 through 40, and so forth. More

than one connection to that unit can come from the same *S*-unit. But the location of the *S*-to-*A* connections, once set, remains fixed throughout the simulations.

For example, suppose the random number generator applied ten times yields .3171, .0295, .3246, .4878, .9135, .7076, .3168, .5040, .0511, and .2607. Then the locations obtained for the S-unit connections will be 127, 12, 130, 196, 366, 284, 127 (again), and 202 (all excitatory), then 125 and 1 (both inhibitory).

The input stimuli are horizontal or vertical bars of width four. The connectivity within the *S* grid is *toroidal*: that is, the top row is considered to be adjacent to the bottom row and the leftmost column to the rightmost column. The topmost horizontal bar activates units 1 through 80, the second horizontal bar activates units 21 through 100, and so on, down to the twentieth and last horizontal bar, which activates the bottom row and the top three rows of units, that is units 381 through 400 and 1 through 60. Likewise, the leftmost vertical bar activates the left four columns of units, that is, units 1, 21, 41, . . . , 381, 2, 22, 42, . . . , 382, 3, 23, 43, . . . , 383, 4, 24, . . . , 384, up through the last vertical bar, which activates units 20, 40, . . . , 400, then 1, 21, 41, . . . , 381. OBJECT: To teach *R* to respond positively to vertical, negatively to horizontal.

If the *i*th *S*-unit is activated, then $s_i^*(t) = 1$, otherwise $s_i^*(t) = 0$. An *A*-unit computes $a_j = \sum_i s_i^* v_{ij}$, and its activity $a_j^*(t) = 1$ if $\alpha_j > 2$, 0 if $\alpha_j \leq 2$, with 2 being set as the threshold for *A*-unit activation.

For example, suppose we are looking at the *i*th *A*-unit, and that the input is the horizontal bar activating rows 5 through 8 of the retina. This means that *S*-units 81 through 160 are activated, so $s_i^*(t) = 1$ for *I* between 81 and 160, and 0 otherwise. The *S*-to-*A* connection weight $v_{ij} = 1$ for $I = 127$ (twice) and 130 (once) and –1 for $I = 105$, because those are locations of excitatory and inhibitory connections. Those being the only *S*-units for which both $s_i^*(t)$ and v_{ij} are nonzero, $a_j = \sum_i s_i^* v_{ij} = 1(1) + 1(1) + 1(1) + 1(-1) = 2$, so $a_j^*(t) = 0$, that is, that *A*-unit is not activated.

At the start of each run, the *A*-to-*R* connection strengths w_j are set to values that are randomly (uniformly) distributed between –1 and 1. Error correction and alpha reinforcement are used. That is, whenever a horizontal line is input and *R* (incorrectly) responds positively, any w_j that are positive while $a_j^*(t) = 1$ are reduced by an amount δ; whenever a vertical line is input and *R* (incorrectly) responds negatively, any *wj* that are negative while $a_j^*(t) = 1$ are increased by the same value δ. *R* in turn responds negatively if $\Phi = \sum_i a_j^* w_j \leq 0$, and positively if $\Phi > 0$. Time delays can all be set to 0.

Definition 23: An *elementary perceptron* is a simple perceptron with simple *R*- and *A*-units, and with transmission functions of the form $c_{ij}^*(t) = u_i^*(t-\tau)v_{ij}(t)$.

The concepts relating to reinforcement are defined precisely in pages 88–92 of Rosenblatt (1962). We list only the definitions for positive, negative, alpha system, and error-correcting reinforcement:

Definition 33: *Positive reinforcement* is a reinforcement process in which a connection from an active unit u_i that terminates on a unit u_j has a value changed by a quantity $\Delta v_{ij}(t)$ (or at a rate dv_{ij}/dt), which agrees in sign with the signal $u_j^*(t)$.

Definition 34: *Negative reinforcement* is a reinforcement process in which a connection from an active unit u_i that terminates on a unit u_j has its value changed by a quantity Δv_{ij} (or at a rate dv_{ij}/dt) that is opposite in sign from $u_j^*(t)$.

(Note: The "active units" u_i in the above definitions could be either *A*-units or *R*-units.)

Definition 37: *Alpha system reinforcement* is a reinforcement system in which all active connections c_{ij} which terminate on some unit u_j (i.e., connections for which $u_i^*(t-\tau)$ is not equal to 0) are changed by an equal quantity $\Delta v_{ij}(t) = \delta$ or at a constant rate while reinforcement is applied, and inactive connections ($u_i^*(t-\tau) = 0$) are unchanged at time *t*.

Definition 41: An *error-correcting reinforcement system* (error correction system) is a training procedure in which the magnitude of δ (see Definition 37) is 0 unless the current response of the perceptron is wrong, in which case the sign of δ is determined by the sign of the error. In this system, reinforcement is 0 for a correct response, and *negative* (see Definition 34) for an incorrect response.

ADALINE Equations

The ADALINE network of Widrow (1962) is a supervised learning network. In the ADALINE model, a set of bipolar (1 or –1) inputs is filtered through a corresponding set of adaptive weights, and the sum of the weighted inputs is then compared with a desired output. Then error-correcting reinforcement is applied.

The ADALINE equations are as follows. Let w_i, $I = 0, \ldots, n$ represent the corresponding weights. Let I_i, $I = 1, \ldots, n$ represent specific inputs, and I_0 a constant ("bias") input equal to 1. Then the actual output, called *y*, is

$$\begin{aligned} 1 \text{ if } S &= \sum_{i=0}^{n} I_i w_i \geq 0 \\ -1 \text{ if } S &= \sum_{i=0}^{n} I_i w_i < 0 \end{aligned} \tag{2.1}$$

Let y_0 be a desired output; for example, if the network is trained to learn the logical "AND" operation, $n = 2$, and the inputs are 1 and 1, the desired output is also 1. Then, at each time step, weights are updated according to the rule

$$\Delta w_i = a(y_o - y)\frac{I_i}{n+1} \tag{2.2}$$

The network runs through all the input vectors in sequence, each time adding to each weight w_i the corresponding value Δw_i shown in Equation (2.2). Weights are changed until the network has learned the desired output to every bipolar input vector; that is, until the error term in (2.2), y_0–y, is 0 for each of the inputs.

Exercises for Chapter 2

○1. Hebb's rule for synaptic modification states that strength of a connection will increase if activities of the two connected units are both high at the same time (with suitable delays, perhaps). Other modelers have proposed alternative rules (sometimes called "differential Hebbian") whereby connection strengths increase when activity of one unit is coupled with *change* in activity of the other unit, or when changes in activities of both units are coupled. Give some possible advantages and disadvantages of Hebbian versus differential Hebbian rules for network models of learning. You may also suggest modifications of either type of rule.

**2. Design a McCulloch–Pitts network with heat and cold receptors and a cell that fires after the sequence "heat–cold" or the sequence "cold–heat" but nothing else. Assume that each cell takes exactly one time step to compute its output, and that cold and heat cannot be simultaneously felt at the same time step.

Design another McCulloch–Pitts network, possibly a modification of the first one, so that the last cell fires after alternating sequences of three – "heat–cold–heat," "cold–heat–cold," "heat–cold–neither," or "cold–heat–neither" – but not after sequences that include repeats – "heat–cold–cold" or "cold–heat–heat."

* 3. Using the definitions given in this chapter and the algorithm shown in Figure 2.9, do six runs of the Rosenblatt elementary perceptron that learns to respond positively to 20-by-4 vertical bars and negatively to 20-by-4 horizontal bars, as described in the box preceding these exercises. Do two runs with each of three different δ values (2.0, .5, and .1). Present each of the horizontal and vertical bars, in turn, in any order, repeatedly. It will probably eventually learn to classify all of them correctly but due to the randomness it is not guaranteed to happen. But see how the learning rate depends on δ. (Again, because of the randomness, there is no guarantee

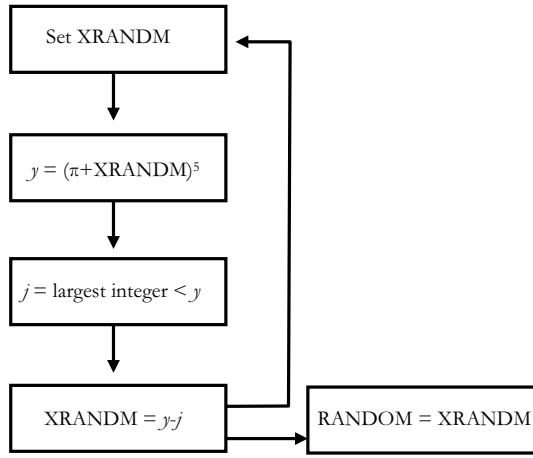


FIGURE 2.9 Generic program segment that, starting from any initial "seed" number between 0 and 1, generates a different random number in that interval on each pass.

about the outcome. But study the tradeoff between the effects of too small or too large a learning rate.)

*4. Do the same simulation as in Exercise 4 but teach the network only half of the bars and then present one it has not learned. No generalization should occur.

○*5. Design a simulation in which a perceptron is trained to discriminate between two types of figures. Examples would be a square of a given size versus a triangle of a given size (that is, translates of a fixed square and a fixed triangle anywhere along the grid) or a square and a diamond. Another example would be to discriminate whether a figure does or does not contain the letter "X."

*6. Simulate the ADALINE network of Widrow (1962) defined by Equations (2.1) and (2.2) and the surrounding text, with the number n of nonbias inputs equal to 2 and the learning rate a to 1. Set the initial values of each of the weights w_0, w_1, and w_2 to random values uniformly distributed between 0 and 1.

(a) Teach the network the logical "AND" operation, which maps

$(1, 1) \rightarrow 1$
$(1, -1) \rightarrow -1$
$(-1, 1) \rightarrow -1$
$(-1, -1) \rightarrow -1$.

Show that this can be learned in two passes through the sequence of four input vectors.

(b) Teach the network the logical "OR," which maps

$(1, 1) \rightarrow 1$
$(1, -1) \rightarrow 1$
$(-1, 1) \rightarrow 1$
$(-1, -1) \rightarrow -1.$

Show that this can be learned in two passes through the sequence.

(c) Teach the network the logical "NAND," which always gives the sign opposite to the one given by the "AND":

$(1, 1) \rightarrow -1$
$(1, -1) \rightarrow 1$
$(-1, 1) \rightarrow 1$
$(-1, -1) \rightarrow 1.$

Show that this can be learned in three passes through the sequence.

(d) Show that the network *cannot* learn the "exclusive OR," which maps

$(1, 1) \rightarrow -1$
$(1, -1) \rightarrow 1$
$(-1, 1) \rightarrow 1$
$(-1, -1) \rightarrow -1,$

by going through the sequence of training inputs and getting an infinite loop. (The exclusive OR can be learned by multilayer nonlinear networks, as will be seen in Chapter 7).

○ 7. The controversy over local versus distributed representations of concepts in the brain, mentioned briefly in Section 2.2.3 of this chapter, is still present in the neuroscience community. Neurons that respond selectively to specific objects are commonly known as *grandmother cells*, from the idea of a neuron that lights up when one recognizes their own grandmother. The existence and/or utility of grandmother cells is still fiercely debated (see, e.g., Bowers, 2017, and Thomas & French, 2017). From a reading of those articles or related ones, come to your own conclusions about this controversy.

Some Additional Sources

Early Neural Network Models (Original Articles or Reviews)

Anderson, Pellionisz, and Rosenfeld (1990); Anderson and Rosenfeld (1988); Palm (1982); Sejnowski (1976); Steinbuch (1961, 1990); Widrow, Pierce and Angell (1961).

Notes

1. In most of the models discussed in this book, network elements are called "nodes" or "units" rather than "cells" or "neurons." The exception is made for the McCulloch–Pitts network because their network is directly inspired by the all-or-none firing properties of neurons.
2. This usage differs from the standard usage of experimental psychologists. In psychology, negative reinforcement refers to a stimulus whose removal is rewarding. What Rosenblatt called negative reinforcement, psychologists call punishment.

3

ASSOCIATIVE LEARNING AND SYNAPTIC PLASTICITY

> The present contains nothing more than the past, and what is found in the effect was already in the cause.
>
> Henri Bergson, *L'Évolution Créatrice*

> The mind is slow in unlearning what it has been long in learning.
>
> Seneca, *Troades*

3.1. Physiological Bases for Learning

Recall from Section 2.1 the contribution of Hebb (1949) to the bridging of psychology and neurophysiology. Hebb proposed on psychological grounds the existence of synaptic modifications during learning, in the absence, then, of any physiological evidence for such modifications. Since that time, it has been experimentally demonstrated that correlated activity at the pre- and postsynaptic cells of many synapses in animal nervous systems alters the efficacy of the synapse in causing action potentials at the postsynaptic cell.

The first widely recognized demonstration of synaptic modifiability (plasticity) in a living animal was the work of Eric Kandel and his colleagues on the sea slug *Aplysia californica*, which led to Kandel receiving the 2000 Nobel Prize in Physiology and Medicine. In particular, Kandel and Tauc (1965) discovered in *Aplysia* a mechanism called *heterosynaptic facilitation*. Heterosynaptic facilitation has been defined (Byrne, 1987) as "a change in synaptic efficacy (or cellular excitability) in one neuron as a result of release of a modulatory transmitter from another neuron" (p. 354). This work initiated a long series of cellular studies of learning in *Aplysia* and other invertebrates, which is still taking place. These studies started with nonassociative facilitation (strengthening) and habituation (weakening) of specific pathways then went

on to different kinds of associative modifications. Hawkins and Kandel (1984) reviewed how invertebrate findings may relate to different forms of conditioning (see Chapter 6 for further discussion).

After the early invertebrate work, the search for synaptic plasticity in mammals began with the hippocampus, the area involved in consolidation of short-term memory into long-term memory. Bliss and Lømo (1973) demonstrated in the rabbit hippocampus the phenomenon of *long-term potentiation* (LTP). LTP is defined (Byrne, 1987) as "a persistent enhancement of synaptic efficacy generally produced as a result of delivering a brief (several seconds) high-frequency train (tetanus) of electrical stimuli to an afferent (incoming) pathway" (p. 389). This potentiation can last up to several hours in an isolated cellular preparation and several days in an intact animal.

The article of Byrne (1987) ended with a statement of four general principles about neural plasticity that largely remain valid:

1. *Plasticity involves changes in existing neural circuits*. This means that cellular correlates of associative learning typically do not, at least in adult animals, involve growth of new synaptic connections but rather changes in the efficacy of existing connections. (There have been more recent findings of neurogenesis, that is, formation of new neurons, in adult mammals including humans, most notably in the hippocampus and the olfactory cortex; see Eriksson et al., 1998, and Zhao, Deng, & Gage, 2008. Yet the role of such new neuron formation in learning has not been established. Nor has it been established whether or not adult neurogenesis occurs in other parts of the brain such as the rest of the cortex; see Gould, 2007.)
2. *Plasticity is not localized to one site or type of neuron*. Evidence for modifiable synapses has been found at motor neurons in some experiments (e.g., eyeblink conditioning in the cat) and at sensory neurons in other experiments (for example, heart-rate conditioning in the pigeon).
3. *Plasticity involves second messenger systems*. Second messengers are particular chemical substances that regulate amounts of available neurotransmitters (see Appendix 2 for more detail). Briefly (see Figure 3.1), action potentials involve characteristic patterns in the transport across nerve membranes of potassium, sodium, and chloride ions. Transmission of impulses across a synapse is mediated by a chemical transmitter that affects the "channels" carrying those ions across the postsynaptic membrane. There are over twenty known neurotransmitters; some of the most common are *glutamate*, *gamma-amino butyric acid* (*GABA*), *acetylcholine*, *norepinephrine*, *serotonin*, and *dopamine*. Transmitter production and release are in turn affected by second messengers. The commonest second messengers are *cyclic AMP*, *cyclic GMP*, the calcium (Ca^{++}) ion, and an enzyme, *mitogen-activated protein kinase* (*MAPK*).

4. *Plasticity at one site involves multiple synergistic processes*. For example, in the sea slug, *Aplysia*, there can be coordinated effects on the release of a chemical transmitter and on a postsynaptic potassium channel affected by that same transmitter.

Bliss and Collingridge (1993) reviewed the physiological properties of long-term potentiation (LTP) in the hippocampus. These authors characterized LTP by three basic properties: *cooperativity*, *associativity*, and *input-specificity*. Cooperativity means that there is a threshold of intensity of electrical stimulation, and thus of numbers of activated nerve fibers, below which LTP cannot occur. Associativity (which Bliss and Collingridge, p. 31, called "a cellular analogue of classical conditioning") means that a weak input to one pathway can be potentiated if it is active at the same time as a strong input to a separate but convergent pathway. Input-specificity means that the process potentiating pathways active at the time of the stimulus does not spread to other pathways.

Bliss and Collingridge stated that induction of associative LTP requires "a molecular coincidence detector, able to respond to the conjunction of activity in afferent fibers and adequate depolarization in target dendrites" (p. 31; see Appendix 2 of this book for definitions of some of these neurophysiological terms). They reviewed evidence that this function of coincidence detection is performed by a specific type of receptor at the neuron membrane for glutamate, which is the brain's primary excitatory neurotransmitter substance (see Appendix 2). This type of receptor is called the *NMDA receptor*, after a compound called N-methyl-D-aspartate (NMDA), which enhances glutamate action and which has a special affinity for that type of receptor. An important variable here is the flux of the calcium (Ca^{++}) ion in the vicinity of the NMDA receptor triggered by the combination of presynaptic glutamate release and postsynaptic depolarization (increase of positive voltage inside the postsynaptic membrane). More recent work has clarified that this Ca^{++} influx in turn leads to the release of *kinases* (enzymes that mediate reactions involving adding phosphorus to molecules) that are essential for LTP to occur; for a review see Byrne, LaBar, LeDoux, Schafe, and Thompson (2014) and Heidelberger, Shouval, Zucker, and Byrne (2014).

Several more recent articles (e.g., Bear, 2003; Collingridge, 2003; Malenka & Bear, 2004) review the history of results that followed the establishment of LTP mechanisms. The discovery of NMDA receptors led to a search for those same receptors elsewhere in the brain, notably in the visual cortex of kittens during development. Also, investigators aware of the effects of monocular deprivation on later vision looked for mechanisms for the opposite of LTP, namely *long-term depression* (LTD). LTP in the visual cortex was found by Kirkwood and Bear (1994), and it was found that a common form of LTD is dependent, like the most common form of LTP, on NMDA receptors.

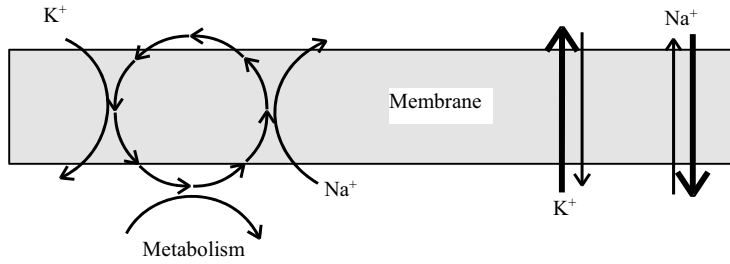


FIGURE 3.1 Schematic diagram of ionic mechanisms for resting and action potentials of the nerve membrane. On the left, ionic "pumps" help to preserve the concentration difference of sodium (Na^+) and potassium (K^+) ions inside and outside the cell membrane during rest. This concentration difference changes during the action potential, as shown on the right. (Adapted by permission from Thompson, 1967.)

Furthermore, evidence was found in support of the hypothesis that "presynaptic activity triggers synaptic depression or potentiation depending on the concurrent level of postsynaptic activity" (Bear, 2003, p. 650).

Yet the NMDA receptors did not tell the whole story. Roles were also discovered for two other types of glutamate receptors (*AMPA* and *metabotropic*), for neurotransmitters that modulate glutamate synapses (acetylcholine, serotonin, dopamine, and norepinephrine), and for GABA, the most common inhibitory transmitter in the brain.

Hence, considerable progress has been made at illuminating biochemical substrates for LTP and LTD in the brain, substrates that vary a great deal across species and across brain regions. The sequence of events in the commonest form of hippocampal LTP is particularly well understood:

First, glutamate binds to AMPA receptors and depolarizes the postsynaptic cell. The depolarization allows glutamate to bind to the N-methyl-D-aspartate (NMDA) class of receptors. Calcium then flows into the cell through the NMDA channel and triggers a host of intracellular events that ultimately result in gene induction and synthesis of new proteins. (LeDoux, 2000, p. 167).

Feldman (2009) reviewed recent results on synaptic plasticity in the mammalian cerebral cortex. Feldman noted that there have been impressive results on the roles of LTP and LTD in the plasticity of sensory maps, particularly in the visual and somatosensory cortices. He noted that the requirement for LTP in adult learning had been verified in the hippocampus and amygdala but not yet in the cortex. There is also suggestive evidence, not yet proven, that long-lasting plastic changes involve genetic modification. Short-term neuronal response modifications in many areas including orbitofrontal cortex and amygdala during conditioning have also been verified: some of that evidence is reviewed in the context of models (discussed in Chapters 6 and 9 of this book) by Frank and Claus (2006) and Dranias, Grossberg, and Bullock (2008).

In summary, there now appears to be a sufficient physiological basis for many if not all of the neural network learning rules that have been suggested on cognitive grounds. Some of these rules incorporate variants of Hebb's postulate or of LTP, whereby synaptic efficacy increases with coordinated presynaptic and postsynaptic activities. Others incorporate variants of LTD or synaptic efficacy decreasing with use, or include the possibility of either an increase or decrease with use. We now proceed to a discussion of different synaptic modification rules used in some of the earliest model networks and the cognitive and behavioral consequences of these rules.

3.2. Rules for Associative Learning

Work on translating an associative rule for synaptic modification into explicit equations essentially began in the late 1960s, as discussed in Section 6 of Levine (1983). The most important early contributors to this effort were Stephen Grossberg, James Anderson, Teuvo Kohonen, and their collaborators.

3.2.1. Outstars and Other Early Models of Grossberg

Grossberg derived his formal rules and related equations from psychological considerations. The networks implementing these rules, in turn, suggested analogs of neural elements and interconnections. The most general form of these equations (Grossberg, 1969b) remains in current use in modeling multilevel adaptive networks with modifiable synapses between levels (see Chapters 6–9 of this book).

The psychological and neurophysiological implications of the theory thus derived were described in Grossberg (1969a). This article posed the question of how an organism learns to produce one sound (say *B*) in response to another (say *A*) after repeatedly hearing them in sequence. The network designed to answer that question was motivated by the following psychological postulates among others:

1. Language appears to be spatiotemporally discrete. That is, a sound like *A* is psychologically treated as an "atom" instead of being subdivided.
2. Such discrete symbols as occur in language are used to represent sensory experience, which is spatiotemporally continuous.
3. Learning changes from continuous to discrete; for example, a child learning to walk must concentrate continuously on his or her movements, whereas an adult walking has an automatic sequence of discrete steps.

The network that Grossberg used to satisfy these postulates consists of discrete elements with time-varying activities that satisfy continuous differential equations. Mathematical results about the equations, showing what is

learned by particular networks, were proved in other articles (Grossberg, 1968a, 1968b, 1969b, 1972a). For the variables defining these equations, Grossberg borrowed from Hull (1943) the notions of *stimulus trace* and *associative strength*. For each stimulus "atom" (i.e., node) such as A, a stimulus trace $x_A(t)$ is defined that measures how active the memory for A is at any given time t. For each pair of nodes A and B, the associational strength $w_{AB}(t)$ measures how strongly the sequential association AB is in the network's memory at time t. In Figure 3.2, the ith stimulus trace x_i is located at the node v_i, and the association w_{ij} between the ith and jth traces is located along the edge e_{ij}. Grossberg drew an analogy between the v_i and cell bodies, the e_{ij} and nerve axons, and the junctions e_{ij}-to-v_j and synapses (see Appendix 2).

Table 3.1 summarizes the effects that Grossberg incorporated into his differential equations. As this table shows, in the case of the sequence AB, it was desired that B should be produced if, and only if, A has been presented *and* the sequence AB is strong in memory. Similarly, the sequence AB should become stronger if, and only if, A is presented and followed by B. Hence, replacing A and B by the ith and jth stimuli in general, the variable x_j should increase if both x_i and w_{ij} are high, which means that the equation for the rate of change of x_j should include a (nonlinear) term like the product $x_i w_{ij}$. Likewise, the variable w_{ij} should increase if both x_i and x_j are high, which means that the equation for w_{ij} should have a term like the product $x_i x_j$. These product terms mean that the network is *nonlinear*: that is, node activities change over a time in a way that is not directly proportional to the influences from other nodes or outside inputs. (See Appendix 1 for an introduction to the general process of incorporating neural or cognitive variables into differential equations.)

The product terms in these equations incorporate a version of Hebb's postulate that synaptic weights will increase with coordinated pre- and post-synaptic activities. Yet, in addition to LTP they include a potential mechanism for LTD, in the form of a slow decay in the weights w_{ij}, causing the weights to decrease with combined pre- and postsynaptic activities that are too small to overcome the decay.

The import of Grossberg's approach to learning can be gleaned from study of a specific type of simple network architecture called the *outstar* (Grossberg, 1968a; see Figure 3.2). In the outstar, one node v_1, called a *source*, projects to arbitrarily many other nodes $v_2, v_3, \ldots, v_n$, called *sinks*. Long-term storage can be interpreted as residing in the *relative* weights of $w_{12}, \ldots, w_{1n}$, that is, in the functions

$$W_{1i} = \frac{w_{1i}}{\sum_{j=2}^{n} w_{1j}} \tag{3.1}$$

where "Σ" denotes summation (in this case, the sum of all weights w_{1j} of connections from source to sinks).

The outstar is affected by an input I_1 to the source node v_1, and a pattern of inputs $I_2, \ldots, I_n$ to the sink nodes $v_2, \ldots, v_n$. (Grossberg sometimes interpreted the source input as a conditioned stimulus, and the pattern of sink inputs as an unconditioned stimulus; see the discussion of conditioning models in Chapter 6.) The activity x_1 of v_1 tends to increase if the input I_1 is present and to decay back toward a baseline (interpreted as 0) in the absence of input. As illustrated in Table 3.1(a), the activity x_i of each v_i also tends to increase if both x_1 (activity of v_1) and w_{1i} (associative strength between v_1 and v_i) are significant, and decay otherwise. Finally, as illustrated in Table 3.1(b), w_{1i} tends to increase if both x_1 and x_i are significant, and decay otherwise.

If x_1 is interpreted as encoding A and x_i as encoding B, the decay of w_{1i} implies that the association between A and B is weakened while the network

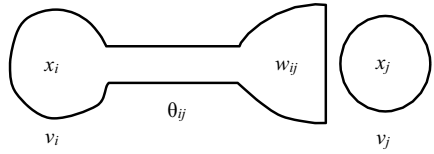


FIGURE 3.2 Schematic of two nodes and one modifiable connection between them, based on Grossberg's 1968 and 1969 articles.

A is presented	*AB has been learned*	*B is expected*
Yes	Yes	Yes
Yes	No	No
No	Yes	No
No	No	No

(a)

A is presented at a given time	*B is presented a short time later*	*AB is learned*
Yes	Yes	Yes
Yes	No	No
No	Yes	No
No	No	No

(b)

TABLE 3.1 Effects incorporated into Grossberg's differential equations.

is not actively hearing A. This property is contrary to intuition: it is to be expected, rather, that the association will decay when A is presented without being followed by B, but remain constant when A is not presented at all. In a modified form of the outstar equations, used in much of Grossberg's later work, w_{1i} decreases only if x_1 is large and x_i is small. This implies that an association such as AB remains intact if neither A nor B is presented, but is weakened if A is presented and not followed by B.

The outstar equations are given at the end of this chapter. Before discussing the major results about those equations, let us briefly consider the implications of this theory for memory. The stimulus traces x_1 and x_i can be considered as analogs of *short-term memory* (often abbreviated STM), while the associative strengths w_{1i} are analogous to *long-term memory* (often abbreviated LTM). It is desired that STM traces should decay quickly after inputs cease, while LTM traces should be relatively stable. Hence, the decay rate for w_{1i} is set much smaller than the decay rates for x_1 and x_i.

Grossberg (1968a) studied the *asymptotic behavior*, that is, behavior as time increases, of the outstar equations given at the end of this chapter. He showed that, for many classes of inputs, the functions $W_{1i}(t)$ defined by (3.1) and the analogous functions

$$X_{1i} = \frac{x_{1i}}{\sum_{j=2}^{n} x_{1j}} \tag{3.2}$$

both approach limits as these equations evolve in time. Moreover, for each j, the limiting values of X_j and W_{1j} are equal. Thus, the same distribution of weights is coded both in the relative stimulus traces and the relative associational strengths at the outstar sinks.

Simulation exercises at the end of this chapter illustrate how the outstar's limiting behavior may relate to different classes of inputs. A particularly important case is where the inputs to the sink nodes form what Grossberg called a *spatial pattern*, that is, where the relative proportions of inputs to the different sink nodes are unchanged over time (see Figure 3.4). The mathematical definition of a spatial pattern input is:

$$\text{For } j > 1,\ I_j(t) = \theta_j I(t),\ \sum_{j=2}^{n} \theta_j = 1 \tag{3.3}$$

In Expression (3.3), the values θ_j represent relative pattern weights, since $\theta_j(t) = (I_j/I)$, while I, which equals the sum of all the I_j, represents the total input to all sink nodes of the outstar. In this case, under suitable conditions on the inputs $I_1(t)$ and $I(t)$, the input pattern weights were shown to be stored in long-term memory, that is, if $\lim_{t\to\infty}$ denotes the limiting value as time goes on,

$$\lim_{t\to\infty} W_{1j}(t) = \lim_{t\to\infty} X_j(t) = \theta_j,\ j = 2,\ldots, n \tag{3.4}$$

The conditions under which (3.4) holds, which are listed at the end of this chapter, can be interpreted as meaning that inputs to both the source and the sink are presented "often enough" for large times. After learning, the learned spatial pattern can be reproduced by activating the source node with an input I_1.

Other articles (e.g., Grossberg, 1968b, 1969b, 1972a) extended the above results to equations describing networks with learning laws similar to the outstar's but different in architecture. In some cases, learning theorems for spatial patterns, analogous to (3.2), can be proved. In other cases, spatial patterns can be shown to be learned for some parameter values and forgotten (leading to asymptotically uniform synaptic weights) for other parameter values. Such variability of dynamics occurs, for example, in the *complete graph with loops*, where every node projects to every node including itself, and the *complete graph without loops*, where every node projects to every other node.

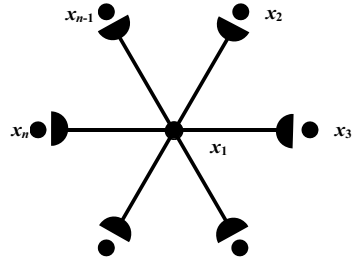


FIGURE 3.3 Outstar architecture.

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

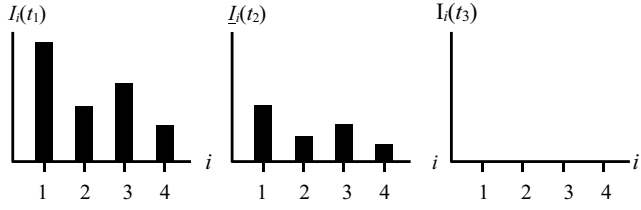


FIGURE 3.4 Example of a *spatial pattern* input, in which absolute inputs may change over time but always remain in the relative proportions θ_i.

Some new psychological properties, separate from the original postulates, emerged from these mathematical results. These properties are:

1. The more often the network hears *B* following *A*, the more likely it is to say *B* after hearing *A*.
2. An isolated network can remember without practicing overtly.
3. Memory can sometimes improve spontaneously on recall trials.
4. If the network has learned the association *AB*, it can be changed to *AC* by sufficient presentation of the new sequence, but the change takes a long time if the association *AB* has become very strong.

Grossberg and his associates made use of these associative pattern learning results in models of more complex cognitive processes, which are discussed in Chapters 6–9. In Section 3.3, we see that Grossberg also developed a class of modifiable synapses where the associative tendency is counteracted by habituative tendencies. The explanation for habituation is that repeated pairing of pre- and postsynaptic activities increases *production* of chemical transmitter at the synapse, but also increases *release* of transmitter.

In mathematical terms, a pattern of inputs or of node activities can be described as a *vector*, or linear array of numbers. The connection strengths between nodes form a square array of numbers, or *matrix* (plural: *matrices*). (Appendix 1 provides more details about the mathematics of vectors and matrices.) These mathematical objects play a large role in the models of Anderson and of Kohonen, which are discussed in the next two subsections. Their models bear some similarity to Grossberg's but their node interactions are more linear.

3.2.2. Anderson's Connection Matrices

Anderson and Kohonen both tackled problems of encoding multiple patterns (vectors) simultaneously in memory. The linear model of learning in Anderson (1972) is based on models of memory storage, retrieval, and recognition by Anderson (1968, 1970). In all these articles, a memory trace is described as a vector,

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

each of whose components is the activity of a single element of the network. Hence, these memory traces are similar to the stimulus traces of Hull (1943) and Grossberg (1969a). If the elements are neurons, this activity is interpreted as instantaneous firing frequency; if the elements are populations of neurons, activity is interpreted as average firing frequency. All of the memory traces present in those elements are summed into a total storage vector

$$\mathbf{s} = \sum_{k=1}^{K} \mathbf{X}_k$$

The model is discrete in time, although synaptic efficacy can take on any of a continuum of values.

The emphasis in Anderson's articles was on developing a simple model that would capture some of the basic properties of memory without resorting to much physiological detail. In Anderson (1968), the problems considered were:

1. *Recognition*: If an input pattern (vector of activities) is given, how can one tell whether a trace similar to that pattern is already stored as part of the storage vector?
2. *Retrieval*: Once recognition has occurred, how can the stored trace be reconstructed?
3. *Association*: Once retrieval has occurred, how can other stored traces be found that are also similar to the input pattern?

Retrieval poses a particular challenge in this model because it involves recovering one term of a vector sum, which is subject to error. The techniques described for solving the recognition and retrieval problems uses various kinds of filters designed so as to minimize the probability of error (, that is, to maximize a certain signal-to-noise ratio.

The mathematical theory of these linear filters was discussed further in Anderson (1970). The input trace is used to construct a *matched filter* whose output is the dot product of the stored array with the input trace, that is,

$$V = \mathbf{s} \cdot \mathbf{x} \sum_{k=1}^{N} s_k x_k$$

where N is the number of nodes in the network. If $\mathbf{s}$ is considered to equal the input $\mathbf{x}$ added to a noise vector $\mathbf{n}$, with components n_k, the signal-to-noise ratio is defined as

$$\frac{\left(\sum x_k^2\right)^2}{\left(\sum n_k x_k\right)_{\text{avg}}^2},$$

the average being taken over all possible storage vectors. Using probability theory, the signal-to-noise ratio was shown to be close to N/K, with N the number of nodes and K the number of traces. The derivation of the signal-to-noise ratio is sketched at the end of this chapter. Thus, increasing the number of nodes makes the network more reliable, as von Neumann (1951) had

discovered in a much earlier model. Also, increasing the number of traces makes recognition more difficult, which is to be expected because of mutual interference. The signal-to-noise ratio achieved is the maximum possible for a linear filter, but can be improved by using a suitable nonlinear filter.

The problem of association was discussed in Anderson (1968), and its relationship with a theory of synaptic connection weights was described in Anderson (1972). In these articles, a model for association was proposed that involves two sets of nodes, α and β. The following assumptions were made for simplicity: (1) both α and β have the same number N of nodes; (2) there is another number M such that M nodes of α project to every neuron of α, and every neuron of α projects to M nodes of β. There was assumed to be an input trace $\mathbf{x} = (x_1, \ldots, x_n)$ at α, a trace $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ at β, and values a_{ij} for the efficacy of the synapse from the ith element of α to the jth element of β. It was assumed that $\mathbf{y}$ should be as close as possible to $\mathbf{Ax}$, where $\mathbf{A}$ is the matrix of connection weights a_{ij}. (Multiplication of matrices is explained in Appendix 1.) Under this assumption, the optimal matrix $\mathbf{A}$ (from the standpoint of signal-to-noise ratio) is obtained by:

$$a_{ij} = cx_i y_i \tag{3.5}$$

for some constant c. In other words, the optimal values of the connection weights reflect the Hebb-like operation of multiplying pre- and postsynaptic firing frequencies.

Equation (3.5) illustrates that optimality considerations lead to a rule similar to the Hebb rule for connection weights. Nass and Cooper (1975) extended Anderson's analysis to develop a network where the matrix of connection weights start with arbitrary values and converge to the optimal weights; algorithms of this sort are still in frequent use. This optimality analysis could be extended to some cases of multiple input traces and multiple output traces. If there are K input traces $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kn})$, $1 < k < K$, and K output traces $\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kn})$, all with equal "power"

$$P = \sum_{i=1}^{N} x_{ki}^2$$

then the Equation (3.5) for the optimal weight matrix generalizes to

$$a_{ij} = \sum_{k=1}^{K} cx_{ki} y_{kj} \tag{3.6}$$

Equation (3.6) indicates that synaptic weights in this model are calculated instantaneously from the inputs. The dynamics of the system over time were

not developed, and decay terms (as in the outstar equations) were not included. Thus, in this model, a trace can only be forgotten if it is interfered with by the memory of another stimulus whose trace involves the same neural elements.

The extension of this work to include temporal dynamics with decay terms was suggested in Anderson (1973) and done explicitly in Nass and Cooper (1975). The 1973 article dealt with the modeling of some psychological data on the learning of lists. In these data, a subject's reaction time to an item is related to the probability of occurrence of the item. This result was modeled by letting the kth trace coefficient (analogous to the factor c in (3.5) and (3.6)) be $c_k = 1 + \Gamma p_k$, where Γ is a weighting constant and p_k is the probability of presentation of the kth item. Anderson (1973) explained his model as follows:

> The probabilistic term due to the immediate past history could easily arise in the following way. Each time the item is tested or rehearsed, the stored trace is slightly strengthened. This can easily occur since the trace has just been present in the correct form. If an exponential decay of trace strength in short-term memory is assumed, the probability dependent term will appear to be of this form.
>
> (p. 429)

He thus suggested, without proposing a formula, that the memory trace of an item decays exponentially while the item is not being presented, and increases while the item is being presented.

Nass and Cooper (1975) showed that an optimal matrix of the form (3.6) arises as the limit of a matrix that is modified at time t according to the rule $a_{ij}(t+1) = \Gamma a_{ij}(t) + \mu x_i y_j$, with Γ and μ positive constants, $\Gamma < 1$ but close to 1. That rule effectively means that while stimuli are not being presented, associative strengths decay spontaneously at a rate $1 - \Gamma$.

The Nass–Cooper model was specifically applied to the development of orientation detecting neurons in the visual cortex, based on results showing that exposure during a critical period to lines of the right orientation is necessary for cats to develop those detecting cells (Blakemore & Cooper, 1970; Hirsch & Spinelli, 1970). Nass and Cooper also considered the issue of selectivity: how different, neighboring neurons might learn to respond maximally to different orientations. Selectivity was accomplished by adding lateral inhibition to the network; the next chapter (Chapter 4) discusses in detail the role of lateral inhibition in visual processing. Nass and Cooper (1975) provided some of the foundation for the later orientation detection model of Bienenstock, Cooper, and Munro (1982), which is discussed in Section 7.1. The Bienenstock et al. model includes a weight modification rule that combines LTP and LTD (cf. Bear, 2003).

3.2.3. Kohonen's Early Work

Anderson's theories of associative memory and category learning are similar to theories of the same phenomena by Kohonen (1977) and Kohonen et al. (1977). Kohonen et al. (1977) used the terms "associative memory" and "associative learning" for two distinct yet interrelated sets of cognitive phenomena. One set of phenomena involves memory or learning of associations that develop, for example, by classical conditioning or serial learning of lists. The other set involves recollection of a total pattern if part of the pattern is perceived (Kohonen et al., 1977):

> The term associative memory is here restricted to denote a process in which a signal pattern is recalled upon the basis of a fraction of it (the key). The signal pattern is composed of the activities in a set of axons, acting as the input to a neural network during a short period of time. These axons may originate within one modality, with the signal pattern thus representing one sensory stimulus, or they may originate within different parts of the central nervous system, and thus represent the simultaneous activity in these structures. Consequently, for instance, both the recall of a visual image from its fraction, and a paired association in the classical conditioning, can be regarded as different aspects in the functioning of the associative memory.
>
> (p. 1065)

He continued to develop this distinction in later work (Kohonen, 1984), where he referred to the above two aspects of associative memory as *autoassociative* and *heteroassociative*; this later work is discussed in Section 3.4.

Kohonen et al. (1977) and Kohonen (1977, Chapter 1) illustrated the recognition process using simulations of human face recognition. The algorithm for these simulations is a good example of Kohonen's general method. The nodes in his model were assumed to be analogs of pyramidal cells in the cerebral cortex (the largest type of cortical neuron), or else of the columns into which cortical neurons are organized (cf. Mountcastle, 1957; Hubel & Wiesel, 1962, 1965).

Kohonen et al. (1977, p. 1069) assumed that each unit modulates the activities of other units in proportion to its activity, and that a nonspecific background activity is present. The variables defining the network are output firing frequencies y_i, which depend on input firing frequencies (spiking frequencies) x_i. If the direct connectivity between input and output is denoted by w_i, the long-range connectivity from unit j onto unit I by w_{ij}, and the background activity by y_b^*, the input–output transformations can be represented as a system of linear equations:

$$y_i = w_i x_i + \left(\sum_j w_{ij} y_j \right) + y_b^* \tag{3.7}$$

The values w_{ij} of the connection weights in Equation (3.7) can be positive for excitatory connections, negative for inhibitory ones, and 0 for units that do not project to the given unit.

In some earlier examples of Kohonen's work, the connectivities w_{ij} were assumed to be constant, and, by methods similar to Anderson's, the optimal weights were found for recall of specific patterns. But Kohonen also incorporated memory effects, modeled by a Hebb-like associative law. The equations on which Kohonen's simulations were based, which are presented at the end of this chapter, combine the interactions of (3.7) with that type of learning law.

The actual simulations of Kohonen et al. (1977) used an approximation to these equations combined with a preprocessing (sharpening) of the pattern using lateral inhibition. As discussed in Chapter 4, lateral inhibition is widely used in neural networks to enhance contrast between different locations within a pattern. Kohonen and his coworkers achieved some success in recovering a recognizable face from a blurred or partially missing image. Like Anderson, they found that the larger the number of stored images, the more difficult it is to get a sharp recollection of each image. It was found that inclusion of lateral inhibition markedly sharpens the recollected image.

Kohonen et al. (1977) also commented that the model could be improved by incorporation of *temporal* as well as spatial contrast enhancement:

> The temporal differentiation of signals exerts an effect of improvement, similar to that of lateral inhibition, since the most relevant information is usually associated with changes in state.
>
> (p. 1075)

A previous article (Kohonen & Oja, 1976) had used temporal difference effects to construct a "novelty filter," which selectively enhances those parts of a pattern that had not previously been present. Learning laws based on various forms of temporal difference are discussed extensively in the next section.

3.3. Learning Rules Related to Changes in Node Activities

In the associative rules discussed in the last section, connection weights are modified by correlated presynaptic and postsynaptic activities. By contrast to rules based on correlated node activities, many other modelers have suggested learning rules in which the crucial variable is *change* in postsynaptic (and also,

in some cases, presynaptic) activity. The error-correcting rules of Widrow and Hoff (1960) and Rosenblatt (1962) (see Section 2.1) fit into this general category. Later modelers employing similar rules, for a variety of reasons, include Klopf (1979, 1982, 1988), Sutton and Barto (1981, 1991), Kosko (1986b), and Rumelhart et al. (1986), among others.

3.3.1. Klopf's Hedonistic Neurons and the Sutton–Barto Learning Rule

Klopf (1982) proposed that a synapse is increased in efficacy if its activity is followed by a net increase in the depolarization (positive stimulation) received by the postsynaptic cell. In other words, he proposed that depolarization acts as *positive reinforcement* for neurons. Klopf's theory was based on an analogy between single neurons and whole brains, both of them being treated as goal-seeking devices. This is the reason for the words "hedonistic neuron" in the title of his book.

The importance of activity change, as opposed to activity itself, was also highlighted in the psychological theory of Rescorla and Wagner (1972), which is not neurally based but has influenced many neural models of classical conditioning. Their theory is based on the results of many classical conditioning experiments indicating that associative learning of a conditioned stimulus can be greatly influenced by the background stimuli present during both training and recall trials. The main tenet of Rescorla and Wagner's theory was that:

> organisms only learn when events violate expectations. Certain expectations are built up about the events following a stimulus complex: expectations initiated by the complex and its component stimuli are then only modified when consequent events disagree with the composite expectation.
>
> (Rescorla & Wagner, 1972, p. 75)

Sutton and Barto (1981) set out to explain classical conditioning with a theory that included elements of both the Rescorla–Wagner and Klopf theories. Their conditioning model includes n stimulus traces $x_i(t)$, an output signal $y(t)$, and n synaptic weights $w_i(t)$, as shown in Figure 3.5. These weights are considered to denote associations between conditioned stimuli (CSs) and a primary reinforcer or unconditioned stimulus (US or UCS).

Sutton and Barto proposed that in addition to traces $x_i(t)$ that denote the duration and intensity of given CSs, there are other traces called $\overline{x}_i(t)$ that are separate from the stimuli and longer lasting. These are the actual memory traces, analogous to those in the outstar equations (see Section 3.2). Sutton and Barto termed them *eligibility traces* because they indicate when a particular synapse is eligible for modification. The existence of these two separate sets

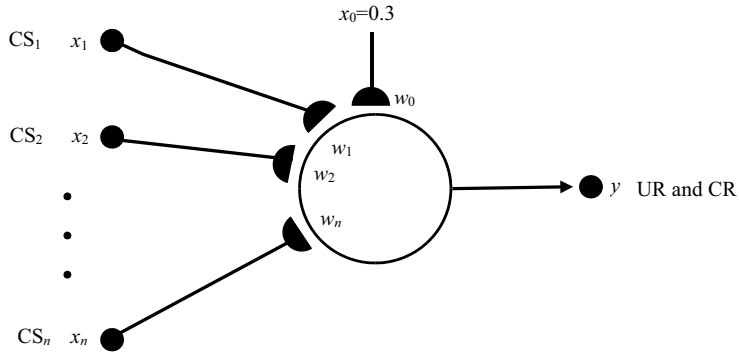


FIGURE 3.5 Adaptive element analog of classical conditioning. This network has n modified conditioned stimulus (CS) pathways, and a pathway with fixed weight w_0 for the unconditioned stimulus (US). The node y represents the unconditioned and conditioned responses (UR and CR).

Source: From Sutton & Barto, *Psychological Review*, *88*, 135–170, 1981. Copyright 1981 by the American Psychological Association. Adapted by permission.

of traces had been previously proposed by Klopf (1972). Sutton and Barto suggested possible cellular mechanisms for eligibility traces, involving calcium ions and cyclic nucleotides. Finally, the current amount of reinforcement, $y(t)$, was compared with the weighted average $\bar{y}(t)$ of values of y over some time interval preceding t. These variables are governed by Equations (3.22) listed at the end of the chapter.

The two innovations in Sutton and Barto's model – including eligibility traces and making learning depend on the comparison term $y(t) - \bar{y}(t)$, rather than simply the postsynaptic activity $y(t)$ – were motivated by experimental results on timing in classical conditioning. In particular, the model can explain the fact that, in many conditioning paradigms, the optimal interstimulus interval (time interval between the two stimuli to be associated) is greater than 0; this explanation is further developed in Chapter 6. Sutton and Barto's network can also simulate other contextual effects in classical conditioning, such as the blocking of formation of associations to a new stimulus if another stimulus that has already been conditioned is simultaneously present.

The synaptic learning law involving change in postsynaptic activity is not the only possible way to simulate timing effects or blocking in classical conditioning. The same data were simulated by Grossberg and Levine (1987) using a form of the earlier Grossberg learning law (see Section 3.2) combined with competitive attentional effects in a larger network.

Yet while the behavioral data studied by Sutton and Barto did not lead to a unique physiological model, they inspired a class of models that later developed into the widely used temporal difference (TD) models, which are discussed extensively in Chapter 6. The TD models have made extensive

contact with physiological data on midbrain dopamine neurons and the basal ganglia.

3.3.2. Error Correction and Back Propagation

Sutton and Barto (1981) noted some formal analogies between their learning rule, where present is compared with previous reinforcement, and other rules whereby actual is compared with desired reinforcement. Rules of the latter sort arose from considerations both in psychology (Rescorla & Wagner, 1972) and in engineering (Widrow & Hoff, 1960). Such error-correcting rules fall into the category of *supervised learning*; that is, they rely on a "teacher" to tell whether particular neural responses are "right" or "wrong." The best known of these supervised learning rules is the *generalized delta rule* used in the back propagation algorithm.

The essentials of the back propagation algorithm were developed by Werbos (1974, 1993), as a procedure for optimizing the predictive ability of mathematical models. LeCun (1985) and Parker (1985) discovered this procedure independently, and Rumelhart et al. (1986) placed it in a widely studied connectionist framework. It is often applied to discrimination or classification of sensory input patterns (see Chapter 8). The principle is as follows. The network is feedforward with three layers, composed of input units, hidden units, and output units (see Section 2.1). A particular pattern of output responses to particular input patterns is desired. To the extent that the actual response to the current input deviates from the desired response, the weights of connections from hidden to output units change. Then those weight changes propagate backward to cause changes in weights from input to hidden units that will reduce the error in the future. The hidden units thereby come to respond to, hence encode, specific patterns of input activities (the *internal representations* in the title of the article by Rumelhart et al., 1986).

The original delta rule (similar to that of Widrow & Hoff, 1960, and many others) was formulated by Rumelhart et al. as a rule for changing weights following presentation of a given pattern, labeled by the index p. If w_{ij} is the weight from the ith input unit to the jth output unit, then:

$$\Delta w_{ij} = \eta \delta_{pj} i_{pj} \tag{3.8a}$$

where η is a learning rate constant, δ_{pj} is a measure of desired change in the jth component of the output response, and i_{pi} is the value of the ith component of the input pattern. In the simplest network with only input and output units and no hidden units, the desired change in the jth output component is

$$\delta_{pj} = t_{pj} - y_{pj} \tag{3.8b}$$

where y_{pj} is the jth component of the actual output response and t_{pj} is the jth component of the desired or "target" response. The work of Rumelhart et al. and their predecessors consisted of generalizing (3.8b) to networks with hidden units. We now give a condensed description of the learning law arising from this generalization.

Rumelhart and McClelland (1986a, Volume I, Chapter 2) had previously demonstrated that there is no advantage to including hidden units if their *activation functions* (outputs as a function of total signal received) are linear. They therefore posited activation functions that are nondecreasing and differentiable but nonlinear. Typically, they used sigmoid functions (see Figure 2.7).

Let f be a sigmoid function, and let f' be its rate of change or derivative (see Appendix 1). Let the net signal received by the jth unit in any given layer of the network be:

$$\text{net}_{pj} = \sum_j w_{ij} y_{pi} \tag{3.9}$$

a linear sum of the outputs y_{pi} from the previous layer weighted by the connections w_{ij}. If j is a hidden unit so that I is an input unit, y_{pi} equals the input component i_{pi}. If j is an output unit so that I is a hidden unit, then y_{pi} equals $f(\text{net}_{pi})$, the activation function f applied to the ith net signal net_{pi}.

Let $L = 0, \ldots, N$ indicate the Lth layer of the network, where $L = 0$ represents the input layer and $L = N$ ($N = 2$ in most applications) represents the output layer. The rule of Rumelhart et al. (1986) for back propagation of errors, which generalizes (3.8b), is:

$$\delta_{pj} = f'(\text{net}_{pj})\,[t_{pj} - y_{pj}] \tag{3.10a}$$

if $L = N$ and

$$\delta_{pj} = f'(\text{net}_{pj}) \sum_k \delta_{pk} w_{jk} \tag{3.10b}$$

if $L \neq N$, where the jth node in layer L can either be an output or a hidden unit, and net_{pj} is defined by Equation (3.9). The sum in (3.10b) is over all units k in the next layer downstream, that is, in layer $L + 1$. The changes in weights of connections to the jth node are in turn calculated from the errors using a generalization of Equation (3.8a), namely

$$\Delta w_{ij} = \eta \delta_{pj} y_{pi} \tag{3.11}$$

The derivation of (3.10b) is given at the end of this chapter.

The back propagation algorithm is an example of a long-established class of mathematical methods known as *steepest descent* (see Duda & Hart, 1973, for details). That is, an expression is found for the total error in the network's response (based on the desired or target response), and the weight changes that cause the sharpest possible decrease in this error measure are computed. Heuristically, (3.10b) says that weight changes are greatest at connections from node activities sending signals net_{pj}, whose values are on the sharpest rising slope of the sigmoid function f. This means that those values are in the intermediate range; that is, the units sending those signals are furthest from an established "yes-or-no" response to the input.

Back propagation of synaptic weights occurs because the changes in input-to-hidden weights are computed via (3.8) from the δ_{pj} values, which in turn are computed via (3.10b) from the changes in hidden-to-output weights. Hence, this scheme allows for *credit assignment*: deciding which connections at an earlier level in the network to alter if the responses of later stages are inappropriate (see also Barto & Anandan, 1985). Chapter 8 discusses convergence properties of the back propagation algorithm.

3.3.3. The Differential Hebbian Idea

Learning rules including changes in postsynaptic activity, as in the Sutton–Barto model (and also, sometimes, changes in presynaptic activity), are called *differential Hebbian* rules (Kosko, 1986b). This term is used to contrast with the term *Hebbian* for rules including a simple cross-correlation of pre- and postsynaptic activities (see the earlier discussion in Section 3.1).

Klopf (1988), building on Sutton and Barto's work, developed a differential Hebbian learning model that he called the *drive-reinforcement model*. In Klopf's model, the synaptic efficacy changes as a function of changes in both presynaptic and postsynaptic activities. But in order to account for the positive optimal interstimulus interval in classical conditioning, the change in post-synaptic activity is delayed in time. Also, the change in efficacy of a given synapse is made proportional to the current efficacy. The purpose of this latter rule is to account for the initial positive acceleration in the S-shaped acquisition curves observed in animal learning (see Figure 3.6).

Klopf noted that Kosko (1986b) had independently been led to a rather similar learning law by philosophical and mathematical considerations. Using this law, Klopf was able to simulate a wide variety of classical conditioning data including effects of stimulus duration, partial reinforcement, and compound stimuli. Chapter 6 considers these conditioning data in more detail, comparing Klopf's models of such phenomena, and those of Barto and his coworkers, with the models of Grossberg and Levine (1987) and Grossberg and Schmajuk (1987). One essential difference between these two sets of models is worth noting now. The Klopf and Barto approaches rely on complex

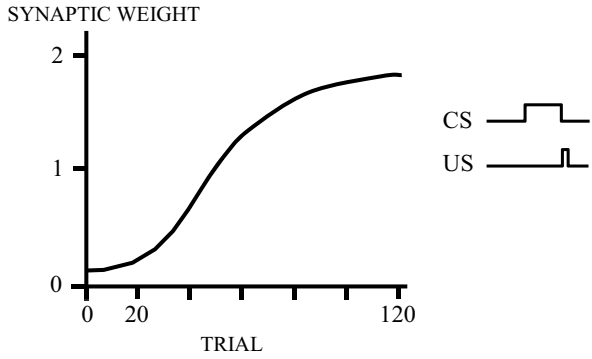


FIGURE 3.6 Synaptic weight between CS and US representations in a simulated classical conditioning experiment using the drive-reinforcement learning rule. The model yields an S-shaped (sigmoid) acquisition curve, consistent with some animal learning data. The CS is shut off at the time of US onset (*delay conditioning*).

Source: Adapted from Klopf, *Psychobiology*, *16*, 85–125, 1988. Permission to reproduce granted by Psychonomic Society, Inc.

learning laws that are suggested to represent processes at the neuronal level. The Grossberg approach, by contrast, relies on simpler learning laws (associative cross-correlation combined with exponential decay) combined with certain characteristic network-level interactions that are discussed next.

3.3.4. Gated Dipole Theory

One of the network interactions used in Grossberg's models of classical conditioning is competition, or lateral inhibition; this is a common feature in the networks of other neural modelers, and is the main topic of Chapter 4. Competition is particularly important to models of attentional effects such as blocking. This type of competition usually involves feedback, or recurrent, network interactions. But feedforward competition is part of another type of network interaction known as *opponent processing*, which is the basis for an architecture called the *gated dipole*. The gated dipole theory, like the differential Hebbian theory, was motivated by an effort to compare current values of stimulus or reinforcement variables with recent past values of the same variables.

Grossberg (1972b, 1972c) introduced gated dipoles to answer the following question about reinforcement. Suppose an animal receiving steady electric shock presses a lever that turns off the shock. Later, in the same context, the animal's tendency to press the lever is increased. How can a motor response associated with the *absence* of a punishing stimulus (shock) become itself positively reinforcing? Clearly, absence of shock is not rewarding per se: if you walk to the back right corner of a room and do not get shocked, that corner

does not become more attractive for you unless you have just been shocked elsewhere in the room. Hence, zero shock must become (transiently) rewarding by contrast with the ongoing shock level.

Figure 3.7 shows a schematic gated dipole, which obeys Equations (3.26) below. The synapses w_1 and w_2, marked with squares, have a chemical transmitter that tends to be depleted with activity, as indicated by the $-y_i w_i$ terms in the differential equations for those w_i values. Other terms in those equations denote new transmitter production. The amount produced is greatest when the transmitter is much less than its maximum.

In Figure 3.7, the input J represents shock, for example. The input I is a nonspecific arousal to both channels y_1-to-x_1-to-x_3 and y_2-to-x_2-to-x_4. While shock is on, the left channel receives more input than the right channel; hence transmitter is more depleted at w_1 than at w_2. But the greater input overcomes the more depleted transmitter, so left channel activity x_1 exceeds right channel activity x_2. This leads, by feedforward competition between channels, to net positive activity from the left channel output node x_3. For a short time after shock is removed, both channels receive equal inputs I but the right channel is less depleted of transmitter than the left channel. Hence, right channel activity x_2 now exceeds x_1, until the depleted transmitter recovers. Again, competition leads to net positive activity from the right channel output node x_4. Whichever channel has greater activity either excites or inhibits x_5, thereby enhancing or suppressing a particular motor response.

The network is called a gated dipole because it has two channels that are opposite ("negative" and "positive") and that "gate" signals based on the amount of available transmitter. Characteristic output of one gated dipole is graphed in Figure 3.8. This graph illustrates the "rebound" in x_4 activity after the cessation of x_3 activity. Grossberg (1972c) discusses in detail the mathematical relationships between J, I, and the other system parameters that are required for such a rebound to occur. Concurrently with Grossberg's work, Solomon and Corbit (1974) developed an opponent processing theory of motivation, arguing that significant events elicit both an initial reaction and a subsequent counter-reaction.

Grossberg and Levine (1987, p. 5027) compared the gated dipole model for measuring temporal differences with the differential Hebbian model. They argued that the dipole model is better at reproducing two important psychological effects. In the context of shock avoidance data, one effect is that the amount of reinforcement from escaping shock is sensitive not only to the shock's intensity but also to its duration. The other effect is that the amount of reinforcement depends on the overall arousal level of the network (or organism).

The gated dipole can also model the absence of a *positive* stimulus acquiring *negative* significance. If the two channels in Figure 3.7 are *reversed in sign* so that the channel receiving input is the "positive" one, the network provides an

explanation for frustration when a positively reinforcing event either is terminated, or does not arrive when expected. The rebounds between positive and negative also explain the *partial reinforcement extinction effect* (PREE), whereby a motor response learned by an animal under intermittent reinforcement is more stable than the same response learned under continuous reinforcement (e.g., Gray & Smith, 1969). According to the gated dipole theory (or the differential Hebbian theory), a reward's attractiveness is enhanced by comparison with an expected lack of reward.

The idea of opponent processing is applicable to many other neural processes. It is an old idea in vision; for example, the retina contains pairs of receptors for opponent colors, such as green and red, and one of the two opponent colors is transiently perceived after removal of the other one. The dipole idea in the sensory domain involves "on cells" and "off cells" responding to presence or absence of specific sensory stimuli. Grossberg (1980) joined on cells and off cells for different stimuli into a network called a *dipole field*. Transient rebounds in such a dipole field were used in that article to model various visual phenomena such as color-dependent tilt aftereffects. Also, gated dipoles have been applied to the modeling of motor systems: Grossberg and Kuperstein (1986/1989) and Bullock and Grossberg (1988) used dipoles to simulate the actions of neuron populations innervating agonist-antagonist muscle pairs.

Indeed, Grossberg (1987d) hinted that a gated dipole can exhibit switching back and forth between opposite responses:

> Spontaneous switching behavior has previously been shown to be a property of a gated dipole field in response to image pairs that create approximately balanced, but competitive, input patterns. . . . Periodic switching occurs because the habituating transmitters within a winning channel weaken the competitive advantage of that channel by causing a decrease in the size of its positive feedback signals. The inhibited channel can then win the competition because its transmitters are able to accumulate while it itself is being inhibited. Then the cycle of rivalry repeats itself, leading to cyclic recovery and habituation of transmitter gates as a given channel periodically loses and wins the competition.
>
> (p. 123)

The on cells and off cells in the gated dipole are reminiscent of the novelty filter developed in Kohonen and Oja (1976) for selectively enhancing those parts of a pattern that have not been seen before. An example of such a novelty filter, used in visual pattern recognition, is shown in Figure 3.9.

To simplify the discussion in this section and the previous sections of this chapter, we have treated the stimuli that are being associated as if they activate single nodes. Other modeling considerations arise when we look instead at the learning of associations between activity patterns that span large numbers of

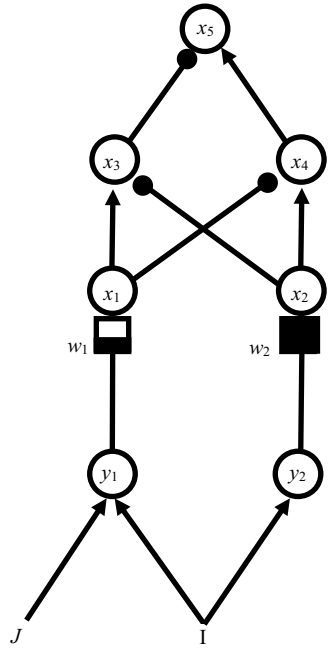


FIGURE 3.7 Schematic gated dipole network. J is a significant input (in the example of Grossberg, 1972b, electric shock), whereas I is nonspecific arousal. Synapses w_1 and w_2 can undergo depletion (as w_1 has in this diagram), as indicated by partial lightening of square boxes. After J is shut off, $w_1 < w_2$ (transiently), so $x_1 < x_2$. By competition, x_4 is activated, enhancing a motor output suppressed by J. Note that *if J is a positive reinforcer such as food instead of shock, the signs downstream are reversed*; that is, x_3 excites x_5 and x_4 inhibits x_5.

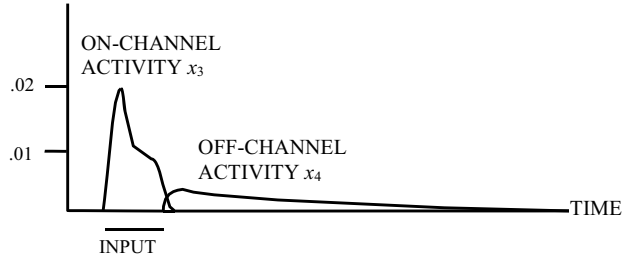


FIGURE 3.8 Typical time course of the channel outputs of a gated dipole.

Source: Adapted from *Neural Networks*, *2*, D. S. Levine & P. S. Prueitt, Modeling some effects of frontal lobe damage: Novelty and perseveration, 103–116, with permission from Elsevier Science.

nodes. Building on some ideas previously introduced in Section 3.2, the next section presents aspects of associative learning of patterns in some influential neural network models.

3.4. Associative Learning of Patterns

Recall from the discussion in Section 3.2 that a sensory pattern can be encoded as a *distribution* or *vector* of activities across different nodes; this theme is taken up again in Chapter 4. Early neural network models yielded some insights into how a single node can learn a particular pattern (e.g., Grossberg, 1968a, 1968b) and how a network can learn to respond with one given pattern to another given pattern (e.g., Anderson, 1970, 1972). In recent years, more sophisticated versions of this type of associative learning have played a large role in applications of neural networks to pattern recognition.

There are likely to be differences between a single node learning an association with a pattern of other node activities, as occurs, for example, in the outstar (see Figure 3.4), and the same node learning an association with a single other node. Figure 3.10 illustrates one possible difference. If the association to be learned is simply between single node activities, a law whereby contiguous presentations increase connection strength (such as that of Hebb, 1949) seems to make sense. But, if a pattern is to be learned, what is important is that the distribution of learned synaptic weights comes to approximate the distribution of original pattern intensities. Under these conditions, contiguous presentation can sometimes lead to increases in some connection strengths and decreases in others.[1]

Section 3.2 discussed the early work of Anderson (1972) on association between vector patterns. Work along these lines was developed further, with both mathematical theory and implementation, by several investigators, most notably Kohonen (1984) and Kosko (1987a, 1987b, 1988). (Another discussion of associations between patterns is found in Rumelhart & McClelland, 1986a, Volume 1, pp. 33–40).

3.4.1. Kohonen Models of Autoassociation and Heteroassociation

Kohonen (1984) defined:

> two types of transformation operations, the *autoassociative recall*, whereby an incomplete key pattern is replenished into a complete (stored) version, and the *heteroassociative recall* which selectively produces an output pattern y_k in response to an input pattern x_k; in the latter case the paired associates x_k and y_k can be selected freely, independently of each other. This operation is a generalization of the simple stimulus-response (S-R) process.
>
> (p. 162)

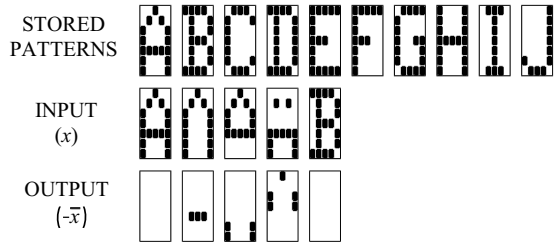


FIGURE 3.9 A demonstration of the novelty filter. The output selectively enhances those parts of the stored patterns that are absent from the current input pattern.

Source: Reprinted from Kohonen & Oja, 1976, with permission of Springer-Verlag.

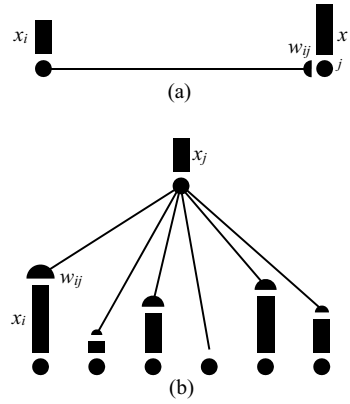


FIGURE 3.10 (a) Hebbian associative learning at single nodes. Correlation between node activities x_i and x_j always increases the LTM trace, or synaptic strength, w_{ij}. (b) Non-Hebbian associative learning of a pattern. Correlation of a spatial pattern of node activities xi with a single node activity x_j enables the LTM traces $w$ij either to increase or decrease to match the spatial pattern.

Source: Adapted from Grossberg & Levine, 1987, with permission of the Optical Society of America.

Autoassociative theory was implemented, for example, in the restoration of human faces from blurred or partially missing images; this process is illustrated in Kohonen, Reuhkala, Makisara, and Vainio (1976), Kohonen et al. (1977), and Chapter 1 of Kohonen (1977). Figure 3.11 shows one face recall simulation. Detailed neural network connections for this process are best presented in Kohonen et al. (1977); the simulations in that article are based on his differential equation for associative learning, which are given at the end of this chapter.

Kohonen's simulations employed an approximation to this associative learning equation combined with a preprocessing (sharpening) of the pattern

using lateral inhibition. In the primary patterns x_i, each "pattern element" x_i was replaced by a numerical value $x\mathrm{N}_I$, which is a weighted sum of itself and its neighboring elements, where the weighting factors λ_{ia}, for a given I, add up to 0. It was further assumed that λ_{ii} is positive for each I.

The heteroassociative theory was developed in Kohonen (1984), Chapter 6. His general theory of optimal linear mapping for paired associations encompasses the autoassociative as well as the heteroassociative case. For if the set of pattern pairs to be encoded is $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, \ldots, n$, then $\mathbf{x}_k$ can be thought of as a key for recovering a desired pattern, and $\mathbf{y}_k$ as the pattern to be recovered. In that framework, the autoassociative case occurs if $\mathbf{x}_k$ is a subpattern of $\mathbf{y}_k$ with some of the components missing (as is true, for example, in Figure 3.11 where $\mathbf{x}_k$ is part of the face and $\mathbf{y}_k$ the whole face. This work forms some of the basis for Kohonen's *self-organizing maps* (Kohonen, 1997), discussed in Chapter 7.

3.4.2. Kosko's Bidirectional Associative Memory

The theory of associations between pattern pairs was further advanced by the development in Kosko (1987a, 1987c, 1988) of the *bidirectional associative memory*, or BAM.[2] Kosko developed a dynamical system of differential equations for a general heteroassociative link between collections of nodes as shown in Figure 3.12. In the case where the activity pattern vectors are binary (consisting of 1s and 0s) or *bipolar* (consisting of 1s and –1s), Kosko (1988) proved that the states of the system converge to a stable equilibrium value denoting the pairing of patterns. (For those unfamiliar with the mathematical notion of equilibrium, or steady state, it is discussed in Section 4.2 and again in Appendix 1.) In the autoassociative case, where the a_i and b_i represent the same patterns, Kosko showed that his system is a generalization of the models of Hopfield (1982, 1984) and a special case of the system for which Cohen and Grossberg (1983) proved a convergence theorem (see Chapter 4). All these proofs have in common that the convergence to a steady state is based on a symmetry assumption in the connection weights: if w_{ij} denotes the strength of the connection from x_i to y_j in Figure 3.12, it is also the strength of the connection from y_j to x_i. This is true both for the nonadaptive case where w_{ij} are constant, and for the adaptive case, studied in Kosko (1987a), where w_{ij} vary over time according to an associative learning rule with decay.

However, strict symmetry of connections is not realistic for most forms of associative learning. Asymmetric associative learning occurs, for example, in Pavlovian conditioning: the animal learns that the bell predicts food, but not that food predicts the bell. As the conditioning models discussed in Section 6.2 make clear, what is typically learned is that the CS *precedes* the US, in fact, predicts it. Section 3.5 discusses asymmetric associative learning models that include explicit spiking at the neuronal level.

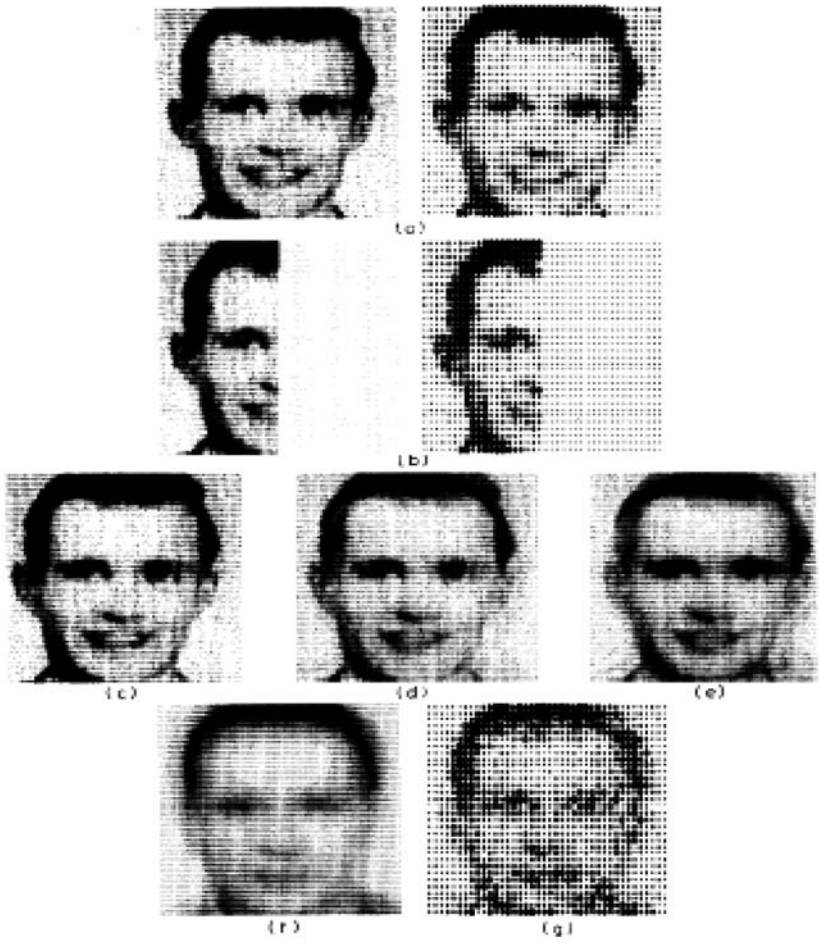


FIGURE 3.11 Face recall by an autoassociative memory network. (a) A stored face, as represented in a network with 5120 and 1280 nodes respectively. (b) Key images tested in recall. (c), (d), (e) Recollections from a 5120-unit network with 16, 160, and 500 stored images respectively. Note that quality gets worse as there are more images. (f) Recollection from a 5120-node network with 16 stored images and no lateral inhibition. (g) Recollection from a 1280-node network with lateral inhibition.

Source: Reprinted by permission of the publisher from Kohonen et al., *Neuroscience*, *2*, 1065–1076. Copyright 1977 by Elsevier Science Publishing Co., Inc.

Grossberg (1969c, 1970a) had previously generalized the learning of spatial patterns in the outstar (see Section 3.2) to the learning of *spatiotemporal patterns*, that is, time sequences of spatial patterns, in a network called the *outstar avalanche*. The avalanche consists of a collection of outstars that share the same sink nodes but whose sources are connected in series. Each source

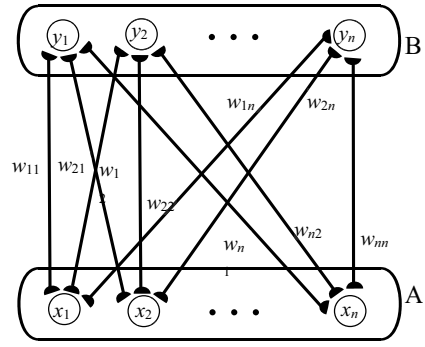


FIGURE 3.12 Bidirectional associative memory. x_i and y_j are nodes whose activity levels are either binary (1 or 0) or bipolar (1 or –1). w_{ij} are their connection weights, which are usually symmetric ($w_{ij} = w_{ji}$). Patterns are vectors of the x_i and y_j activity levels; the w_{ij} determine what pairs of patterns at A and B the network will learn to associate.

Source: Adapted from Kosko, 1988, copyright 1988 IEEE; reprinted by permission.

has learned, via synaptic weights, a different pattern of sink node activities. Hence, the spatial patterns determined by these weights are activated in sequence. In this manner, for example, a ritualistic sequence of movements or musical notes can be learned.

Just as Kosko (1988) interpreted his BAMs as reciprocal outstars, he generalized the BAMs to nonsymmetric dynamical systems known as TAMs (for *temporal associative memories*) that can be interpreted as outstar avalanches. One example of a TAM is a network that can learn a repeating temporal pattern, say (A_1, A_2, A_3, A_1). In that case, where the symmetry assumption is no longer valid, the system can be shown to converge not to an equilibrium but to an oscillatory solution or limit cycle (see Section 4.2 for a discussion of these mathematical terms).

Kosko (1987b) added competitive, or lateral inhibitory, interactions (see Chapter 4) within a level (a_i or b_i) to the adaptive BAMs. He showed that the system still converges to an equilibrium, and the equilibrium behavior approximates that of the adaptive resonance theory (ART) network of Carpenter and Grossberg (1987a). ART is a network designed particularly for pattern classification, and is discussed extensively in Chapter 7.

The multiplicity of learning laws in the neural network literature reflects an immense variation in both biological capabilities and cognitive tasks. Yet, in spite of this variability, a few basic types of laws are widely repeated. A similar repeatability is seen in the laws for neural competition discussed in the next chapter. These two sets of laws provide most of the "building blocks" needed for the models of larger-scale processes discussed in Chapters 6–9.

3.4.3. The Distributed Outstar

Carpenter (1994) generalized the outstar network (see Section 3.1) to a network with arbitrarily many source nodes as well as sink nodes, which she called the *distributed outstar*. The activity pattern at the source field of the distributed outstar can be arbitrarily distributed or restricted to one or a few nodes. Learning occurs via decrease of weights in pathways that are inactive for any length of time. Carpenter also challenged the prevailing practice (used in the original outstar of Grossberg, 1968a, and numerous networks designed by other authors) of multiplying presynaptic signal by synaptic weight to compute the influence on postsynaptic activity. This *product rule*, it was found, can lead to catastrophic forgetting in the distributed outstar if the pattern at the source field is highly distributed. This forgetting could be eliminated by replacing this rule by one where there is a signal threshold that decreases linearly as the weight increases, and the threshold is subtracted from the presynaptic signal.

3.5. Spike Timing–Dependent Plasticity

Most of the associative learning models presented in Sections 3.2–3.4 treat node activities as average firing frequencies across a large number of neurons. These models do not use information about the exact timing of spikes in presynaptic and postsynaptic neurons. Starting in the mid-1990s a number of researchers have developed models that explicitly or implicitly include spikes and that include learning laws based on the timing of those spikes.

Inclusion of spikes makes the model more biologically realistic, if the architecture of the network follows realistic design principles. Another reason for including spikes and spike timing was to come closer to the spirit of Hebb's (1949) intentions when he formulated his synaptic modification rule (see Chapter 2). According to many of the nonspiking network learning rules, synaptic weights can increase based simply on correlation of high levels of presynaptic and postsynaptic activities. Yet Hebb's rule contains the phrase "When the axon of cell *A* is near enough to excite a cell *B* and repeatedly or persistently takes part in firing it." Gorchetchnikov, Versace, and Hasselmo (2005) noted that this phrase suggests causation and not just correlation. Causation implies that the firing of neuron *A* should predict the firing soon after of neuron *B*, by analogy with the behavioral sequence in classical conditioning. These results suggest learning rules that are temporally asymmetric between presynaptic and postsynaptic elements.

Indeed, several investigators starting with Levy and Steward (1983) found that the timing between the firings of pre- and postsynaptic neurons influenced whether there was long-term potentiation (LTP) or long-term depression (LTD). Levy and Steward studied connections between two parts of the hippocampal system in the rat, the entorhinal cortex and dentate gyrus. The

connections from entorhinal to dentate that go from one hemisphere of the brain to the other (*crossed* connections) are normally much weaker than connections from entorhinal to dentate that stay in the same hemisphere (*ipsilateral* connections). Yet, if activation of the crossed pathway precedes activation of the ipsilateral pathway, LTP occurs and the crossed pathway is strengthened. If crossed pathway activation occurs after ipsilateral pathway activation, LTD occurs and the crossed pathway is weakened. Markram, Lübke, Frotscher, and Sakmann (1997) confirmed this result in cortical pyramidal neurons and showed that there is a sharp transition between potentiation and depression as the time interval changes. Several studies in the 1990s, on both isolated hippocampal and cortical slices and intact animals, pinpointed the precise timing of spikes for optimal strengthening of pathways (see Bi & Poo, 2001, for a review).

These results led to the development of computational neural network models of *spike timing–dependent plasticity* (*STDP*) (see the book edited by Maass & Bishop, 1999). One of the first of these models, by Gerstner, Kempter, van Hemmen, & Wagner (1996, 1999), was inspired by data on the auditory system of the barn owl. The owl catches prey that it detects via precise time difference in the auditory signals to the two ears. Associative learning in the model network leads to a phenomenon called *phase locking*, whereby the time delay between spikes at two different auditory nuclei becomes regular, and this delay can be shorter than the time constants of individual neurons.

Several modelers have proposed different mechanisms for spike timing–dependent plasticity based on biophysics of neurons. Gorchetchnikov and Hasselmo (2005) and Gorchetchnikov et al. (2005) noted that the model of Gerstner et al. (1999) had the desirable property that learning is dependent only on local events at the pre- and postsynaptic neurons and the synapse between them. However, the Gerstner et al. model relies on events over a long enough time interval that the information about these events is not accessible at the present. Gorchetchnikov and his colleagues developed a learning rule that is "temporally local" as well as "spatially local." In the spirit of Hebb's rule, weight changes are proportional to a product of presynaptic and postsynaptic terms. The presynaptic term is based on synaptic conductance. The postsynaptic term has a positive part dependent on transmembrane voltage and a negative part dependent on the after-hyperpolarization that follows a spike. Figure 3.13 illustrates the STDP that emerges from summing the positive and negative parts.

Since the hippocampus is a major learning site, spike timing–dependent modeling has been applied since the 1990s and early 2000s to a variety of hippocampal functions. This includes involvement of the hippocampus in navigational learning (Gerstner & Abbott, 1997), sequence learning (Minai & Levy, 1993), and autoassociative memory encoding (Lengyel, Kwag, Paulsen, & Dayan, 2005). The autoassociative model of Lengyel et al. (2005) is based

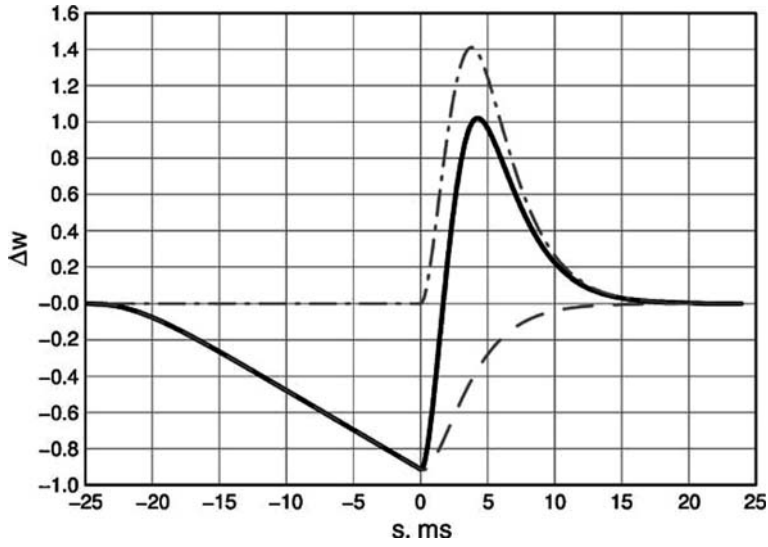


FIGURE 3.13 Example plot of an STDP curve. The postsynaptic excitatory component is shown by the dot-dashed line, the postsynaptic inhibitory component by a long-dashed line, and their sum by a bold black line.

Source: Reprinted from Gorchetchnikov et al., 2005, with the permission of Elsevier Science, Ltd.

on probabilistic encoding and optimal recovery of a pattern from a memory trace on which other patterns are superimposed, which is reminiscent of the work of Anderson and Kohonen, discussed in Section 3.2.

Another common application of STDP is to the development of maps in the cortex (especially visual cortex), typically involving columns of neurons that learn during development to respond to similar inputs. Song and Abbott (2001) noted that STDP biases competition between synapses by strengthening synapses that receive correlated inputs. This allows some neurons that become selective for particular stimuli (e.g., visual orientations) to make other nearby neurons selective for the same stimuli. Section 7.1 discusses other models of map formation in the visual cortex.

More recently, a range of models of learning processes in different parts of the brain have included explicit spiking and STDP. An example is the categorization and attentive learning model of Grossberg and Versace (2008), discussed in Section 7.3, which added thalamocortical interactions to a previously developed adaptive resonance network.

Equations for Networks in Chapter 3

Detailed Description: The Gated Dipole

In Figure 3.7, suppose that the shock input J to the "negative" channel y_1 6 x_1 6 x_3 is on for a period of, say, 100 time units. At time 0, both of the depletion synapses w_1 and w_2 are "filled" to maximum efficacy. During those 100 time steps, y_1 receives an input equal to $I + J$, where I is nonspecific arousal, whereas y_2 receives an input equal to J. Since $I + J$ is larger than J, this means y_1 becomes more activated than y_2. The synapse w_1 is depleted at a rate proportional to the product $y_1\ w_1$, whereas w_2 gets depleted at a rate proportional to $y_2\ w_2$, so w_1 (as Figure 3.7 shows) becomes depleted faster than w_2.

The crucial element in the behavior of the gated dipole network is which channel output, x_3 of the "negative" channel or x_4 of the "positive" channel, becomes active if either. This is turn determined by *which of* x_1 *and* x_2 *is larger at a given time*. x_1 grows in proportion to the product of its input y_1 and synaptic weight w_1, and x_2 in proportion to the product of y_2 and w_2. During the 100 time units that J is on, the advantage of y_1 over y_2 overcomes the advantage of w_2 over w_1. This means that x_1 is larger than x_2. As a result, x_2 which receives an input proportional (in different variants of the equations) to either $x_2 - x_1$ or its positive part, becomes activated, and x_4 does not (see the graph in Figure 3.8).

After J is shut off, y_1 and y_2 are both receiving equal inputs I, so in a short period of time y_1 and y_2 become nearly equal. Since w_1 is still much more depleted than w_2, $y_1 w_1$ is smaller than $y_2 w_2$, and so x_1 becomes smaller than x_2. This activates x_4, which in this version of the network is the output of the "positive" or "shock off" channel, representing relief from electric shock.

Finally, many more than 100 time units later, w_1 has recovered from depletion. After that time, x_1 and x_2 are equal, so that neither channel output is activated. Then the network "feels" no fear (from shock) or relief.

But Figure 3.7 is only one possible schema for a gated dipole. There could be, for example, an input J to the "positive" side y_2 of the dipole instead of to the "negative" side y_1. In that case, the sequence of events outlined earlier in the box goes through *except* with all the channels and subscripts reversed. The transient aftereffect of turning off J is then not relief but frustration, from the loss of a positive input.

Or there could be simultaneous inputs to both channels. That case was studied in the gated dipole network for decision-making by Grossberg and Gutowski (1987). In the Grossberg–Gutowski network, a *prospect* consists of different probabilities of gains and losses of specific magnitudes. This is

represented as a mixture of positive and negative inputs, usually of different intensities. The network compares a prospect B with another prospect A for desirability by first inputting A to the dipole network (on both the positive and negative sides), then letting the transmitter weights w_1 and w_2 adjust to reflect A's influence, and finally inputting B to the network and observing which channel output becomes active.

Equations for Grossberg's Outstar

A form of the outstar equations is as follows. As discussed earlier in Section 3.2, the activity x_1 of the source node is affected positively by the source node input I_i, and negatively by exponential decay back to a baseline rate (interpreted as 0). Recalling that the rate of change of x_1 as a function of time can be described by its derivative, dx_1/dt, this leads to a differential equation of the form

$$\frac{dx_1}{dt} = -ax_1 + I_1 \tag{3.12}$$

where a is a positive constant (the decay rate). The activities x_i of the sink nodes, $i = 2, \ldots, n$, are affected by inputs and decay in the same manner as is x_1. Each sink node is also affected by source node activity, weighted by the strength w_{1i} of the synapse from the source node. Hence

$$\frac{dx_i}{dt} = -ax_i(t) + bx_1(t-\tau)w_{1i}(t) + I_i(t), \quad i = 2,\ldots, n \tag{3.13}$$

where b is another positive constant and τ is a transmission time delay.

In one version of the theory, the synaptic weights (long-term memory traces) w_{1i} at the source-to-sink synapses undergo a passive decay which is counteracted by correlated activities of x_1 (with a time delay) and x_i. Hence:

$$\frac{dw_{1i}}{dt} = -cw_{1i} + ex_1(t-\tau)x_i$$

Synaptic weights, since they encode long-term memory, are assumed to decay much more slowly than potentials; hence $c << a$. In some examples of this theory, a threshold term is subtracted from x_1 in Equations (3.13) and (3.14).

A modification of the outstar equations is suggested by learning considerations and used in much of Grossberg's later work on different network architectures. The modification is to make the synaptic weights decay only if the source node is activated and not followed by sink node activation. This is achieved by replacing (3.14) (if τ is set to 0) with:

$$\frac{dw_{1i}}{dt} = x_1(-cw_{1i} + ex_1) \tag{3.15}$$

so that w_{1i} remains unchanged while $x_1 = 0$ but decreases while $x_1 > 0$ and $x_i = 0$.

Grossberg (1968a) studied the large-time (asymptotic) behavior of a general system of equations which includes both (3.12)–(3.14) and {(3.12), (3.13), (3.15)} as subcases. This general system is:

$$\begin{aligned} \frac{dx_i}{dt} &= a(t)x_i(t) + b(t)w_i(t) + I(t) \\ \frac{dw_i}{dt} &= c(t)w_i(t) + d(t)x_i(t) \end{aligned} \tag{3.16}$$

with the restrictions that $a(t)$, $b(t)$, $c(t)$, and $d(t)$ are continuous functions, and that $b(t)$ and $d(t)$ are nonnegative. (The function $x_1(t)$ is incorporated into $b(t)$ and $d(t)$.)

Recall, from Section 3.2, Grossberg's definition of a spatial pattern input as a vector of inputs $I_i(t)$ such that:

$$\text{For } j > 1,\ I_j(t) = \theta_j I(t),\ \sum_{j=2}^{n} \theta_j = 1 \tag{3.3}$$

Recall also his definition of the *relative* node activities:

$$X_i = \frac{x_i}{\sum_{j=2}^{n} x_j} \tag{3.2}$$

and the relative synaptic weights

$$W_i = \frac{w_i}{\sum_{j=2}^{n} w_j} \tag{3.1}$$

(In (3.1), the weights were doubly subscripted as w_{1i}, but the "1" is dropped for the more general system. Also, the sums were previously taken from 2 to n instead of 1 to n, when the source node x_1 was included in the equations.)

The outstar learning theorem says that, for a network obeying Equations (3.16a, b), the relative synaptic weights defined by (3.1) converge to the relative weights θ_i of the input pattern, under certain technical conditions on the inputs. These conditions are such as to guarantee that inputs are presented to both the source and the sinks for arbitrary large times. Mathematically, in the case of an

outstar obeying (3.12)–(3.14), this means that there exist two positive constants r and t_0 such that for all times $t \geq t_0$,

$$\int_\tau^t e^{-a(t-v)} I_1(v)dv \geq r \quad \text{and} \quad \int_\tau^t e^{-a(t-v)} I(v)dv \geq r$$

The basic method of proof of the outstar learning theorem involves transforming (3.16) into a system of equations in the relative activities and relative weights. The system derived from (3.16), (3.1), (3.2), and (3.3) is:

$$\begin{aligned} \frac{dX_i}{dt} &= A(t)(W_i - X_i) + B(t)(\theta_i - X_i) \\ \frac{dW_i}{dt} &= C(t)(X_i - W_i) \end{aligned} \tag{3.17}$$

where $A(t)$, $B(t)$, and $C(t)$ are nonnegative functions (specifically,

$$\text{if } x(t) = \sum_{j=1}^n x_j(t) \quad \text{and} \quad w(t) = \sum_{j=1}^n w_j(t),$$

$$\text{then } A(t) = \frac{b(t)w(t)}{x(t)},\ B(t) = \frac{I(t)}{x(t)},\ C(t) = \frac{d(t)x(t)}{w(t)}.$$

The second equation in (3.17) shows that, as t increases, each $W_i(t)$ moves closer to $X_i(t)$, whereas (3.17a) shows that $X_i(t)$ moves closer to $\theta_i(t)$. Hence, the relative node activities converge to the relative input pattern weights and then bring the relative synaptic weights toward themselves. This ultimately causes the relative synaptic weights also to converge to the θ_i.

Derivation of the Signal-to-Noise Ratio for Anderson's Linear Filter

As discussed in Section 3.2 above, Anderson (1970) considered the retrieval of a single vector pattern $\mathbf{x}$ from a stored trace $\mathbf{s} = \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is regarded as noise. Putting the trace with a *matched linear filter*, that is, taking its dot product with $\mathbf{x}$, yields an output of

$$\begin{aligned} V &= \mathbf{s} \cdot \mathbf{x} \\ &= \mathbf{x} \cdot \mathbf{x} + \mathbf{n} \cdot \mathbf{x} \end{aligned} \tag{3.18}$$

If N is the number of nodes in the network, therefore of components in each trace, the *mean* of a trace $\mathbf{x}$ is defined as the average of its components $x_1, \ldots, x_n$. For simplicity, and also to prevent biases in favor of one stored trace over others, Anderson assumed that all stored traces had the same mean m. Hence, if K is the number of traces other than the one to be retrieved, and if $\mathbf{I}$ is defined to be the

vector (1, 1, . . ., 1), then $\mathbf{x}$ and $\mathbf{n}$ can be expressed as $\mathbf{x} = \mathbf{x}_0 + m\mathbf{I}$, $\mathbf{n} = \mathbf{n}_0 + m\mathbf{I}$ for vectors $\mathbf{x}_0$ and $\mathbf{n}_0$ with mean 0. If P_0 is defined as $x_0 \cdot x_0$ (the *power* of the trace $\mathbf{x}_0$), then from (3.18), the signal component of the output V is:

$$\begin{aligned}\mathbf{x}\cdot\mathbf{x} &= (\mathbf{x}_0 + m\mathbf{I})\cdot(\mathbf{x}_0 + m\mathbf{I}) \\ &= \mathbf{x}_0\cdot\mathbf{x}_0 + 2m(\mathbf{I}\cdot\mathbf{x}_0) + m^2(\mathbf{I}\cdot\mathbf{I}) \\ &= P_0 + Nm^2\end{aligned} \tag{3.19a}$$

and the noise component is

$$\begin{aligned}\mathbf{n}\cdot\mathbf{x} &= (\mathbf{n}_0 + mK\mathbf{I})\cdot(\mathbf{x}_0 + m\mathbf{I}) \\ &= \mathbf{n}_0\cdot\mathbf{x}_0 + mK(\mathbf{I}\cdot\mathbf{x}_0) + m^2K(\mathbf{I}\cdot\mathbf{I}) \\ &= \mathbf{n}_0\cdot\mathbf{x}_0 + NKm^2\end{aligned} \tag{3.19b}$$

The actual signal-to-noise ratio is calculated as the ratio of the *squared* signal (to make all terms positive) to the average squared noise over all possible stored traces. This is $(S/N)_0 = (\mathbf{x}\cdot\mathbf{x})^2/[\mathbf{n}\cdot\mathbf{x}]^2_{\text{avg}}$, which by (3.19a,b) is equal to:

$$\frac{(P_0 + Nm^2)^2}{[(\mathbf{n}_0\cdot\mathbf{x}_0)^2]_{\text{avg}} + N^2K^2m^4}$$

Since $\mathbf{n}_0 \cdot \mathbf{x}_0$ has mean 0, the average value of its square is equal to the variance of $\mathbf{n}_0 \cdot \mathbf{x}_0$. Letting n_{0i} and x_{0i} denote components of $\mathbf{n}_0$ and $\mathbf{x}_0$, respectively, that variance is:

$$\text{var}\left(\sum_{i=1}^{N} n_{0i}x_{0i}\right) = \sum_{i=1}^{N} x_{0i}^2 \text{ var } (n_{0i}) \tag{3.20}$$

Each n_{0i} is the sum of K random variables with mean 0 and equal variance; if it is assumed that those traces have, on the average, the same power as $\mathbf{x}_0$, each of those variances becomes P_0/N, yielding a total variance of KP_0/N for each n_{0i}. This, combined with (3.20) and the definition of P_0, yields KP_0^2/N for the average value of $\mathbf{n}_0 \cdot \mathbf{x}_0$. Hence, the signal-to-noise ratio $(S/N)_0$ equals:

$$\frac{\left[P_0 + Nm^2\right]^2}{\left[K\dfrac{P_0^2}{N} + N^2K^2m^4\right]} \tag{3.21}$$

By (3.21), if $m = 0$, this ratio equals N/K, the number of nodes divided by the number of traces (see the discussion in Section 3.2). It can also be shown, using elementary calculus, that the maximum value of $(S/N)_0$ occurs for $m^2 = P_0/(N^2K)$ and equals $N/K(1 + (1/NK))$, which is not appreciably larger than N/K.

Equations for Sutton and Barto's Learning Network

Recall the network of Sutton and Barto (1981) shown in Figure 3.5, including n stimulus traces $x_i(t)$, an output signal $y(t)$, and n synaptic weights $w_i(t)$ representing connections between x_i and y. The conditioning model is based on the existence of two additional sets of variables. One of these sets of variables consists of the nonstimulating or eligibility traces $\bar{x}_i(t)$ for each sensory stimulus. The value of $\bar{x}_i(t)$ is assumed to be large when the x_i-to-y synapse is "eligible" for modification. The other is the ongoing activity level $y(t)$ of the output node, which represents a weighted average of its past activities.

All these effects (eligibilities, weighted averages, and delta rule for synaptic modification) are incorporated in the following equations for the changes in these variables from time t to time t+1:

$$\begin{aligned}
\bar{x}_i(t+1) &= \alpha\bar{x}_i(t) + x_1(t) \\
\bar{y}(t+1) &= \beta\bar{y}(t) + (1-\beta)y(t) \\
w_i(t+1) &= w_i(t) + c(y(t) - \bar{y}(t))\bar{x}_i(t) \\
y(t) &= f\left[\sum_{i=1}^{n}\left(w_j(t)x_j(t) + w_0(t)x_0(t)\right)\right]
\end{aligned} \tag{3.22}$$

where $y(t)$ is bounded to remain in the interval [0, 1] (that is, replaced by 1 if it gets above 1); α and β are constants between 0 and 1; f is a sigmoid function (see Figure 2.7); c is a positive constant determining the rate of learning; and $x_0(t)$ and $w_0(t)$ are the activity and associative strength of the US trace.

Klopf's Differential Hebbian Rule

The basic learning rule of Klopf's (1986, 1988) differential Hebbian model is inspired by the heuristic that synaptic efficacy changes as a function of changes in both presynaptic and postsynaptic activities. In addition, as discussed in Section 3.3, he made change in postsynaptic activity delayed in time, in order to simulate interstimulus interval data, and made change in efficacy of a given synapse proportional to the current efficacy, in order to simulate S-shaped animal learning curves. All these heuristics led to the equations

$$w_i(t+1) = w_i(t) + \Delta y(t-1)\sum_{j=1}^{\tau} c_j \left|w_i(t-j)\right| \Delta x_i(t-j-1) \tag{3.23}$$

where, for any given time T, $\Delta x_i(T) = x_i(T+1) - x_i(T)$ and $\Delta y(T) = y(T+1) - y(T)$; the expressions of c_j denote weighting constants for the influences of different past times of presynaptic input activity; and two vertical bars denote the absolute value.

Derivation of Rumelhart, Hinton, and Williams' Back Propagation Algorithm

As discussed in Section 3.3, Rumelhart et al. (1986) assumed that unit j (whether hidden or output) receives a signal equal to the linear sum of the outputs y_{pi} from the previous layer weighted by the connections w_{ij}. (If j is a hidden unit so that i is an input unit, y_{pi} equals the input component i_{pi}. If j is an output unit so that i is a hidden unit, then y_{pi} equals the activation function f applied to the ith net signal net_{pi}.) Hence, the signal it receives is

$$\text{net}_{pj} = \sum_j w_{ij} y_{pi} \tag{3.9}$$

If f is the activation function,[3] then the output of unit j is:

$$y_{pj} = f(\text{net}_{pj}) = f\left(\sum_j w_{ij} y_{pi}\right) \tag{3.24}$$

Now let the measure of the total error in the pth output pattern be:

$$E_p = \frac{1}{2}\sum_j \left(t_{pj} - y_{pj}\right)^2 \tag{3.25}$$

Then if (3.8b) holds, the response change δ_{pj} is simply the negative derivative of the total error E_p with respect to y_{pj}; in other words, it is a measure of how much the jth unit contributes to the incorrectness of the response.

In the case where there are hidden units and nonlinear activation functions, it is desired therefore to compute δ_{pj} by taking the derivative of E_p, from (3.24), with respect to the jth signal net_{pj}. Using the expressions (3.24) and (3.25) and the chain rule for derivatives (see Appendix 1), this gives us the transformed learning rule:

$$\delta_{pj} = f'(\text{net}_{pj})\ (t_{pj} - y_{pj}) \tag{3.10a}$$

if the jth unit is an output unit. If the jth unit is instead a hidden unit, then again using the chain rule, we obtain from (3.24) through (3.26) ("M" denoting partial derivative) that

$$\delta_{pj} = -\frac{\partial E_p}{\partial \text{net}_{pj}} = f'(\text{net}_{pj})\ \frac{\partial E_p}{\partial y_{pj}}$$

If k is the generic index of output units that receive projections from hidden unit j, we obtain, again by the chain rule and (3.10a), a value for the above expression in brackets, namely:

$$\frac{\partial E_p}{\partial y_{pj}} = \sum_k \left[\frac{\partial E_p}{\partial \text{net}_{pk}}\right]\left[\frac{\partial \text{net}_{pk}}{\partial y_{pj}}\right]$$
$$= \sum_k \left[\frac{\partial E_p}{\partial \text{net}_{pk}}\right] w_{jk}$$
$$= -\sum_k \delta_{pk} w_{jk}$$

Combining the above two expressions, we obtain finally that, if unit j is a hidden unit,

$$\delta_{pj} = f_j'(\text{net}_{pj}) \sum_k \delta_{pk} w_{jk} \tag{3.10b}$$

Grossberg's Gated Dipole

The general equations for the gated dipole of Figure 3.7 are given in Grossberg (1972c). The simplified form of those equations, with thresholds and time delays set to 0, is

$$\frac{dy_1}{dt} = -ay_1 + I + J \qquad \frac{dy_2}{dt} = -ay_2 + I \tag{3.26}$$
$$\frac{dw_1}{dt} = b(c - w_1) - ey_1 w_1 \qquad \frac{dw_2}{dt} = b(c - w_2) - ey_2 w_2$$
$$\frac{dx_1}{dt} = -fx_1 + gy_1 w_1 \qquad \frac{dx_2}{dt} = -fx_2 + gy_2 w_2$$
$$\frac{dx_3}{dt} = -hx_3 + k(x_1 - x_2) \qquad \frac{dx_4}{dt} = -hx_4 + k(x_2 - x_1)$$
$$\frac{dx_5}{dt} = -mx_5 + (x_3 - x_4)$$

where a, b, c, e, f, g, h, k, and m are all positive constants. Equations (3.26) reflect a symmetry between the "positive" and "negative" channels. Both have the same set of activity decay rates (a, f, and h), the same transmitter depletion rate (e), the same transmitter recovery rate (b), the same maximum amount of depletable transmitter (c), and the same coefficients for signal transmission between levels (g and k). I is *tonic* (active all the time), whereas J is *phasic* (on for the duration of a stimulus, such as electric shock).

Equations (3.26) were written under the assumption that the "negative" channel is the one receiving the phasic input J, as in the case of relief from electric shock. If instead the "positive" channel receives the phasic input, then the term $I + J$ in (3.26a) is replaced by I, and the term I in (3.26b) is replaced by $I + J$.

In a variant of (3.26) (see Exercise 4 below), the signals received by x_3, x_4, and x_5 from lower levels are constrained to be positive or zero. Hence, each of the terms $x_1 - x_2$ in (3.26g), $x_2 - x_1$ in (3.26h), and $x_3 - x_4$ in (3.26i) is replaced by 0 if it becomes negative.

Kosko's Bidirectional Associative Memory (BAM)

In the simplest form of the bidirectional associative memory, described in Kosko (1988), there are two *fields* or collections of nodes, A and B. The aggregate activation of the ith node in A is denoted by x_i and the activation of the jth node in B by y_j. These variables can either be binary (taking on the values 0 or 1), bipolar (taking on the values 1 or –1), or analog (taking on any of a continuous range of values). In the continuous case, the x_i and y_j are governed by the system of equations

$$\begin{aligned} \frac{dx_i}{dt} &= -x_i + \sum_j f(y_j) w_{ij} + I_i \\ \frac{dy_j}{dt} &= -y_j + \sum_j f(x_i) w_{ij} + J_j \end{aligned} \tag{3.27}$$

where f is a sigmoid function, the w_{ij} denote the (symmetric) interfield connection weights, and I_i and J_j denote the (constant) inputs to the ith and jth cells. In the adaptive version of the BAM, as described in Kosko (1987a), the w_{ij} obey associative learning equations of the form

$$\frac{dw_{ij}}{dt} = -w_{ij} + f(x_i) f(y_j) \tag{3.28}$$

Kosko (1987c) extended the ideas of (3.27)–(3.28) to include competition. For the competitive BAM, in addition to connections between A and B described by coefficients w_{ij}, there are interactions within A, described by coefficients r_{ij}, and within B, described by coefficients s_{ij}. The competitive nature of these connections is ensured by the rules $r_{ii} > 0$, $s_{ii} > 0$, and for i and j unequal, $r_{ij} = r_{ji} < 0$ and $s_{ij} = s_{ji} < 0$. Equations (3.27) are then replaced by equations of the form

$$\begin{aligned} \frac{dx_i}{dt} &= -x_i + \sum_j S(y_j) w_{ij} + \sum_k S(x_k) r_{ik} + I_i \\ \frac{dy_j}{dt} &= -y_j + \sum_j S(x_i) w_{ij} + \sum_k S(y_k) s_{ik} + J_j \end{aligned}$$

The coefficients r_{ik} and s_{ik} can either be constant or obey learning equations similar to (3.28). In either case, Kosko showed, the system converges to a solution corresponding to a set of associations between patterns in A and patterns in B.

Kosko (1987d) demonstrated a simplified form of the BAM that replaces Equations (3.27) and (3.28) by an algorithm which combines difference equations and linear threshold mappings. The network may be simultaneously taught to associate several pairs of binary vector patterns.

The BAM algorithm used is as follows:

Step 1: For all i, j, reset w_{ij}, a_i, and b_j to 0.

Step 2: Get the binary inputs into the A and B arrays for an association to be learned.

Step 3: (a) For each i, let $x_i = 2a_i - 1$. (Hence, the x_i vector is bipolar, taking on values of 1 or –1.)
(b) For each j, let $y_j = 2b_j - 1$. (Hence, the y_j vector is also bipolar.)
(c) For each pair I, j, let $w_{ij} = w_{ij} + x_i y_j$.

Step 4: If there is another association between different A and B vectors to learn, return to Step 2. The final result is that the weight matrix is the sum of the cross-correlation matrices with entries $x_i y_j$ for all pairs of patterns [A B] to be associated.

Step 5: Input binary A and B vectors to be run on the network. These may or may not be the same vectors as any of the ones input in Step 2 that determined the weights.

Step 6: Run the A-to-B iteration of the network. For each j,

(a) The new $b_j = 1$ if $\Sigma_i a_i w_{ij} > 0$;
(b) The new $b_j = 0$ if $\Sigma_i a_i w_{ij} < 0$;
(c) The new b_j equals the current value of b_j if $\Sigma_i a_i w_{ij} = 0$.

Step 7: Using the new B vector found in Step 6, run the B-to-A iteration of the network. For each i,

(a) The new $a_i = 1$ if $\Sigma_j b_j w_{ij} > 0$;
(b) The new $a_i = 0$ if $\Sigma_j b_j w_{ij} < 0$;
(c) The new a_i equals the current value of a_i if $\Sigma_j b_j w_{ij} = 0$.

Step 8: Repeat steps 6 and 7 until there are no changes in the A and B vectors.

Kohonen's Autoassociative Maps

Of the articles by Kohonen and his colleagues on autoassociative maps, the one in which the neural network connections were best developed was Kohonen et al. (1977). The simulations in that article were based on the difference equations that combine a linear input–output transformation with a Hebb-like associative learning law. Recall from Section 3.2 above that the output firing frequencies y_i

depend on input spike frequencies x_i, direct input–output connectivities y_i, and inter-unit connectivities w_{ij} in a manner shown by the linear equations

$$y_i = w_i x_i + \left(\sum_j w_{ij} y_j \right) + y_b^* \tag{3.7}$$

The w_{ij} in turn are assumed to obey the associative learning law $dw_{ij}/dt = ay_i(y_j - y_b)$, where y_b, like y_b^*, is a measure of baseline activity. In actual simulations, this differential equation is approximated by a stepwise solution of the form

$$w_{ij}(t) = \alpha(\Delta t) \sum_j \sum_k \left[y_j(t_k) - y_b \right] + w_{ij}(0)$$

where Δt is the step size, m the number of steps from time 0 to time t, and t_k the time at the end of the kth step. This difference equation is in turn substituted into (3.7) to yield

$$\begin{aligned} y_i(t) = x_i^*(t) + \alpha(\Delta t) \sum_j \sum_k y_i(t_k) \left[y_j(t_k) - y_b \right] y_j(t) \\ + \sum_j w_{ij}(0) y_j(t) + y_b^* \end{aligned} \tag{3.29}$$

where $x_i^*(t)$ denotes the ith *effective input excitation* $w_i x_i(t)$, and the values t_k denote all times previous to t.

The simulations of face recognition in Kohonen et al. (1977) combined an approximation to Equation (3.29) with a preprocessing of the pattern using lateral inhibition. In the primary patterns x_i at time t_k, each "pattern element" x_i is replaced by a numerical value x_i^* which is a weighted sum of itself and its neighboring elements. Hence, the equation $x_i^* = w_i x_i$ used above is replaced by:

$$x_i^* = \sum_a \lambda_{ia} x_a \tag{3.30}$$

where the weighting factors λ_{ia}, for a given i, add up to 0. It was further assumed that λ_{ii} is positive. Then the recollections of the patterns, weighted by past experience, are given by

$$\hat{x}_i(t) = \sum_{k=1}^{m} \Gamma_i(t, t_k) x_i(t_k) \tag{3.31a}$$

with the Γ_i defined for each given unit I and time t_k by

$$\Gamma_i(t, t_k) = \sum_j x_j^*(t) x_j^*(t_k) \tag{3.31b}$$

The sum in (3.31b) (which is a correlation between the pattern at time t and the pattern at a previous time tk) is taken over those units j that are assumed to have connections with unit i.

Exercises for Chapter 3

○1. Can you think of a way, using the framework of the 1970 and 1972 Anderson articles, to model selective attention to one trace rather than another (say, because of motivational significance)? If you believe that is difficult to do in Anderson's framework, why?

*2. Consider a Grossberg outstar whose source node has activity x_1 and whose four sink nodes have activities x_2, x_3, x_4, and x_5. Let w_2, w_3, w_4, and w_5 be the corresponding synaptic strengths as shown in Figure 3.14.

The equations defining the network are

$$\frac{dx_1}{dt} = -5x_1 + I_1$$
$$\frac{dx_i}{dt} = -5x_1 + x_1 w_1 + I_1, \quad i = 2, 3, 4, 5$$
$$\frac{dw_i}{dt} = x_1(-.1w_i + x_i), \quad i = 2, 3, 4, 5$$

(Time delays have been set to 0 for ease of implementation. Other parameters have been set to obey a boundedness criterion – see Grossberg, 1970b). The inputs I_i, I = 1 to 5, are defined below, differently for two subcases.

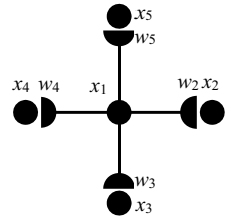


FIGURE 3.14 Outstar network used in the simulation of Exercise 2.

Solve these equations numerically using the simple Euler method (see Appendix 1) with step size .1 or less, or some other differential equation solving algorithm. The two cases are:

(a) Set I_1 = 2 for two steps on every tenth time step, starting with the first, and 0 on all other time steps. I_2, I_3, I_4, I_5 form a spatial pattern $I_i = \theta_{iI}$,

where $\theta_2 = .4$, $\theta_3 = .3$, $\theta_4 = .2$, $\theta_5 = .1$; $I = 2$ for the two time steps directly *after* those times when $I_1 = 2$, and 0 on other time steps, as shown in Figure 3.15. (Hence $I_2 = .8$, $I_3 = .6$, $I_4 = .4$, and $I_5 = .2$ when they are not zero.)

The starting value of x_1 is 0; x_i, $I = 2$ to 5 start at positive numbers of your own choosing but *not* proportional to .4, .3, .2, .1, and so do w_i.

Run up to 10,000 or more time steps.

Define $x = x_2 + x_3 + x_4 + x_5$, $w = w_2 + w_3 + w_4 + w_5$, and for each I, $I = 2, 3, 4, 5$, define

$$X_i = \frac{x_i}{x},\ W_i = \frac{w_i}{w}$$

Graph (for each i, $i = 2, 3, 4, 5$) X_i, W_i, and θ_i on the same axes, showing values at every 100th time step. (Graphing may be done either by hand or by computer. By the outstar learning theorem, these three variables should get closer together as time increases.)

(b) Do the same as in (a) except that I_2, I_3, I_4, and I_5 are a mixture of two spatial patterns. After times where I_1 is nonzero, the other I_i become nonzero for two time steps, but alternate on different presentations between

$I_2 = .8$, $I_3 = .6$, $I_4 = .4$, $I_5 = .2$

$I_2 = .5$, $I_3 = .5$, $I_4 = .3$, $I_5 = .7$.

In Part (b), the X_i should be graphed at times directly *after* the pattern presentation times in order to observe their oscillations.

*3. (a) Do the simulation in Exercise 2(a) for a variety of different initial conditions. Show that convergence is fastest when the initial values of X_i and W_i are closest to θ_i.

(b) From the result of Part (a), an outstar that has come close to learning one spatial pattern will be slow to learn another, vastly different spatial pattern (see the quote from Seneca at the start of this chapter). Confirm this by running an outstar simulation with the two patterns from Exercise 2(b) presented in succession, each for 5000 time steps.

* 4. Do a series of simulations of a slight modification of Grossberg's gated dipole equations, (3.26). The modification is that, in the equations for x_3 and x_4, the quantities in parentheses ($x_1 - x_2$ or $x_2 - x_1$) are replaced by 0 whenever they are negative. Use the following parameter values: the decay rates a, f, h, and m are all 5; c, e, and k are 1; $b = .5$; $g = 10$. Set the "shock" J to 2 units, and keep it on for a length of time that varies between runs. Study the maximum over time of the rebound x_4 as a function of

(a) intensity of arousal I;

(b) time duration of shock J.

Make tables of the results. In the right range of values that maximum rebound should increase steadily as shock increases, but should increase as arousal increases up to a point and then decrease. Hint: use 100 time units for the entire run. For the duration of J, vary between 20 and 80 time units. For the intensity of I, vary between 2 and 5.

○5. Consider some phenomenon from experimental psychology that involves response to a change in stimulation. One example is *extinction*: a response learned by classical conditioning is weakened if the conditioned stimulus is presented and not followed by reinforcement. Another would be *conditioned inhibition*: first a stimulus CS_1 is associated with a US and thereby conditioned to a response followed by a second stage where a combination of two stimuli CS_1 and CS_2 is associated with absence of the US. As a consequence, CS_2 subsequently leads to suppression of the same response when it is associated with other stimuli.

Whichever psychological phenomenon you choose, give a network interpretation of it using

(a) the differential Hebbian model

(b) the gated dipole model.

Does this suggest advantages or disadvantages of either model?

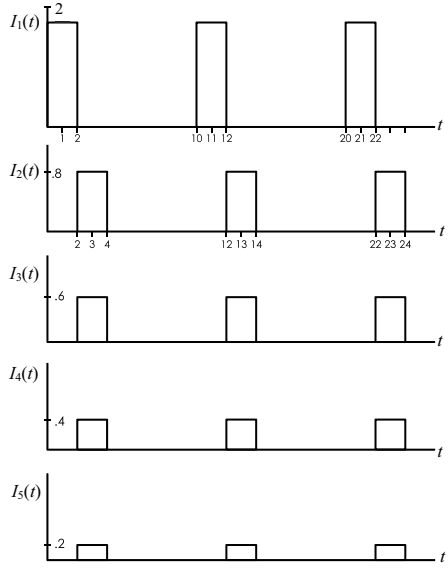


FIGURE 3.15 A spatial pattern input to the outstar shown in Figure 3.14.

* 6. Run the following simulations of Kosko's bidirectional associative memory or BAM, using the simplified algorithm described in Steps 1 through 8 of the equations section. There are six nodes at the *A* level of Figure 3.13 and four nodes at the *B* level. The network may be simultaneously taught to associate several pairs of binary vector patterns, but Kosko (1987d) states that if the number of pattern pairs is larger than the minimum of the numbers of nodes at the two levels (in this case, four), the network may not be able to learn all these associations.

The BAM algorithm used goes through the following steps in sequence:

Step 1: For all I, j, reset w_{ij}, a_i, and b_j to 0.

Step 2: Get the binary inputs into the A and B arrays for an association to be learned.

Step 3: (a) For each I, let $x_i = 2a_i - 1$. (Hence, the x_i vector is bipolar, taking on values of 1 or –1).
(b) For each j, let $y_j = 2b_j - 1$.
(c) For each pair I, j, let $w_{ij} = w_{ij} + x_i y_j$.

Step 4: If there is another association to learn, return to Step 2.

Step 5: Input binary A and B vectors to be run on the network.

Step 6: Run the A-to-B iteration of the network. For each j,

(a) The new $b_j = 1$ if $\Sigma_i a_i w_{ij} > 0$.
(b) The new $b_j = 0$ if $\Sigma_i a_i w_{ij} < 0$.
(c) The new b_j equals the current value of b_j if $\Sigma_i a_i w_{ij} = 0$.

Step 7: Run the B-to-A iteration of the network. For each I,

(a) The new $a_i = 1$ if $\Sigma_j b_j w_{ij} > 0$;
(b) The new $a_i = 0$ if $\Sigma_j b_j w_{ij} < 0$;
(c) The new a_i equals the current value of a_i if $\Sigma_j b_j w_{ij} = 0$.

Step 8: Repeat Steps 6 and 7 until there are no changes in the A and B vectors.

(a) Input the patterns

$$A^1 = (1, 0, 1, 0, 1, 0),\ B^1 = (1, 1, 0, 0)$$

and then the patterns

$$A^2 = (1, 1, 1, 0, 0, 0),\ B^2 = (1, 0, 1, 0),$$

using Step 2 twice, to set the weights.

(b) After the weights from (a) are established, input

$$A^1 = (1, 0, 1, 0, 1, 0),\ B^3 = (0, 0, 0, 0)$$

via Step 5. Then run through Steps 6–8 as often as needed and show that the network converges to the pair (A^1, B^1).

(c) With the same weights used in parts (a) and (b), input

$$A^3 = (1, 0, 1, 0, 0, 0),\ B^3 = (0, 0, 0, 0)$$

and see what vectors the network converges to. What does this say about possible steady states of the network?

(d) Add to the associations (a) one pattern pair at a time, and study how the number of time steps to convergence increases. If the number of associations increases beyond 4, the minimum of the numbers of nodes in the two levels, the network may in fact be unable to learn all the associations simultaneously.

*7. Kosko's simplified algorithm is very similar to the algorithm for Kohonen's *correlation matrix memory* (see p. 183 of the 1988 edition of Kohonen, 1984). For this algorithm, if the vector pairs $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, \ldots, n$, are to be associated, an optimal matrix $\mathbf{W}$ is chosen for that purpose, and

$$\mathbf{W} = \sum_{k=1}^{n} \mathbf{y_k} \mathbf{x_k^T}$$

where T denotes the transpose.

For example, if $\mathbf{x}_k = (1, 0, 0)$, and $\mathbf{y}_k = (1, 1, 0)$, then

$$\mathbf{y}_k \mathbf{x}_k^T = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

using standard matrix multiplication. If the $\mathbf{x}_k$ that are encoded are *orthogonal*, that is, the dot product of any two of them is 0, then $\mathbf{W}\mathbf{x}_k = \mathbf{y}_k$ for each k. The response of the system to any pattern is obtained by multiplying the vector encoding that pattern by the matrix $\mathbf{W}$.

If $\mathbf{x}_k = \mathbf{y}_k$ for each k, the correlation matrix memory is called auto-associative; otherwise it is called heteroassociative. Now consider the following set of binary vectors:

$\mathbf{A} = (1, 0, 0, 0, 0)$ $\quad$ $\mathbf{E} = (1, 0, 1, 0, 0)$
$\mathbf{B} = (0, 1, 0, 0, 0)$ $\quad$ $\mathbf{F} = (1, 0, 0, 0, 1)$
$\mathbf{C} = (0, 0, 1, 0, 0)$ $\quad$ $\mathbf{G} = (1, 1, 1, 0, 0)$
$\mathbf{D} = (0, 0, 0, 1, 0)$

(Note that the vectors **E**, **F**, and **G** can each be considered as noisy versions of one of the vectors **A**, **B**, **C**, and **D**, or of some sum of these vectors. Note also that **A**, **B**, **C**, and **D** are orthogonal.)

(a) Simulate an autoassociative, correlation matrix memory for the vectors **A**, **B**, **C**, and **D**. Then list and discuss the response of the system to each of the above seven patterns.
(b) Do the same as (a) with the vector **E** added to the correlation matrix memory.
(c) Simulate a heteroassociative correlation matrix memory that associates each of the vectors **A**, **B**, **C**, and **D**, to their bitwise complements, **P**, **Q**, **R**, and **S**, respectively:

P = (0, 1, 1, 1, 1)
Q = (1, 0, 1, 1, 1)
R = (1, 1, 0, 1, 1)
S = (1, 1, 1, 0, 1)

Then list and discuss the response of the system to **P**, **Q**, **R**, and **S**.

Some Additional Sources

General Models of Associative Learning

Abbott and Dayan (1999); Sejnowski and Tesauro (1989).

Models of Differential Hebbian Learning

Rao and Sejnowski (2000); Roberts (1999).

Models of Spike Timing–Dependent Plasticity

Kepecs, van Rossum, Song, and Tegner (2002); Porr, Saudargiene, and Wörgötter (2004); Rao & Sejnowski (2001); Song, Miller, and Abbott (2000).

Notes

1. The outstar discussed in Section 3.2, which incorporates non-Hebbian associative learning, in fact includes both Hebbian and error-correcting or delta rule elements. For Equation (3.16) at the end of this chapter includes a "Hebbian" term that combines pre- and postsynaptic activities to determine the weight change between them. However, the Equations (3.18) for *relative* weights and activities incorporate "correction" of these weights in the direction of a desired spatial pattern θ_i.
2. The order of the dates in Kosko's three articles seems to be an accident of journal scheduling, since the most basic of these articles is the one that is dated 1988!
3. Rumelhart et al. (1986) demonstrated this rule with a separate activation function f_j for each node index j. Since this does not affect the demonstration herein, and since most of their actual simulations used the same activation function for all nodes, we are using a single f for simplicity.

4

COMPETITION, LATERAL INHIBITION, AND SHORT-TERM MEMORY

> Victory at all costs, victory in spite of all terror, victory however long and hard the road may be; for without victory there is no survival.
>
> Winston Churchill (Speech, May 13, 1940)

> O for a life of Sensations rather than Thoughts!
>
> John Keats (*Letter to Benjamin Bailey*)

Inhibitory connections in neural networks serve a variety of purposes. The discussion of random nets in Section 2.2 noted that inhibition can facilitate the stabilization of network activity levels. Also, the discussion of network principles in Chapter 1 noted that inhibition can provide a mechanism for making choices. These choices might be, for example, between input patterns for short-term memory storage, between categories for classification of a single input pattern, or between drives for activation. It must be added, however, that the choices are not always all-or-none.

Both the stabilization and choice properties have been achieved in neural networks using mechanisms suggested by sensory (particularly visual) anatomy and physiology. The next section provides some history of the ideas behind those mechanisms.

4.1. Contrast Enhancement, Competition, and Normalization

The systematic study of visual perception was advanced in the middle to late nineteenth century by the noted physicists Helmholtz and Mach. In particular, both of these scientists observed that edges or contours between light and dark portions of a scene tend to be enhanced relative to the light or dark interiors

of the scene. They explained this phenomenon by means of networks of retinal cells, each excited by light within a central area and inhibited by light within a surrounding area. Receptive fields with that structure were later found experimentally, in the compound eye of the horseshoe crab *Limulus* (Hartline & Ratliff, 1957) and in the vertebrate retina (Kuffler, 1953). This kind of structure is sometimes called *on-center off-surround*. In neural networks, on-center off-surround interactions are typically modeled by *lateral inhibition*, that is, mutual inhibition between neurons or nodes at the same level of processing (the competition of the thought experiment in Chapter 1).

Figure 4.1 shows schematic pictures of two types of lateral inhibitory architectures used in pattern processing models: *nonrecurrent* or feedforward, and *recurrent* or feedback inhibition. The principle of lateral inhibition generalizes to networks where nodes *both* excite and inhibit each other, but inhibition operates over a greater distance than excitation, as shown in Figure 4.2.

This chapter explores the functions of lateral inhibition in the transformation and short-term storage of patterns in model neural networks. There is some experimental evidence that in actual mammalian nervous systems, the lateral inhibitory principle is operative at central as well as peripheral areas. Several neuroscientists have found that the largest neurons in the cerebral cortex, which are called *pyramidal cells*, typically excite smaller neurons which in turn project to and inhibit other, nearby pyramidal cells (e.g., Feldmeyer, Egger, Lubke, & Sakmann, 1999; Hirsch & Gilbert, 1991; Krimer & Goldman-Rakic, 1991; McGuire, Gilbert, Rivlin, & Wiesel, 1991). Similar kinds of interactions between large and small cells occur in subcortical areas such as the hippocampus (Sloviter & Brisman, 1995) and cerebellum (Dizon & Khodakhah, 2011; Eccles, Ito, & Szentagothai, 1967). There is also some evidence for longer-range lateral inhibition in the cortex mediated by pathways connecting the cortex to subcortical brain areas such as the thalamus and basal ganglia (see Figure A2.5 of Appendix 2 for locations of these areas, and Taylor & Alavi, 1993, 1995, for a summary).

In the models discussed in this chapter, the functional units or nodes are frequently collections of neurons sharing some common response properties. This idea, previously suggested in Section 2.2, has a partial physiological basis in the organization of the visual cortex (Hubel & Wiesel, 1962, 1965) and somatosensory cortex (Mountcastle, 1957) into columns of cells with the same preferred stimuli. Moreover, columns that are close together also tend to have preferred stimuli that are close together. Evidence for such columnar organization has also appeared in the medial temporal area of visual cortex (Fujita, Tanaka, Ito, & Cheng, 1992) and in prefrontal multimodality association cortex (Fuster, Bauer, & Jervey, 1982; Goldman-Rakic, 1984; Rosenkilde, Bauer, & Fuster, 1981; Sawaguchi, 1996). Because the evidence from vision

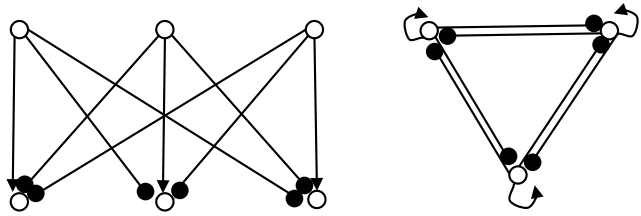


FIGURE 4.1 Examples of two kinds of lateral inhibitory networks: (a) nonrecurrent (feedforward); (b) recurrent (feedback).

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

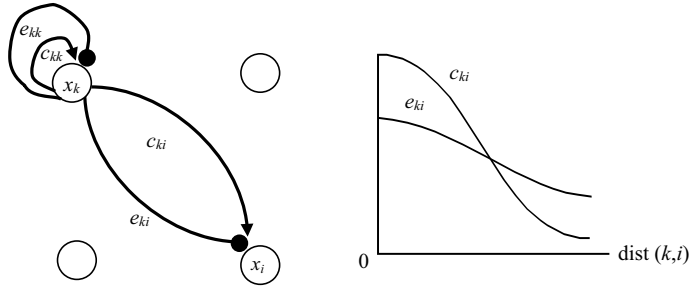


FIGURE 4.2 Generalization of the recurrent lateral inhibition in Figure 4.1(b). Each node x_k sends *both* excitation and inhibition to itself and to all other nodes x_i. Excitatory interaction strengths c_{ki} and inhibitory interaction strengths e_{ki} both decrease with distance, reflecting fewer synaptic connections as distance increases, but c_{ki} decreases more rapidly, as seen in the graph.

is the most compelling so far, many of the models in this chapter are inspired by visual data, though the modeling principles they incorporate may be more broadly applicable.

Some general themes about the functions of lateral inhibition emerged in early modeling studies from the late 1960s and early 1970s. Many of the lateral inhibitory networks studied at that time did not include learning, but were later embedded in multilevel architectures that include learnable connection weights between levels (see Chapters 6–9).

4.1.1. Hartline and Ratliff's Work and Other Early Visual Models

Hartline and Ratliff (1957) modeled inhibition in the horseshoe crab eye by means of a pair of simultaneous algebraic equations for two mutually inhibiting receptors, as follows:

$$x_1 = \left(I_1 - k_{12}(x_2 - \theta_2)^+\right)^+$$
$$x_2 = \left(I_2 - k_{21}(x_1 - \theta_1)^+\right)^+$$

Here x_i denotes the impulse frequency in the axon of cell i, and I_i denotes the excitation of cell i by an external stimulus. θ_2 and θ_1 are threshold frequencies that each cell has to exceed before it can exert inhibition, and k_{12} and k_{21} are coefficients of inhibitory action. Finally, for any real number x, the quantity x^+ denotes x if x is positive, and 0 if x is negative or 0; for example, if $x_2 \leqslant \theta_2$, then $(x_2 - \theta_2)^+ = 0$, whereas if $x_2 > \theta_2$, then $(x_2 - \theta_2)^+ = x_2 - \theta_2$.

The linear Hartline–Ratliff equations proved effective in the modeling of a variety of experimental data, and several other early lateral inhibitory models used extensions of these equations. But other effects, many of them nonlinear, had to be introduced to model some additional complexities of vertebrate vision. For example, Sperling and Sondhi (1968) developed a lateral inhibitory model of effects in the mammalian retina in order to explain certain data on luminance and flicker detection. Their model includes both feedback and feedforward stages. In the feedback stage, as shown in Figure 4.3, the jth node is excited by the $(j-1)$st node, for $j > 1$, and inhibited by feedback from the nth node.

The type of inhibition exerted by the feedback stage in Sperling and Sondhi's model is *shunting* rather than *subtractive*. In subtractive inhibition, the incoming signal is linearly weighted, and an amount proportional to that signal is subtracted from the activity (or firing frequency) of the receiving node. In shunting inhibition, the amount subtracted is also proportional to the activity of the receiving node. Thus the inhibiting node acts as if it *divides* the receiving node's activity by a given amount, that is, as if it "shunts" a given fraction of the node's activity onto another, parallel pathway.

In addition to shunting (multiplicative) inhibition, recent lateral inhibitory models often include *shunting excitation*, whose strength is proportional to the

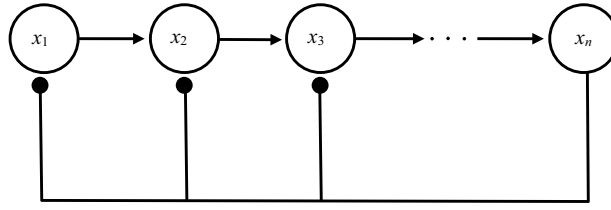


FIGURE 4.3 Schematic feedback connections in the flicker detection model of Sperling and Sondhi (1968). (Their actual diagrams used an electrical analogy with resistors and capacitors.)

difference of a node's activity from its maximum possible level. This is in contrast to *additive* excitation, the opposite of subtractive inhibition, which simply adds an amount proportional to excitatory signal to the activity of a receiving node. Shunting interactions in neural networks have been suggested by experimental results on the effects of a presynaptic neuron on the conductances of various ions across the postsynaptic membrane (see Grossberg, 1973; Hodgkin, 1964; Appendix 2 of this book). Additional evidence for shunting interactions in actual neurons has been summarized in Blomfield (1974) and Freeman (1983).

Sperling and Sondhi (1968) described the effect of shunting inhibition as "reducing dynamic range." In other words, while sensory inputs can be arbitrarily intense, the response of network nodes to these inputs has an upper limit. But, while lateral inhibition can reduce distinctions between input intensities at extreme ranges, it can enhance such distinctions at intermediate ranges. Variants of the latter effect have been called "contour enhancement" (Grossberg, 1973; Ratliff, 1965), "input-pattern sharpening" (Morishita & Yajima, 1972), and "contrast enhancement" (Ellias & Grossberg, 1975; Grossberg & Levine, 1975). The latter term is the main one we use in this chapter.

Contrast enhancement is an outgrowth of decision or competition between inputs. Competition can be biased in favor of either more intense or less intense inputs by nonlinear interactions. In multilevel networks (Chapters 6–9), competition can also be biased in favor of motivationally significant inputs.

Also, similar competitive mechanisms can exist at many levels in the brain. Whereas different sensory inputs compete for storage in short-term memory, for example, different drives or gross modes of behavior can also compete for activation, as in a model by Kilmer et al. (1969) of the midbrain reticular formation. Sections 4.2 and 4.3 of this chapter concentrate mainly on the dynamics of competition between inputs, particularly at the level of sensory areas of the cerebral cortex. This includes a sketch of the growing body of mathematical results on competitive neural systems. Section 4.4 returns briefly to competition at other cognitive levels: between drives, between categories, and between behavior sequences.

Besides contrast enhancement, another common property of lateral inhibitory networks is *pattern normalization* (see Figure 4.4). Normalization (e.g., Grossberg, 1970a) means that a pattern of node activities $x_1, x_2, \ldots, x_n$ at one ("input") level is replaced by activities $y_1, y_2, \ldots, y_n$ at the next ("output") level that are proportional to the x_i's but independent of the total intensity (sum of x_i values). Normalization has often been used in lateral inhibitory networks to keep the total network activity within bounds, so it will not get large enough to damage neurons. This concept is reminiscent of Sperling and Sondhi's reduction of dynamic range.

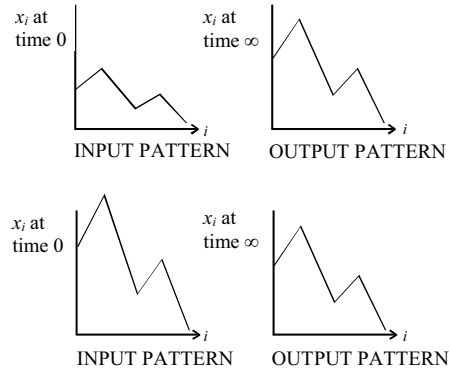


FIGURE 4.4 Example of pattern normalization. Output pattern has same *relative* activities as input pattern, but is independent of *absolute* input activities.

Source: Adapted from *Mathematical Biosciences*, *66*, D. S. Levine, Neural population modeling and psychology: A review, 1–86, copyright 1983, with permission from Elsevier Science.

4.1.2. Nonrecurrent versus Recurrent Lateral Inhibition

In early models involving lateral inhibition, nonrecurrent (feedforward) and recurrent (feedback) inhibition were preferred for different purposes and used to model different processes. Recall from Section 2.1.2 that recurrent (reverberating) loops have been used in neural models since the 1940s to extend the duration of a stimulus trace. Grossberg (1970, 1972d), modeling pattern discrimination in the retina, chose nonrecurrent rather than recurrent lateral inhibition in order to shorten the duration of pattern representations. This was done because the retina is designed to encode a fairly accurate representation of ongoing visual events. The visual cortex, by contrast, is designed to encode both present events and memories of recent past ones. Hence, for modeling of pattern processing at the cortical level, it is important to keep patterns active in sensory memory for longer periods. Therefore recurrent lateral inhibition tends to be preferred in cortical models (e.g., Grossberg, 1973; Wilson & Cowan, 1972). Differences between actual cellular architecture in the retina and the cortex generally reflect this functional difference.

An example of the "retinal" level of modeling is shown in Figure 4.5, adapted from Grossberg (1970, 1972d). Two stages of nonrecurrent inhibition are constructed so that a particular node fires in response to one and only one *space-time pattern*, that is, to one time-varying input distribution. Aspects of this network's anatomy are reminiscent of particular layers of the vertebrate retina.

In the next section, we concentrate on the "cortical" level of modeling. In particular, we consider the use of networks with recurrent lateral inhibition (and, sometimes, lateral excitation) to model short-term storage of sensory patterns.

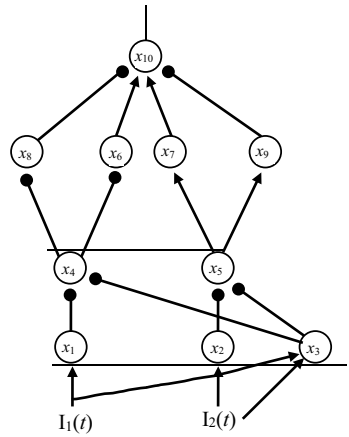


FIGURE 4.5 Network to recognize a single space-time pattern (consisting of inputs $I_i(t)$, two of which are shown). Subnetwork x_1 through x_5 does *low-band filtering* (filtering out patterns with activity levels lower than those in the desired pattern) and pattern normalization (see Figure 4.4). Subnetwork x_4 through x_{10} does *high-band filtering* (filtering out patterns with activities higher than those in the desired pattern). x_i corresponds roughly to retinal cell layers: x_1 and x_2 to receptors, x_3 to horizontal cells, x_4 and x_5 to bipolar cells, x_6 through x_9 to amacrine cells, x_{10} to ganglion cells.

Source: Adapted from Grossberg, 1970b, with permission of Academic Press; see that article for details.

4.2. Lateral Inhibition and Excitation between Sensory Representations

Short-term memory in recurrent lateral inhibitory networks has been modeled since the early 1970s, particularly by Wilson and Cowan, Grossberg, Amari, and their colleagues. Typically, an input pattern is regarded as the initial state of a mathematical *dynamical system*, which can roughly be defined as the movement through time of the solutions of a system of differential equations for interacting variables (see Appendix 1 for further discussion). This solution is described by a vector composed of the values of all the variables in the system at any given time (see the discussions in Section 3.2). The equations describe the transformation of this pattern and its storage in short-term memory; the stored pattern is then regarded as a limiting vector to which the system converges as time gets large.

Lateral inhibitory architectures, as stated in Section 4.1, tend to enhance contrasts between pattern intensities. The inhibitory connections mean that larger activities tend to suppress smaller ones; thus, after a certain time, some subcollection of nodes becomes, and remains, dominant. As a consequence, dynamical systems defined by such networks often, but not always, converge to an *equilibrium* state (also called a *steady* state). An equilibrium is a state

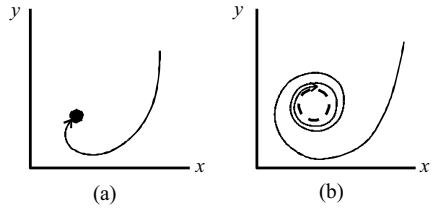


FIGURE 4.6 If a dynamical system has two variables (say x and y in these graphs), the changes in these variables over time can be shown by a curve. Arrows denote direction of flow as time increases. The system can either approach an *equilibrium point* (Part (a)) or else oscillate toward a limiting curve or *limit cycle* (the inner dashed circle in Part (b)).

where the system interactions are in "balance," so that, once the system reaches that state, it will not be perturbed from it. (See Appendix 1 for a more precise mathematical definition of an equilibrium.) The study of pattern transformations by analysis of the equilibrium behavior of a dynamical system was common in early neural network models (e.g., Cohen & Grossberg, 1983; Geman & Geman, 1984; Golden, 1986; Hopfield, 1982; Hopfield & Tank, 1985, 1986; White, 1987). System vectors do not always converge to an equilibrium, however. The next section includes examples of networks where the state vector converges instead to a *limit cycle*, that is, to a periodic (oscillatory) solution, as shown in Figure 4.6.

4.2.1. Wilson and Cowan's Work

One of the first published mathematical studies of a neural network emulating cortical lateral inhibition was done by Wilson and Cowan (1972). This work was further elaborated in Wilson and Cowan (1973) (where the lateral inhibitory principle is most explicit) and Ermentrout and Cowan (1980).

The network of Wilson and Cowan (1972) is composed of neuron populations whereby intrapopulation connections are random and interpopulation connections deterministic. This is reminiscent of the notion of "randomness in the small and structure in the large" (Anninos et al., 1970) discussed in Section 2.2. The distributions of different cell thresholds are averaged into sigmoid functions (see Figure 2.7b) describing input–output relations at the node (i.e., population) level.

The model of Wilson and Cowan (1972) includes excitation as well as inhibition between nodes. The network is described in terms of the two functions $x_E(t)$ and $x_I(t)$, which denote the proportions of excitatory and inhibitory cells, respectively, firing per unit time at time t. The change over time of the excitatory activity $x_E(t)$ reflects a combination of three influences (incorporated later in Equations (4.10)):

(a) decay back to a baseline activity level (assumed to be zero, for simplicity);
(b) a signal that linearly combines excitation from excitatory cells, inhibition from inhibitory cells, and excitatory inputs from outside the network, and then is transformed by a sigmoid signal function;
(c) refractory periods of the excitatory cells themselves (see Section 2.2), so that only cells that have not fired within a recent time interval can be excited by the signal described in (b).

The inhibitory node is subject to the same set of influences, though the interaction coefficients and sigmoid functions are different from those for the excitatory node.

Mathematically, the effect of refractory periods on the excitatory node x_E is described by a term that decreases linearly with excitatory signal strength, that is, a term of the form $a_E - b_E x_E$ for some positive numbers a_E and b_E. As a consequence, any given input signal has the strongest effect on an inactive node, and no effect on a node already at a saturation point of activity (namely a_E/b_E). This term is multiplied by the strength of the incoming signal. Similarly, the signal to the inhibitory node is multiplied by a decreasing linear factor $a_I - b_I x_I$. Grossberg (1973) noted that the factors $a_E - b_E x_E$ and $a_I - b_I x_I$ are equivalent to terms for shunting (multiplicative) excitation in passive membrane equations for a single neuron (see the discussion of Sperling and Sondhi in Section 4.1); this point is addressed in the next subsection.

A second article by Wilson and Cowan (1973) described a two-dimensional network for representing an area of the cerebral cortex or thalamus. (The thalamus is an area of the brain just below the cortex, much of it connected to the cortex in a one-to-one fashion and providing a "relay" to the cortex from lower brain areas; see Appendix 2.) This network has properties similar to the network of Wilson and Cowan (1972) with the addition of distance-dependent interactions. The two variables $x_E(t)$ and $x_I(t)$ are replaced by variables $x_E(s,t)$ and $x_I(s,t)$ that depend on distance as well as time, and excitation falls off more sharply with distance than does inhibition (see Figure 4.2).

Hence, different positions in the visual field, or different line orientations, can be represented at different cortical or thalamic locations. The network variables represent averaged activities of excitatory and inhibitory neurons at these locations.

The distance-dependent networks of Wilson and Cowan (1973) have the same range of limiting behavior as those of their earlier article, despite greater mathematical complexity. The large-time dynamics of Wilson and Cowan's equations include the possibility of *hysteresis*, whereby if the amount of external stimulation is changed, the dynamics are dependent on the past history of stimulation. These equations can also, for some parameters, exhibit limit cycles (see Figure 4.6). Wilson and Cowan saw limit cycles as possible analogs of the reverberatory loops between the cerebral cortex and the

thalamus. These loops have often been suggested as a physiological substrate for short-term memory (see Section 2.1).

The network of Wilson and Cowan (1973) reproduced various phenomena of visual psychophysics. This included characteristic responses to different spatial frequencies; *metacontrast*, or perceptual masking of a brief stimulus by a second, subsequent, stimulus presented elsewhere in the visual field; and a hysteresis phenomenon found in stereopsis. (Stereopsis, or three-dimensional binocular vision, is discussed further in Section 4.3.) Ermentrout and Cowan (1980), studying a more abstract version of Wilson and Cowan's network, proved the existence of periodic solutions, and discovered that these solutions had properties in common with some simple visual hallucinations.

4.2.2. Work of Grossberg and Colleagues

Another series of articles on lateral inhibition between cortical populations was initiated by Grossberg (1973), who used shunting interactions but combined excitatory and inhibitory influences in a different manner than did Wilson and Cowan. Later articles in this series include Grossberg and Levine (1975), Ellias and Grossberg (1975), Levine and Grossberg (1976), Levine (1975), Grossberg (1978a), Cohen and Grossberg (1983), and Cohen (1988).

The implementation of shunting recurrent lateral inhibition in Grossberg (1973) was motivated by some heuristics relating to shunting nonrecurrent inhibition. The need for inhibition in his model arose in turn from consideration of a shunting network without inhibition, as defined by differential equations. The left-hand side of a differential equation for a variable denotes the variable's rate of change, using the symbol dx/dt, while the right-hand side describes the interactions causing it to change; see Appendix 1 for more details. The equations for Grossberg's network are:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_i \tag{4.1}$$

for the activity x_i of the ith node as a function of time, where I_i are outside inputs, B is the maximum possible activity of each node, and A is a decay rate. Equation (4.1) says that shunting (multiplicative) excitation, proportional to the difference of x_i from its maximum activity B, is supplied by outside inputs, whereas shunting inhibition, proportional to x_i itself, is supplied only by spontaneous decay.

But if the inputs I_i to the nodes defined by (4.1) form a spatial pattern (see Figure 3.4), the shunting term $B - x_i$ in (4.1) causes a distortion of relative pattern weights. (This assertion is justified in Exercise 1 of this chapter.) The distortion is removed (also see Exercise 1) by making the ith input not only excite the ith node but inhibit all other nodes (see Figure 4.7). Equations (4.1)

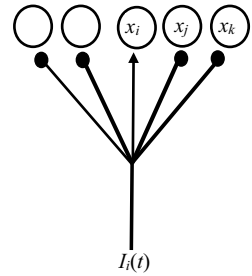


FIGURE 4.7 Nonrecurrent on-center off-surround interactions.
Source: Adapted from Grossberg, 1976a, with permission of Springer-Verlag.

are then modified so that shunting inhibition is from a combination of spontaneous decay and inputs to other nodes:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_i - x_i \sum_{k=i} I_k \tag{4.2}$$

where the "Σ" is standard notation for a sum (in this case, taken over all inputs exciting nodes other than the *i*th).

In many versions of Grossberg's recurrent network, the same nodes are both excitatory and inhibitory. (The exception is Ellias & Grossberg, 1975). Shunting inhibition (proportional, in general to x_i minus its minimum activity C_i) and excitation (proportional to the maximum *i*th node activity B_i minus x_i) are now supplied partly by outside inputs and partly by the node itself and other network nodes. Hence, (4.2) is replaced by:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)\left(\sum_{k=1}^{n} f(x_k)c_{ik} + I_i\right) - (x_i - C_i)\left(\sum_{k=1}^{n} f(x_k)e_{ik} + J_i\right) \tag{4.3}$$

where each node might possibly receive an "excitatory input" I_i and an "inhibitory input" J_i. Expression (4.3), while appearing as a single equation, actually represents a system of equations for the activities of arbitrarily many nodes with identical connection properties. In these equations, f is a *signal function* reflecting input–output transformations at the single-cell level. The function f might or might not be sigmoid (a point to which we return later) but must be increasing with x_k. The positive constants c_{ik} and e_{ik} are non-modifiable excitatory and inhibitory interaction coefficients, and $I_i(t)$ is the input to the *i*th node.

Equations (4.3) are similar to Wilson and Cowan's, except that in (4.3) the signal function is computed separately for the excitatory and inhibitory inputs. Therefore, excitatory and inhibitory inputs combine not linearly as in Wilson

and Cowan's network, but nonlinearly via shunting terms. Grossberg (1973) interpreted these shunting interactions as separate excitatory and inhibitory *gain control*, allowing for automatic tuning of network sensitivity in response to fluctuating inputs. Also, if these equations reflect averaging on the node level of the single-neuron equations, B_i and C_i can be interpreted as equilibrium voltages for sodium and potassium ions respectively (see Appendix 2 for discussion of the physiological significance of sodium and potassium). In this interpretation, the factors under the summation signs are conductances of those ions across the neuron membrane.

Grossberg (1973) and succeeding articles include mathematical proofs that, as time gets large, the variables x_i in (4.3) always converge to steady-state values (limits) for broad classes of functions f. These steady-state values can either be zero for all nodes, or nonzero for one or more nodes. The nonzero limiting node activities, and their relative sizes, were interpreted as reflecting the network's choice as to what parts, if any, of a pattern are to be stored in short-term memory. Periodic or chaotic patterns of the system variables are thereby prevented (see Appendix 1 for discussion of these alternatives in mathematical dynamical systems).

Computer simulations of Grossberg's network show that the system reaches activities close to its steady-state values quickly (in a few minutes if decay rates are chosen to reflect neuronal time constants). Hence, this long-time behavior actually approximates short-time sensory pattern transformations such as occur in short-term memory.

In Grossberg (1973), the dynamics of Equations (4.3) were studied, with all maximum activities B_i equal. The values x_i at time 0 reflect the input pattern, and pattern transformations after time 0 reflect the recurrent interactions. The connectivity of the network is pure on-center off-surround; that is, the excitatory coefficients c_{ik} are set to 1 if $i = k$ and 0 otherwise, with the reverse true for the inhibitory coefficients e_{ik}.

In the network of Grossberg (1973), the steady state approached by the system depends on the function f; in particular, it depends on whether f grows linearly with x_i, faster than linearly, slower than linearly, or in a sigmoid fashion (i.e., faster than linearly for small x_i and slower for large x_i; see Figure 4.8). If f is linear in x_i, the output values, $x_i(\infty)$ (x_i at the limiting or "infinite" time) are proportional to the input values $x_i(0)$; in the case of a visual pattern, for example, relative reflectances are conserved. Such proportionality might initially appear to be a good property for a sensory memory system. But Grossberg argued that perfect proportional storage is undesirable because it means that insignificant network noise is stored along with significant signal traces. A better outcome, he stated, is contrast enhancement with noise suppression; that occurs if f is a suitable sigmoid function. For such functions, $x_i(\infty)$ is proportional to $x_i(0)$ if $x_i(0)$ is above a threshold value and equal to 0 if $x_i(0)$ is below that threshold.

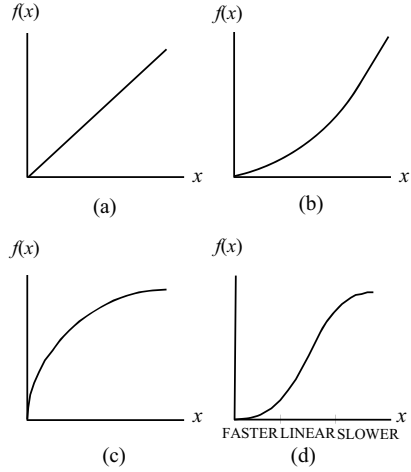


FIGURE 4.8 Schematic graphs of four types of signal functions: (a) linear; (b) faster than linear; (c) slower than linear; (d) sigmoid.

Grossberg and Levine (1975) generalized many of these results to (4.3) with unequal maximum activities B_i, representing biases in the competition between nodes for storage of their preferred sensory features. Possible sources of such biases are: (1) some stimuli occur more often than others during development, causing unequal growth of relevant feature detectors (see Blakemore & Cooper, 1970, or Hirsch & Spinelli, 1970); (2) some stimuli are attended to more than others during adult life; or (3) neuron populations exhibit random inequalities of cell distribution. The results of Grossberg (1973) on contrast enhancement with noise suppression remain true within each *subfield* (defined as the subnetwork of all nodes with a given value of B_i). A choice is made *between* subfields, and typically the activities of all but a small number of subfields are suppressed as t becomes large. The subfields chosen are the winners of a "tug-of-war" between those with largest inputs and those favored by network biases (see Figure 4.9).

Later articles (Cohen & Grossberg, 1983; Grossberg, 1978a) further extended these theorems on global existence of limits. To date, the most general theorem of this sort is Cohen and Grossberg's. They proved the existence of limits for Equations (4.3) in the case of self-excitation with distance-dependent inhibition. Self-excitation is defined as the situation where $c_{ik} = 1$ when $i = k$ and 0 otherwise; distance-dependent inhibition is defined as the situation where e_{ik} values are arbitrary as long as $e_{ik} = e_{ki}$.

4.2.3. Work of Amari and Colleagues

One of the pioneers in mathematical study of random neural networks, as noted in Section 2.2, was Amari. Like Wilson and Cowan's work, Amari's work evolved from consideration of randomly connected networks to study of networks with connections exhibiting lateral inhibition and lateral excitation. Amari (1977a) and Amari and Arbib (1977) studied networks of the latter type, based on nonlinear distance-dependent interactions that are additive rather than shunting. This work became the basis for a theory of categorization developed by Amari and Takeuchi (1978); see Section 7.1. (Hirsch, 1990, showed that the Amari–Takeuchi system is equivalent to that of Hopfield, 1984.)

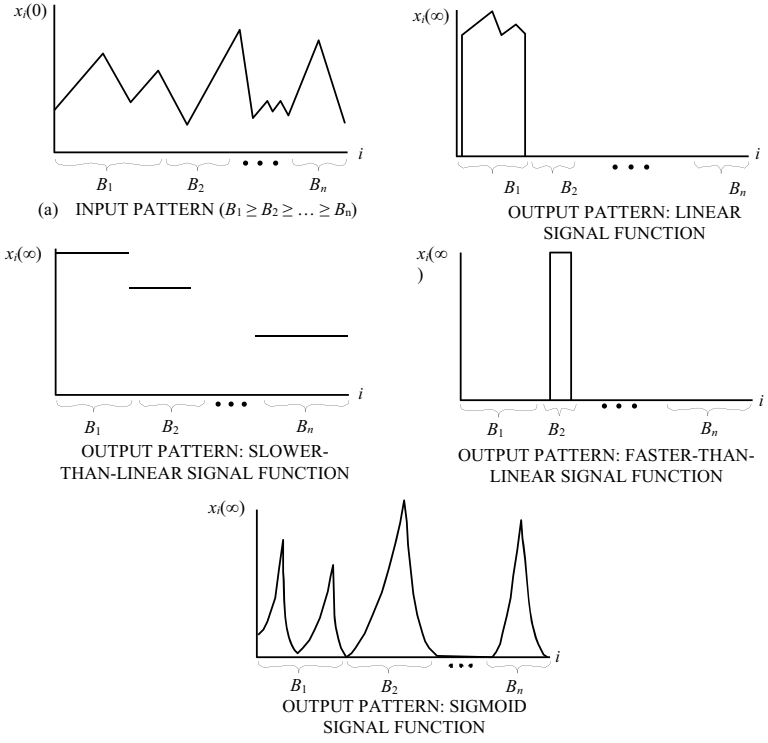


FIGURE 4.9 Pattern storage with unequal B_i. (a) Input pattern at time 0. (b) Output pattern at time 4 if the function f of (4.3) is linear. Inputs to nodes with the largest B_i (called B_1) are stored in proportion to input activities $x_i(0)$, and suppress others. (c) Output pattern for f slower than linear. The pattern becomes uniform within each subfield. Activities are largest for nodes with largest B_i, regardless of $x_i(0)$. (d) Output pattern for f faster than linear. Only inputs to nodes with one B_i value (not always the largest) survive, and only those with largest $x_i(0)$ in that subfield. (e) Output pattern for f a suitable sigmoid. Inputs with $x_i(0)$ below a threshold are suppressed. Inputs with $x_i(0)$ above the threshold, at least in some subfields, are enhanced.

Source: Adapted from Grossberg & Levine, 1975, with permission of Academic Press.

Amari (1977a) studied neural populations arranged in a "field" in the sense used in physics. That is, unlike the network described in Equations (4.3) that consists of a finite number of distinct nodes, his network is mathematically described by an activity variable that depends continuously on both time and location. The equations describing the dynamics of this variable use separate excitatory and inhibitory weighting functions, with inhibition decreasing less sharply with distance than excitation as in Figure 4.2(b). He found that the system typically (but not always) approaches an equilibrium state in which some part of the field remains active in short-term memory. Depending on various system parameters, the part that remains active could either be the entire field, a wide range, or a narrow range.

Amari and Arbib (1977) extended the earlier model to a neural field with two dimensions. The design of that model was motivated by previous work of Dev (1975) on modeling binocular vision, which is discussed in Section 4.3. Briefly, some results on binocular vision were based on detection of disparity between the positions of an image on the right and left retinas. The Amari–Arbib model includes inhibition between detectors of different disparities at the same position, and excitation between detectors of the same disparity at different positions.

Hence, the neural field developed by Amari and Arbib consisted of nodes indexed by the two dimensions of disparity and distance. The resulting network was termed *competitive-cooperative*, a term that has come into common usage in the field (see, e.g., Amari & Arbib, 1982). In general, if lateral inhibition is interpreted as competition between different percepts for encoding in short-term memory, then by the same token, lateral excitation can be interpreted as *cooperation* between related or compatible percepts.

4.2.4. Energy Functions in the Cohen–Grossberg and Hopfield–Tank Models

The theorem of Cohen and Grossberg (1983) relies on a common construction from mathematical dynamical systems theory, namely, a *Lyapunov function* or "energy" function. A Lyapunov function is some function of the system's state variables whose value decreases as the system's state changes over time. An equilibrium point of the dynamical system corresponds to a local minimum of the energy function (see Figure 4.10). In dynamical systems derived from physics, this function frequently represents an actual energy; in dynamical systems for neural networks, the energy function is more abstract.

The Lyapunov function for the Cohen–Grossberg system is discussed in the equations at the end of this chapter. We now discuss the Lyapunov or energy function for the related but simpler system of equations introduced in Hopfield (1982). The Hopfield networks do not always include lateral inhibition but are

discussed in this section because they deal with issues of short-term pattern storage and obey equations formally similar to Cohen and Grossberg's. Hopfield's 1982 article and more recent articles by Hopfield and Tank were later applied to optimization problems such as the traveling salesman problem, whereby a person traveling through several cities has to find the path of shortest total distance between the cities.

In the simplest form of the network of Hopfield (1982), the ith node has two possible states: $x_i = 0$ ("not firing") or $x_i = 1$ ("firing at the maximum rate"). Hence, the instantaneous state of the system is a binary vector whose number of components is equal to the number of nodes. There are also connection strengths w_{ij} from node j to node i, for all pairs where $i \neq j$. (Not all pairs of nodes have to be connected; for those that are not connected, $w_{ij} = 0$. Also, the w_{ij} can be positive or negative; in fact, in examples they tend to be negative more often than not.) Later versions of Hopfield's network include the possibility that connection strengths might change over time as a result of associative learning. But the energy function formulation first arose in the context of unchanging connection strengths.

The state changes over time are governed by a linear threshold algorithm reminiscent of some of those used in Rosenblatt (1962). If $x_i(t)$ denotes the state of the ith node at time t (t an integer), this state readjusts at random times, according to the rule

$$\begin{aligned} x_i(t+1) &= 1 \text{ if } \sum_{j \neq i} w_{ij} x_j(t) > 0 \\ x_i(t+1) &= 0 \text{ if } \sum_{j \neq i} w_{ij} x_j(t) < 0 \end{aligned} \tag{4.4}$$

where "Σ" denotes summation. (In later articles of Hopfield and Tank, the thresholds of 0 in Equations (4.4) were replaced by more general threshold

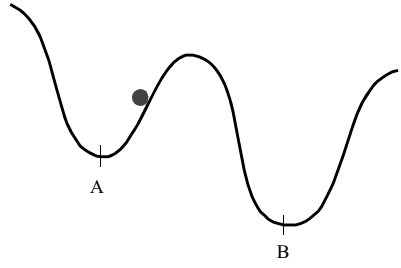


Figure 4.10 The path of the variables in a typical competitive dynamical system is analogous to the path of a ball bearing along a curve (representing the system "energy function"). Like the ball bearing, the system eventually reaches a local minimum state of the energy (either *A* or *B* in this figure).

Source: Reprinted from Rumelhart & McClelland, 1986a, Vol. 1, with permission of MIT Press.

parameters. The energy function formulation easily generalizes to that version; see the equations at the end of this chapter.)

The algorithm described by (4.4) means that each time x_i changes, the increment Δx_i (using the character Δ for change in any variable) is either 0 or 1 if $\Sigma_{j \neq i} w_{ij} x_j(t) > 0$ and is either 0 or –1 if $\Sigma_{j \neq i} w_{ij} x_j(t) < 0$. Hence Δx_i, if not 0, always has the same sign (positive or negative) as $\Sigma_{j \neq i} w_{ij} x_j(t)$. Now impose the condition that the weights are symmetric,[1] that is, $w_{ij} = w_{ji}$. If we then consider the Lyapunov or energy function,

$$E = -\frac{1}{2} \sum_i \sum_j w_{ij} x_i x_j \tag{4.5}$$

each change Δx_i in a given node activity leads (by (4.4) and (4.5)) to an energy change of

$$\Delta E = -\Delta x_i \left(\sum_{j \neq i} w_{ij} x_j \right) \tag{4.6}$$

Since Δx_i is 0 or of the same sign as $\Sigma_{j \neq i} w_{ij} x_j$, Expression (4.6) means that the energy change $\Delta E < 0$ at all times; that is, energy always decreases over time.

Some extensions of the above energy function, in both discrete and continuous models, were made in Hopfield (1984) and Hopfield and Tank (1985, 1986). In Hopfield (1984), the discrete-time system of the 1982 article was extended to include external inputs to each node and arbitrary thresholds for node activation. Also, a generalization of the algorithm defined by (4.6) to the continuous-time case was introduced in Hopfield (1984) and developed further in Hopfield and Tank (1985, 1986). In this work, the output V_i of the ith node is treated as a function (usually sigmoid) of the input, as in the articles of Wilson and Cowan (1972) and Grossberg (1973). All these systems have Lyapunov functions similar to (4.5).

The systems studied by Hopfield and Tank become laterally inhibitory (competitive) when all the w_{ij}, $i \neq j$, are set to be negative. (One of the examples studied by Hopfield and Tank, 1986, a network that can convert computational data from analog to binary, exhibits this lateral inhibitory structure.) Also, the theorem of Cohen and Grossberg (1983), showing the convergence of the system defined by (4.3) to some equilibrium state, does not depend on the values e_{ij} (which correspond to the $-w_{ij}$ in Hopfield and Tank's systems) being positive. Hence, that theorem covers the Hopfield–Tank systems as subcases. (This mathematical point is discussed more fully in the article of Grossberg, 1987a, which also shows that this theorem applies to the Boolean model of McCulloch and Pitts, 1943, and to the "brain-state-in-a-box" model of Anderson, Silverstein, Ritz, & Jones, 1977.)

4.2.5. The Implications of Approach to Equilibrium

A theorem stating that a system must approach a system equilibrium point does not specify *which* equilibrium is approached. In particular, if an energy function is always decreasing, the system will approach one of several states that are local minima of the energy function (see Figure 4.10). In some applications, the global minimum corresponds to the optimal state, and schemes have been studied for escaping from local minima that are not globally optimal. The most widely used of these schemes is *simulated annealing*, introduced by Kirkpatrick, Gelatt, and Vecchi (1983) and incorporated into neural networks by Hinton and Sejnowski (1986), Smolensky (1986), and others. Simulated annealing is an introduction of random noise that perturbs the network if it is at or close to a nonoptimal equilibrium, thereby increasing the probability that it will ultimately move toward the global minimum. Hence, the results of Grossberg, Amari, Hopfield, and others indicate that competitive neural systems often arrive at a choice as to what to store in short-term memory, but the nature of the choice is heavily influenced both by network parameters and by outside inputs.

The choice-making property of competitive or competitive-cooperative networks has been a valuable guide for models of natural or artificial pattern recognition. Since network theories of this type have made extensive contact with psychological and neurophysiological data in vision, we devote the next large section to visual modeling. Some network mechanisms used to model specific visual data are introduced to give the reader a "hands-on" sense of how data can suggest theory. In particular, we consider models of early visual processes that do not include learning or reference to previously stored templates (although the parameters defining these processes could have been influenced by development). The combination of early processing effects with learning in models of multilevel visual processes, such as coding, is discussed in Chapter 7. More recent visual models that have made more contact with neuroanatomy and neurophysiology are discussed in Chapter 9.

The type of short-term memory found in these networks has been described (e.g., Cohen, 1988; Hopfield, 1984) by the term *content-addressable memory*, familiar from computer science. This term signifies that each node ("address") is distinguished by what input events it encodes.

4.2.6. Networks With Synchronized Oscillations

Yet there are some significant exceptions to the equilibrium (choice-making) property of competitive networks. Cohen (1988) found an example of a system obeying Equations (4.3), with $n = 2$, whose solution approaches a limit cycle (oscillates) instead of approaching an equilibrium (see Figure 4.8). In his example, the assumption that the excitatory coefficients $c_{ik} = 0$ for $i \neq k$ is

removed, but the symmetry assumption $e_{ik} = e_{ki}$ still holds. Ellias and Grossberg (1975) found oscillations in certain examples of the *unlumped* on-center off-surround network, where excitatory and inhibitory cell populations occupy separate nodes. (The unlumped case is probably closer to real neuroanatomy than the *lumped* case, including the other articles of Grossberg's group, where the same nodes are excitatory and inhibitory. As was discussed early in this chapter, the largest cell type in the cerebral cortex, pyramidal cells, typically excite smaller interneurons which in turn appear to inhibit other pyramidal cells.) Approach to limit cycles was also found in some subcases of the network of Amari (1977a).

Oscillatory neural responses are widely thought to play their own particular role in visual pattern perception. Specifically, it has been suggested by many investigators (starting with Milner, 1974) that one mechanism for seeing an object as a unified percept is synchronization of oscillatory firing in many neurons responding to the object. For as discussed in later sections of this chapter, early processing stages in the visual cortex fire to *features* of an object, such as color, orientation, or position. The different features of the same object may be processed at different rates by the visual system. Eventually, there must be a mechanism for "binding" these features so that the correct ones are mapped together into representations of objects. This suggestion has been confirmed by the neurophysiological findings of Eckhorn et al. (1988), Gray et al. (1989), and Gray and Singer (1989). These investigators found stimulus-evoked synchronized oscillations in the cat visual cortex at a frequency of 40 to 60 Hertz (cycles per second).

This type of synchronization, in response to visual stimuli such as long moving bars, was simulated by Grossberg and Somers (1991) in a network that extended the "unlumped" excitatory–inhibitory interactions of Ellias and Grossberg (1975). In the Grossberg–Somers model, synchronization of the phases of several oscillators (that is, excitatory–inhibitory node pairs) was achieved by means of long-range cooperation among the excitatory nodes. (Such long-range cooperation had previously been proposed by Grossberg and Mingolla, 1985a, 1985b, to explain visual illusion data; see the next section.) A variety of architectures was used for such cooperation, of which the most successful was the one that employed a node analogous to a cortical cell called the *bipole cell*, which had been predicted by other visual models by the same authors and later found experimentally. The bipole cell is characterized by responsiveness to stimuli along two independent flanks.

Yet synchrony across an entire network carries the danger of what psychologists call *illusory conjunctions* (e.g., Intraub, 1985). That is, if more than one object is present in a visual scene, it is possible to perceptually bind features from different objects to one another: for example, if a blue circle and a green square are presented, under some conditions one can see a green circle and a blue square. Several neural network models (e.g., Grossberg &

Grunewald, 1997; Horn, Sagi, & Usher, 1992; Wang, Buhmann, & Malsburg, 1990) have dealt with the problem of illusory conjunctions by connecting oscillators in such a fashion that different parts of the network oscillate with their peak responses occurring at different times. This is accomplished through spatial or temporal segregation, or both, of the inputs to the network.

Binding via synchronized oscillations has also been applied to inference processes. Shastri and Ajjanagadde (1993) developed a neural network called SHRUTI that performs what they called *reflexive reasoning*. This is defined as the type of reasoning that most speakers of a language do automatically when they hear particular words or phrases. For example, when English speakers hear "John gave Mary the book," they infer without thinking that Mary now owns the book and can sell it. Reflexive reasoning requires a combination of rule encoding and binding of entities to general concepts that denote roles. An example of binding occurs when one hears "John gave Mary the book": the entities "John," "Mary," and "the book" are bound respectively to the concepts of "giver," "recipient," and in Shastri and Ajjanagadde's term, "give-object." Such binding is hard to achieve in neural networks based on Hebbian learning (see Chapter 3). This is because in a continuous dynamical system the nodes for all the various entities and concepts described would remain active at about the same time. Hence, Hebbian connections between these nodes are likely to lead to spurious associations (e.g., in the above example between "John" and "recipient.")

Shastri and Ajjanagadde solved the binding problem using nodes that respond to concept inputs by means of oscillatory pulses. Such regular oscillations are probably a quite crude approximation to biological mechanisms. As the authors noted, oscillations that are less regular (perhaps including chaotic patterns) are likely to be closer to actual concept encodings. Also, Shastri and Ajjanagadde did not say how particular nodes *learn* to encode particular concepts. Still, their network is a step toward a more neurally realistic theory of a common type of inference. Shastri (2001, 2002) later elaborated this binding network to model formation of episodic memories in the hippocampus and cerebral cortex.

4.3. Visual Pattern Recognition Models

4.3.1. Visual Illusions

The dynamics of the visual system can be illuminated by studying some of its characteristic illusory percepts. Several early network models (Grossberg & Mingolla, 1985a, 1985b; Levine & Grossberg, 1976; Wilson & Cowan, 1973) incorporate the notion that such illusions are by-products of a lateral inhibitory network designed to correct for irregularities in the luminance data that reaches the retina. Models of visual illusions typically involve both competition and

cooperation, sometimes along different dimensions (as in the work of Dev, 1975) and sometimes within the same dimension.

Levine and Grossberg (1976) simulated some illusions in angle perception. Their model incorporated the findings of Hubel and Wiesel (1962, 1965) that most cells in the primate or cat visual cortex respond preferentially to lines or bars of some particular orientation; a characteristic tuning curve for orientation in a visual cortical cell is shown in Figure 4.11. The Levine–Grossberg model is based on a recurrent competitive-cooperative network in which each node represents a specific line orientation. The interaction coefficients of Equations (4.3) decrease with distance between the orientations coded by given nodes, but the c_{ik} decrease faster with distance than the e_{ik}; that is, cooperative interactions are short-range and competitive interactions long-range. Percepts of one or two lines are denoted by inputs that are nonzero to the nodes corresponding to those lines. The perceived orientation of a line is interpreted as being the orientation corresponding to that node whose activity is largest after the inputs have been transformed by recurrent interactions.

If all maximal node activities (B_i in Equations (4.3)) are equal, the Levine–Grossberg network can reproduce the experimental result (Blakemore, Carpenter, & Georgeson, 1970) that an acute angle is seen as up to one degree larger than it really is. In the network, an acute angle stimulus excites two orientation-sensitive network nodes. Recurrent inhibition from either of those two nodes shifts the local peak of activity away from the other to a different node corresponding to an angle a degree or two out from it. This causes a shift in each line's perceived location. If instead the nodes responsive to vertical or horizontal orientations have larger B_i than others, the same network can reproduce the result (Gibson & Radner, 1937) that a fixed near-vertical tilted line appears closer to vertical after prolonged viewing. Recall from the discussion in Section 4.2 that differences in B_i could be influenced by experience during development. In fact, there is evidence that cultural experience affects the bias toward vertical and horizontal in humans (Annis & Frost, 1973) as well as other human visual illusions (Deregowski, 1973). (Section 9.2 mentions another network that has modeled the tilt aftereffect data, that of Bednar & Miikkulainen, 2000.)

In some other competitive-cooperative neural networks, each node is interpreted as a receptor for a given visual field position rather than a line orientation. For example, Wilson and Cowan (1973) and Levine (1975) both simulated results of Fender and Julesz (1967) on the perceived location of vertical lines viewed stereoptically. In the Fender–Julesz experiments, two parallel vertical lines are shown simultaneously, each seen only by one eye; the lines are pulled apart and then slowly pushed back together. During the stage when they are being pulled apart, the two lines are seen as one for a considerable distance (2 degrees of arc) until they suddenly appear to jump apart. But, while they are being pushed back together, these lines are seen as

two until the distance between them is much shorter (.1 degrees of arc), at which point they appear to fuse together. Hence, hysteresis, or memory, occurs: the perceived location of the two lines depends partly on their past locations as well as their present ones.

Full understanding of the binocular hysteresis effect depends on understanding of binocular depth perception, which is discussed later in this section. But the Wilson–Cowan and Levine networks simulate the Fender–Julesz result without using depth perception, by treating the two lines as inputs to nodes representing different positions in the visual field, and interpreting perceived location as corresponding to the (one or two) nodes with largest activity.

Orientation and position are not the only two variables that are coded by cell populations in the visual cortex. Table 4.1 sums up information about these and other key visual variables, such as spatial frequency; disparity of the images on the right and left retinas, which is a measure of depth; color; and ocularity (cells may have a preference for one or another eye or else respond equally to inputs from either eye).

Some neural networks used to simulate visual data combine two or more of these variables. For example, both orientation and position information are used in the networks of Grossberg and Mingolla (1985a), which simulate some illusory percepts of visual contours. An example is the white square perceived in Figure 4.12, by Kanizsa (1976), whose corners are formed from the boundaries of four black (or two black and two gray) "Pac-Man" figures.

In Figure 4.12, two white line segments that are actually present and of the same orientation are perceptually joined together by an illusory longer line segment. That insight among others suggested Grossberg and Mingolla's modeling scheme. In their competitive-cooperative network, boundaries are

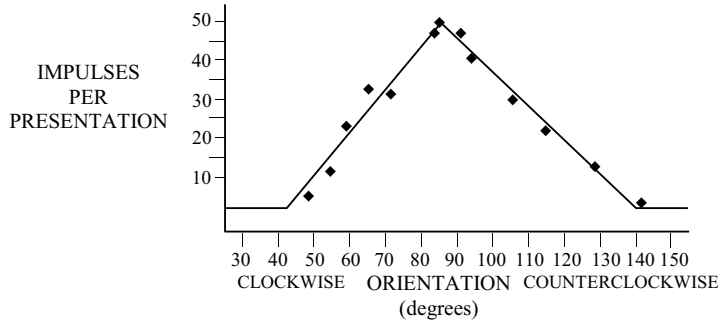


FIGURE 4.11 Example of the tuning curve of a complex cell (the second of three layers of visual cortical cells described by Hubel and Wiesel). Each point is the mean response, above the cell's spontaneous firing rate, for ten sweeps of a moving bar of light across an oscilloscope during a three-second interval.

Source: Adapted from Rose & Blakemore, 1974, with permission of Springer-Verlag.

Variable	*Source of experimental findings*
Position	Hubel and Wiesel (1962, 1965)
Orientation	Hubel and Wiesel (1962, 1965)
Ocularity (left or right eye)	Hubel and Wiesel (1962, 1965)
Disparity (between left and right retinal images)	Barlow, Blakemore, and Pettigrew (1967)
Spatial frequency	Robson (1975)
Color	Hubel and Wiesel (1962, 1965)

TABLE 4.1 Stimulus variables to which single cells in the visual cortex can be differentially responsive.

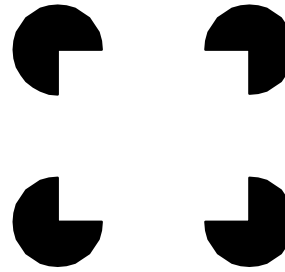


FIGURE 4.12 Illusory white square induced by four black "Pac-Man" figures.
Source: From Kanizsa, Gaetano, Subjective contours. Copyright by Jerome Kuhl. All rights reserved.

perceived as signals "sensitive to the orientation and amount of contrast at a scenic edge, but not to its direction of contrast" (Grossberg & Mingolla, 1985a, p. 176).

Figure 4.13a illustrates insensitivity to contrast direction. Each node of the network responds to lines of a particular orientation at a particular position in the plane. There are two forms of competition, between receptors for like orientations at nearby positions (Figure 4.13b), and between receptors for widely different (especially mutually perpendicular) orientations at the same location (Figure 4.13c). The scheme of Levine and Grossberg (1976) is reversed: short-range competition is supplemented by long-range cooperation (Figure 4.13d). Such long-range cooperation enables continuous contours to form by linking together separated lines of the same orientation. One of the benefits to the organism of this linkage of contours is compensation for discontinuities (caused by blind spots) in the image on the retina.

All visual phenomena – color, brightness, and form detection, motion detection, and binocular integration – include characteristic illusory percepts.

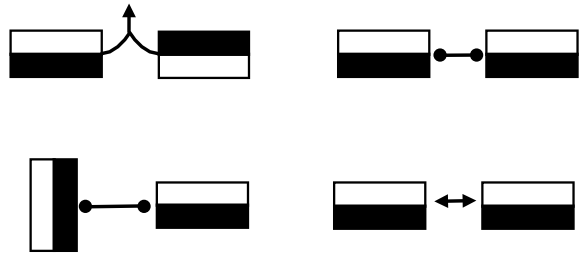


FIGURE 4.13 (a) *Boundary contour* signals sensitive to orientation and amount of contrast at the edge of a scene, but not to direction of contrast. (b) Like orientations compete at nearby perceptual locations. (c) Different orientations compete at each perceptual location. (d) Once activated, aligned orientations cooperate across a larger visual domain to form "real" and "illusory" contours.

Source: Grossberg & Mingolla, *Psychological Review*, *92*, 173–211, 1985. Copyright 1985 by the American Psychological Association. Reprinted by permission.

Most of these illusions are part of *preattentive* vision, that is, direct processing of visual inputs by the retina and cortex before the evaluation of their significance or their relationship to other modalities such as hearing and touch. This processing includes grouping or segmentation of the visual environment. The computational models discussed in the next few subsections of form, color, brightness, depth, and motion detection are partly inspired by illusion data, some of which is discussed further in Section 5.2.

4.3.2. Boundary Detection versus Feature Detection

The importance of edge detection for understanding the form of visual percepts has been emphasized by many vision theorists. For example, Marr and Poggio (1979) and Marr and Hildreth (1980) described boundaries between light and dark areas of a scene as points of zero curvature (or inflection points) of the curve for luminance as a function of distance, as shown in Figure 4.14. This is a simple mathematical representation of the point of sharpest transition in the luminance value.

But the mechanism for perceiving boundaries must be supplemented by another mechanism for perceiving the form of what is *inside* those boundaries, a so-called *feature contour mechanism*. The feature contour mechanism, unlike the boundary contour mechanism, should be sensitive to the direction of contrast.

One possible combination of boundary and feature contour mechanisms, using both lateral inhibition and excitation, is discussed in Grossberg (1983). In the feature contour mechanism, the excitatory and inhibitory spread coefficients c_{ik} and e_{ik} (as in (4.3)) determine structural scales of the on-center off-surround network. The network's recurrent interactions transform structural scales into functional scales.

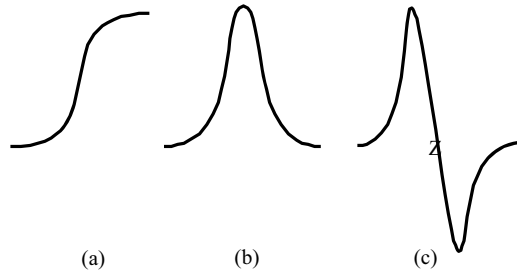


FIGURE 4.14 The notion of a zero-crossing. (a) A transition (edge) between dark and light regions is shown by a sharp rise in the graph of luminance as a function of distance. (b) The first derivative of this function has a peak. (c) The second derivative of this function has a *zero-crossing* (transition from positive to negative) at the point *Z*.

Source: Adapted from Marr, 1982, with permission of MIT Press.

Figure 4.15 describes one scheme for functional scaling. A linear nonrecurrent mechanism that can only generate boundaries (Figure 4.15b) is contrasted with a nonlinear recurrent mechanism that can generate perceptions of both boundaries and interiors (Figure 4.15c). Initially, all nodes excited by the rectangular input of Figure 4.15a receive equal inputs. Since the inhibitory interaction coefficients e_{ik} are distance-dependent, once recurrent inhibition has time to be established, nodes excited by the part of the rectangle near its boundary receive less inhibition than those nodes nearer the rectangle's center. As time goes on, those selectively enhanced boundary nodes inhibit other nodes whose preferred positions are immediately contiguous to those boundaries but closer to the center. This in turn disinhibits some nodes still nearer to the center, leading to the wavelike pattern shown in Figure 4.15c. The distance between peaks of the wave ("functional scale") is dependent in a complex nonlinear way on the excitatory and inhibitory interaction coefficients c_{ik} and e_{ik}; see Grossberg (1983, p. 646) for details.

Figure 4.15 and the accompanying mathematics also provide one possible explanation for the experimental result (e.g., Robson, 1975) that many visual cortical neurons fire preferentially to some specific spatial frequency. From this result, many theorists have concluded that spatial frequency is one of the primitives of the visual system, or, more speculatively, that the visual system performs Fourier analysis of patterns into frequency components (e.g., Graham, 1981; Pribram, 1991; Wilson & Bergen, 1979). In Grossberg's scheme, by contrast, Fourier analysis (which is a linear transformation) does not occur, and spatial frequency detection is a by-product of more fundamental nonlinear interactions.

The interaction of the feature and boundary contour systems provided the basis for a theory of visual object recognition. In contrast to Marr's view that boundaries are what we mainly see, Grossberg (1987c) advances the "radical

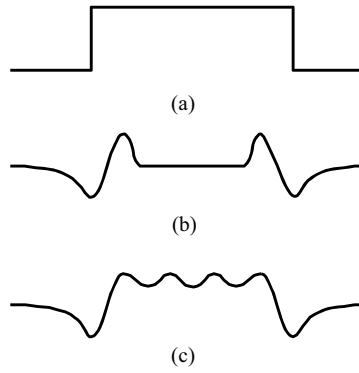


FIGURE 4.15 (a) Input pattern whereby a region is activated uniformly and the rest of the visual field not at all. (b) Response of a feedforward competitive network to pattern (a); edges of the activated region are enhanced and its interior suppressed. (c) Response of a feedback competitive network to pattern (a); the interior is activated in a spatially periodic fashion.

Source: Reprinted from Grossberg, 1983, with permission of Cambridge University Press.

claim that *all* boundaries are invisible until they can support different filled-in featural contrasts within the FC [feature contour] System" (p. 108). Within the feature contour system in this theory, the apparently separate modules that neurophysiologists have discovered in the cerebral cortex for processing form, color, and depth (Livingstone & Hubel, 1984) are seen as part of a unified whole that is called *FACADE*, for "form and color and depth" (Grossberg, 1994).

Grossberg's theory involves a hierarchy of networks, some competitive, some cooperative, some involving opponent processing (see Section 3.3.4). One of the purposes of all these feedback and feedforward interactions is to compensate for some imperfections in the uptake of visual stimuli by the retina. Some of these imperfections result from blind spots or blood vessels in the eye. Others result from possible distortions of relative brightness or color relationships in the scene by the ambient light; hence one of the functions of the cortical networks is to *discount the illuminant*, that is, calculate color or brightness of the actual scene rather than what impinges directly on the retina. Some possible brain mechanisms for all this, involving several different parts of the visual cortex and the lateral geniculate body, which is a processing station between the retina and cortex, are discussed in Section 9.2.

These networks used to model biological vision have also been utilized in devices for machine vision, leading to some novel solutions to computer vision problems that have long been studied by artificial intelligence researchers. Carpenter, Grossberg, and Mehanian (1989), for example, developed a network architecture for invariant recognition of cluttered scenes.

It combines a preprocessing stage based on the boundary contour system of Grossberg and Mingolla (1985a, 1985b) with the adaptive resonance network (see Chapter 6) for high-level processing. Cruthirds et al. (1992) and Grossberg, Mingolla, and Williamson (1995) applied both the boundary and feature contour systems to processing of synthetic radar images. Bradski and Grossberg (1995) applied a similar architecture to the processing of three-dimensional objects from multiple two-dimensional views.

Further elaborations of the cortically inspired networks for boundary–feature interactions have included mechanisms for binocular vision and for detecting visual motion. These are discussed in the next two subsections.

4.3.3. Binocular and Stereoscopic Vision

The relationship between disparity of the two images and depth of the actual object viewed is illustrated in Figure 4.16. The existence of cells in the visual cortex responding preferentially to given disparities was demonstrated experimentally by Barlow, Blakemore, and Pettigrew (1967). Typically, these models are based on cooperation between detectors of the same disparity at different positions, and competition between detectors of different disparities at the same position. (Amari & Arbib, 1977, discussed the mathematical dynamics of systems that combine competition and cooperation in this fashion.)

Modeling of binocular vision has been advanced by the study of *random-dot stereograms*, introduced by Julesz (1960, 1971). These are pairs of patterns presented separately to the two eyes, each of which alone consists of incoherent random dots but which together can lead to sensations of depth. As described by Dev (1975):

> The two patterns are identical to each other in some regions and differ in others. In the regions where they differ, the difference simply consists of a lateral shift of one pattern with respect to the other. . . . [T]he region

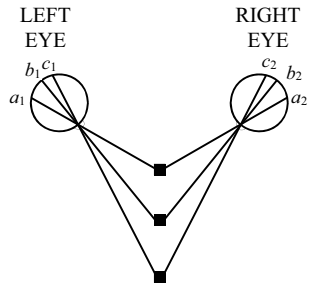


FIGURE 4.16 Depth of a binocularly viewed object is encoded by disparity of positions along the circles denoting the two retinas (a_1 to a_2 for the nearest object, b_1 to b_2 for the next nearest, c_1 to c_2 for the farthest).

> of the pattern that requires lateral shift is perceived as at a depth other than the depth at which the observer is fixated.
>
> (p. 524)

For example, if the laterally shifted region is square-shaped, the observer sees a square with some of the random dots lying above the rest of the figure. Dev (1975) developed a computational procedure, involving cooperation and competition between disparity detectors as described earlier, to analyze the perception of depth surfaces from these stereograms (see Figure 4.17). This computational procedure has since been refined by Marr and Poggio (1977a), among others.

Marr and Poggio (1979) also summoned evidence to show that retinal image disparity measures as in the earlier models are insufficient to compute perceived depth. Hence, achieving a coherent three-dimensional percept (whether of a random-dot stereogram or a natural, binocularly viewed scene) involves integrating disparity information with orientation and spatial frequency information. In Marr and Poggio's model, a three-dimensional scene is filtered through channels ("masks") that select particular orientations. Boundaries can be located by taking the image through given orientation masks and locating the edges at zeros of the second derivative of perceived luminance (see Figure 4.15). Similar filtering is done through spatial frequency channels.

Given all the disparity, orientation, and spatial frequency information, Marr and Poggio showed how to construct a coherent three-dimensional

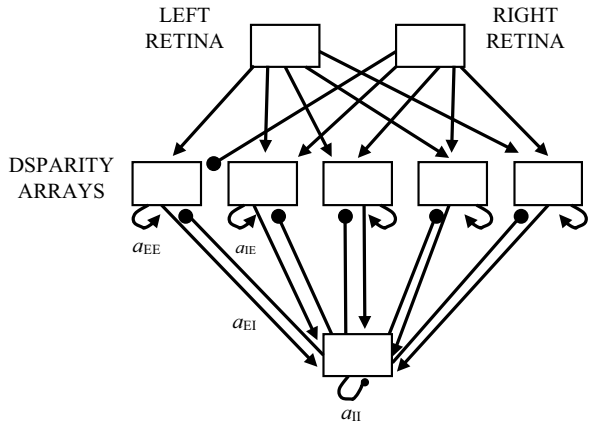


FIGURE 4.17 Network for simulation of depth perception in random-dot stereograms. Each "disparity array" is a group of nodes sensitive to illumination of a point on the left retina combined with illumination of the point on the right retina that is laterally displaced by the given amount (see Figure 4.16). a_{EE}, a_{EI}, a_{IE}, a_{II} are functions describing interactions among node groups.

Source: Adapted from Dev, 1975, copyright 1975 IEEE, with permission of the publishers.

approximation of a given 3-D scene preceding the stage of binocular integration. They called this approximation the *2½-D sketch* of the scene. An example of a 2½-D sketch is shown in Figure 4.18; the concept was developed further in Marr's (1982) book on vision.

Another visual model involving random-dot stereograms is that of Hinton and Becker (1990). These researchers combined perception with learning to discover planar or curved surfaces in stereograms. Their learning procedure was designed to maximize the coherence of information from spatially adjacent patches of the images.

A different approach to binocular vision has been developed by Cohen and Grossberg (1984) and Grossberg (1987d). In the networks of Cohen and Grossberg, there is extensive feedback between monocular and binocular representation areas, each with its own separate on-center off-surround network (see, e.g., Figure 4.19) and, in later versions of the network, including extensive opponent processing. In contrast to Marr and Poggio's idea of the prebinocular 21/2-D sketch, Grossberg and Cohen developed a theory in which binocular integration is nearly inseparable from the processing of other visual information such as color and form. For example, in the boundary contour system, mechanisms of stereopsis are related to mechanisms of boundary completion. And the interactions between boundary contour and feature contour systems in the theory explain why boundary completion and segmentation become binocular at an earlier processing stage than do color and brightness perception.

4.3.4. Visual Motion

Marr and Ullman (1981) set out to explain how we perceive visually that objects move. As with other visual phenomena, this has an illusory as well as a veridical component. Since the days of the early Gestalt psychologists of the late nineteenth century, it has been known that apparent motion can be generated by, for example, two separate flashes of light in different locations at particular time intervals. Their explanation was based on a network that combined different nodes with *sustained* and *transient* responses to stimuli. The sustained units respond to particular contrast and orientation patterns which persist even if their location in the visual field shifts slightly. The transient units respond to changes in light intensity, color, and the like at particular locations.

Marr and Ullman based the visual responses in their network on the zero crossings previously described by Marr and Poggio (1979), which represent transition points or boundaries. Grossberg and Rudd (1989, 1992) combined the Marr–Ullman idea of sustained and transient detectors with the feature and boundary contour systems (Grossberg & Mingolla, 1985a, 1985b) achieved by shunting on-center off-surround interactions. Grossberg and Rudd (1992) saw visual motion as related to rather than separate from other visual percepts:

> . . . psychological literature that exists on the topic of apparent motion – and the more general category of motion perception – indicates a complex interdependency between such stimulus variables as contrast, size, duration, color, and figural organization in determining the perceived motion.
>
> (p. 82)

They described how this approach to theory leads to simulating a wide range of data on both the direction and speed of real or apparent visual motion not covered by the Marr–Ullman theory. The *motion-oriented contour system* they developed includes a set of nodes combining signals from both sustained and transient units, and with properties analogous to the visual motion area of the cortex variously known as *V4* (with V1 and V2 representing earlier visual processing) or *MT* (for medial temporal).

The motion-oriented contour system was developed further in the articles of Grossberg and Mingolla (1993) and Chey, Grossberg, and Mingolla (1997). The motion-oriented boundary contour system is further subdivided into a motion-oriented contrast filter for preprocessing moving images and a motion cooperative-competitive feedback loop for generating boundary segmentations of the filtered signals.

This combination of architectures enables the model of Grossberg and Mingolla (1993) to simulate such widely studied motion phenomena as the *aperture problem* (Wallach, 1976), the *barber pole illusion*, and *motion capture*. The aperture problem denotes the fact that the perceived motion of a straight edge or grating is influenced by the shape of the aperture through which it is viewed. If a line or grating is viewed through a circular aperture it will be seen as moving in a direction perpendicular to its orientation. If it is viewed through a rectangular aperture it will be seen as moving in a direction corresponding to the longer side of the rectangle. An offshoot of the aperture problem is the barber pole illusion, whereby a set of parallel rotating colored stripes seen though a glass cylinder appear to be moving upward when they are moving horizontally.

Another effect of the aperture problem is that sometimes with complex moving patterns and different arrangements for viewing those patterns, some parts of the pattern are clearly moving in a particular direction, the signals denoting direction of motion are ambiguous in other parts. (This is particularly true of so-called *plaid patterns*, which include lines of different orientations that cross one another.) Motion capture means that the signals from those parts where motion is unambiguous dominate the ambiguous signals from other parts, leading to a percept that the entire pattern is moving in the direction of the unambiguous signals. Francis and Grossberg (1996) discussed how motion perception is integrated with form perception to capture the percept of moving forms, including forms that involve illusory contours.

The dynamics of MT neurons influenced another visual motion model by Nowlan and Sejnowski (1994, 1995) and Sejnowski and Nowlan (1995). This model is designed to combine spatial integration of signals from neighboring regions of the visual field and sensitivity to small velocity differences. The model includes two types of units with properties similar to some MT neurons. One type of unit in the model integrates information about the direction of motion to estimate the local velocity, with competition between velocity detectors. The other type of unit selects regions of the visual field where the velocity estimates are most reliable. The result is a distributed segmentation of the image into patches that support distinct objects moving with a common velocity. Unlike Grossberg and his colleagues, Nowlan and Sejnowski optimized some of the parameters in their network in order to fit the motion data.

4.3.5. Nonlinear Feedback or Modularity?

Grossberg (1983) compared his approach to visual (including stereopsis) modeling with the approach of Marr and other members of his school. His article includes commentaries by two members of that school, Grimson (1983) and Stevens (1983). Some of this dialogue is summarized here because of its general implications for neural modeling.

In both sets of models, disparity information is influenced by orientation and spatial scale information. But Grossberg's model involves nonlinear feedback mechanisms, whereas Marr and Poggio's model is linear and feedforward. Grossberg argued that nonlinear and feedback mechanisms are necessary for accurate representation of many kinds of visual information, such as reflectances. In response, Grimson (1983) posed the following question:

> Can early visual processing be considered as a system of roughly independent modules which interact loosely to create a global perception, or is the processing so tightly interconnected that the simplest possible description of the process is in terms of its interactions?
>
> (p. 666)

Grimson answered his own question on the side of the first, "modular" approach. Citing the large number of psychophysical predictions made by Marr and Poggio (1979), he said that tightness of interaction as posited by Grossberg is not needed to account for psychophysical data.

Grimson's comment appears to be influenced by the approach of traditional artificial intelligence, which tends toward separate heuristic programs for separate tasks. This heuristic programming flavor also permeates some comments by Stevens (1983):

> As Marr and Poggio (1977a) eloquently argue, complex information processing requires satisfactory descriptions at several levels, of which a mechanism description is but one. They distinguish the *computational theory* (What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?), the level of *representation and algorithm* (What is the representation for the input and output, and what is the algorithm for the transformation?) and the level of *implementation* (How can the representation and algorithm be realized physically?).
>
> (p. 675)

Stevens went on to say:

> Grossberg's descriptions of visual computations . . . are primarily at the level of mechanism, of patterns of neural activity within networks. There is no notion, for instance, of symbolic information processing. For those of us interested in understanding vision, the real problems seem to lie here.
>
> (p. 675)

In Grossberg's approach, the computational theory and algorithm levels are not apparent in the network at hand but are assumed to be represented by other nodes and connections, outside and interacting with the network for early visual processing.

Related issues arise in the modeling of knowledge. While the connectionist or neural network approach has had a major impact on cognitive science, there is a school of cognitive scientists arguing against connectionism and in favor of the symbolic representations from "classical" artificial intelligence. This school (e.g., Fodor & Pylyshyn, 1988) contends that connectionist models cannot be used in the understanding of some levels of a cognitive task – typically, those levels involving purposes and goals. But their argument appears to be based on an overly narrow sense of what constitutes a connectionist model. Modeling purposes and goals is still at the forefront of neural network modeling, but some progress has been made in that direction as shown by some of the models discussed in Section 9.6.

The visual system models described in the last two sections are primarily described in cognitive terms with rather imprecise locations for network segments in the brain. Yet, starting in the late 1990s there has been a growth of visual models that have built on and refined the models described therein and located their subnetworks more precisely in areas including the retina, lateral geniculate, up to four layers of visual cortex, and prefrontal cortex. These more "brain-based" visual models are discussed in Section 9.2.

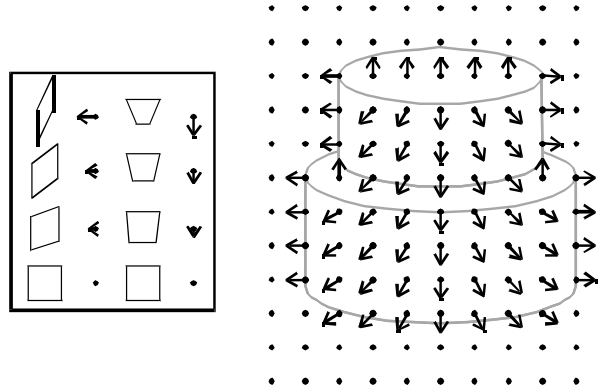


FIGURE 4.18 (a) Representation of surface orientation. Orientation of arrows is determined by the projection of the surface perpendicular to the image plane, and length of arrows represents the dip out of that plane. (b) 2½-D sketch including surface orientations and their discontinuities.

Source: Reprinted from Marr and Poggio, 1977b, with author's permission.

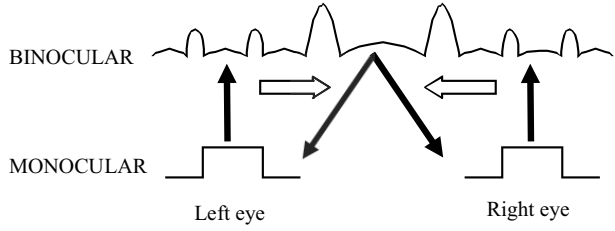


FIGURE 4.19 Monocular processing of patterns through feedforward competitive networks is followed by binocular matching of the two transformed monocular patterns. Pooled binocular edges are then fed back to both monocular representations.

Source: Reprinted from Cohen & Grossberg, 1984, with permission of Lawrence Erlbaum Associates.

4.4. Uses of Lateral Inhibition in Higher Level Processing

Kilmer et al. (1969) developed an early computational model of the reticular formation, consisting of connected modules, each receiving independent sensory information. Based on this information, each module "votes" for inclining the organism toward one of several gross behavioral modes (eating, sex, exploration, etc.). The model of Kilmer et al. includes competition, between nodes that represent different drives, and cooperation, between modules that incline the organism toward satisfying the same drive. Montalvo (1975) noted that the structural features of such a two-dimensional competitive-cooperative network of drive representations are similar to those of Dev's (1975) competitive-cooperative network of binocular disparity detectors.

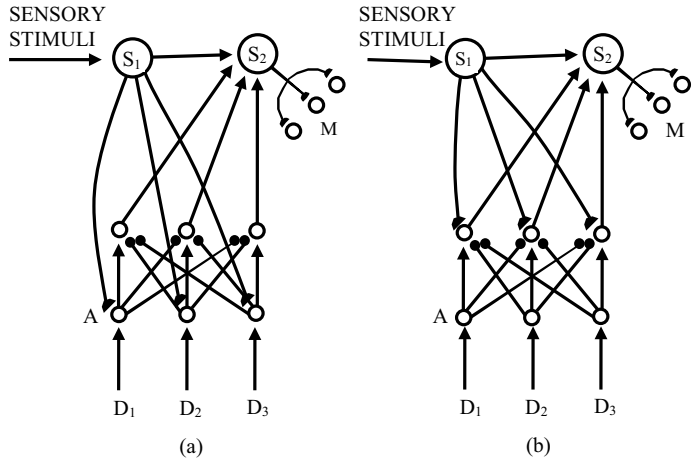


FIGURE 4.20 Competition between drives, denoted *Di*. In a conditioning network (see Chapter 5 for details), sensory stimulus representations, collectively denoted S_1, are conditioned to drive arousal sources A. The strongest drive sends feedback to S_2, allowing S_1 nodes to be conditioned to motor responses at M. (a) *Sensory–drive heterarchy*: the "winning" drive is determined by a combination of internal drive level and compatible cues. (b) *Drive heterarchy*: the "winning" drive is determined only by internal drive level.

Source: Adapted from Grossberg, 1975, with permission of Academic Press.

The idea that competition between drive representations is biased by the current sensory environment also appears in Grossberg (1975). Grossberg proposed a *sensory–drive heterarchy* (see Figure 4.20) in which each drive representation is activated by a combination of internal drive level and external sensory inputs compatible with the given drive. (In one variant, drive and sensory influences combine multiplicatively rather than additively, so that neither can activate the representation without the other.) In this way, even if one drive is strongest, another drive can be satisfied if cues compatible with the first drive are unavailable; for example, one can eat meals in spite of prolonged absence of a sexual partner.

Competitive neural processes, as noted in Chapter 1, are also likely to be important in category formation, which is discussed in Chapters 7 and 8. In order to make choices when classifying a sensory pattern, many modelers (e.g., Bienenstock, Cooper, & Munro, 1982; Carpenter & Grossberg, 1987a, 1987b; Rumelhart & Zipser, 1985) use an on-center off-surround field among cell populations within a level of nodes that detects categories (presumably at intramodal or intermodal association areas of the cerebral cortex).

Finally, there can be competition between representations of time sequences of sensory stimuli or motor actions. That idea was propounded by Grossberg (1978b) in a model of goal-directed behavior. In this model, competition

between sequence representations is biased in favor of longer over shorter time sequences. This bias causes the network to respond not just to the most recent events but to a longer sequence of events and to the consequences of its own recent past actions. The representation of stimulus or motor sequences is developed further in the masking field model of Cohen and Grossberg (1987).

Hence, the on-center off-surround or competitive-cooperative network is one of the most versatile and most widely used neural architectures. Along with associative synaptic modification, it is a prime component of networks that replicate complex cognitive processes, as discussed in the second section of this book (Chapters 6–9).

Equations for Networks in Chapter 4

Detailed Description: Shunting Lateral Inhibition with Faster-than-Linear Signal Function

The Grossberg version of the shunting on-center off-surround field, captured mostly by Equations (4.3), includes a wide variety of subcases based on (1) presence or absence of biases that select some of the nodes over others; (2) values of the excitatory and inhibitory coupling coefficients; (3) presence or absence of outside inputs; and (4) nature of the signal function or activation function f. For definiteness let us consider the case where (1) all biases, or maximum activities, B_i are equal to the same number B; (2) each node excites itself and uniformly inhibits all the others, so that the excitatory coefficients c_{ik} are 1 if $i = k$ and 0 otherwise, and the inhibitory coefficients are the reverse, namely 0 if $i = k$ and 1 otherwise; (3) excitatory inputs I_i and inhibitory inputs J_i are shut off, with the values of the node activities x_i at time 0 roughly reflecting the pattern of previous inputs; (4) the activation function f is a function that grows faster than linearly, specifically, $f(x) = x^2$. This leads to the following special case of Equations (4.3):

$$\frac{dx_i}{dt} = -ax_i + (b - x_i)x_i^2 - x_i \sum_{k \neq i} x_k^2.$$

The term x_i^2 can be added into the sum of x_k^2 over all k not equal to i to make the sum of x_k^2 over all k including i, yielding the equations

$$\frac{dx_i}{dt} = -ax_i + bx_i^2 - x_i \sum_k x_k^2 \tag{4.7}$$

Now suppose the nodes whose activities are described by (4.7) are indexed so that they are in descending order of initial (time = 0) activities, that is:

$$x_1(0) \geq x_2(0) \geq x_3(0) \geq \ldots \geq x_n(0),$$

with n being the total number of nodes. What do the interactions described by Equations (4.7) (and specifically the faster-than-linear nature of the activation function) tell us about the dynamics of the system over time?

From (4.7), the ratio of dx_i/dt, the rate of change of the activity of node i to the node activity x_i itself, is:

$$-a + bx_i^2 - \sum_k x_k^2 \tag{4.8}$$

Equation (4.8) yields the relative "growth rate" of the ith node activity. Since the term $-a$ and the sum have the same value for all nodes i, only the term bx_i differs between nodes, and it is larger for those for which x_i is larger (namely, those with smaller indices i). This means that if the relationship $x_i > x_j$ obtains at time 0 (or any later time), it also obtains for all later time. This is the consequence of using x^2 as the activation function, and would be true for any faster-than-linear function. In more poetic terms, "the rich get richer and the poor get poorer": a very (excessively, from the point of view of perceptual psychology) strong form of contrast enhancement. This is because Grossberg (1973) provides that in fact the activities of nodes with nonmaximal activities not only decrease but decay to 0 as time gets large.

Equations of Sperling and Sondhi

Sperling and Sondhi (1968) developed a lateral inhibitory model of effects in the mammalian retina, in order to explain certain data on luminance and flicker detection. Their model includes both feedback and feedforward stages. In the feedback stage, as shown in Figure 4.5, the jth node is excited by the $(j-1)$st node, for $j < n$, and inhibited by feedback from the nth node. Hence, the activity x_j of the jth node is described by the differential equation

$$\frac{dx_j}{dt} = -x_j(1 + x_n) + x_{j-1} \tag{4.9}$$

The nonlinear term $-x_j x_n$ in Equation (4.9) corresponds to multiplicative (shunting) inhibition exerted by the nth stage of cells on the jth stage. This type of inhibition has strength proportional to the present activity of the area being inhibited. Shunting inhibition, and the related process of shunting excitation,

whose strength is proportional to the difference of a cell's activity from its maximum possible level, can be related to the effects of a presynaptic neuron on various postsynaptic ionic conductances.

Equations of Wilson and Cowan

The variables used in Wilson and Cowan (1972) are $E(t)$ and $I(t)$, the proportion of excitatory and inhibitory cells active at time t. The equations incorporate refractory periods of individual cells, in which a cell is prevented from becoming active a short time after firing (see Section 2.2). This leads to time-integral terms that are then removed by averaging the node activities over a suitable time interval. The refractory period is figured in as a linear factor that decreases as the number of active cells increases. If $I_E(t)$ and $I_I(t)$ are the inputs to excitatory and inhibitory nodes, respectively, the resulting differential equations are of the form

$$\begin{aligned} \Gamma_E \frac{dx_E}{dt} &= -x_E + (k_E - r_E x_E) f_E(c_1 x_E - c_2 x_I + I_E) \\ \Gamma_I \frac{dx_I}{dt} &= -x_I + (k_I - r_I x_I) f_I(c_3 x_E - c_4 x_I + I_I) \end{aligned} \tag{4.10}$$

where f_E and f_I are sigmoid functions that transform the linearly combined excitatory and inhibitory signals, and Γ_E and Γ_I are positive constants (reciprocals of decay rates).

The shunting interactions in Equations (4.10) place bounds on the network's activity. For, if x_E and x_I are positive, and x_E reaches the value k_E/r_E, then (4.10) shows that dx_E/dt will be negative; hence, x_E can never exceed the value k_E/r_E. Similarly, x_I can never exceed the value k_I/r_I.

The model defined by (4.10) was extended in Wilson and Cowan (1973) to include distance-dependent interactions. In the later article, the variables $x_E(t)$ and $x_I(t)$ of Equations (4.10) are replaced by $x_E(s,t)$ and $x_I(s,t)$, the average strength of excitation and inhibition at location s at time t. Otherwise, the equations are essentially the same as (4.10) with the terms for x_E and x_I inside the sigmoid functions being replaced by the spatial convolutions of x_E and x_I with distance-dependent connectivity functions. (Convolution of two functions is the operation "*", defined by

$$(f * g)(x) = \int f(x')\, g(x - x')\, dx'$$

which provides a moving average of one function weighted by another.) This leads to integrodifferential equations that are not shown here. These equations include four different connectivity functions – excitatory-to-excitatory, excitatory-to-inhibitory, inhibitory-to-excitatory, and inhibitory-to-inhibitory.

Equations of Grossberg and Coworkers: Analytical Results

Grossberg (1973) considered a pure on-center off-surround network, with shunting interactions, in which all nodes have the same maximum activity B and minimum activity 0. The equations for the node activities x_i in this network are

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i) f(x_i) - x_i \sum_{k \neq i}^{n} f(x_k) + I_i \tag{4.11}$$

As in (4.10), shunting interactions cause a bound in each node's activity, since (4.11) shows that if all $x_k > 0$ and some $x_i = B$, then that dx_i/dt is negative; hence, no x_i can exceed B.

An important subcase of Equations (4.11) is the case where the inputs I_i are all equal to 0. This was interpreted to mean that the inputs are encoded by the starting values $x_i(0)$ for the node activities. The pattern that is ultimately stored in short-term memory, after transformation by the recurrent interactions, was interpreted as $x_i(\infty)$. This limiting, or equilibrium, pattern was shown to exist in a wide variety of cases; later (in Grossberg, 1978a) limits were shown to exist for a more general system that included as subcases the systems of both Grossberg (1973) and Grossberg and Levine (1975).

What the limiting pattern is, that is, which parts of the input pattern are stored in short-term memory, depends on the choice of the function f. That function, which has to be monotone increasing, is a signal function representing input–output transformations at the neuronal level, reminiscent of similar constructions in Wilson and Cowan (1972, 1973).

The classes of functions f that were studied closely are the ones shown in Figure 4.10. The dynamics of the network for these functions are as shown in Figure 4.11:

(a) f linear. In this case, the limiting pattern has a *fair distribution*: the values $x_i(\infty)$ are proportional to the values $x_i(0)$, thus representing faithful storage of the original pattern.
(b) f grows slower than linearly. The limiting pattern has a *uniform distribution*: all $x_i(\infty)$ are equal, regardless of the distribution of the $x_i(0)$.
(c) f grows faster than linearly. The limiting pattern has a *0–1 distribution*: for those x_i where $x_i(0) = \max\{x_i(0): i = 1, \ldots, n\}$, $x_i(\infty) = 1$. For all other nodes, $x_i(\infty) = 0$.
(d) f sigmoid. Since a sigmoid function is linear, faster than linear, and slower than linear over different ranges of its argument, the limiting distribution combines fair, uniformizing, and 0–1 tendencies. Inequalities were found which prevent uniformization from occurring. Hence, fair and 0–1 effects combine into an effect described as *contrast enhancement with noise suppression*. The limiting $x_i(\infty)$ are proportional to $x_i(0)$ if $x_i(0)$ is above a threshold value (*quenching threshold*) and equal to 0 if $x_i(0)$ is below that value.

The equations in Grossberg and Levine (1975) are the same as (4.11) except that B is replaced by B_i, which might be different for each i. The set of nodes with the same B_i, called a *subfield*, can be interpreted as the set of neuron populations responsive to a particular sensory feature (such as the color red or the vertical orientation). The results of Grossberg (1973) generalized to distributions that were fair, uniform, 0–1, or contrast enhancing *within each subfield*, as shown in Figure 4.9.

The equations of Grossberg and Levine (1975), without outside inputs, can be rewritten

$$\frac{dx_i}{dt} = -Ax_i + B_i f(x_i) - x_i \sum_k^n f(x_k),$$

the sum now being over all k *including* i (see Exercise 2 of this chapter). Those equations are a subcase of the more general system

$$\frac{dx_i}{dt} = -A_i(x)\left[B_i(x_i) - C(\mathbf{x})\right] \tag{4.12}$$

where $\mathbf{x}$ denotes the vector $(x_1, x_2, \ldots, x_n)$. Grossberg (1978a) proved that every solution of the system of Equations (4.12) approaches an equilibrium, the only restrictions on the functions in the equation being that A_i are nonnegative, B_i are bounded, and C is nondecreasing with respect to each x_i. Since $C(x)$ has a negative influence on the growth of x_i, the latter condition ensures that each node will tend to inhibit other nodes. The proof of approach to equilibrium did not use Lyapunov functions; rather, competitive interactions were shown to restrict the possible number of changes in which variable is growing fastest at a given time. This technique had been used before in Grossberg and Levine (1975) and was carried further in a general mathematical study of competitive dynamical systems by Hirsch (1982, 1984).

But Equation (4.12) generalizes only the form of Grossberg's equations in which inhibition is distance-independent. It does not encompass the distance-dependent equations

$$\frac{dx_i}{dt} = -Ax_i + (B_i - x_i)\left(\sum_{k=1}^n f(x_k)c_{ik} + I_i\right) - (x_i - C_i)\left(\sum_{k=1}^n f(x_k)e_{ik} + J_i\right) \tag{4.3}$$

which were mentioned in Section 4.2. As Cohen (1988) proved, solutions of Equations (4.3) do *not* always converge to an equilibrium, even when $n = 2$. But Cohen and Grossberg (1983) showed that these equations do converge to an equilibrium in the case where there is no internode cooperation, that is, excitatory interaction coefficients c_{ik} are 1 when $i = k$ and 0 otherwise, and inhibitory interaction coefficients obey the symmetry condition $e_{ik} = e_{ki}$.

The Cohen–Grossberg theorem does, however, apply to a system that is more general than (4.12). That system is

$$\frac{dx_i}{dt} = a(x_i)\left[(B_i(x_i) - \sum_{k=1}^{n} c_{ik} d_k(x_k)\right] \tag{4.13}$$

where the interaction coefficients c_{ik} are symmetric ($c_{ik} = c_{ki}$), the functions a_i are nonnegative, b_i are arbitrary, and d_k are differentiable with nonnegative derivative (which indicates the competitive nature of the system). Equation (4.13) includes as subcases not only (4.3) but also the continuous form of the equations by Hopfield (1984) and Hopfield and Tank (1985, 1986), as will be seen in the next subsection. Cohen and Grossberg showed that the Lyapunov function

$$V(\boldsymbol{x}) = -\sum_{i=1}^{n} \int_0^{x_i} b_i(y) d_i'(y) dy + \frac{1}{2} \sum_{j,k=1}^{n} c_{jk} d_j(x_j) d_k(\boldsymbol{x}_k) \tag{4.14}$$

is nonincreasing along trajectories of the system (4.13). We do not give their demonstration here, but in the next subsection we offer an analogous demonstration for the Hopfield–Tank network.

Equations of Hopfield and Tank

Recall from Section 4.2 above that Hopfield (1982) developed a linear threshold algorithm in which nodes had states that can take on the value 1 or 0. The ith node readjusts its state, at random moments in time, according to the rule

$$\begin{aligned} x_i(t+1) &= 1 \text{ if } \sum_{j \neq i} w_{ij} x_j(t) > 0 \\ x_i(t+1) &= 0 \text{ if } \sum_{j \neq i} w_{ij} x_j(t) < 0 \end{aligned} \tag{4.4}$$

Then Hopfield considered the energy (Lyapunov) function

$$E = -\frac{1}{2} \sum_i \sum_j w_{ij} x_i x_j \tag{4.5}$$

He then showed that, if energy changes by an amount ΔE, whenever x_i changes by Δx_i, then

$$\Delta E = -\Delta x_i \left(\sum_j w_{ij} x_j \right) \tag{4.6}$$

Since Δx_i is 0 or of the same sign as $\Sigma_{j \neq i} w_{ij} x_j$, (4.6) implies that $\Delta E \leqslant 0$ at all times.

Some extensions of the above energy function, in both discrete and continuous models, were made in Hopfield (1984) and Hopfield and Tank (1985, 1986).

In Hopfield (1984), the discrete system of the 1982 article was extended to include external inputs I_i to each node and thresholds θ_i that were not necessarily equal to 0. Hence, (4.4) was replaced by the criterion

$$\begin{aligned} x_i(t+1) &= 1 \ \text{ if } \sum_{j \neq i} w_{ij} x_j(t) + I_i > \theta_i \\ x_i(t+1) &= 0 \ \text{ if } \sum_{j \neq i} w_{ij} x_j(t) + I_i < \theta_i \end{aligned} \tag{4.15}$$

Under the condition (4.15), there will always be a decrease in the Lyapunov function

$$E = -\frac{1}{2} \sum_i \sum_j w_{ij} x_i x_j - \sum_i I_i x_i + \sum_i \theta_i x_i \tag{4.16}$$

A generalization of the algorithm defined by (4.15) to the continuous-time case was introduced in Hopfield (1984) and developed further in Hopfield and Tank (1985, 1986). In this work, the output x_i of the ith node was treated as a function (usually sigmoid) of the input u_i, as in the articles reviewed earlier in this section by Wilson and Cowan (1972) and Grossberg (1973). The equation for u_i is then

$$C_i \left(\frac{du_i}{dt} \right) = \sum_j w_{ij} x_i - \frac{u_i}{R_i} + I_i \tag{4.17}$$

where $x_i = g_i(u_i)$ for some increasing, differentiable functions g_i. Since g_i is increasing, it has an inverse function g_i^{-1}; hence, one can write $u_i = g_i^{-1}(x_i)$. C_i and R_i are analogs of capacitance and resistance across the membrane of a single neuron (for more details on membrane electrical flows, see Katz, 1966).

The system (4.17) also has a Lyapunov function similar to (4.16). This function is

$$E = -\frac{1}{2} \sum_i \sum_j w_{ij} x_i x_j + \sum_i \frac{1}{R_i} \int_0^{x_i} g_i^{-1}(V) dV - \sum_i I_i x_i \tag{4.18}$$

If the matrix of weights is symmetric ($w_{ij} = w_{ji}$), then differentiating (4.18) with respect to t yields

$$\frac{dE}{dt} = -\sum_i \frac{\partial E}{\partial x_i} \frac{dx_i}{dt} = -\sum_i \frac{dx_i}{dt} \left[\sum_j w_{ij} x_j - \frac{u_i}{R_i} + I_i \right] \tag{4.19}$$

But the expression in brackets on the right-hand side of (4.19) is just the right-hand side of (4.17). Hence

$$\frac{dE}{dt} = -\sum_i C_i \frac{dx_i}{dt}\frac{du_i}{dt} = -\sum_i C_i \frac{dx_i}{dt}\left[\left(\frac{d}{dt}\right) g_i^{-1}(x_i)\right]$$

$$= -\sum_i \frac{C_i\left(\frac{dx_i}{dt}\right)}{g_i'\left(g_i^{-1}(x_i)\right)}$$

which means that $dE/dt \neq 0$, because C_i is a positive constant and g_i, being an increasing function, has a positive derivative.

The system (4.17) arises from the Cohen–Grossberg system (4.13) with the following substitutions: $a_i(u_i)$ in (4.13) is the constant function equal to $1/C_i$; the coefficients c_{ij} in (4.13) are the *negatives* of the coefficients w_{ij} in (4.17); the functions d_j and g_j of the two respective systems are identified; and the $b_i(u_i)$ are set equal to $-(u_i/R_i) + I_i$. With these substitutions, and the Hopfield–Tank identity $x_i = g_i(u_i)$, it can be seen that the Lyapunov function (4.14) reduces to (4.18).

Equations of Amari and Arbib

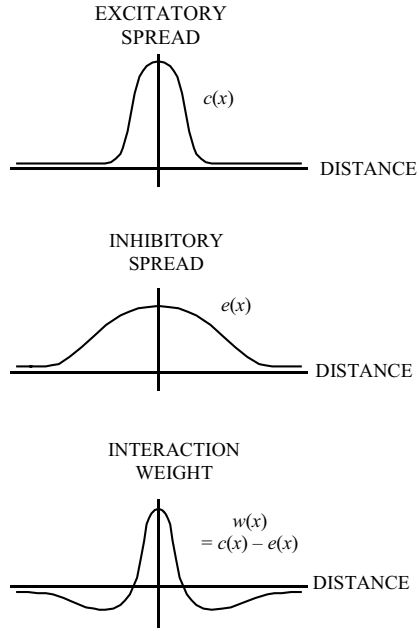


FIGURE 4.21 If excitatory and inhibitory spreads as in Figure 4.4 combine subtractively, a distance-weighting function arises like the one in the bottom graph. This function is called a *difference of Gaussians* (DOG) because the functions $c(x)$ and $e(x)$ can arise from a Gaussian (normal) probability distribution. The DOG is used in model equations by Amari (1977a) and many others.

Amari (1977a) developed an equation for a single-layer neural field of lateral inhibition type. His variable is an activity $u(s,t)$ that depends both on (unidimensional) visual field location s and time t. This equation is

$$\tau \frac{\partial u(s,t)}{\partial t} = -u + \int_y w(s-y) f\left[u(y)\right] dy + h + I(s,t) \tag{4.20}$$

In Equation (4.20), $w(x)$ is a distance-dependent weighting function, one that typically combines short-range excitation and long-range inhibition in an additive fashion, as shown in Figure 4.21. The function f is a step function (1 for u above a threshold, 0 for u below) which is, of course, an approximation of a sigmoid. The constant h denotes baseline activity level, and $I(s,t)$ denotes outside inputs.

In Amari and Arbib (1977), Equation (4.20) was elaborated into equations for a two-dimensional competitive-cooperative field. The two dimensions are position and binocular disparity. There are separate excitatory and inhibitory weighting functions that combine multiplicatively and separate excitatory and inhibitory nodes. (Only excitatory and not inhibitory activity is disparity-dependent.)

Exercises for Chapter 4

**1. Consider Grossberg's differential equation for shunting without lateral inhibition:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_i \tag{4.1}$$

Let the inputs I_i form a constant spatial pattern, $I_i = \theta_i I$ with $\Sigma_i \theta_i = 1$ so that $\Sigma_i I_i = I$. The steady-state solution of a system of differential equations is obtained by setting the derivatives equal to 0. Hence, in (4.1), at the steady-state values $x_i(\infty)$,

$$\begin{aligned} 0 &= -Ax_i(\infty) + (B - x_i(\infty))I_i \\ &= -Ax_i(\infty) + (B - x_i(\infty))\theta_i I \end{aligned}$$

so that, by algebra, $x_i(\infty) = B\theta_i I/(A + \theta_i I)$. This leads to a distortion of the relative pattern weights θ_i. Such distortion has been called the *noise-saturation problem* (Dalenoort, 1983; Grossberg, 1973), because insignificant inputs ("noise") are amplified, while distinctions between intense inputs are blurred ("saturated").

(a) Show that in the shunting equations *with* lateral inhibition,

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_i - x_i \sum_{k \neq i} I_k \tag{4.2}$$

with $I_i = \theta_i I$, the noise-saturation problem disappears. That is, the steady-state values $x_i(\infty)$ are proportional to θ_i, the relative pattern weights.

(b) Find the steady-state values of x_i if (4.2) is replaced by the same equation with minimum activities equal to $C < 0$ instead of 0, namely

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_i - (x_i - C)\sum_{k \neq i} I_k \tag{4.21}$$

Show that for a network defined by (4.21), the $x_i(\infty)$ are proportional to the $\theta_i - K$ for K a constant.

(c) Find the steady-state values of x_i if (distance-dependent) excitatory and inhibitory interaction coefficients are included, i.e.,

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)\sum_{k=1}^{n} I_k c_{ki} - (x_i - C)\sum_{k=1}^{n} I_k e_{ki} \tag{4.22}$$

Note that for the network defined by (4.22), steady-state values are no longer proportional to θ_i, or any linear function of θ_i. For $n = 5$, $B = 2$, $C = .5$, choose values of c_{ki} and e_{ki} that are symmetric and decrease with distance between k and i, with c_{ki} decreasing faster, and see how the steady-state values vary with the total intensity I.

*2. The following problem deals with simulation of the shunting recurrent on-center off-surround equations with attentional biases

$$\frac{dx_i}{dt} = -Ax_i + B_i f(x_i) - x_i \sum_{k=1}^{n} f(x_k) \tag{4.23}$$

as studied by Grossberg and Levine (1975). Note: Equation (4.23) can also be written

$$\frac{dx_i}{dt} = -Ax_i + (B_i - x_i) f(x_i) - x_i \sum_{k \neq i} f(x_k)$$

which is the subcase of (4.3) with

$c_{ik} = 1$ for $i = k$ and 0 for $i \neq k$
$e_{ik} = 0$ for $i = k$ and 1 for $i \neq k$
$I_i = J_i = 0$

Let $n = 3$.

(a) Let $A = 1$, choose values for the B_i such that $B_1 > B_2 > B_3$, and let the signal function $f(x)$ be x^2. Note this means f is "faster than linear"; see Section 4.2 and Figure 4.13. Choose different sets of initial conditions such that the x_i are in an order opposite to that of the B_i, that is, $x_1(0) < x_2(0) < x_3(0)$, and such that for each $i = 1$, 2, or 3, $x_i(0) < B_i$. Verify that the values of x_i reach nearly steady-state values after several hundred or fewer iterations. Vary the ratios between the initial values of x_i and study how that affects which node "wins" the competition. (Hint: Equations (4.23) can be used to give information on the relative sizes

of $(dx_i/dt)/x_i$), which is the "growth rate" of x_i, for different values of x_i and B_i.) It is also possible that no node will win, that is, all three x_i values converge to 0.

(b) Do the same as part (a) with a sigmoid signal function, and with a slower-than-linear signal function such as $f(w) = aw/(b + w)$ for a suitable a and b. The sigmoid should be chosen so that $f(0) = 0$ and so that its inflection point occurs at a positive x value; initial node activities should then be chosen within the range where f is faster than linear or nearly linear.

**3. To simplify Equations (4.4)–(4.6): the Hopfield net consists of nodes with activities x_i equal to 0 or 1 and fixed connection weights that satisfy $w_{ij} = w_{ji}$ when $i \neq j$ (there is no w_{ii}). At random times, activities change *one at a time* according to the following rule:

Let $S_i = \Sigma_{j \neq i} w_{ij} x_j$. Then if $S_i > 0$, x_i becomes 1, whatever its current value is. If $S < 0$, x_i becomes 1, whatever its current value is. If $S_i = 0$, x_i stays at its current value.

For the Hopfield net with 3 nodes and weights $w_{12} = 4$, $w_{13} = 3$, and $w_{23} = -5$, find

- all possible transitions between the 8 possible states;
- all equilibrium states;
- the state at which the Lyapunov function $E = \frac{1}{2}\Sigma_{j \neq i} w_{ij} x_i x_j$ is at a global minimum

○4. One of the most studied visual illusions is the *Muller–Lyer* illusion. In this illusion the perceived length of a line segment can be influenced by the directions of arrowheads to its side; for example, the line segment in (a) looks longer than the line segment in (b), even though objectively they are of the same length (see Figure 4.22). Construct a neural network that explains the Muller–Lyer illusion. Use a recurrent competitive-cooperative network of position and orientation detectors, as discussed in Section 4.3 above.

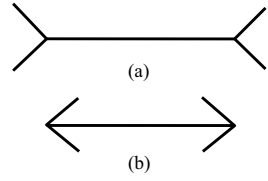


FIGURE 4.22 Muller–Lyer illusion (see text for details).

○5. Using the idea of synchronized oscillations (Section 4.2.6), and any other network ideas from this chapter, design a network model of the perceptual binding process. Present the network with a green circle and a blue square.

The network with the set of connection parameters representing normal cognition should bind the green to the circle and the blue to the square so that it sees the correct objects. Change some of the parameters to represent either attentional overload or parietal lobe damage, and the network should now sometimes incorrectly see a green square and/or a blue circle.

○6. Read the article of Dev (1975), from which Figure 4.21 is taken. Build a network according to the lines she suggested with competition and cooperation in the position dimension, and competition in the disparity dimension. Attempt to reproduce the results of Figure 7 of that article, where a random-dot stereogram is partitioned into segments in which different disparities are detected. (Note: inhibition between detectors of different disparities is mediated by "inhibitory arrays" at another level of Figure 4.21.)

*7. This exercise is designed to simulate the process of boundary completion, as in the illusory square of Figure 4.12. Specifically, the object is to qualitatively reproduce the graphs in Figure 4.23, which is based on Grossberg and Mingolla (1985a).

The "*Y* Field" and "*Z* Field" in Figure 4.23 correspond to two different layers of nodes that respond to a given orientation (say, horizontal) at different visual field positions. They are part of a competitive-cooperative feedback loop, most of which does not need to be reproduced to get the desired effects.

Index the nodes of the lower level (*Y* field) from 1 to 40. Let nodes 15 and 25 receive sustained inputs I_i and not the others. The equations for feedback between higher and lower levels are:

$$\frac{dy_i}{dt} = -By_i + (C - y_i)\left(E[z_i - \delta]^+ + I_i\right)$$

$$\frac{dz_i}{dt} = -Az_i + \frac{.5(S_i^{\text{left}})^2}{\gamma^2 + (S_i^{\text{left}})^2} + \frac{.5(S_i^{\text{right}})^2}{\gamma^2 + (S_i^{\text{right}})^2}$$

where the signals from left and right neighbors are

$$S_i^{\text{left}} = y_i + D\sum_{j=u}^{i-1} y_j, \quad u = \max(i-6,\ 1)$$

$$S_i^{\text{right}} = y_i + D\sum_{j=i+1}^{v} y_j, \quad v = \min(i+6,\ 40)$$

Remember that the superscript "+" in the dy_i/dt equation means replacing the quantity in brackets by 0 if it is negative and keeping that quantity is it is positive. For $i = 1$ the summation in the S_i^{left} equation is 0, and $i = 40$ the summation in the S_i^{right} equation is 0.

Let $A = B = 1$, $D = 10$, and $\gamma = \delta = 0.5$. Find settings of the parameters C and E that will yield the "filling-in" process described by the two graphs *while simulating the differential equations in with time step 0.1*. The separate curves on each graph denote the activities of the nodes in the appropriate field at successive times, with the higher curves occurring at later times. Hints: (1) The simulations that have worked best have used a very high value for the feedback parameter E, a value 20 to 70 times as high as those of the other parameters. (2) When $y_i = C$, The right-hand side of the dy_i/dt equation is negative because $C–y_i = 0$. This means that on the computer generated graphs, y_i should never be larger than the value you have chosen for C! If it goes higher this is due to numerical instability of the program. The time step was chosen in part to prevent that.

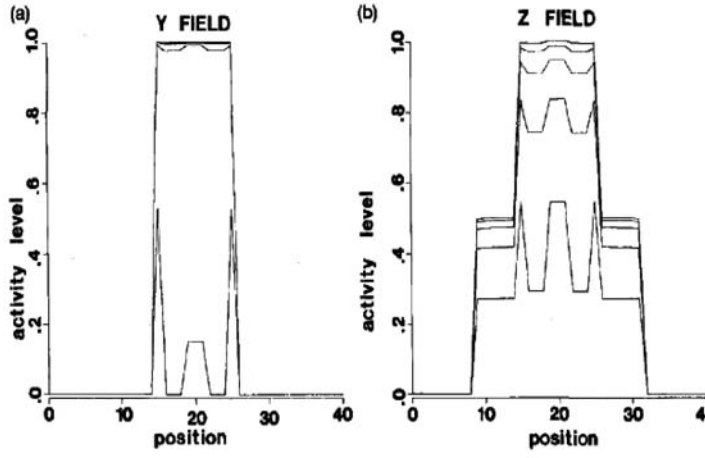


FIGURE 4.23 Activities of two layers of nodes in response to lines of a given orientation; see text for details.

Source: From Grossberg & Mingolla, *Psychological Review*, *92*, 173–211, 1985. Copyright 1985 by the American Psychological Association. Adapted by permission.

Some Additional Sources

Lateral Inhibition in Cortex and Other Brain Areas

Experimental data

Chevalier & Deniau (1990); LaBerge (1990).

Network Models

Coultrop, Granger, and Lynch (1992); Koch and Ullman (1985); LaBerge, Carter, and Brown (1992); Reeves and Sperling (1986); Reggia, D'Autrechy, Sutton, and Weinrich (1992); Tsotsos et al. (1995); Walley and Weiden (1973).

Review of Roles in Cognitive Psychology

Levine and Brown (2007).

Synchronized Oscillations

Ghose and Freeman (1997); Horn & Opher (1996); Levine, Brown, & Shirey (2000); Malsburg & Schneider (1986).

Modeling Visual Form, Color, and Depth Detection:

Grossberg and Mingolla (1987); Grossberg and Todorovi_ (1988); Takebe, Nakauchi, and Usui (1996); Usui, Nakauchi, and Miyake (1994).

Modeling Visual Motion Detection

Carandini and Heeger (1994); Chey, Grossberg, and Mingolla (1998); Francis and Grossberg (1996); Heeger, Simoncelli, and Movshon (1996); Lange and Lappe (2006); Marshall (1990); McKinstry, Seth, Edelman, and Krichmar (2008); Ö_men and Gagné (1990); Tlapale, Dosher, and Lu (2015); Wurbs, Mingolla, and Yazdanbakhsh (2013); Xue and Liu (2014).

Note

1 This symmetry assumption is made for mathematical convenience and is not likely to be biologically realistic. On this point, the reader should refer back to the discussion of Kosko's BAM in Section 3.4.

PART II

Computational Cognitive Neuroscience

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

PROGRESS IN COGNITIVE NEUROSCIENCE

> Too much sanity may be madness and the maddest of all, to see life as it is and not as it should be.
>
> Miguel de Cervantes

> The true method of knowledge is experiment.
>
> William Blake

5.1. The Emergence of Cognitive Neuroscience

Since the publication in 2000 of the second edition in this book, the growth of cognitive neuroscience has been explosive. While functional magnetic resonance imaging (fMRI) has been employed as a research tool since 1991, transcranial magnetic stimulation (TMS) since 1985, and positron emission tomography (PET) since the early 1980s, gradual improvement in the functioning and availability of all those devices has greatly increased the number of noninvasive human studies in this century. In addition, over the same period there has been ever increasing knowledge of the single-cell properties of both monkeys and rodents involved in behavioral and cognitive tasks. Also, there has been a wide range of studies involving a host of other techniques including the traditional electroencephalography (EEG) and event-related potentials (ERP); different imaging techniques such as near-infra-red spectroscopy (NIRS, which is primarily cortical) and single photon emission computed tomography (SPECT); and gene manipulation.

The term "cognitive neuroscience" was coined by the psychologist George Miller and the neuroscientist Michael Gazzaniga in the late 1970s. At that time there was still little contact between neuroscientists and either experimental

psychologists or cognitive scientists. In the 1980s things began to change: leading psychologists such as Amos Tversky started to deliver keynote addresses at the annual meeting of the Society for Neuroscience, and large psychology conferences included many brain imaging and single-neuron studies. The growth of the connections between neuroscience and cognitive science can be attributed mainly to the noninvasive brain mapping techniques and to the presence in every researcher's office and laboratory of computers capable of handling what are now called "big data." The development of more brain-faithful neural network modeling was also both a cause and an effect of this development and occurred concurrently with it.

The growth of studies has enabled us to parcel cognitive neuroscience into related subspecialties depending on the processes being studied. For example, the widely read collection edited by Gazzaniga (2009) includes sections, each with its own section editor, on the following topics: development and evolution; plasticity; attention; sensation and perception; motor systems; memory; language; the emotional and social brain; higher cognitive functions; and consciousness. In addition, the widely read collection of neuroimaging studies of cognition edited by Cabeza and Kingstone (2006) contains the following sections: attention; skill learning; semantic memory; language; episodic memory; working memory; executive functions; early cognitive development; cognitive aging; emotion and social cognition; neuropsychologically impaired patients.

While cognitive neuroscience is a young field, a number of studies have become classic enough to be part of the canon of neuroscientists, psychologists, and neural modelers alike. This chapter, organized by subsections, will attempt to review some of these classic studies. It is inevitable that we will omit some results that some readers consider equally important in the field's development. For those readers needing additional background in neuroscience, particularly in the location and function of specific brain regions, Appendix 2 of this book provides some of this background and cites other books that give greater detail about those regions.

5.2. Cognitive Neuroscience of Conditioning and Reinforcement Learning

5.2.1. Dopamine and Reward

A series of studies, starting in the 1980s, by Wolfram Schultz and his colleagues on the responses of neurons in two midbrain nuclei, the *substantia nigra pars compacta* (SNc) and *ventral tegmentum* (VT), made significant contributions to understanding the neuroscience of conditioning (e.g., Ljungberg, Apicella, & Schultz, 1992; Mirenowicz & Schultz, 1994; Schultz, 1986; Schultz,

Apicella, & Ljungberg, 1993). Those two nuclei are the brain's main sources of the neurotransmitter dopamine, the transmitter most implicated in reward perception and prediction. The experiments of Schultz et al. were inspired by many other investigators' work implicating dopamine in the actions of addictive drugs and in brain self-stimulation.

The experiments of Schultz's group involved monkeys performing behavioral tasks and receiving rewards, usually fruit juice or a piece of apple. In the early stages the majority of dopamine neurons increased their firing rates for short periods of time when the reward was presented, regardless of the type of reward. After that, visual or auditory cues were repeatedly paired with the reward. When the animals had been trained, the dopamine neurons shifted their phasic burst of activation to the time of delivery of the visual or auditory conditioned stimulus, and no longer responded directly to the reward itself. Hence at all times these neurons responded to the earliest predictor of reward.

Schultz and his colleagues also found that the midbrain dopamine neurons were sensitive to extinction of the conditioned response. Once training had occurred, the animals expected a reward at a particular time interval after the presentation of the CS. If the reward was not delivered, the dopamine neurons showed a corresponding decrease in firing rate. Hence, dopamine neurons are sensitive to reward prediction error, that is, either more or less reward delivered than expected.

Fiorillo, Tobler, and Schultz (2003, 2005) found that these same dopamine neurons are also sensitive to the uncertainty of reward. In a conditioning paradigm wherein reward is obtained on some trials, in addition to the phasic activity in response to an unexpected reward the dopamine neurons show sustained activity in response to the CS. This sustained activity does not occur if the probability of a trial being rewarded is 1, but gradually increases as the reward probability is lowered to .5. If the reward probability is brought lower than .5 the sustained activity decreases again until it is absent at reward probability 0.

The results of Schultz et al. led other researchers to investigate the exact function dopamine performs in the process of learning about rewarding stimuli. Berridge and Robinson (1998) and Berridge (2007) reviewed evidence that dopamine antagonists did not interfere with the affective pleasure from rewards. These researchers concluded that dopaminergic reward signals strengthen the "wanting" of a reward, that is, the motivation to work for the reward in a current context, not the "liking" or affective enjoyment which is more related to other brain systems (including opioid receptors and the amygdala).

The results on dopamine function from Schultz's laboratory and several others consistently show that dopamine is involved in learning and prediction errors for affectively positive and rewarding stimuli. As for affectively negative or punishing stimuli, some studies show dopamine neuron involvement but many others do not. Another neurotransmitter, serotonin (5HT), seems to be

involved in avoidance responses and often seems to counteract the effects of dopamine. For that reason, Daw, Kakade, and Dayan (2002) and Boureau and Dayan (2011) conjectured that 5HT neurons code prediction errors for punishment in the same way that DA neurons do for reward. Yet that conclusion is hard to verify because of the multitude of receptor types for 5HT (even greater than it is DA), and the data so far do not clearly support or refute the notion of a punishment prediction error.

5.2.2. Some Roles of Striatum, Orbitofrontal Cortex, and Amygdala

The dopamine neurons send signals to wide areas of the brain but have a particularly close association with the striatum, both its dorsal parts (caudate and putamen) and ventral parts (nucleus accumbens). Learnable synapses from prefrontal cortex to striatum are particularly important in reward learning. There is now considerable evidence that plasticity in corticostriatal synapses depends on intact dopamine systems, which is not true of plasticity elsewhere in the brain.

Aosaki, Graybiel, and Kimura (1994) recorded electrically from tonically active neurons (TANs) in the dorsal striatum of monkeys who were being classically conditioned to associate a sound with a reward. They found that in normal monkeys these TANs increased their responsiveness to the click CS in the course of the conditioning task. However, monkeys given a dopaminergic neurotoxin, and thereby depleted of dopamine inputs to the striatum, did not show that increase in responsiveness. More recent results on dopamine and corticostriatal plasticity are reviewed in Calabresi, Picconi, Tozzi, and DiFilippo (2007) and Wickens (2009). These articles summarize evidence that dopamine mediates both long-term potentiation (LTP) and long-term depression (LTD) at corticostriatal synapses, through molecular mechanisms that have not been completely specified. Of these two processes, LTD is more prevalent, which may be consistent with the suggestion of Wickens and Kotter (1995) that "dopamine increases output from the most active cells (which are in the minority) and decreases output from the less active cells" (p. 192). There have also been suggestions that the balance between LTP and LTD in the striatum is mediated by the dopaminergic system's interactions with other neurons in the striatum that use acetylcholine as a transmitter (Bullock, Tan, & John, 2009).

The orbital prefrontal cortex (OFC) has strong projections to the ventral striatum, as well as reciprocal connections with the amygdala, and orbitofrontal damage in humans impairs making decisions based on expected positive or negative consequences of actions. Hence it was natural to look for OFC activity related to reward and behavioral control. Tremblay and Schultz (2000a, 2000b) studied monkeys on a *go–no go* task, that is, one where they either

performed or withheld a reaching movement and received either a liquid reward or a sound that could indicate future rewards. Tremblay and Schultz discovered that neurons in the OFC mainly respond to the liquid reinforcer itself, regardless of whether the monkeys have been trained to execute or withhold the movement at that time.

Other results have shown that OFC neurons respond to the *relative* values of either rewards or punishments. Tremblay and Schultz (1999) first found that monkeys have a distinct order of preference among food rewards; for example, they prefer raisins to apples and apples to cereal. With the preference order of foods labeled A for most preferred, B for next most, and C for least preferred, Tremblay and Schultz divided a conditioning experiment into trial blocks where the reward could either be A or B, and other trial blocks where the reward could either be B or C. They found that the same OFC neurons that were more active before A than before B when those were the alternatives were just as active before B when B and C were the alternatives. Several years later, in a human imaging experiment, Elliott, Agnew, and Deakin (2008) found that OFC codes relative and not absolute value of financial rewards. Finally, Blair et al. (2006) found that both OFC and *anterior cingulate cortex* (*ACC*) respond differently to the comparison of two desirable outcomes versus the comparison of two undesirable outcomes.

Both the OFC and amygdala play roles in the encoding of the positive or negative consequences of actions or values of stimuli, and several neuroscientists have investigated the subtle differences between the functions of those two regions. Among the first were Bechara, Damasio, and their colleagues, who developed and studied a human decision task called the *Iowa gambling task* (IGT) (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Damasio, Damasio, & Lee, 1999; Bechara, Damasio, & Damasio, 2003). In the IGT the participant goes through trials (usually 100) where he or she must draw a card from one of four decks of cards shown on a computer screen. Each deck provides different gains and losses of play money. In the most common version, two of these decks (decks A and B) have higher short-term payoffs (say, $100 per card as opposed to $50 for decks C and D). However, the decks with higher short-term gains also lead to long-term expected losses.

Bechara and his colleagues tested subjects with and without damage to either the OFC or amygdala. They found that the subjects without brain damage begin with selections from one of the risky decks but gradually begin to shift toward advantageous decks as the task progresses. On the other hand, patients with damage to either the OMPFC or amygdala never learn the advantageous strategy.

Bechara et al. also studied their subjects' skin conductance responses (SCRs). Subjects without brain damage not only generate SCRs to positive or negative feedback but gradually develop anticipatory SCRs during the interval

preceding a risky choice. Patients with amygdalar damage do not generate SCRs to negative feedback. Patients with OFC damage, on the other hand, do show SCRs to negative feedback but not anticipatory SCRs.

The main connections of OFC to amygdala are to an area called the *basolateral nucleus*: these connections are bidirectional but there are more fibers going from amygdala to OFC than the reverse (Ghashghaei & Barbas, 2002). The basolateral amygdala in turn is connected to the *central nucleus*, which is connected to the hypothalamus and autonomic nervous system.

Thus both the OFC, with its connections to higher-order goal representations, and the amygdala, with its connections to primary emotional areas, are essential for the flexible representation of the positive and negative emotional consequences of events. Rolls (e.g., 2000, 2004) has found both OFC and amygdalar neurons that respond to reinforcing stimuli for all the different sensory modalities (vision, hearing, touch, taste, and smell). Rolls showed that when reinforcement contingencies change, for example when a food is no longer valued owing to satiety for that particular taste or when a conditioned stimulus is no longer rewarded, amygdalar neuron responses to the stimuli are much slower to change than OFC neuron responses. Yet Schoenbaum, Setlow, Saddoris, and Gallagher (2003) found that in animals with basolateral amygdala lesions, the OFC neurons do not respond to reinforcing stimuli in a normal manner.

As humans have developed complex societies, some social reinforcers have come to activate the same brain regions as biological reinforcers. This is particularly true of financial reinforcers. An fMRI study by O'Doherty, Kringelbach, Rolls, Hornak, and Andrews (2001) of a probabilistic visual association task found that different parts of OFC (medial and lateral) showed activations that correlated with the amount of money gained or lost.

There have also been many studies of the neuroscience of two specific animal conditioning paradigms. One of these is fear conditioning, whereby the animal learns to associate a particular stimulus such as a tone with an aversive stimulus such as electric shock. The other is eyelid conditioning, also called the nictitating membrane response or eyeblink conditioning, whereby the animal learns to associate a tone with an air puff to the eye or a tap to the forehead that induces eyelid closure.

5.2.3. Fear Conditioning

There have been a large number of studies of auditory fear conditioning involving areas of cortex, thalamus, and amygdala (see, e.g., LeDoux, 2000). Typically, the subject is presented with a tone followed by a brief electric shock. After several pairings the tone by itself begins to elicit any or all of several fear-related responses. These responses could include freezing, autonomic responses such as changes in heart rate or skin conductance,

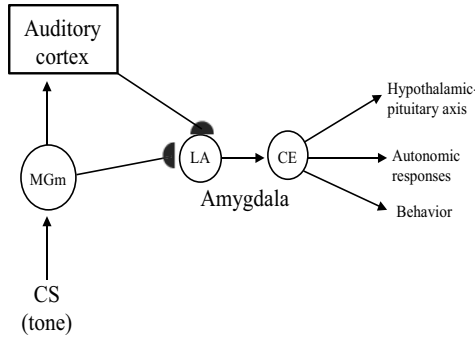


FIGURE 5.1 Neural pathways involved in fear conditioning of an auditory CS. MGm = medial division of the medial geniculate body, the auditory area of the thalamus. LA is the lateral nucleus and CE the central nucleus, both of the amygdala.

Source: Adapted from LeDoux, 2000, with the permission of Annual Reviews.

endocrine responses such as release of cortisol, and enhancement of some reflexes such as startle and eyeblink reflexes.

Figure 5.1 shows some of the pathways involved in fear conditioning to a tone. There are plastic connections, involving NMDA receptors (see Chapter 3), both from auditory cortex and from the medial geniculate, the main auditory areas of thalamus, to the amygdala, which terminate on the amygdala's lateral nucleus. The cortical connection allows for fine distinctions among auditory stimuli, whereas the thalamic connection is thought to allow for faster but less differentiated responses (but see Pessoa & Adolphs, 2010, for a different view). The lateral nucleus in turn sends projections, both directly and indirectly via other nuclei (basal and accessory basal), to the amygdala's central nucleus. The central nucleus projects both to the autonomic nervous system and to the hypothalamus, generating internal responses such as heart rate, blood pressure, and skin conductance increases and pituitary stress hormones.

In addition to the tone, animals in the fear conditioning paradigm develop fear responses to the apparatus itself (the stimulus context). The contextual associations are mediated by connections from the hippocampus to the basal and accessory basal nuclei of the amygdala (Byrne et al., 2014; LeDoux, 2000).

5.2.4. Eyelid Conditioning

The behavioral properties of eyelid conditioning in the rabbit (which has a nictitating membrane, unlike humans) were studied by Smith, Coleman, and Gormezano (1969). These investigators found that the CS has to precede the US by at least 50 msec for successful conditioning to occur, with the optimal CS–US interval being several hundred milliseconds. Conditioning can occur at CS–US intervals up to four seconds in *delay conditioning*, whereby the CS

is still present at the time of US onset. In *trace conditioning*, whereby the CS is removed before US onset, conditioning cannot occur at intervals above two seconds.

Studies of cerebellar lesions showed that both the cortex and deep nuclei of the cerebellum are required for the nictitating membrane response. In particular, lesions of the lateral anterior interpositus nucleus of the cerebellum have been found to prevent occurrence of the eyelid conditioned response (CR) but have no effect on the unconditioned response (UR) (e.g., McCormick & Thompson, 1984; Yeo, Hardiman, & Glickstein, 1985a). Lesions of the cerebellar cortex also have various disruptive effects on eyelid conditioning, particularly on the timing of the CR (McCormick & Thompson, 1984; Perrett, Ruiz, & Mauk, 1993; Yeo, Hardiman, & Glickstein, 1985b).

The lesion studies have also been supported by single-cell studies of Purkinje neurons in the cerebellar cortex during eyelid conditioning. In particular, Berthier and Moore (1986) found inhibition of Purkinje cell spiking that preceded the CR. Since the connections from Purkinje cells to deep cerebellar nuclei are also inhibitory, this meant that interpositus firing increased during the conditioning paradigm. Purkinje cells have two major inputs: the *climbing fibers* from the inferior olive and the *parallel fibers* from other cells in the cerebellar cortex. Ito, Sukurai, and Tongroach (1982) found that conjoint activation of those two sets of fibers leads to long-term depression (LTD, see Chapter 3) in parallel fiber synapses, suggesting that climbing fibers could be a pathway for the CR and parallel fibers for the UR.

Another brain region that is particularly important for eyeblink conditioning paradigm (as it is for conditioning in general) is the hippocampus. The hippocampus does not seem to be necessary for eyeblink delay conditioning but it is necessary for eyeblink trace conditioning (e.g., Berger & Thompson, 1978; Green & Woodruff-Pak, 2000).

The role of the cerebellum in mediating timing of the conditioned response is complemented by a role for the hippocampus in encoding the timing of *stimulus* arrivals. In particular, Berger, Berry, and Thompson (1986) found that during the NMR and one other conditioning paradigm (conditioned jaw movement) the pattern of neuron responses in the largest type of hippocampal neurons (*pyramidal neurons*) mimics the time course of the conditioned response. This time course is called *adaptive* because it fits the learned timing of US arrival.

These adaptively timed cell responses are from a subregion of hippocampus called *CA3*. The CA region receives inputs from different types of cells in another region of hippocampus called the *dentate gyrus*. These dentate cells are "time-locked" to the CS; that is, each cell exhibits an increase in firing rate starting at a fixed time interval after the CS. Hence the hippocampal network has to convert an array of fixed time delays into adaptive timing.

5.3. Cognitive Neuroscience of Categorization

Knowlton and Squire (1993) did a behavioral study on amnesic patients and non-amnesic participants using dot patterns. They presented patterns that were either small or large distortions of the prototype of a category of dot patterns or random patterns that were not similar to the prototype of that category. The amnesics were not significantly worse than the non-amnesics at the classification task, which involved placing the prototype-like patterns in the category and the random patterns outside the category. Yet the amnesics were significantly worse than the non-amnesics at remembering specific category members that they had seen.

From these behavioral results Knowlton and Squire concluded that the brain regions involved in categorization decisions were different from those involved in declarative memory of category members; specifically, the hippocampus and other medial temporal lobe structures involved in declarative memory were less important for categorization. This supposition was supported by the fMRI studies of Reber, Stark, and Squire (1998a, 1998b) and Reber, Wong, and Buxton (2002). In particular, with a categorization task such as that of the earlier Knowlton–Squire article, the posterior occipital cortex exhibited *less* activity for dot patterns that were placed in the category than for noncategorical dot patterns. In some but not all of the studies the categorical patterns elicited greater activity in some prefrontal regions. On a recognition task involving the same dot patterns, the recognized members of the category elicited greater activity in posterior occipital than other patterns.

Reber and his colleagues explained the categorization data by positing that members of a category were processed visually in a less effortful manner than category nonmembers. This involves a sense of *familiarity*, which is widely recognized by cognitive psychologists to be separate from explicit memory of previous events (also known as *recollection*; see also Eichenbaum, Yonelinas, & Ranganath, 2007). Hence, the neuroscience of categorization is closely related to the neuroscience of distinguishing familiar from novel events.

A partial review of brain mechanisms involved in detecting novel events is given by Ranganath and Rainer (2003). These authors review literature implicating a variety of areas in novelty detection, including parts of the lateral prefrontal cortex, orbital prefrontal, anterior insular and anterior temporal cortex, temporo-parietal cortex (brown), perirhinal and posterior parahippocampal areas, much of the hippocampus, and amygdala and cingulate cortex. Two neurotransmitters are particularly involved in novelty detection: acetylcholine, which is implicated in attention, and norepinephrine, which is implicated in arousal.

In addition to fMRI, novelty has been studied by means of event-related potentials, particularly on auditory tasks. That line of inquiry began with the discovery by Sutton, Braren, Zubin, and John (1965) of a positive potential

around 300 milliseconds after presentation of a novel or improbable stimulus, known as the *P300*. In a typical study of this kind, one tone is presented repeatedly but a different tone is interspersed with the dominant tone less often, and the P300 is elicited by the different (sometimes called "oddball") tone. In some studies by Banquet and his colleagues, reviewed in Banquet and Grossberg (1987), the oddball also elicits some other, earlier ERP components, such as a negativity at processing, positivity at 120 ms, and negativity at 200 ms.

Differences between familiarity and recollection in neural processes in the hippocampal formation are reviewed by Yonelinas (2002) and Eichenbaum et al. (2007). These articles discuss different roles in memory for the hippocampus itself and three areas of older cortex contiguous to the hippocampus in the medial temporal lobe area, namely the *perirhinal*, *parahippocampal*, and *entorhinal* cortices. Familiarity is associated with memories of items and recollection with a memory that includes the context (spatial and temporal) of those items. As Eichenbaum et al. (2007) review, memory for items is concentrated in the perirhinal cortex and lateral entorhinal areas, whereas memory for context is concentrated in the parahippocampal and medial entorhinal areas. Inputs from the cortex to both of those general regions converge in the hippocampus, which represents items in context. Hence the hippocampus is crucial for recollection but not for familiarity, which is consistent with the Knowlton and Squire (1993) amnesia data.

Yet, some fMRI results of Nosofsky, Little, and James (2012) challenge the notion that categorization and recognition memory involve separate brain systems. Nosofsky and his colleagues used categorization and recognition tasks with the same types of random-dot patterns as Reber and his colleagues, but included a version of the recognition task with a lax criterion. That is, their instructions emphasized the importance of not missing any previously seen patterns, but downplayed the importance of avoiding false alarms to patterns that had not been seen. With the lax instructions, the activation patterns accompanying recognition became much more similar to those involved in categorization. This suggested to the authors that, rather than separate memory systems, the two tasks might involve different parameter settings within the same neural system.

5.4. Cognitive Neuroscience of Vision and Visual Attention

5.4.1. Visual Cortical Cell Properties and Lamination

The modern era of visual neuroscience began with the work of Hubel and Wiesel (e.g., 1962, 1963, 1965, 1968). These Nobel laureates set out initially to find what stimuli were most effective at eliciting neuronal responses in the primary visual, or striate, cortex in cats and monkeys. Previous investigators

had found that circular spots of light were the most effective at eliciting neuronal responses in the retina. Yet Hubel and Wiesel discovered that the preferred stimuli in the visual cortex tended to be oriented bars or slits of light. Moreover, cells responsive to particular orientations tended to cluster together in columns, analogous to what Mountcastle (1957) had previously found in the somatosensory cortex. They also found cells selective for other attributes besides orientation, including retinal position, color, and ocularity (left or right eye); later investigators found cells that were selective for spatial frequency (Robson, 1975) and for disparity between left and right visual images, a measure of depth (Barlow, Blakemore, & Pettigrew, 1967).

Hubel and Wiesel also discovered within the visual cortical areas a hierarchy of different types of cells with varying levels of complexity of the descriptions of their preferred stimuli. These hierarchical levels were termed *simple*, *complex*, and *hypercomplex cells*. For example, the area of space in which a simple cell responds to objects, termed its *receptive field*, often has an oval or rectangular shape, with the cell being excited by light within that oval or rectangle and inhibited by light just outside it. Complex cells often require the bar of light to move in a particular direction in order to evoke a cell response. Hypercomplex cells in addition tend to respond to bars of a particular length and be inhibited by longer bars, which led to the later designation of a large class of those neurons as *end-stopped cells* (Hubel & Livingstone, 1987).

A series of subsequent articles by Hubel and Livingstone (e.g., Hubel & Livingstone, 1987; Livingstone & Hubel, 1984) further illuminated the functional significance and receptive properties of different neurons within a hierarchy of connected visual regions from lateral geniculate to primary visual cortex (Brodmann area 17, also known as visual area 1 or V1) to secondary visual cortex (Brodmann area 18, also known as V2). These investigators reviewed data from cells in two species of monkeys (macaques and squirrel monkeys) and came up with the following picture of those regions, with some interspecies variation. In the lateral geniculate, the main thalamic gateway from retina to cortex, inputs from the retina are divided into two distinct sets of layers that differ in their cell sizes so are called *magnocellular* and *parvocellular*. In V1, labeling of cells with a stain called *cytochrome oxidase* reveals a split of the parvocellular pathway into blobs that are sensitive to color and regions between blobs ("interblobs") sensitive to shape. Then cytochrome oxidase staining reveals three different domains within V2. These regions consist of thick stripes that are orientation-selective and receive inputs from a particular layer of V1; pale stripes that are also orientation-selective, and more than half hypercomplex, and reciprocally connected with interblobs; and thin stripes that are color-sensitive and reciprocally connected with blobs.

DeYoe and Van Essen (1988) reviewed these processes and added two further visual cortical processing areas called V3 and V4. These authors also incorporated previous work showing that visual processing in the cortex

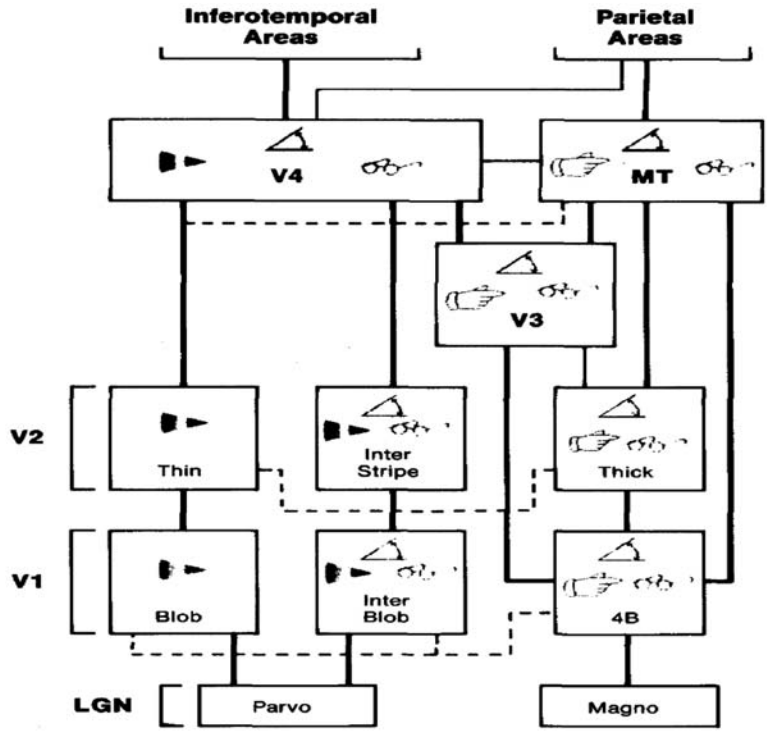


FIGURE 5.2 Schematic of connections between lateral geniculate (LGN), areas V1, V2, and V4, going up to inferotemporal and parietal cortex.

Source: Reproduced from DeYoe & Van Essen, 1988, with the permission of Elsevier Science.

could be divided into a "What" pathway mapping properties of objects and a "Where" pathway mapping their locations (Ungerleider & Mishkin, 1982). After both pathways go through different yet overlapping visual cortical areas, the What pathway ends up at the inferotemporal cortex, and the Where pathway at the parietal cortex. They come together ultimately in the prefrontal cortex. A rough schematic of these connections, developed by DeYoe and Van Essen (1988), is shown in Figure 5.2. A more complete description of all the brain's visual areas and their connections is found in Felleman and Van Essen (1991; see Figure 5.3).

5.4.2. Visual Attention

We have noted that most of the connections in the visual system are reciprocal (see Figure 5.3). These connections provide a pathway for selective attentional influences on the perceptions of color, form, orientation, and other features described in the last section. As with the What and Where pathways, the

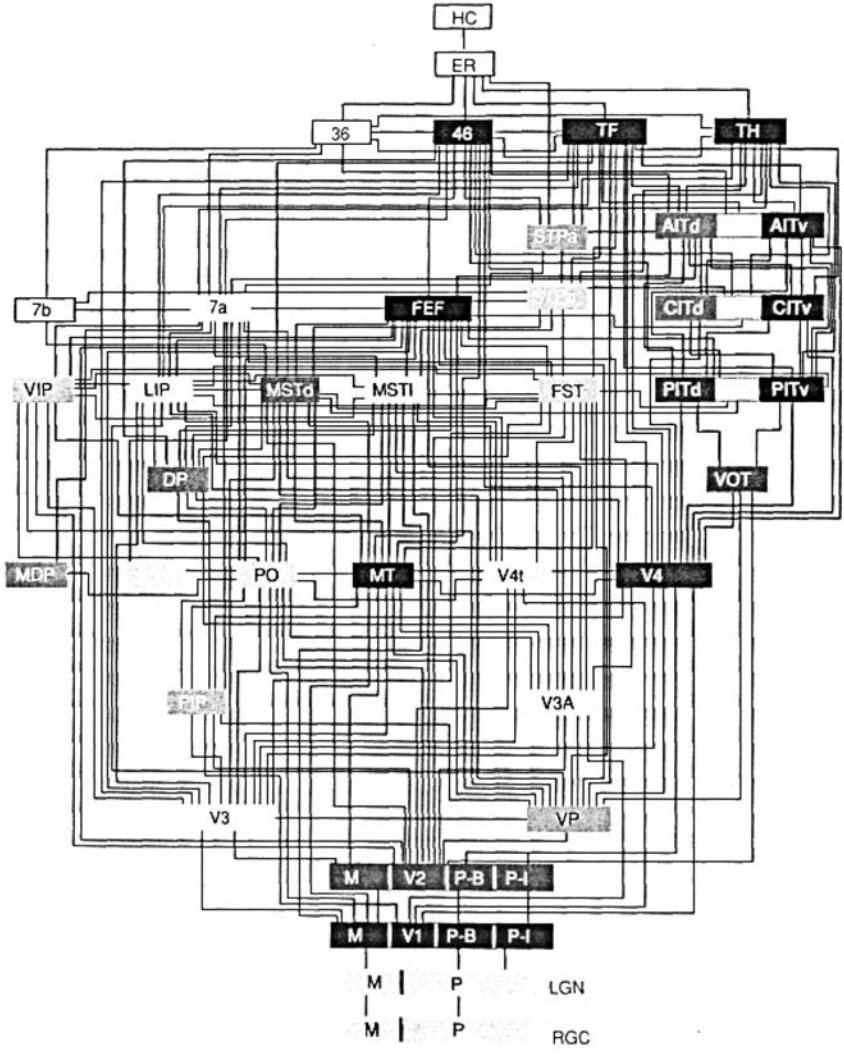


FIGURE 5.3 Complete diagram of interconnections among over 30 visual areas of the brain. RGC = retinal ganglion cells, LGN = laternal geniculate nucleus. Most of the pathways have been verified to have reciprocal pathways. A few nonvisual areas (e.g., somatosensory cortex and perirhinal cortex) are included.

Source: Felleman & Van Essen, 1991, with the permission of Oxford University Press.

selective attention can either be to particular locations or to particular features such as colors or shapes. Also, selective attention can be influenced from below by stimulus properties or from above by goals or tasks.

Selective attention in both humans and monkeys has been found to operate by biasing signals, mainly from association areas of cortex, that influence

processing by visual cortex. As Kastner and Ungerleider (2000) note, these biasing signals can influence responses of visual cortical neurons in a variety of ways:

> (a) the enhancement of neural responses to attended stimuli; (b) the filtering of unwanted information by counteracting the suppression induced by nearby distracters; (c) the biasing of signals in favor of an attended location by increases of baseline activity in the absence of visual stimulation; and (d) the increase of stimulus salience by enhancing the neuron's sensitivity to stimulus contrast.
>
> (p. 332)

A series of articles by Desimone's laboratory (e.g., Moran & Desimone, 1985; Desimone & Duncan, 1995; Reynolds, Chelazzi, & Desimone, 1999) describes studies of attentional influences on monkey cell responses in V1, V2, and V4. In the first of these studies (Moran & Desimone, 1985), the animal was trained to attend to stimuli at two different locations at different times. The receptive fields and effective stimuli were determined for cells in V4 and inferotemporal cortex. It was found that if there were an effective and an ineffective stimulus (for that cell) both located in its receptive field, and the animal were attending to the location of the ineffective stimulus, its firing rate was less than half of what it had been when the animal attended to the location of the effective stimulus. If the ineffective stimulus was outside the cell's receptive field it did not have the same suppressive effect on the cell's response. Variations on the behavioral paradigm led to similar attentional effects on responses in V2 and V1 cells in later work (Motter, 1993; Reynolds et al., 1999), but the effects on the earlier visual processing stages tended to be smaller than those on V4 and inferotemporal cortex.

These results led Desimone and Duncan (1995) to develop the theory that "objects in the visual field compete for limited processing capacity and control of behavior" and that "competition is biased in part by bottom-up neural mechanisms that separate figures from their background (in both space and time) and in part by top-down mechanisms that select objects of relevance to current behavior" (p. 216). Yet, at the time these authors wrote, there were few results that verified the location of these biasing signals. Subsequent fMRI studies on humans verified that different parts of parietal and prefrontal cortex played dissociable roles in selective attention (e.g., Giesbrecht, Kingstone, Handy, Hopfinger, & Mangun, 2006; Greenberg, Esterman, Wilson, Serences, & Yantis, 2010; Kastner & Ungerleider, 2000; Taylor, Rushworth, & Nobre, 2008), although a complete network of regions involved in attention and their specific roles has not yet been established.

Attentional influences of V2 on V1 were further investigated in several monkey studies by Bullier and his colleagues, notably Bullier, Hupé, James,

and Girard (1996). They found that inactivation of V2 feedback to on-center neurons in V1 increased those neurons' responses to stimuli in the surrounds of their receptive fields while decreasing or leaving unchanged their responses to stimuli in their receptive field centers. This suggests a sort of on-center off-surround organization of feedback pathways from V2 to V1. Analogously, the work of Sillito, Jones, Gerstein, and West (1994) suggests an on-center off-surround organization of feedback from V1 to LGN. Sillito et al. presented stimuli with orientations preferred by specific V1 neurons in cats, and found that these led to synchronized firing of those LGN cells whose receptive fields were encompassed by those stimuli.

The role of parietal and prefrontal areas in visual attention was suspected when it was found that lesions in these areas can cause neglect of part of the visual field (e.g., Posner & Petersen, 1990). Different parietal and prefrontal regions are involved in the three stages described by Giesbrecht et al.:

> . . . when a subject is cued to move his or her spatial attention to a new location, attention must first be *disengaged* from its current location within the visual field. Following disengagement, attention must then be *moved* from its initial location to the new location. Finally, once the attentional spotlight has been moved to the new location, attention must then be *engaged* with whatever stimuli are in the new location.
>
> (pp. 88–89, authors' italics)

As Kastner and Ungerleider (2000) noted, the higher-order component of attention often works in the absence of current visual stimulation, to direct the organism's mental processes toward a particular spatial location or a particular attribute before it is seen. This is anticipation of future stimulation. Similarly, attention is closely connected to the memory of past stimulation. Kastner and Ungerleider also note the strong connections between attention and working memory, with some of the same brain areas (particularly prefrontal areas) involved in both processes. Working memory is discussed further in Section 5.6.

Attentional influences on the early stages of visual processing can cause responses of V1 neurons to reflect any of a wide diversity of vision-related instructions given the (human or monkey) subject. An example occurs in the monkey curve tracing experiment of Roelfsema, Lamme, and Spekreijse (1998). If monkeys are trained to attend to a target curve and ignore a distractor curve, the V1 neuronal responses to locations along the target curve are enhanced even if the distractor curve crosses the target curve.

5.4.3. Visual Filling-In

Paradoxically, while attention leads us to ignore parts of the visual world that are physically present, we also often preattentively fill in parts of the visual

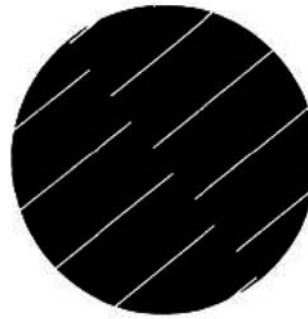


FIGURE 5.4 Example of a stimulus with an illusory contour. Line segments in the grating point up and 45 degrees to the right, causing an illusory line segment to be perceived that points down and to the right. V2 neurons sensitive to the down-and-right orientation also respond to that illusory line segment.

Source: Reproduced from von der Heydt & Peterhans, 1989, with the permission of the Society for Neuroscience.

scene that are physically absent. As discussed in Chapter 4, visual filling-in allows us to perceive continuity in objects in spite of either occlusion by other objects or blind spots on the retina. A neural basis for visual filling-in was found in the 1980s by von der Heydt, Peterhans, and their colleagues, who showed that there are neurons in the visual cortical area V2 (but not in the primary visual cortex, V1) of monkeys that respond to illusory contours.

For example, von der Heydt and Peterhans (1989) found that close to half the neurons in V2 that responded to lines of a given orientation also responded to illusory lines of the same orientation that were made by the ends of line segments in a grating of the orthogonal orientation, such as the one shown in Figure 5.4. Peterhans and von der Heydt (1989) found that neurons responsive to dark bars in their receptive fields also responded to an illusory bar created by moving notches in two rectangles just above and below the cells' receptive fields. These illusory contour results show that the higher levels of visual processing are necessary for the filling-in effects that create continuity in visual perception.

5.5. Cognitive Neuroscience of Sequence Learning and Performance

Many important tasks in the real world require learning and later performing repeatable sequences of behaviors or actions. These tasks include, for example, speaking, typing, playing a musical instrument, and reproducing items from a presented list. Sequence learning and reproduction involves a huge number of brain regions, in particular the cerebellum; several areas of basal ganglia; motor cortex, including regions of Brodmann area 6 known as the *supplementary*

motor area (SMA) and *pre-SMA*; prefrontal cortex; and hippocampus. The cognitive neuroscience of sequences has a relatively recent history, starting with single-cell studies on monkeys learning sequences in the mid-1990s and later supplemented by human imaging studies.

A series of monkey single-cell studies by Tanji, Shima, and their colleagues found a diversity of cell responses in the SMA and pre-SMA. For example, Mushiake, Inase, and Tanji (1990) and Mushiake, Masahiko, and Tanji (1991) trained monkeys to push four buttons on a touch pad in a particular order. The task was either executed under a visually guided condition (three of the buttons were illuminated in sequence followed by a GO signal) or a memory condition (GO signal only). There were neurons in the SMA that were active only in the memory condition and only before one specific sequence. Shima and Tanji (2000) and Tanji and Shima (1994) trained monkeys to carry out a sequence, the elements of which were one of three possible movements: a push, a pull, or a turn of a manipulandum. There were visually guided and memory conditions for the task and neurons in both the SMA and pre-SMA selective for memory and for a particular sequence. But there was also a second type of neuron, more common in the pre-SMA than in the SMA, which was selective for the movement's position in a sequence, regardless of which specific movement was being performed. A third type of neuron, more common in the SMA than in the pre-SMA, was active selectively at transitions between two particular movements – for example, after a push and before a pull, but not after a push and before a turn, or after a turn and before a pull.

Nakamura, Sakai, and Hikosaka (1998) recorded in SMA and pre-SMA during monkeys' learning and performance of what they call the 2-by-N task, which involves learning a sequence ("hyperset") of several sequences of two actions (see also Hikosaka et al., 1999). The responses of individual neurons were preferentially related to either the acquisition of new hypersets or the performance of previously learned hypersets. Neurons related to learning of new sets were predominantly located in pre-SMA, whereas the SMA appeared to contain a roughly equal distribution of neurons related to new and learned sets.

What is the role of subcortical regions such as cerebellum and basal ganglia in sequence/learning? Hikosaka et al. (1999) reviewed a variety of evidence that the interactions of cerebellum with the cortical regions discussed above are important for precise timing of sequential behaviors. The interactions of basal ganglia with cortex, on the other hand, are important for learning the connections between sequential movements and rewards.

These single-cell studies of different cortical and subcortical regions did not address whether neural representationof a sequence being learned is a serial or a parallel process. Many early psychologists regarded sequence learning as a serial process consisting of chains of associations; that is, to learn, say, the sequence ABCD one needs to strengthen the connections A-to-B, B-to-C, and

C-to-D. Yet, Lashley (1951) argued that sequence learning cannot be based on associative chaining but instead is based on parallel representations of all the sequence elements, excited to different degrees at a given time. Lashley's argument was based on the characteristic errors in sequence learning, which involve reproducing the correct sequence elements but in the wrong order. Later cognitive psychologists have verified this argument by showing that typical errors on many tasks involve the "filling-in" of elements that were previously skipped; that is, if the sequence to be learned is ABCD and the participant has produced A and then C, he or she is most likely to go next to B, which has been skipped, rather than D, which follows C, as an associative chaining model would predict. Lashley concluded that neural representations of all sequence elements are present at the start of the sequence reproduction task.

Neurophysiological support for Lashley's parallel representation hypothesis came with the work of Averbeck, Chafee, Crowe, and Georgopoulos (2002, 2003) and Averbeck, Crowe, Chafee, and Georgopoulos (2003) on neurons in an area of the prefrontal cortex (a part of Brodmann area 46). These researchers trained monkeys to draw a set of geometric shapes including a triangle, square, trapezoid, and upside-down triangle. After the monkey held a joystick for one second, a template (geometric form) appeared on the right half of the screen, and the monkey was free to draw on the left half. If the monkey executed a complete drawing trajectory, while keeping the cursor within invisible limits that defined acceptable form, it received a juice reward. Shapes were drawn in blocks of consecutive trials of the same shape, enabling the monkey to anticipate the appropriate shape in the subsequent trial, on all trials except the first trial of a block. Analysis of the monkey's hand movements showed that the continuous trajectory was composed of a sequence of individual segments. While the monkeys carried out this task, ensembles of individually isolated single area 46 prefrontal neurons were recorded. Neural activity patterns were defined based upon the average ensemble neural responses that occurred during the drawing of individual segments of the geometric shapes, and these activity patterns were considered neural correlates of each segment of the shape. During the period before drawing the segments, it was found that the neural correlates of all the segments were active (see Figure 5.5). As that figure shows, the relative strength of the representation of each segment corresponded to the serial position of the segment, such that, prior to the execution of the sequence, the first segment had the strongest representation, the second had the second strongest representation, etc.

Given that we do form sequential associations, current models predict that there should be serial as well as parallel processes in sequence learning. The location of serial processes is not as well established but one likely candidate is the cerebellum. A variety of evidence from patient studies led Rhodes and Bullock (2002) to include in their sequence learning model a cerebellar controller that can learn interitem transitions as well as sequence chunks.

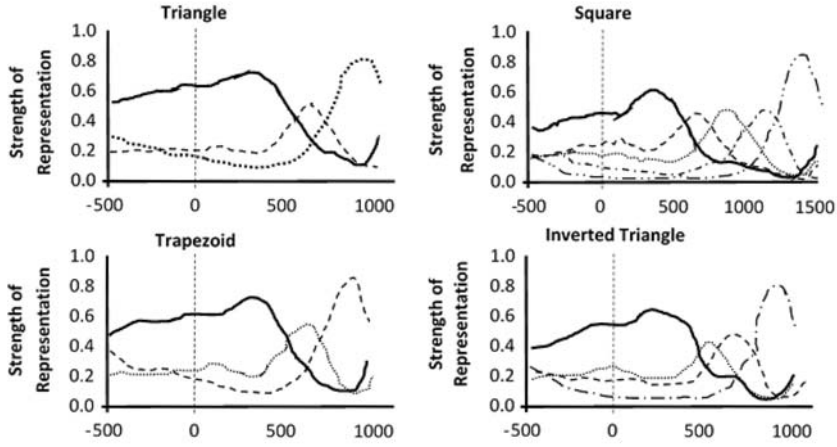


FIGURE 5.5 Strength of neural representations (defined in the text) of different shapes that a monkey had to draw, as a function of time. Time 0 indicates the onset of the template, while time bins during the hold period and RT are 25 ms. Length of segments were normalized to permit averaging across trials. Segment 1 is indicated in solid lines; Segment 2 in green; Segment 3 in dashed lines; Segment 4 in dotted-dashed lines; and Segment 5 in dashed–double dotted lines.

Source: Reproduced from Averbeck et al., 2002, with the permission of the National Academy of Sciences.

5.6. Cognitive Neuroscience of Executive Function and Cognitive Control

The term *executive function* was developed in the 1960s and 1970s to describe behavior that is not automatic or following fixed patterns but adaptable to current tasks and short- or long-term goals. Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) distinguished between *automatic* and *controlled* processes. Controlled processes require active mental manipulation and are employed when the consequences of specific actions are not known for sure. Automatic processes do not require active manipulation and are employed when the consequences of actions are invariant. Processes that start out being controlled often become automatic after they have been well learned: common examples are driving a car and speaking a language.

Much of our early knowledge of the cognitive neuroscience of executive function was based on lesion data, specifically data about humans or monkeys with damage to one or another part of the prefrontal cortex (PFC). One of the first examples was the famous nineteenth-century patient Phineas Gage, who lost the ability to plan and order behavior after damage to a part of the PFC; a modern reconstruction (Damasio, 1994) showed that the main locus of his damage was the OFC. Milner (1964), Nauta (1971), and Pribram (1973) found

evidence that prefrontal lesions could lead to many types of executive deficits, including distractibility, disorganization, impulsiveness (or over-deliberateness), perseveration in formerly rewarding behavior, and excessive attraction to novelty.

Norman and Shallice (1986) posited two interacting neural systems for generating automatic and controlled cognitive processes. They called the generator of automatic processes the *contention scheduling system* and called the generator of controlled processes the *supervisory attentional system* (*SAS*). These authors enumerated the following conditions for the SAS to become involved:

1. Those that involve planning or decision-making.
2. Those that involve error correction or troubleshooting.
3. Situations where responses are not well learned or contain novel sequences of actions.
4. Dangerous or technically difficult situations.
5. Situations that require the overcoming of a strong habitual response or resisting temptation.

(Norman & Shallice, 1986, paraphrased by Wikipedia)

Starting in the 1960s and 1970s, several investigators observed the errors of frontally damaged humans and monkeys on tasks such as the Wisconsin card sorting test and Stroop test. On the Wisconsin card sorting test, participants are asked to classify a special deck of cards which differ by three criteria: color, shape, or number of the designs on the face of the cards (Figure 5.6).

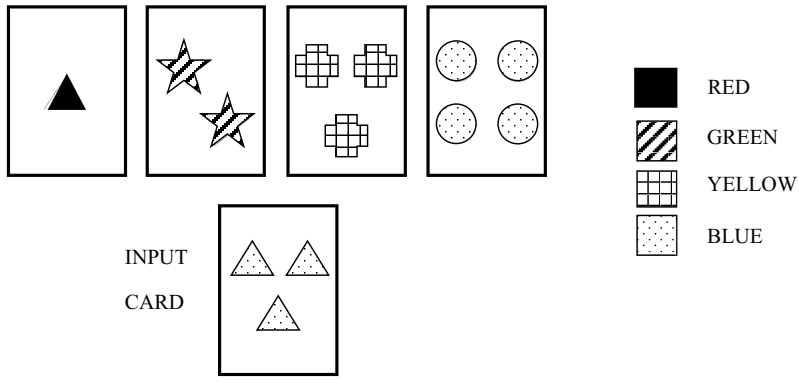


FIGURE 5.6 Cards used in the Wisconsin card sorting test; the input card is matched to one of the four template cards above it.

Source: Reprinted from *Neural Networks*, *2*, D. S. Levine & P. S. Prueitt, Modeling some effects of frontal lobe damage: Novelty and perseveration, 103–116, copyright 1989, with permission from Elsevier Science.

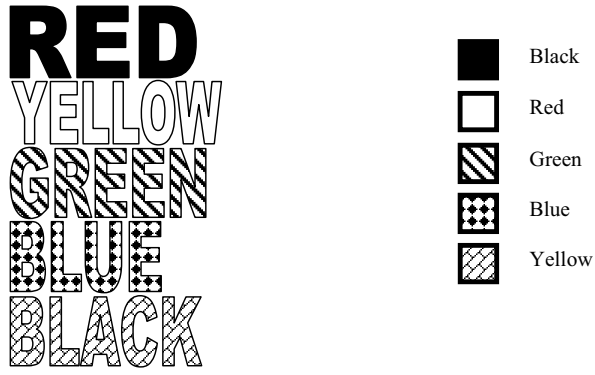


FIGURE 5.7 Example of stimuli used in the Stroop test.

The experimenter says "right" or "wrong" without giving a reason, but changes the criterion used after the participant makes ten consecutive correct classifications. Patients with damage to the *dorsolateral prefrontal cortex* (*DLPFC*) tend to learn the first criterion used but not be able to switch. On the Stroop test (Figure 5.7) the participant sees a word representing a color in a display of either the same color or another color, and is asked to say what color the word is, suppressing his or her natural tendency to simply read the word. Frontal patients have more trouble than those with intact frontal lobes at saying the right color.

Studies of such neuropsychological tests led to the conception that the prefrontal cortex exerts top-down control on other brain regions via representations of goals. This theory was further reinforced in the 1980s by results from monkey single-neuron studies, particularly those of Fuster and his colleagues (for summaries see Fuster, 1985a, 1985b, 1997). For example, Fuster and his colleagues found that, while monkeys performed cognitive tasks such as delayed-matching-to-sample (whereby the monkey was reinforced for approaching an object identical to one it had previously seen), neurons in the prefrontal cortex became selectively responsive to various aspects of the task (e.g., the stimulus, the delay, and the responses). Quintana and Fuster (1992) found that lateral prefrontal neurons can code the degree of association between a cue and a response (see also Asaad, Rainer, & Miller, 1998). Other lateral PFC neurons have been found to code associations between cues and rewards (e.g., Watanabe, 1992).

In order to direct the organism toward goals, the PFC neuronal activity needs to be sustained throughout the task performance period. Numerous investigators, starting with Fuster (1973) and Kubota and Niki (1971), have found that PFC neurons remain active during the delay period between a cue and a response. Often the delay period activity is specific to a particular type of information, including either the "what" or "where" of a stimulus, its sequential

association within a series, an expected reward, or an associated response: all these different types of specificity have been found in various studies reviewed by Miller and Cohen (2001, p. 180).

Recent human neuroimaging results are consistent with the idea that prefrontal activity is required for tasks that require control and not for repeated or routine tasks. Jansma, Ramsey, Slagter, and Kahn (2001) found that practice on a working memory task involving letters led to decreased task-induced activation of various areas previously shown to be involved in working memory tasks, namely DLPFC, right frontopolar cortex (Brodmann area 10), and a broad area that includes both ACC and SMA. These investigators did not find that different cortical areas increased activity with more practice, but other investigators studying other cognitive tasks found increased activation of different areas with practice (e.g., Petersen, van Mier, Fiez, & Raichle, 1998, using a word production task).

Co-activation of prefrontal areas, particularly DLPFC, and ACC has been found in various executive and working memory tasks by several investigators (e.g., Duncan & Owen, 2000; Smith & Jonides, 1999). Yet, MacDonald, Cohen, Stenger, and Carter (2000) found a double dissociation between activations of DLPFC and ACC, which suggested a division of labor between those two regions. Specifically, on the Stroop task the ACC was more active than the DLPFC for incongruent stimuli, where a word for one color was shown in another color, than for congruent stimuli, where a word for a color was shown in the same color. The DLPFC and not the ACC, by contrast, was more active when the participant was accurately naming the color than when she or he was reading the word. This led MacDonald et al. to posit that the ACC is involved in detecting conflict and the DLPFC in control of the responses to that conflict. That theory is further elaborated by Miller and Cohen (2001):

> The demands for control are associated with an increase in PFC activity; tasks demanding greater control elicit stronger activity within the PFC; and the ACC responds selectively to conflict in processing. However, further work is needed to establish the causal relationship between detection of conflict within the ACC and the augmentation of control by the PFC.
>
> (pp. 190–191)

The general picture Miller and Cohen (2001) drew of the interrelated roles of ACC and DLPFC in cognitive control is still widely accepted. Yet the exact functions of these regions are still undetermined and remain an active focus of research. Botvinick, Braver, Barch, Carter, and Cohen (2001) review several findings that implicate the ACC in response to the occurrence of conflict. Some of their data involve a requirement to override a prepotent response, as in the Stroop test. Other data involve what is called *underdetermined responding*:

the participant is required to choose from a set of responses without strong indications in favor of one of those responses. Yet, Brown and Braver (2005) obtained data suggesting to them that rather than detecting conflict, ACC responds to a prediction of the likelihood of making an error. These two closely related interpretations of ACC function (conflict versus error) are still in dispute (e.g., Aarts, Roelofs, & van Turennout, 2008; Yeung & Nieuwenhuis, 2009).

More recent studies of control circuits involved in high-level function have looked at different regions of PFC besides the dorsolateral and orbital/ventromedial parts. One of these is the *frontopolar cortex* (Brodmann area 10), the furthest forward part of the PFC. Christoff and Gabrieli (2000) reviewed several neuroimaging studies of reasoning and episodic memory. They found that DLPFC is involved in evaluation of externally generated information, whereas internally generated information requires serial activation of DLPFC and then frontopolar cortex. Later studies (Badre & D'Esposito, 2007; Christoff, Keramatian, Gordon, Smith, & Mädler, 2009; Koechlin & Hyafil, 2007) point to a hierarchy of abstraction, both of stimuli and intended behaviors, encoded by the PFC as one moves forward (from ventrolateral to dorsolateral to frontopolar).

Another area of prefrontal cortex that has emerged in recent imaging studies is the *ventrolateral* (VLPFC; Brodmann areas 44, 45, and 47). Bunge (2004) reviewed evidence that the use of rules to guide human behavior involves a division of labor between VLPFC and DLPFC, whereby VLPFC and its interactions with the temporal cortex are required for rule retrieval, whereas DLPFC is required for rule-based response selection.

Braver and Ruge (2006) reviewed some more recent work on neuroimaging of executive functions and cognitive control. These authors identify seven different categories of executive functions which different tasks require to different degrees. The categories are (1) strategic control of memory, (2) stimulus–response interference, (3) response inhibition, (4) underdetermined responding; (5) performance monitoring, (6) task management, and (7) higher cognition. In their review of fMRI findings they found that each category of functions tends with some variation to activate specific classes of brain regions, as follows:

- Task conditions involving the preparatory cuing of attention or the use of attentional control to resolve interference tend to activate posterior and inferior regions of lateral PFC.
- Task conditions requiring the temporary suppression of ongoing responses tend to activate right inferior PFC regions.
- Task conditions requiring rapid shifting of attention to different dimensions or reconfiguration of task sets reliably engage the superior parietal cortex.

- Task conditions involving the free selection of potential response alternatives engage superior medial frontal areas near the SMA.
- Task conditions involving the processing of internal or external feedback related to the outcome of generated actions reliably engage the ACC and nearby medial frontal areas.
- Task conditions that require the tracking of changing stimulus–response contingencies elicit activation in the OFC.
- Complex cognitive activities (such as planning, analogy verification, and controlled episodic retrieval) that involve the evaluation or integration of abstract dimensions maintained in working memory tend to engage the anteriormost regions of the PFC (BA 10).

(p. 330)

5.7. Cognitive Neuroscience of Decision-Making

Since the groundbreaking behavioral studies of Tversky and Kahneman (e.g., Tversky & Kahneman, 1974, 1981) it has been widely recognized that human decision-making is not typically constrained by executive control but is heavily influenced by heuristics which are partly automatic. In other words, decision-makers sometimes override prepotent responses with task-relevant responses but at least as often employ task-inappropriate prepotent responses instead. Recently, some fMRI studies of classical decision paradigms have shed light on some differences in brain activation patterns between controlled (deliberative) and automatic (heuristic) decision processes on the same task.

One of the classic decision problems by Tversky and Kahneman has been termed the *Asian disease problem*. It asks participants to decide between two public health measures to combat a disease expected to come to the United States and kill 600 people: one measure will save 200 for sure; the other has a 1/3 probability of saving 600 and 2/3 of saving none. The framing of the choices makes a strong difference in the decisions: if they are framed in terms of people saved, participants prefer the safer option of saving fewer for sure, but if they are framed in terms of people dying, participants prefer the risky option that might lead to no deaths. DeMartino, Kumaran, Seymour, and Dolan (2006) conducted an fMRI study of a monetary decision task analogous to the Asian disease problem and compared activation patterns for choices that conformed to the traditional framing effect (risk seeking for losses or risk averse for gains) with choices that violated the framing effect (risk seeking for gains or risk averse for losses). These investigators found larger OFC and ACC activity on choices violating the framing effect and larger amygdala activity on choices conforming to the framing effect.

DeNeys, Vartanian, and Goel (2008) instructed their participants that, within a specific group of people, there were a certain number of lawyers and

a certain number of engineers, with the numbers varying across participants. The researchers then identified a person as a member of that lawyer–engineer group, gave the participants a description closely related to stereotypes of either the law or engineering profession, and asked the participants to estimate the probability of that person being a lawyer or engineer. The prepotent response is to judge the probability of being in each profession solely from the description: this is called *base rate neglect* because the appropriate response is to consider the base rates (relative frequencies in the population as a whole) as well as the description. In the *incongruent case*, the group was dominated by members of one of the professions but the description fit the stereotype of the other profession (e.g., the group consisted of 9990 lawyers and 10 engineers, but the description fit the engineer stereotype). In the *congruent case*, the majority of the group consisted of people in the same profession whose stereotype was fit by the description. DeNeys et al. (2008) found that virtually all participants showed greater ACC activation for the incongruent case compared with the congruent case, consistent with participants detecting conflicting information. Yet those participants who used the base rate showed greater DLPFC activation than participants who neglected the base rate.

Many recent brain imaging studies have investigated the neural bases for emotional reactions to monetary gains and losses. One of the key areas for representing gains is the ventral striatum (or *nucleus accumbens*), an area much studied in relationship to other kinds of rewards including food, sex, and drugs. Knutson and Peterson (2005) found that the ventral striatum is activated by the anticipation of increasing monetary gains, in a manner roughly proportional to the magnitude of those gains and accompanied by positive emotional arousal. Yet, many years of behavioral studies note that humans to varying degrees show loss aversion, in that the negative value of the loss of a certain amount of money is greater than the positive value of an equivalent gain. Tom, Fox, Trepel, and Poldrack (2007) investigated the neural basis of loss aversion with fMRI studies of people who were deciding whether to accept or reject gambles with a 50% chance of gaining some amount of money and losing a different (smaller, usually about half) amount of money. They found that anticipated gains activated the nucleus accumbens and other reward areas including dopaminergic midbrain nuclei and ventromedial PFC, whereas anticipated losses were coded by decreased activity in those same regions. Moreover, the amount by which activity in reward areas decreased in individuals correlated with behavioral loss aversion. There were no brain areas that were more active with anticipated losses than gains.

Also, there are several imaging studies suggesting that many of the emotion-related brain areas are sensitive to the comparison of obtained positive or negative outcomes with alternative, unobtained possible outcomes. One of the first of these studies was by Breiter, Aharon, Kahneman, Dale, and Shizgal (2001), who devised a gambling game where the participants were each

presented with one of three spinners composed of three sectors each labeled with a monetary value. There was a good spinner, where the lowest value was $0 and the others positive; an intermediate spinner, where the medium value was $0 and the others of opposite signs; and a bad spinner, where the highest value was $0 and the others negative. In addition to differential responses to the spinners themselves, two reward areas – the nucleus accumbens and sublenticular extended amygdala – were most activated by the same $0 outcome when it was part of the bad spinner, so that the $0 was considered good relative to the unobtained losses. Henderson and Norris (2013) informed their participants of two possible outcomes of a card they might select: for one card either a larger or smaller gain, for another either a larger or smaller loss. Hence there were four possible outcomes, which they called "outright win," "disappointing win," "relieving loss," and "outright loss." Several of the emotional regions of the basal ganglia, limbic system, and OFC were sensitive both to the gain versus loss distinction and the type of gain or loss. Also, a fronto-parietal network was sensitive to the outcomes that caused mixed emotion (a disappointing win that was not as good as it might have been, and a relieving loss that was not as bad as it might have been).

5.7.1. Social Decision-Making

Since the start of this century there have been an increasing number of imaging and lesion studies of social interactions, as reviewed by Rilling and Sanfey (2011). These studies have implicated many of the same brain regions involved in decision-making and reward processing in general.

Some of these studies have involved three games that measure the amount of trust and altruism between two players. The first of these games is the *prisoner's dilemma* (PD). The scenario of the PD is that the prosecutor talks separately to each of two people involved in a crime and tells them how many years they will get behind bars if they defect (confess to the crime) or cooperate (not confess in order not to implicate their partner), numbers that also depend on what their partner does. Neither of the two prisoners knows what the other will do, and the numbers are set up so that each individually will get fewer years if they defect regardless of what the other does, but both get fewer years if they both cooperate than if they both defect. In the *ultimatum game* (UG), one player proposes a division of a specified sum of money between the two. The other player has to decide whether to accept the division, and if so the sum is divided as proposed. If the other player rejects the division, neither player receives any money. The *dictator game* (DG) is like the UG except that the second player must accept the offer.

The PD is used as a measure of trust between the two partners, with reciprocated cooperation being considered the most desirable outcome. In fMRI

studies of the PD and related games, reciprocated cooperation is associated with activation of two brain regions involved in reward processing, the caudate nucleus and OFC (Rilling et al. 2002; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004), indicating the social rewards of mutual trust. Rilling et al. (2002) also showed that the strength of response in the caudate predicts the degree of future cooperation. The effects of the decisions to cooperate or defect also interact with personality differences, particularly on a test of prosocial behavior (valuing the welfare of others). In the nucleus accumbens (nACC), an area commonly associated with reward, prosocial participants show more activity when choosing to cooperate than when choosing to defect, whereas less prosocial participants show the opposite pattern of nACC activity (van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2009).

The UG is also a measure of the altruism of the player is who is dividing the money (how much he or she decides to be fair) and of what offers the other player perceives to be fair. Most players offer the other player more than their own minimum acceptable offer, but that is not true of players who have lesions in the ventromedial PFC that includes OFC (Krajbich et al., 2009). It was also found that the level of generosity could be influenced by hormone levels: *oxytocin*, a peptide hormone that has many prosocial effects in both animals and humans, increases generosity whereas testosterone decreases generosity but only in men (Zak, Stanton, & Ahmadi, 2007; Zak et al., 2009). Interestingly, oxytocin and testosterone do not have any effect on generosity in the DG, which is a purer measure of altruism. That result suggests that these two hormones may be related to the ability to both empathize with the partner and predict the partner's behavior, prediction that is irrelevant in the DG where the other player's response is forced (Zak et al., 2009). As for the recipient, receiving what the player perceives to be an insufficiently generous offer activates an area of the limbic cortex called the *insula*, which is implicated both in negative affect and in empathy (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Other results have shown perceived unfairness to lead to activity in the amygdala, which seems to mediate the fear of betrayal.

Finally, the ACC (at least its dorsal part), which has been implicated in monitoring other types of conflict (Botvinick et al., 2001), reacts to violations of social norms (see Rilling & Sanfey, 2011, for review). Examples of events that trigger dorsal ACC activity are breaches of a promise, deviations from a group opinion, and envy of others.

Some Additional Sources

Cognitive Neuroscience of Conditioning and Reinforcement Learning

Dayan and Berridge (2014); Maia (2009); O'Doherty et al. (2004); Solway and Botvinick (2012).

Cognitive Neuroscience of Categorization

Braunlich, Gomez-Lavin, and Seger (2015).

Cognitive Neuroscience of Visual Illusions

Donohue, Green, and Woldorff (2015); Kogo and Wagemans (2013); Maertens and Pollmann (2007); Weidner, Boers, Mathiak, Dammers, and Fink (2010); Wokke, Vandenbroucke, Scholte, and Lamme (2013).

Cognitive Neuroscience of Visual Attention

Baldauf and Desimone (2014); Fries, Womelsdorf, Oostenveld, and Desimone (2008); Johnson and Johnson (2009); Van der Stigchel et al. (2009); Wu et al. (2013).

Cognitive Neuroscience of Sequence Learning and Performance

Ashe, Lungu, Basford, and Lu (2006); Herd et al. (2013); Hitch, Flude, and Burgess (2009); Orban et al. (2010); Reithler, van Mier, and Goebel (2010); Segawa, Tourville, Beal, and Guenther (2015).

Cognitive Neuroscience of Executive Function

Botvinick and Braver (2015); Chevalier, Martis, Curran & Munakata (2015); Menon and Uddin (2010); Nyhus and Barceló (2009); Provost and Monchi (2015); Schubotz (2011); Seeley et al. (2007).

Cognitive Neuroscience of Decision-Making

Chang & Sanfey (2013); Feng, Luo, and Krueger (2015); Naqvi, Shiv, and Bechara (2006); Newell and Shanks (2014); Suter, Pachur, Hertwig, Endestad, and Biele (2015).

6

MODELS OF CONDITIONING AND REINFORCEMENT LEARNING

> What reinforcement we may gain from hope;
> If not, what resolution from despair.
>
> John Milton (*Paradise Lost*)

> Learning is not attained by chance, it must be sought for with ardor and diligence.
>
> Abigail Adams

Section 5.2 of the last chapter reviewed recent progress in the neuroscience of conditioning and reinforcement learning. In the last 20 years there has been a concurrent development of models that have captured some of the roles for brain regions such as dopaminergic midbrain nuclei, basal ganglia, amygdala, hippocampus, and orbital prefrontal cortex in learning processes. Yet these models are strongly rooted in earlier neural and psychological models starting in the late 1960s that predated the recent neuroscientific results.

6.1. Early Network Models of Classical Conditioning

6.1.1. Brindley and Uttley

The first neural networks for Pavlovian conditioning were developed in the 1960s within the framework of all-or-none neuronal models. Brindley (1967, 1969) modeled some conditioning data using the all-or-none, symbolic logic framework of McCulloch and Pitts (1943), with the addition of modifiable synapses as proposed by Hebb (1949).

Brindley (1967) discussed ten types of modifiable synapses and the logic of their operation. Of the ten, the most important is the "Hebb synapse," whose facilitation depends on correlated pre- and postsynaptic activities. Also, it is possible for two neurons in the network to be connected by two different synapses, one modifiable and one unmodifiable. This was an early approximation to a connection between continuous neural elements where synaptic strength can take on a range of values.

In spite of oversimplifications resulting from the all-or-none framework, Brindley's networks for classical and operant conditioning have structural details in common with later models. For example, strengthening a particular synapse in his classical conditioning network enables the conditioned stimulus (CS) to activate a cell that at first needed the unconditioned stimulus (US) to be activated. This takes place through the action of a network of interneurons, one of which is *polyvalent*, that is, responding to a combination of conditioned and unconditioned stimuli. The ideas of polyvalence and of the CS "gaining control" of a US-activated arousal area were prominent in the continuous model of Grossberg (1971), which formed the basis for several later articles. Brindley (1969) also extended some of these ideas to the learning of sequences of three words.

Another early set of theories relevant to classical conditioning was developed by Uttley (1970, 1975), using mathematical information theory. These articles were based on a pattern discrimination model of Uttley (1966), which was in turn based on a linear threshold network (see Chapter 2) with binary inputs x_i, and an output $\Sigma_{i=1}^{m} w_i x_i + \Gamma$ where the w_i are synaptic connection strengths and Γ is the negative of a response threshold. The output cell responds if and only if the output is positive; its binary signal is called y. The connection strengths w_i are in turn calculated from relative probabilities of the cooccurrence of events. Hence, synaptic weights increase with cooccurrence of pre- and postsynaptic events, in accordance with Hebb's postulate (see Chapters 2 and 3).

Uttley (1975), however, found this Hebbian learning could cause connection weights to increase without bound. He solved this problem by reversing the sign of the synaptic change, making conductivities proportional to the *negative* of a function based on probability of cooccurrence of x and y. Uttley used a class of networks with such negative feedback to model classical conditioning. In the case of a single CS–US connection, the CS, called A, excites a pathway of variable weight w_A, and a US, called U, excites a pathway of fixed weight w_U. The output of the network which includes these pathways is the conditioned response (CR). Owing to the negative information synapses, it was shown that, if A is always reinforced by U, the steady-state equation for the weights reduces to $w_A + w_U = 0$. Since $w_A > 0$, because A develops a positive associational strength, this means that $w_U < 0$, that is, the pathway from U to the CR is inhibitory.

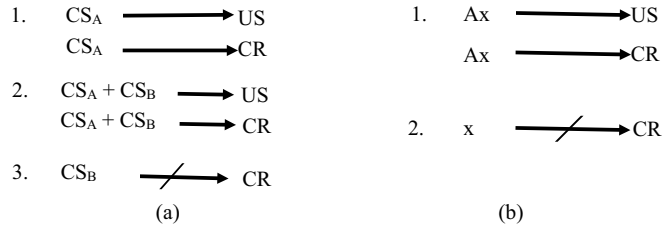


FIGURE 6.1 Schematic of experimental stages in blocking (a) and overshadowing (b); see text for explanations.

Hence, in Uttley's model, there is a given response, and different stimuli compete for ability to be associated with that response. Enhanced associational strength of one stimulus tends to weaken the associational strength of another. This is true not only in the case of a CS and US, as discussed earlier, but also in the case of two CSs. Thus his theory can account for such classical conditioning data as the *blocking* and *overshadowing* paradigms (Kamin, 1969), which are illustrated in Figure 6.1. In blocking (recall the discussion in Section 3.3), the animal is first given many presentations of one CS (called A), each followed by a US at a given time interval. The CS A is then presented many times in combination with another stimulus X, each pair followed by the US at the same time interval as before. On recall trials, the animal has developed a conditioned response (CR) to A alone or to the AX combination, but not to X alone. In overshadowing, the US is associated with the AX combination but cue A is more *salient* than cue X – that is, either more intense or more important to the organism's survival. Again, no CR has been learned to X alone. The competition for associability between different, previously neutral stimuli provides a compact explanation for both of these effects.

The inhibitory US-to-CR connection, however, seems paradoxical because the US reliably evokes the unconditioned response (UR). A series of later articles (Uttley, 1976a, 1976b, 1976c) illuminated this inhibitory connection further. In the networks of these articles, the signals from the various CSs were balanced by an additional input called Z with a binary signal $F(Z)$ of fixed weight w_Z. The weight w_Z was constant and negative. Uttley (1976a) called the pathway from Z a *classifying* pathway. He explained: "The function of $F(Z)$ must be to signal whether the total stimulus to all the variable pathways is a member or non-member of some class" (p. 28). In other words, correlated stimuli tend to become negatively associated unless the particular stimulus possesses a preassigned significance that overrides the negative feedback at the synapses.

Uttley noted that his model is mathematically similar to the learning model of Rescorla and Wagner (1972), which is described in psychological rather than neural terms. It was noted in Chapter 3 that Rescorla and Wagner based their

theory on the general principle that learning only occurs when events violate expectations. Hence, both the pioneers discussed in this subsection developed models that anticipate significant aspects of conditioning models that are in contemporary use.

6.1.2. Rescorla and Wagner's Psychological Model

Rescorla and Wagner expressed their psychological principle as a system of difference equations. Their variables are CS–US associative strengths as defined in Hull (1943; see Section 2.1). Let the CS be labeled A, and the associative strength between A and the US be labeled w_A. Then the equation for the change in w_A over time is

$$\Delta w_A = \alpha_A \beta \left(w^{\max} - w_{AB} \right) \tag{6.1}$$

where α_A is the intensity of the CS; β is a learning rate associated with the given US; S refers to one or more stimuli present along with A; and $w^{\max}$ is the asymptotic value of associative strength, which is a function of the current reinforcement strength of the US ($w^{\max}$ is 0 if reinforcement by the US does not occur). The compound associative strength w_{AS} is assumed to equal $w_A + w_S$.

Rescorla and Wagner used their equations to explain blocking (see Figure 6.1). Their model predicts that if A is strongly connected with the US, such as shock, the associative strength between A and the US reaches its asymptote. In terms of Equation (6.1), while A alone is being conditioned to the US, $w_{AS} = w_A$ becomes very close to $w^{\max}$. If S consists only of a second CS, labeled B, the analog of (6.1) for B is

$$\Delta w_B = \alpha_B \beta \left(w^{\max} - w_{AB} \right) \tag{6.2}$$

Since at the time of presentation of the compound stimulus $AS = AB$, $w_{AB} = w_A$ is already nearly equal to the value of $w^{\max}$ for the given US, (6.2) says that Δw_B will be nearly 0; hence, B will not become significantly conditioned to the US.

The Rescorla–Wagner model is still widely used by psychologists because it explains many of the basic conditioning paradigms with a few simple equations. Yet, as noted before, it is not a genuine neural network model. Also, it is not a real-time model; changes in variables are all-or-none for each *trial* (presentation of one or more CSs followed by a US) and ignore temporal relationships within a trial. As will be seen later in this chapter, several other modelers incorporated insights of Rescorla and Wagner into real-time neural network models (some also inspired by invertebrate neurophysiological data). These modelers include Sutton and Barto (1981, 1990, 1998), Klopf (1982,

1988), Hawkins and Kandel (1984), Montague, Dayan, and Sejnowski (1996), Suri and Schultz (1998, 1999, 2001), and Dayan (2001).

6.1.3. Grossberg: Drive Representations and Synchronization

Rescorla and Wagner noted that blocking, and overshadowing as well, could also be explained by an alternative model. The alternative model is based on attentional competition: A more salient cue, or one that has already strengthened an association to the US, can receive more attention than a less salient cue and thereby inhibit learning of associations to the less salient cue. This section and the next include developments of this attentional interpretation of blocking in several network models. These models include qualitative network analyses by Grossberg (1975) and quantitative computer simulations by Grossberg and Levine (1987). These articles built on a theory of conditioning first developed in Grossberg (1971), which in some ways sharply contrasts with the Rescorla–Wagner theory.

Grossberg (1971) set out to provide a unified mechanism for both classical (Pavlovian) and operant (instrumental, Skinnerian) conditioning in the framework of his earlier articles on spatial pattern learning (Grossberg, 1968a, 1968b, 1969a, 1969b; see Chapter 3). His fundamental postulate was: after a time period wherein a CS repeatedly precedes a US, the CS must be able to generate a motor response previously associated with the US. But if the US is treated as a spatial pattern, a difficulty arises if the CS–US time lag is variable, as illustrated in Figure 6.2. Unless the network is carefully designed, the CS could become associated not with the US but with a noisy mixture of the US and other patterns experienced during the same time period (see Exercise 2(b) of Chapter 3).

Before discussing Grossberg's approach to the problem of variable time lags, also known as the *synchronization problem* (Grossberg & Levine, 1987), it is important to remark that his analysis assumed to a first approximation that the conditioned response (CR) is the same as, or similar to, the unconditioned response (UR). This outlook is known in psychology as *stimulus substitution theory*. Mackintosh (1983, pp. 67–74) discussed whether stimulus substitution theory is correct. The experimental literature on this point is quite varied: Sometimes the CR and UR are similar, and sometimes the CR involves only a small part of the usual responses to the unconditioned stimulus. As an example of the latter, "Pavlov's dogs salivated to the CS signalling food, they did not routinely lick, chew, bite, or swallow it" (Mackintosh, 1983, p. 70). Moreover, the same CS elicits a variety of orientation and approach responses not elicited by food itself. Yet, for modeling purposes, the idea that learning causes the CS to elicit a response previously made to the US has been a useful starting point.

Figure 6.3 shows a class of networks designed to address the synchronization problem. This network includes three sets of nodes. The first set, collectively called *S* for sensory, consists of short-term representations for particular sensory stimuli, both CSs and USs. These *S* nodes are connected into a competitive-cooperative subnetwork like those discussed in Chapter 4, leading both to short-term memory and to selective attention. The second set, collectively called *M* for motor, becomes activated after the US or the conditioned CS is presented, leading to particular responses that could either be skeletal (as in the rat's lever pressing) or autonomic (as in the dog's salivation). The third set of nodes, collectively called *A* for arousal, is of particular interest for later theory.

The synchronization problem was solved by having a strong preexisting connection from the US representation to a particular arousal locus, and then allowing repeated CS–US pairing to strengthen the ability of the CS to activate that same arousal locus. In this model, other patterns do not weaken the development of CS–US pairing unless they are associated with the same arousal locus; for example, intervening presentation of a sexual stimulus does not interfere with a bell–food pairing.

The *A* loci include representations of specific drives such as hunger, thirst, and sex. Activation of a given drive representation in this network requires a combination of internal drive level *and* compatible sensory stimuli; for example, the hunger representation is activated by a combination of hunger

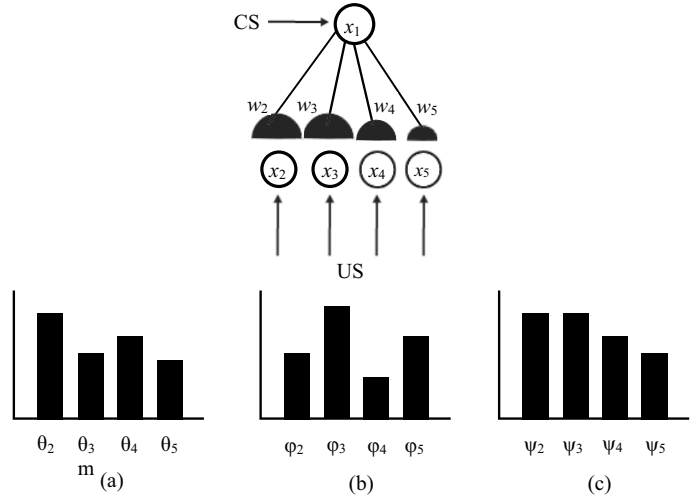


FIGURE 6.2 Schematic of difficulty arising, for example, in an outstar network, when CS–US time lag is variable. (a) Spatial pattern $\{\theta_i\}$ representing the US to be learned. (b) Spatial pattern $\{\Phi_i\}$ presented randomly at times between US presentations. (c) Noisy mixture $\{\psi_i\}$ of $\{\theta_i\}$ and $\{\Phi_i\}$, which is learned by the CS node x_1 (at the synapses w_i; see Exercise 2(b) of Chapter 3).

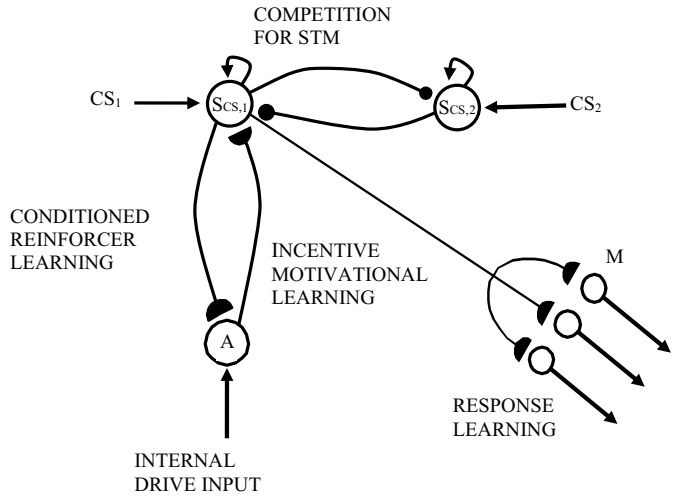


FIGURE 6.3 Schematic conditioning network. Conditioned stimuli (CS_i) activate sensory nodes ($S_{CS,i}$) that compete for short-term memory (STM) storage. Activated $S_{CS,i}$ send conditionable signals both to drive nodes (A) and to motor nodes (M). Conditioned reinforcer learning refers to S-to-A connections, whereby a CS repeatedly paired with a US becomes a secondary reinforcer. *Incentive motivational* learning refers to A-to-S connections (activated by internal drive combined with sensory inputs), which enhance approach to or avoidance of given stimuli.

Source: Reproduced from Grossberg & Levine, 1987, with permission of the Optical Society of America.

and the presence of food. Later articles (Grossberg, 1972a, 1972b), in order to model negative as well as positive reinforcement, expanded the concept to include "negative" arousal loci for aversive stimuli such as electric shock.

Drive representations that are separate from the sensory representations capture some of the functions of subcortical regions such as the hypothalamus and amygdala, which are explicitly considered in more recent models discussed in Section 6.3. Also, long before the development of quantitative models, Hebb (1955) argued that every sensory event has two different effects: its *cue* function, which selectively guides behavior, and its *arousal* function, which energizes behavior. This distinction corresponds in the network of Figure 6.3 to the distinction between the representation of the US or CS at S nodes and the effect of that representation, via fixed or modifiable synapses, on appropriate A nodes.

The existence of drive representations in neural networks is compatible with data on the reinforcement associated with brain stimulation. A long series of studies, starting with Olds (1955) and Olds and Milner (1954), showed that rats can learn to perform motor responses leading to stimulation of certain specific loci within the limbic system, hypothalamus, and midbrain. These results led to the notion that those brain sites are reward loci. Rats can also

learn to avoid stimulation of certain other loci within those brain areas, loci that are presumably associated with punishment.

The theory of Grossberg (1971) can also account for *secondary reinforcement* or *secondary conditioning*. This means that if a stimulus CS_1 is repeatedly followed by a US, until CS_1 evokes a conditioned response, then CS_1 can itself become a US for a new conditioned stimulus, say CS_2. In Grossberg's theory, this occurs because CS_1 acquires the ability to excite the same arousal source (see Figure 6.3) excited by the US. If CS_2 is repeatedly followed by CS_1, then CS_2 will come to evoke the same response.

6.1.4. Aversive Conditioning and Extinction

Thus far we have discussed conditioning that increases the likelihood of a specific behavior. Two conditioning paradigms that decrease the likelihood of a behavior are *aversive conditioning* and *extinction*. Aversive conditioning is the suppression of a particular motor behavior by punishment. For example, the frequency of a particular motor response can be reduced by pairing that response with electric shock. Extinction is the suppression of a motor behavior by nonoccurrence of an expected reward. For example, a response that had previously been paired with a reward such as food can be suppressed by frustration if the expected food is absent. The psychological fact that punishment and frustration have similar suppressive effects on behavior suggests that aversive conditioning and extinction might be described using similar neural mechanisms (Grossberg, 1982b, pp. 335–339).

Some controversy has existed among psychologists as to whether extinction is a passive or an active process, but the latter belief is currently more favored. If a dog learns to salivate in response to a bell after the bell has been repeatedly paired with food, the dog seems not to simply "forget" the salivation response if food is no longer given but rather to countercondition the bell to the aversive experience of frustration. That extinction is not simply passive forgetting, a return to a naive state, is suggested by the fact that reacquisition of a response by an extinguished animal is faster than initial acquisition by an untrained animal (Pavlov, 1927; Ricker & Bouton, 1996).

Hence, extinction is usually considered to occur as a consequence of the disconfirmation of an expectation of reward. This recalls the statement of Rescorla and Wagner (1972) that "organisms only learn when events violate expectations" (p. 75). That principle leads us to seek a general mechanism for processing disconfirmed expectations in the motivational domain, whether in a positive or negative direction. For an example of the latter, a motor response associated with disconfirmation of expected *punishment*, such as a lever press that unexpectedly turns off an ongoing electric shock, can become rewarding.

6.1.5. Differential Hebbian Theory versus Gated Dipole Theory

Processing disconfirmation of expectations involves comparing present with prior values of neural variables. Recall from Chapter 3 the discussion of two alternative methods for such a comparison. One of these methods is based on the gated dipole theory of Grossberg (1972a, 1972b), using habituation of repeatedly presented stimuli (see Figure 3.7). The other is based on the differential Hebbian learning theory (Kosko, 1986b; Klopf, 1986, 1988), using a rule whereby synapses change in strength as a function of cross-correlated *changes* in presynaptic and postsynaptic activities (see Figure 6.4).

Neither the differential Hebbian rule nor the gated dipole rule has yet been verified in actual nervous systems. Contreras-Vidal and Stelmach (1995), based on partial neurophysiological evidence, posited a gated dipole (opponent processing) network in parts of the basal ganglia. Dranias, Grossberg, and Bullock (2008) and John, Bullock, Zikopoulos, and Barbas (2013) cite possible evidence for opponent interactions in the hypothalamus and amygdala. Experimental tests of these two sets of rules are likely to be as much psychological as physiological, and to be based in part on their ability to be embedded in larger networks that perform interesting cognitive tasks. Gated dipoles have been used as components of larger networks that also include associative learning rules and on-center off-surround fields (e.g., Dranias et al., 2008; Grossberg, Bullock, & Dranias, 2008; Grossberg & Schmajuk, 1987; Levine,

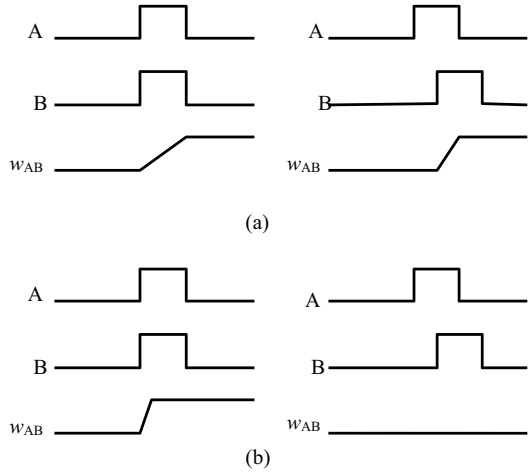


FIGURE 6.4 Schematic diagram of the distinction between (a) a Hebbian learning rule and (b) a differential Hebbian learning rule. In (a), the associative strength w_{AB} between A and B increases with simultaneous occurrence of inputs A and B. In (b), w_{AB} increases with simultaneous increases (e.g., onsets) of A and B. (Note, however, that the differential Hebbian model of Klopf, 1988, obtained learning with nonsimultaneous onsets by manipulating network time lags.)

2012; Levine & Prueitt, 1989; Ricart, 1992). Differential Hebbian learning rules evolved into the temporal difference rules that are widely used in models that partially explain responses of dopamine neurons to unexpected rewards or unexpected absences of rewards (e.g., Montague et al., 1996; Suri & Schultz, 1998, 1999, 2001). Both types of models are discussed later in this chapter.

An explicit version of the differential Hebbian rule (also known as the *drive–reinforcement* rule) was developed by Klopf (1988); equations for Klopf's model are shown at the end of this chapter. Klopf was led to such a rule by his earlier "hedonistic neuron" theory (Klopf, 1982) in which neurons themselves were goal-seeking devices.

Klopf's network simulates a wide variety of classical conditioning data. These data include blocking, secondary conditioning, extinction and reacquisition of an extinguished response, effects of interval between CS and US occurrences, effects of stimulus durations and amplitudes. (However, as seen from the equations given at the end of this chapter, the simulations of CS–US interval effects depend on some weighting factors for time delays, factors that were chosen specifically to match those data. Klopf did not suggest an underlying mechanism for generating those weighting factors.) Figure 6.5 shows the basic neuronal network of Klopf's model; note the similarity to Figure 3.5 from Sutton and Barto (1981).

Klopf's network also simulated *conditioned inhibition*. The conditioned inhibition paradigm consists of a first stage where a CS_1 is associated with a US and thereby conditioned to a CR, followed by a second stage where a combination of two stimuli CS_1 and CS_2 is associated with absence of the US. As a consequence, when CS_2 subsequently is associated with other stimuli that would otherwise evoke the CR, that CR is suppressed.

Further discussion of these alternative conditioning models will be placed in the context of data on the attentional modulation of conditioning. Such

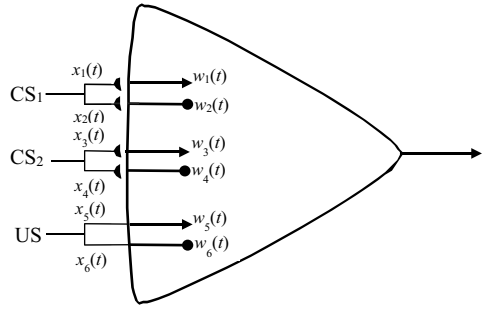


FIGURE 6.5 Schematic of the drive-reinforcement neuronal model. Each CS or US is represented by an excitatory and an inhibitory synapse. Synaptic weights are variable for synapses that mediate CSs and fixed for synapses that mediate USs.

Source: Adapted from Klopf, *Psychobiology*, *16*, 85–125, 1988. Permission to reproduce granted by Psychonomic Society, Inc.

attentional modulation, including the blocking paradigm discussed earlier and other, more complex multistimulus experiments, is the subject of the next section.

6.2. Attention and Short-Term Memory in Conditioning Models

Although neural network models of Pavlovian conditioning differ in their architectures, they share some common heuristic themes. Recall once more Rescorla and Wagner's principle that "organisms only learn when events violate expectations." Similarly, Grossberg (1975), discussing blocking and similar experiments, said that "learning subjects act as minimal adaptive predictors; they enlarge the set of cues that control their behavior only when the cues that presently control their behavior do not perfectly predict subsequent events" (p. 266).

These heuristics, however, have led different modelers to different conclusions, which in turn have different implications for the predictions of other experimental data. For example, Pearce and Hall (1980), developing a nonneural psychological model that is a refinement of Rescorla and Wagner's, stated that "stimuli that fully predict their consequences will be denied access to the processor. . . . A stimulus is likely to be processed to the extent that it is not an accurate predictor of its consequences" (p. 538). Grossberg (1982a, p. 530) argued, however, that Pearce and Hall's statement is violated by the fact that a US is an excellent predictor of its consequences and yet is almost always processed. By way of reconciliation, he proposed that there are separate, interacting systems for processing expected events and for processing unexpected events, and that the architectures of these two systems are different. In that article and elsewhere (e.g., Carpenter & Grossberg, 1987a; Grossberg, 1975; see Chapter 7), these two systems are called the *attentional* and *orienting* systems.

6.2.1. Grossberg's Approach to Attention

A theory for the structure of the attentional system was proposed in Grossberg (1975), incorporating the type of competitive mechanism discussed in Chapter 4 for the short-term storage of patterns. This attentional mechanism is based on the network of Figure 6.3, with the additions of competition between drive representations of different drives, and modifiability of feedback connections from drive to sensory representations. (Kilmer et al., 1969, had previously used competition between drive representations to model decisions between gross behavioral modes.) One version of the resulting, more complex network is shown in Figure 6.6.

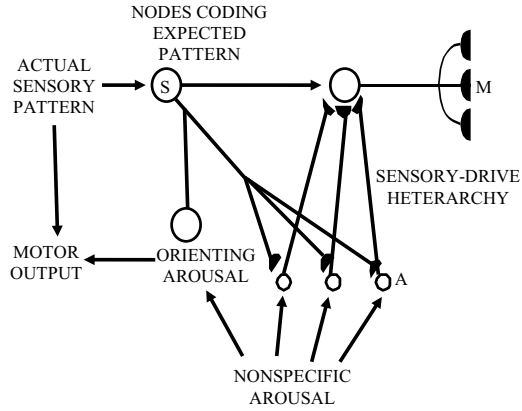


FIGURE 6.6 The sensory–drive heterarchy of Figure 4.22 (not shown in full) is embedded in a larger network with attentional and orienting subnetworks. An orienting ("what is it?") response is instinctive to all sensory cues, and is shut off only by expected cues; details of orienting system architecture are omitted. Expected cues are compatible with some drive, so activate the heterarchy. "*M*" stands collectively for motor responses that can be conditioned to drive-related stimuli; "*S*" and "*A*" are as in Figure 6.3.

Source: Adapted from Grossberg, 1975, with permission of Academic Press.

The network of Figure 6.6 is built on the sensory–drive heterarchy of Figure 4.20 (see the discussion in Section 4.4) in which each drive representation is activated by a combination of internal drive level and external sensory inputs compatible with the given drive. This enables the organism to focus attention on the particular cues that are compatible with whatever drives are relatively active at a given moment. The heterarchy is combined with an orienting system that causes motor responses to new cues in the environment, *unless* those cues are known to be associated with satisfaction of an active drive.

The attentional system in Figures 4.20 and 6.6 not only allows for the modeling of attentional effects in conditioning but also prevents cross-conditioning of stimuli compatible with one drive to irrelevant drives. A graphic example of the consequences of such cross-conditioning is given by Grossberg (1975). In his example, suppose you are eating roast turkey for dinner with your lover. Because you are repeatedly scanning both turkey and lover, it might be expected that each sensory cue would become associated in your mind with the drive compatible with the other cue; in fact, you might learn to want to have sex with turkeys and eat your lover. Grossberg showed heuristically how such an absurd outcome is prevented by competition both between sensory loci and between drive loci.

The network explanation of attention enabled Grossberg and Levine (1987) to simulate the blocking paradigm of Kamin (1969). Grossberg and Levine (1987) phrased the relevant modeling issues as follows:

> How does the pairing of CS_1 with US in the first phase of a blocking experiment endow the CS_1 cue with the properties of a conditioned, or secondary, reinforcer? How do the reinforcing properties of a cue . . . shift the focus of attention toward its own processing? How does the limited capacity of attentional resources arise, so that a shift of attention toward one set of cues . . . can prevent other cues . . . from being attended? How does withdrawal of attention from a cue prevent that cue from entering into new conditioned relationships?
>
> (p. 5016)

These questions can be summarized by asking how an organism predicts the environment so as to maximize (optimize) positive reinforcement and minimize negative reinforcement. This kind of prediction, in natural or artificial neural networks, is called *reinforcement learning* (Hinton, 1987; Werbos, 1988; see Section 5.2).

The first of Grossberg and Levine's "four questions" is answered by associative learning of a connection from the CS_1 representation, not to the US representation itself but to the drive representation. The second and third questions are answered by competition within the on-center off-surround subnetwork of sensory representations. Within that subnetwork, competition favors nodes corresponding to stimuli that have become conditioned reinforcers. Hence (with properly chosen parameters), the activities of other sensory nodes, such as the CS_2 node in the blocking paradigm, are suppressed, reducing the ability of those nodes to form conditioned associations. That effect causes CS_2 to be blocked, in answer to the fourth question.

6.2.2. Sutton and Barto's Approach: Blocking and Interstimulus Interval Effects

The attentional (i.e., lateral inhibitory) effects in the network of Figure 6.3 provide explanations for the blocking and overshadowing results of Kamin (1969). The model of Sutton and Barto (1981), and related models, have typically not addressed the question of which of two stimuli is more likely to attentionally overshadow the other. Barto and Sutton (1982), discussing their own model, stated: "The model clearly does not address higher order modulatory influences such as those produced by attentional or stimulus salience factors" (p. 232). Mackintosh (1975), while proposing a model related to that of Rescorla and Wagner, likewise noted that such models cannot account for the fact that a more salient stimulus can block a less salient one but not vice versa.

The simulation of blocking by Sutton and Barto (1981) is based on a different mechanism, relying on a special synaptic modification rule. In addition to blocking, Sutton and Barto simulated results on time intervals between stimuli

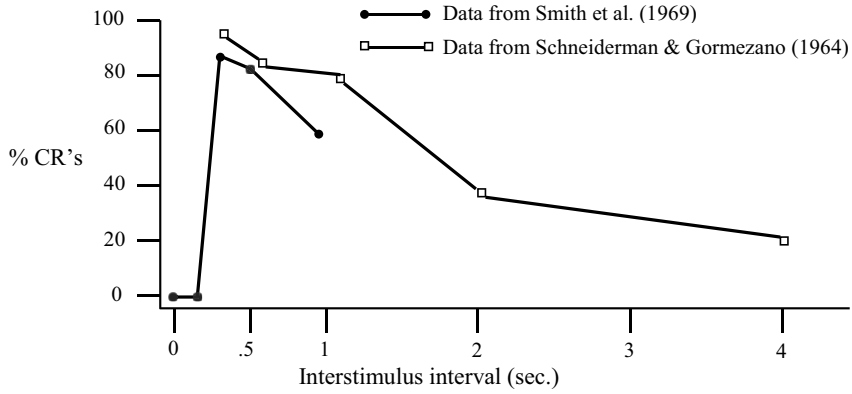


FIGURE 6.7 Asymptotic associative strength (measured as percentage of recall trials on which a conditioned response occurs) as a function of interstimulus interval in conditioning of the rabbit nictitating membrane response.

Source: Sutton & Barto, *Psychological Review*, *88*, 135–170, 1981. Copyright 1981 by the American Psychological Association. Reprinted by permission.

in classical conditioning. They noted that conditioning is typically strongest when the CS precedes the US, usually by 200 to 500 milliseconds, rather than when the CS and US are presented simultaneously. An example occurs in the rabbit nictitating membrane response (eyeblink conditioning) data of Schneiderman and Gormezano (1964) and Smith, Coleman, and Gormezano (1969), shown in Figure 6.7. Sutton and Barto used such data to argue against a "Hebbian" learning law in which actual presynaptic and postsynaptic activities are correlated. For with Hebbian learning, all other things being equal, the optimal *interstimulus interval* (ISI)[1] should be 0 rather than 200 to 500 milliseconds.

Sutton and Barto solved the problem of simulating the ISI effect by introducing the *eligibility traces* defined by Equations (3.22) of Chapter 3. They proposed that each conditioned stimulus, in addition to having a short-term memory trace x_i, has an additional trace which grows more slowly. As shown in Figure 6.8, the time course of this additional trace accounts qualitatively for the time of the optimal ISI. Sutton and Barto gave a single-cell interpretation of this trace, calling it an eligibility trace because it can be regarded as a chemical marker indicating how "eligible" the synapse is for modification. Yet, if Sutton and Barto's Equations (3.22) are compared with Grossberg's outstar equations (3.12–3.15), the eligibility traces $\bar{x}_i$ correspond quite closely to what Grossberg calls short-term memory traces. Sutton and Barto's short-term memory traces correspond to Grossberg's external inputs. This suggests that, in a larger network, the two sets of traces could plausibly be interpreted as being located in different brain areas, with stimulus traces being more peripheral and eligibility traces more central.

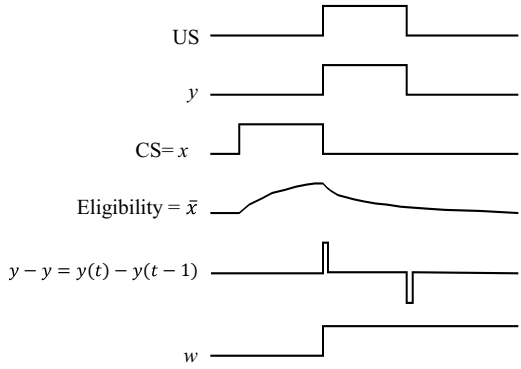


FIGURE 6.8 Time courses of variables in Equations (3.22) for a single trial in which a neutral CS ($w = 0$ at the start) is followed by a US.

Source: From Sutton & Barto, *Psychological Review*, *88*, 135–170, 1981. Copyright 1981 by the American Psychological Association. Adapted by permission.

Sutton and Barto explained blocking by means of a synaptic modification rule whereby presynaptic activity is correlated with change in postsynaptic activity; this rule is formally analogous to the rule of Rescorla and Wagner (1972). Because the introduction of the CS_2, in this model, leads to no increase in US node activity, the CS_2 does not become conditioned to the US. This change in postsynaptic activity is multiplied by presynaptic eligibility, as shown in Equations (3.22). Some results of simulating this network are shown in Figures 6.9 and 6.10.

Hawkins and Kandel (1984) proposed an explanation for blocking quite similar to those of Rescorla and Wagner and of Sutton and Barto. The Hawkins–Kandel model, unlike the others mentioned, is based on neurophysiological data from the sea slug *Aplysia* (see the discussion of Kandel & Tauc, 1965, in Chapter 3). These data (Hawkins, Abrams, Carew, & Kandel, 1983; Walters & Byrne, 1983) suggest that each US activates a given neuron, called a *facilitator neuron*, which influences pathways activated by CSs, and that the growth of associative strengths depends on simultaneous activation of a CS pathway and a facilitator neuron. Within this context, Hawkins and Kandel (1984) proposed that blocking is due to a kind of fatigue effect: "the output of the facilitator neurons decreases when they are stimulated continuously" (p. 385). Thus, after many trials in which a CS_1 is paired with a US, the fatigue of the US's facilitator neuron prevents that US from forming associations with another stimulus CS_2.

The model of Sutton and Barto (1981) was further elaborated in several later articles, particularly Blazis, Desmond, Moore, and Berthier (1986), Moore et al. (1986), and Sutton and Barto (1990). A main thrust of this later work was to fit the model to various quantitative details of the nictitating membrane

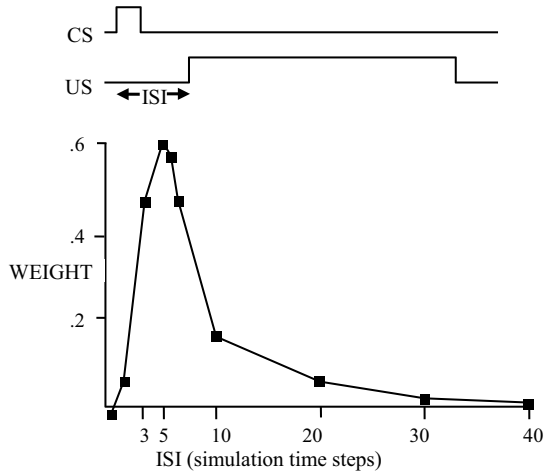


FIGURE 6.9 Asymptotic connection weight as a function of interstimulus interval in a simulation of Equations (3.22). CS is on for three time steps, US for 30. Trials last for 50 time steps.

Source: Sutton & Barto, *Psychological Review*, *88*, 135–170, 1981. Copyright 1981 by the American Psychological Association. Reprinted by permission.

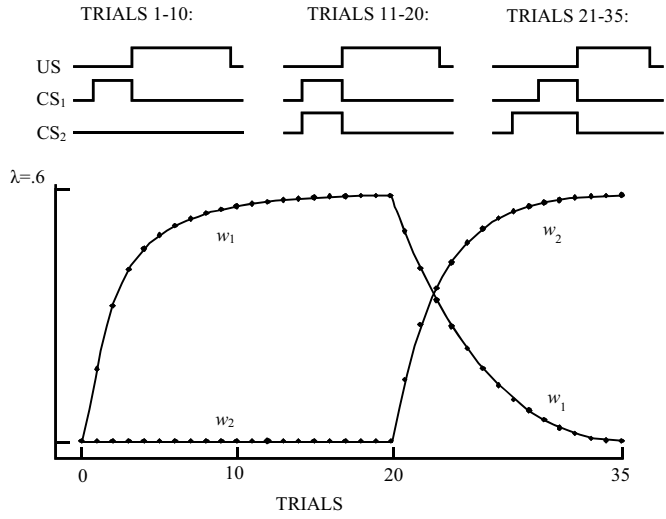


FIGURE 6.10 Blocking simulation. Connection weights w_1 of CS_1 and w_2 of CS_2 are shown at the end of each trial. Trials 0–10: input of CS_1 alone followed by US leads to increase in w_1. Trials 11–20: CS_1 and CS_2 presented together followed by US produces no change; that is, blocking occurs. Trials 21–35: CS_2 begins before CS_1. The output node responds to the earlier predictor and ignores the later.

Source: From Sutton & Barto, *Psychological Review*, *88*, 135–170, 1981. Copyright 1981 by the American Psychological Association. Adapted by permission.

response. The Sutton and Barto (1990) article was particularly influential because it developed the widely used *temporal difference* (*TD*) model. The temporal difference model, and the contact it has made over many years with data on the neurophysiology of dopamine neurons (see Section 5.2), are discussed in Section 6.3.

6.2.3. Some Contrasts between the Grossberg and Sutton–Barto Approaches

Grossberg and Levine (1987) suggested that the Hawkins–Kandel fatigue model for blocking failed to explain the complementary phenomenon of *unblocking* (Kamin, 1969). Unblocking means that if CS1 is first trained to a US, and then the compound stimulus CS1 + CS2 is associated with a US of a different level than before – for example, in the case of an electric shock US, the compound stimulus is associated with either a more intense or a less intense shock than is CS1 alone – then blocking of CS2 does not take place, and the animal conditions normally to CS2. CS2 is conditioned either to fear or relief, depending on whether the shock is more or less intense with CS1 + CS2 than with CS1 alone.

Grossberg and Levine (1987) also discussed the invertebrate data in terms of two interrelated psychological concepts: *conditioned reinforcement* and *incentive motivation*. Conditioned reinforcement means that a previously neutral stimulus becomes a reinforcer through learning. This is represented, for example, by the *S*-to-*A* pathways in Figure 6.3. Incentive motivation means there are incentives to approach or avoid particular stimuli based on their reinforcement value, represented by the A-to-*S* pathways in the same figure. Incentive is a concept that was developed in the 1950s and 1960s by psychologists studying drive (see Bolles, 1975, or Cofer, 1972) and has been defined as "an external event, object, condition, or stimulus in the environment that induces a state of arousal that energizes behavior" (Reeve, 1997, p. 32). The Grossberg–Levine article compared the facilitator neurons found in *Aplysia* by Walters and Byrne (1983) to incentive motivational pathways.

According to Grossberg's version of incentive motivation theory, secondary conditioning occurs because CS presentation, after conditioning has taken place, leads to an increase in *A*-to-*S* pathway activity; hence, the CS becomes a reinforcer in its own right. However, Walters and Byrne (1983) had not demonstrated an analogous increase in activity of facilitatory pathways in *Aplysia*. Hence, the exact mechanism for secondary conditioning in that species, if it does occur, is still in question.

Grossberg and Levine (1987) simulated blocking in the context of attentional competition between stimuli as suggested by Figure 6.3. In response to the objections to "Hebbian" learning by Sutton and Barto (1981), they also simulated the ISI data of Figure 6.7 in an attentional context. With CS and US

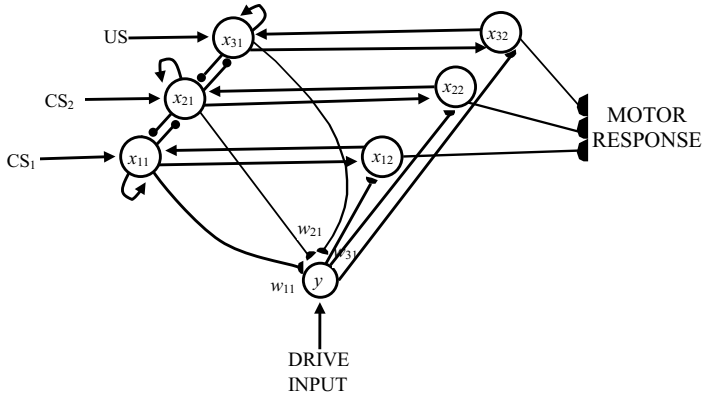


FIGURE 6.11 Network used to simulate blocking and ISI effects. Each CS or US sensory representation has two stages with STM activities x_{i1} and x_{i2}. Activation of x_{i1} generates unconditioned signals to x_{i2} and conditioned reinforcer signals to a drive node y. Conditioned incentive motivational signals from y activate the second sensory stage x_{i2}, which sends feedback to x_{i1}.

Source: Adapted from Grossberg & Levine, 1987, with permission of the Optical Society of America.

presented simultaneously, Grossberg and Levine stated, attentional competition occurs between those two stimuli, with a bias in favor of the US because it has a strong, unconditional association with the drive representation. Hence, the ISI phenomenon can be seen as a form of blocking, with the US taking the place of the CS_1 in the experiment of Kamin (1969), and the CS taking the place of Kamin's CS_2.

Figure 6.11 shows the actual network used by Grossberg and Levine (1987) to simulate the blocking and ISI effects; this network is an elaboration of the one shown in Figure 6.3. Note that the sensory representations in Figure 6.3 are now split into two parts. In later articles such as Grossberg and Seidman's (2006), the first stage is interpreted as sensory cortex and the second as prefrontal cortex, with drive representations interpreted as part of the amygdala.

Some results from simulation of this network are shown in Figure 6.12. These results verified the efficacy of a conditioning model based on associative learning combined with attentional competition, but left open some issues in timing (see Section 6.3.4). In order to achieve attentional blocking of the CS by the US when the two are presented simultaneously, the threshold for an *S* node to increase the efficacy of an *S*-to-*A* synapse has to be set higher than the threshold for the same *S* node to excite activity of the corresponding *A* node (and thereby of the *A*-to-*S* feedback pathway). The separation of conditions for change of synaptic efficacy from conditions for excitation of node activities is reminiscent of Sutton and Barto's separation of eligibility traces from stimulus traces.

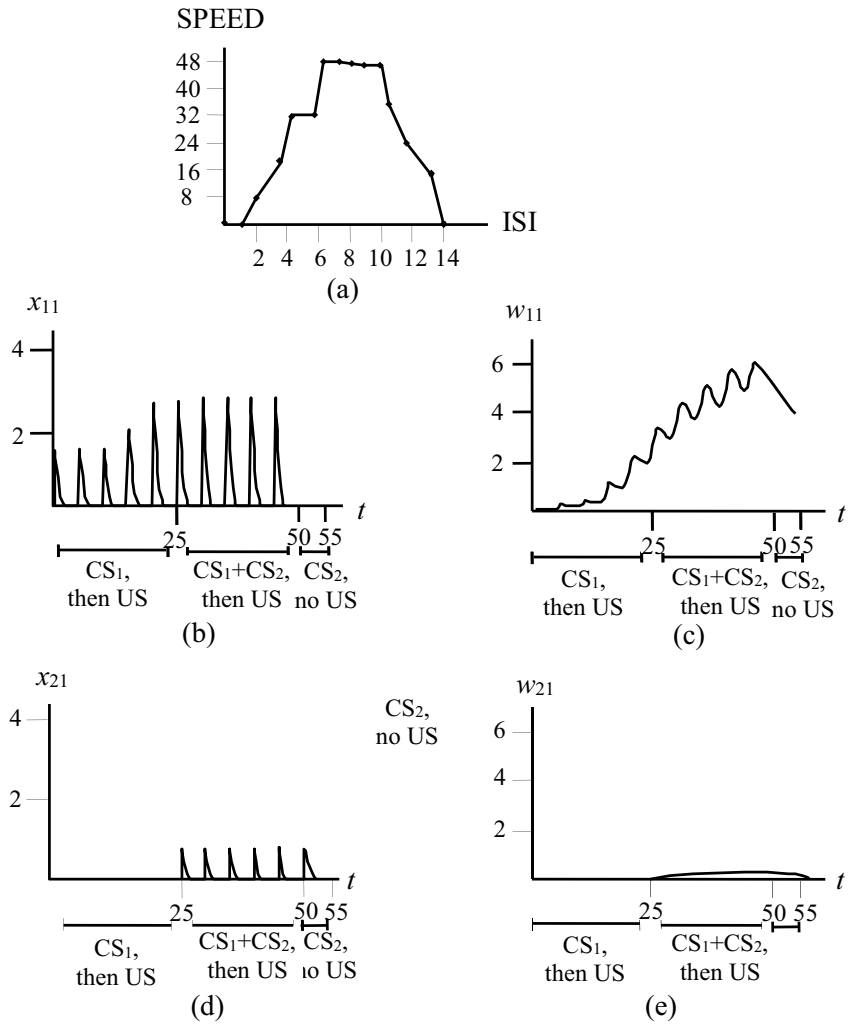


FIGURE 6.12 (a) Plot of R acquisition speed as a function of ISI, in the network of Figure 6.11. A CR is said to occur when x_{12} exceeds some threshold. Speed is computed by the formula 100 × (time units per trial)/(time units to first CR). (b)–(e) are plots of variables through time in a blocking simulation, with five trials (each 50 time units long) of CS_1–US pairing, five of (CS_1+CS_2)–US pairing, one of CS_2 presented alone, ISI = 6. (b) CS_1 STM trace x_{11}; (c) x_{11}-to-y LTM trace w_{11}; (d) CS_2 trace x_{21}; (e) x_{21}-to-y LTM trace w_{21}.

Source: Adapted from Grossberg & Levine, 1987, with permission of the Optical Society of America.

6.3. Computational Cognitive Neuroscience of Conditioning

Both of the streams of conditioning models discussed in the last section – the models by Sutton, Barto, and their colleagues and those by Grossberg and his colleagues – have undergone gradual development since the early 1990s under the influence of results from both invertebrate and vertebrate neurophysiology. In the 1990s there were several models that incorporated specific cellular data from invertebrate studies. More recent models have been more influenced by vertebrate studies, specifically by the studies of both dopamine neurons and nictitating membrane response discussed in Section 5.2.

6.3.1. Some Models Based on Invertebrate Neurophysiology

For the sake of brevity, this book has generally concentrated on modeling at the level of large aggregates of neurons, more than at the level of single neurons. Yet, classical conditioning is an area where the two levels of modeling have increasingly blended. Results on the interactions of electrical potentials, transmitters, and second messengers have been incorporated into many network models of associative learning. When competing models are equally plausible from the psychological viewpoint, neurophysiological data can provide additional constraints that facilitate choosing between models.

Sutton and Barto (1981) developed rough analogies between their eligibility variables and postsynaptic second messengers (calcium ion and cyclic AMP). Such analogies were developed further in the neuronal model for associative learning in Gingrich and Byrne (1987), based on study of conditioned withdrawal reflexes in *Aplysia*. Their model was built on a previous model of *nonassociative learning* (Gingrich & Byrne, 1985). Nonassociative learning is defined as the strengthening or weakening of certain neuronal pathways without dependency on contingent stimulation of other pathways.

Gingrich and Byrne (1987) modeled cell-level phenomena analogous to classical conditioning; these phenomena had been experimentally discovered by Walters and Byrne (1983). (Recall from Chapter 3 the cautionary notes about whether such single-cell mechanisms in fact approximate mechanisms responsible for conditioned behavior of whole organisms.) In the studies of Walters and Byrne (1983), shock to the tail was used as an aversive US, while the CS was direct electrical stimulation of any one of several sensory neurons responsive to stimulation of nontail skin areas. Stimulation of the nontail area became aversive as a result of learning. These authors proposed a mechanism for such cell-level conditioning whereby the US nonspecifically releases a chemical modulator that strengthens synaptic pathways from sensory neurons to output areas. They called this type of mechanism *activity-dependent neuromodulation*.

Gingrich and Byrne (1985) quantitatively simulated the roles of calcium ions and cyclic nucleotides (the commonest second messengers; see Chapter 3), along with enzymes regulating those substances, in nonassociative forms of learning. Gingrich and Byrne (1987) integrated these studies with the concept of activity-dependent neuromodulation, using equations that are discussed at the end of this chapter, to model associative learning.

The work of Gingrich and Byrne was extended by Buonomano, Baxter, and Byrne (1990) to a multineuronal network that included two sensory neurons and a facilitatory neuron, the latter corresponding to a source of reinforcement. These authors found that this neuronally based network exhibited a tradeoff in a crucial parameter. The network could simulate either second-order conditioning or blocking, depending on the value of this parameter, but not both phenomena with the same parameter value. A more recent network of Goel and Gelperin (2006) can simulate both second-order conditioning and blocking. Goel and Gelperin describe their model as based on the learning of the land mollusk *Limax*. Their model uses integrate-and-fire neurons (see Appendix 1 for a definition of that term), with modifiable connections between CS representations and both facilitator and motor neurons, but does not include the biochemical details of the earlier Gingrich and Buonomano models.

6.3.2. Temporal Difference Models

After Sutton and Barto (1981), several other conditioning models employed variations of the idea that reinforcement is based on the time derivative of combined US and CS associations. These time-derivative models included Gelperin, Hopfield, and Tank (1985); Hawkins and Kandel (1984); Klopf (1988); Kosko (1986a); Moore et al. (1986); and Tesauro (1986). Sutton and Barto (1987, 1990) noted several difficulties that time-derivative models had encountered in reproducing specific temporal data on conditioning of the nictitating membrane response. Some of the data involved the most typical conditioning paradigm, delay conditioning whereby CS onset precedes US onset (that is, ISI > 0) but CS offset is simultaneous with, or after, US onset. Other data involved *fixed-CS conditioning*, whereby CS duration is fixed and independent of ISI; this includes trace conditioning whereby CS offset precedes US onset, so that only a memory ("trace") of the CS is paired with the US. The remedies they developed led to the temporal difference (TD) framework, which is still making extensive contact with the neuroscience literature on dopamine cell responses, as well as being used in many engineering control applications.

The temporal difference model of Sutton and Barto (1990) is based on the idea, tracing back to Rescorla and Wagner (1972), that conditioning embodies an attempt to predict reward, that is, predict US arrival. Hence the amount of reinforcement is proportional to the difference between predicted and actual

US levels. Or, to put it another way, the reinforcement is proportional to the differences between predictions of reward at successive time intervals – which, when time is made continuous, puts it in the family of time-derivative models. Predictions of future rewards are reduced by a factor that measures how much the future is discounted relative to the present. As in the earlier Sutton–Barto model, reinforcement is then multiplied by the eligibility trace of a particular CS to obtain an increment of the associative strength of that CS with the US. One version of the TD model network is shown in Figure 6.13.

When the temporal difference model was first developed, the reward prediction error of the TD model was not identified with a particular area of the brain. A brain locus for prediction error emerged in the 1990s with the results of Wolfram Schultz and his colleagues, discussed in Section 5.2, on the

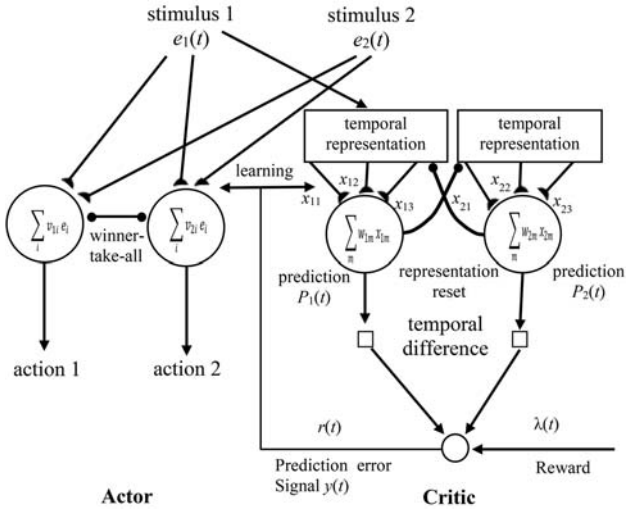


FIGURE 6.13 A version of the TD network. Actor and Critic receive input stimuli 1 and 2 which are coded as functions of time. The Critic computes the effective reinforcement signal $r(t)$ which modifies the weights v_{nl} of the Critic and the weights w_{lm} of the Actor. The Critic associates input stimuli 1 and 2 with Signal $r(t)$. Every stimulus l is represented as a series of components x_{lm} of different durations, each of which influence the reward prediction signal according to its own adaptive weight w_{lm}. The prediction $P_1(t)$ is computed as the weighted sum of these components. Winner-take-all competition between predictions $P_1(t)$ of different stimuli sets all but one representational component to zero. The change in the prediction of a stimulus l is computed by taking the temporal difference between successive time steps. The temporal difference is summed over all stimuli and added to the primary reinforcement signal input $\gamma(t)$ coding the reward, leading to the effective reinforcement signal which codes reward prediction error. The Actor learns to associate stimuli with behavioral actions.

Source: Adapted from Suri and Schultz (1999) with the permission of Elsevier Science, Inc.

responses of dopamine neurons in the midbrain during conditioning episodes. In particular, recall that in the early stages of conditioning the dopamine cells have a burst of activity to the US, but after the CS has become a reliable predictor of the US the dopamine burst occurs to the CS but not to the US. Also recall that there is a dip in dopamine cell activity if an expected US does not arrive. Montague, Dayan, and Sejnowski (1996) used a variant of the TD model to simulate data of Ljungberg et al. (1992) on dopamine cell responses on a reaction time task and of Schultz et al. (1993) on several spatial tasks, the most complex of which involved a delayed response.

The Sutton and Barto (1990) and Montague et al. (1996) versions of TD did not include specific roles for relevant brain regions such as the dopamine nuclei of the midbrain, basal ganglia, and prefrontal cortex. More explicit simulations of brain regions were gradually included in later extensions of the TD model, starting with Suri and Schultz (1998, 1999, 2001). The simulations using TD in the Suri–Schultz articles encompassed sequence learning, delayed response, and anticipatory dopamine neuron activity. The 1999 article noted that previous versions of the TD model had predicted that dopamine cell activity would be depressed not only if reward is delivered later than expected, which is supported by data, but also if reward is delivered *earlier* than expected, which is not supported by data. Suri and Schultz (1999) prevented this anomalous occurrence by adding to the TD network lateral inhibition between traces of different events, including the CS and US, as had been present in the differently structured network of Grossberg and Levine (1987).

The dopamine (DA) neurons have strong reciprocal connections with the striatum, the input layer of the basal ganglia. The striatum is divided into regions called *striosomes* that are surrounded by other areas called the *matrix* (Graybiel, 1991). Striosomes and matrix are different in their biochemistry and in their connection patterns, so different roles for the two should be expected. Several research groups, starting with Houk, Adams, and Barto (1995), have proposed functional roles for DA–striosome connections that exhibit striking analogies with the TD model.

In the model of Houk et al. (1995), the striosomes integrate inputs from three sources: convergent pattern related input from the cerebral cortex; excitatory input from the DA neurons; and signals from the lateral hypothalamus related to primary reinforcement.[2] Each striosomal module in turn inhibits the corresponding DA neuron via a direct pathway and also excites the DA neuron via an indirect pathway ("sideloop") through the subthalamic nucleus. The sideloop is responsible in the model for the DA neuron's response to a CS, and the inhibitory direct pathway to the neuron's subsequent nonresponse to the US.

Further insights into TD and the basal ganglia were developed in a number of later articles. In particular, Joel, Niv, and Ruppin (2002), Niv, Duff, and Dayan (2005), and Niv (2009) integrated neuroscientific data with the

actor/critic framework (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998). In the actor/critic framework, one module, called the *adaptive critic*, sends positive or negative reinforcement prediction errors to another module, called the *actor*. The actor is unaware of which of its actions contributed to the error, so it must modify its action plans (*policies*) in response to the error signal alone.

As to which areas of the brain correspond to the actor and critic, Joel et al. (2002) point to a resemblance between DA neuron activities and the temporal difference prediction error signal in the critic. The standard actor–critic model also includes a prediction error signal in the actor, which Joel et al. (2002) identified with DA-dependent long-term plasticity in the striatum. Niv et al. (2009, p. 145) pointed to evidence that the two main dopaminergic midbrain nuclei are each involved in different parts of actor–critic operations. Ventral tegmentum affects ventral striatum and prefrontal in the critic, whereas substantia nigra affects dorsal striatum in the actor.

More recent articles by TD-oriented modelers have tended to concentrate on details of reinforcement learning in the basal ganglia without integrating these details into simulations of conditioning data (e.g., Diuk, Tsai, Wallis, Botvinick, & Niv, 2013), or else developed models of conditioning data based on statistical optimality considerations not closely tied to neuroscience (e.g., Gershman & Niv, 2012). The TD approach has become popular because it is built around a single and easily understandable teleological principle, namely maximization of predicted future reward. Yet its very simplicity, with the implication that there is a unique locus in the brain that controls Pavlovian learning, limits the predictive applicability of this approach unless it is extended to incorporate principles that suggest roles for regions such as amygdala and prefrontal cortex. One of the first attempts to model conditioning using extended TD is the work of Dayan (2001), which combines TD with some ideas inspired by optimal control.

Another limitation of the TD modeling approach, which could possibly be overcome by proper setting of a network parameter, was found by many researchers including Brown et al. (1999) and Pan, Schmidt, Wickens, and Hyland (2005). The TD models, the authors said, predicted that, as conditioning takes place, dopamine bursts should gradually propagate backwards in time at intermediate steps from the arrival of the US to that of the CS. Instead, the data show that dopamine bursts occur only at the two times at the beginning and end of the CS–US interval, and never at intermediate times. Pan et al. (2005) showed how this gradual back propagation of the dopamine signal could be eliminated by setting the eligibility decay parameter (δ in Equation (6.5) below) close to 0, indicating that CS value, indicating that CS value learning could be influenced by previous CS presentations over a substantial time period.

6.3.3. Another Approach to Reinforcement Learning Via the Cog/EM Model

A different approach to reinforcement learning and complex conditioning data has been developed in a series of articles by Bullock, Grossberg, and their colleagues (Brown, Bullock, & Grossberg, 1999; Dranias, Grossberg, & Bullock, 2008; Grossberg, Bullock, & Dranias, 2008; Tan & Bullock, 2008b). This approach could be considered an alternative to the TD formulation but incorporates many of the same principles that motivated the TD approach, namely the general notion of reward prediction error and the implications of the data of Schultz and his colleagues on dopamine neurons and their connections with basal ganglia and prefrontal cortex.

The foundations of the Bullock–Grossberg alternative to temporal difference are based on the gated dipole model of opponent processing (see Section 3.3) and on some earlier models that capture data on adaptive timing of conditioned responses such as the nictitating membrane response. The reinforcement learning models are described in Section 6.3.5; first, the next section reviews some of those earlier foundational models (Grossberg & Schmajuk, 1987, 1989). Grossberg and Schmajuk (1987) modeled the emotionally positive and negative sides of conditioning, and in Grossberg and Schmajuk (1989) a collection of neurons with a range of time delays was involved in timing a conditioned response. A series of subsequent models related their insights to cerebellar and hippocampal data (Bullock, Fiala, & Grossberg, 1994; Fiala, Grossberg, & Bullock, 1996; Grossberg & Merrill, 1992, 1996; see also Gluck & Myers, 1993, and Myers & Gluck, 1994, for other models relating hippocampus to eyeblink conditioning). This interrelated set of conditioning models has more recently been given the name *CogEM* for "cognitive, emotional, and motor" (e.g., Grossberg & Seidman, 2006).

6.3.4. Gated Dipoles, Aversive Conditioning, and Timing

Grossberg and Schmajuk (1987) continued the work of Grossberg and Levine (1987) on quantitative study of attentional effects in conditioning. These authors included simulations of aversive conditioning via negative reinforcement as well as *appetitive* conditioning (the opposite of aversive) via positive reinforcement. To attentional networks such as appear in Figures 6.6 and 6.11 they added the gated dipole mechanism schematized in Figure 3.7. A loop was added to the gated dipole to allow for secondary inhibitory conditioning (see Grossberg, 1975, for an explanation of why this was needed.) As described in Grossberg and Schmajuk (1987):

> Secondary inhibitory conditioning consists of two phases. In Phase 1 CS_1 becomes an excitatory conditioned reinforcer (e.g., source of conditioned

> fear) by being paired with a US (e.g., a shock). In Phase 2, the offset of CS_1 can generate an off-response which can condition a subsequent CS_2 to become an inhibitory conditioned reinforcer (e.g., source of relief). (p. 197)

All of these considerations led Grossberg and Schmajuk (1987) to design a network called *READ* (*recurrent associative dipole*), as shown in Figure 6.14. The equations for the READ circuit, combining the associative learning and gated dipole equations from Chapter 3, are listed at the end of this chapter. For an appropriate range of parameter values, the network can simulate both primary and secondary excitatory and inhibitory conditioning.

Grossberg and Schmajuk (1987) went on to discuss qualitatively how their network can model extinction and conditioned inhibition. Extinction is treated as an active, not a passive, process, resulting from conditioning a CS to the "off" channel of Figure 6.14 if an expected CS does not occur. This effect is not simulated explicitly in the Grossberg–Schmajuk article because it involves interaction of the READ network with another network that can measure the degree of match or mismatch of an actual with an expected stimulus. Such a match-sensitive network, based on *adaptive resonance theory* (*ART*), is

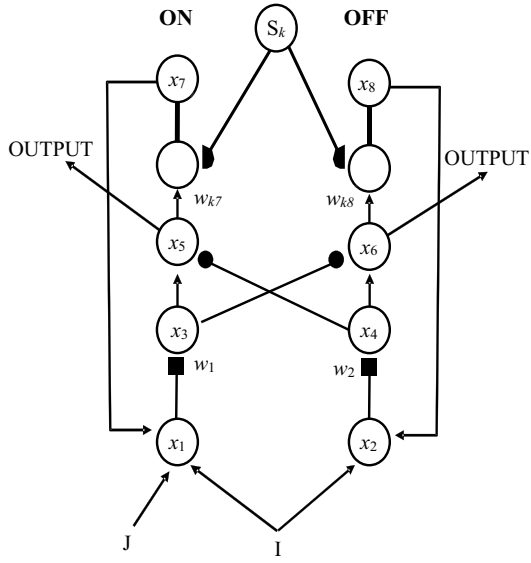


FIGURE 6.14 A READ (recurrent associative dipole) network, joining a recurrent gated dipole with associative learning. Learning occurs at the synapses w_{k7} and w_{k8}, from the sensory nodes S_k to the on-channel and off-channel, respectively, of the motivational gated dipole.

Source: Adapted from Grossberg and Schmajuk, *Psychobiology*, *15*, 195–240, 1987. Permission to reproduce granted by Psychonomic Society, Inc.

discussed in Chapters 7 and 8, within the context of categorization and coding models. The ART network combines mechanisms of association, competition, and opponent processing with an additional design principle for adaptive interactions between network levels.

The READ model of conditioning was extended in Grossberg and Schmajuk (1989) to include a mechanism for timing. Accurate timing is very important for conditioning because suppose an animal has repeatedly received food reinforcement at some given time interval after a particular CS. Then the animal needs to know *when* to expect the food so that it will not get frustrated by the expected) nonoccurrence of food before that time interval has elapsed. Grossberg and Schmajuk's technique for accomplishing this timing function consists of a network with a large number (80 in their first simulation) of gated dipoles, each becoming activated and habituated at a different rate. The authors called this device *spectral timing* because it includes a "spectrum" of possible activation rates, thereby enabling the network to learn to expect stimuli or perform responses at specific time delays after the CS.

The article of Grossberg and Schmajuk (1989) was one of the first to relate conditioning to high-order cognitive processes such as categorization. The theory described therein includes an orienting response to a mismatch between expected and observed stimuli, which is part of the ART categorization network (Carpenter & Grossberg, 1987a, 1987b; see Section 7.2). Also, the spectral timing network includes hypotheses about the roles in reinforcement of the hippocampus and two common neurotransmitters, acetylcholine and dopamine.

A series of later articles (Bullock, Fiala, & Grossberg, 1994; Fiala, Grossberg, & Bullock, 1996; Grossberg & Merrill, 1992, 1996) discuss details of the neuroscience of both the hippocampus and cerebellum that are compatible with the spectral timing hypothesis. These articles focus specifically on modeling neural interactions in the nictitating membrane response. The hippocampus is involved in the timing of when the US is expected to arrive, which makes it important for reinforcement learning and attention to motivationally relevant cues, whereas the cerebellum is involved in the timing of the response itself (Thompson et al., 1984, 1987; Section 5.2.4 of this book). Motor timing is required because without it the animal could respond prematurely when it is expecting a US.

Grossberg and Merrill (1996) highlighted the similarity between the cerebellar and hippocampal timing mechanisms despite their different roles in the conditioning process:

> Both [cerebellar and hippocampal mechanisms] control an inhibitory gate that modulates another learning process, and both occur on dendrites whose summer output across a spectrum of rate-sensitive cell sites determines the collective timed response.
>
> (p. 272)

They suggest that the two systems could undergo similar cellular events (at the hippocampal dentate cells and the cerebellar Purkinje cells) during conditioning. The article by Fiala et al. (1996) discusses much of the detailed biochemistry of the cerebellar system including the Purkinje cell LTD induced by conditioning.

The recent article of Franklin and Grossberg (2017) goes much further in developing the role of the hippocampus in conditioning and in memory consolidation, and the interactions of hippocampus with amygdala, thalamus, and neocortex in those processes. The model of that article, known as *nSTART*, simulates a variety of data on the differential effects of lesions to all these areas at different phases of learning. Among these data are results showing that hippocampus is needed for trace conditioning but not for delay conditioning, and that hippocampal lesions during trace conditioning impair recent but not temporally remote learning. As for amygdala, the model simulates results showing that amygdalar lesions made before or just after training, but not those made later, show down conditioning. It also simulates data on thalamic, sensory cortical, and orbitofrontal lesions.

The Franklin–Grossberg model is built on the earlier Grossberg–Schmajuk and Grossberg–Merrill models involving spectral timing in the hippocampus. Section 9.2 of Chapter 9 discusses how the nSTART model posits that the hippocampus and cortex learn different things, in contrast with other models positing a unitary memory trace that is first learned by the hippocampus and then transferred to the cortex (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2000). Franklin and Grossberg (2017) discuss how spectral timing is part of an overall theory of entorhinal–hippocampal interactions encoding context that could be both spatial and temporal (see, e.g., Pilly & Grossberg, 2012, for the spatial part of the theory).

6.3.5. Modeling Dopaminergic Involvement in Conditioning

Brown, Bullock, and Grossberg (1999) modeled the Schultz laboratory's data on dopamine cell responses using an adaptive timing mechanism in yet another brain region, the striatum (see Figure 6.1). The TD models of Montague et al. (1996) and Suri and Schultz (1999) also include effects similar to adaptive timing.

The model of Brown et al. (1999) includes a reward prediction error, but differs from TD models in that the excitatory and inhibitory components of the prediction error are spatially separated. Their model is based on two parallel learning pathways from the limbic cortex to the dopaminergic nucleus known as substantia nigra pars compacta (SNc), both traversing through the striatum. One pathway, through the ventral striatum (matrix cells), ventral pallidum, and PPTN, is devoted to excitatory conditioning. The other pathway, through the striosomes, is devoted to adaptively timed inhibitory conditioning.

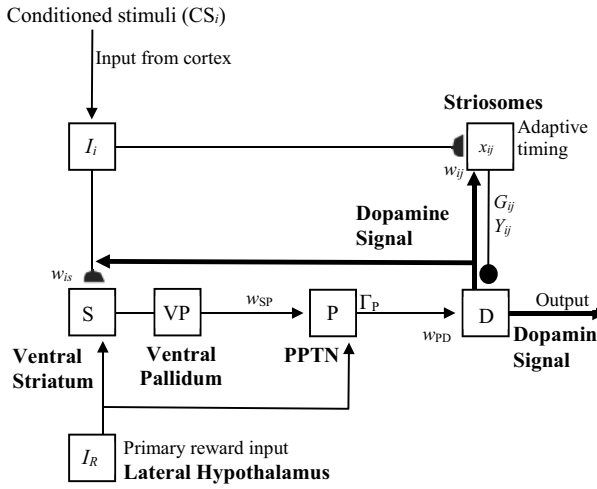


FIGURE 6.15 Network of Brown et al. (1999). Cortical inputs (I_i) excited by CSs learn to excite the SNc (D) via the path through ventral striatum, ventral pallidum, and pedunculopontine tegmental nucleus (PPTN). When PPTN activity exceeds a threshold Γ_P, it excites the dopamine cell. Striosomes learn to generate adaptively timed signals that inhibit the dopamine cell. Primary reward signals from the lateral hypothalamus both excite the PPTN and act as training signals to the ventral striatum.

Source: Adapted from Brown, Bullock, & Grossberg, 1999, with the permission of the Society for Neuroscience.

The excitatory pathway is what generates dopamine bursts in response to predictable reward-related signals and causes dopamine dips when expected rewards are not received. The inhibitory learning uses an intercellular spectrum of timed responses that is similar to timing mechanisms in the hippocampus and cerebellum.

The adaptive timing function of the inhibitory striosomes is implemented by metabotropic glutamate receptor-mediated Ca^{++} spikes that occur with different delays in striosomal cells (a mechanism also used by Fiala et al., 1996, to model the timing function of the cerebellum in the nictitating membrane response). Brown et al. (1999) claimed that implementing reward prediction error by the combination of different excitatory and inhibitory loci avoided a prediction of TD models that did not fit the data. This was the prediction mentioned in Section 6.3.2 that dopamine bursts should gradually propagate backwards in time at intermediate steps from the arrival of the US to that of the CS.

Another point of difference between the two sets of models concerns the responses of dopamine neurons to the uncertainty of reward. Recall from Section 5.2 the results of Fiorillo et al. (2003, 2005) showing that the same dopamine neurons that respond phasically to unexpected rewards also exhibit

what those authors called an *uncertainty response*. That is, on conditioning paradigms wherein the CS is only sometimes rewarded, these neurons show a sustained response that is maximal when the probability of reward is about one half. Niv et al. (2005) explained the uncertainty response as an artifact of averaging across trials based on back propagation of dopamine responses through time. Yet, Tan and Bullock (2008a) noted that such back propagation does not actually occur, and moreover that the uncertainty response occurs robustly on single trials. Tan and Bullock found an alternative explanation for the uncertainty response, based on co-release by SNc-to-striosome synapses of the common inhibitory transmitter GABA and a neuropeptide called *substance P*.

6.3.6. Models of Fear Conditioning

The models discussed in the last three sections have either simulated data on appetitive, reward-based conditioning or on conditioning in general. Yet, fear is the emotion whose neural basis has so far been studied the most thoroughly (e.g., LeDoux, 2000; see Section 5.2). Hence it is natural that there has been significant modeling of the involvement of brain regions such as amygdala, hippocampus, thalamus, and prefrontal cortex in fear conditioning (e.g., Armony, 2005; Armony, Servan-Schreiber, Cohen, & LeDoux, 1995, 1997; John et al., 2013; Moustafa et al., 2013; Vlachos, Herry, Lüthi, Aertsen, & Kumar, 2011). Most of these models have focused on auditory fear conditioning, for which the most data is available.

The model of Armony and his colleagues includes both cortical and subcortical influences on the amygdala, both of whose weights are modified by fear conditioning (that is, by pairing of an auditory CS with a noxious US). As reviewed in LeDoux (1996), there are parallel cortical and subcortical (thalamic) pathways that reach the primary emotional processing areas of the amygdala. The thalamic pathway is faster than the cortical, but the cortex performs finer stimulus discrimination than does the thalamus. This suggests that the two pathways perform complementary functions: the thalamic pathway is the primary herald of the presence of potentially dangerous stimuli, and the cortical pathway performs more detailed evaluations of those stimuli.

Armony et al. (1995, 1997) reproduced many detailed experimental results on the responses of single neurons in different brain regions during fear conditioning experiments. In a follow-up article by Armony, Servan-Schreiber, Romanski, Cohen, and LeDoux (1997) they noted one surprising prediction of the model that later was experimentally verified. Because of the poor discriminative ability of neurons in the auditory thalamus compared with neurons in the auditory cortex, it had been expected that auditory cortex lesions would cause the learned fear responses to generalize more broadly to auditory stimuli other than the original CS. Yet, when the auditory cortex of

the model network was "lesioned," the network's generalization properties were unaffected. This surprising lack of effect on generalization was later confirmed by behavioral studies in lesioned rats. Armony et al.'s explanation of this finding was that, even if individual thalamic neurons are poor discriminators, the total population of thalamic neurons can discriminate among stimuli with sufficient accuracy to compensate for the loss of the cortex.

The idea of fast subcortical and slow cortical signals is popular in the neuroscience community but not universally embraced. Pessoa and Adolphs (2010) argued that, in the visual system, there is no distinction of fast and slow pathways and emotional stimuli do not have a privileged access relative to other types of stimuli. These authors suggested that the fast/slow dichotomy could be specific to audition, particularly in rodents.

The Armony et al. models mainly involve amygdala (treated as a unit), auditory thalamus, and auditory cortex, with conditioning affecting the frequency tuning curves of some cortical neurons. Those articles discuss the differentiation of amygdala as stimuli progress from the lateral to the basal to the central nuclei of the amygdala but the model does not differentiate functionally between amygdalar subregions. Nor do these articles explicitly include the roles of hippocampus, which sets the context, or medial prefrontal cortex. The roles of all these regions in fear conditioning are considered in later models by other researchers.

Moustafa et al. (2013) explicitly considered roles for amygdalar subregions, hippocampus, and medial prefrontal cortex in fear conditioning. Their model is based on the TD concept of prediction error. The central nucleus of the amygdala, the part with direct connections to the hypothalamus and autonomic nervous system, learns two types of fear responses (heart rate changes and freezing) in response to stimuli that have been paired with shock. The CR learning at the central nucleus is modulated by two separate signals. One of the signals is from basal and lateral amygdala and denotes a positive prediction error. The other signal is from ventromedial prefrontal cortex (vmPFC) via *intercalated cells*, which are GABAergic inhibitory neurons that synapse on the central nucleus. The second signal denotes a negative prediction error and thereby is necessary for extinction of the fear response. The hippocampal input to both vmPFC and basal and lateral amygdala is necessary for context-specificity of both acquisition and extinction of the fear responses. Unlike the Armony models, the model of Moustafa et al. does not include sensory cortex or thalamus, and therefore does not deal with the altering of receptive fields by conditioning.

Vlachos et al. (2011) took a different approach to fear conditioning, focusing on one brain region – the basal nucleus of the amygdala (BA) – and presenting a neurophysiologically detailed model of its functions in conditioning, using both excitatory and inhibitory leaky-integrate-and-fire neurons. The excitatory BA neurons integrate CS input from the lateral amygdala and contextual input

from the hippocampus or medial prefrontal cortex. Learning occurs with temporal coincidence of CS and context inputs. From the simulations emerge two distinct neuron populations relating to fear conditioning and extinction, and those populations in turn influence the output in the central amygdala.

The model of John et al. (2013) relies heavily on prefrontal connections to the amygdala intercalated cells, like that of Moustafa et al. (2013). In addition, John et al. distinguished two types of intercalated cells that they called ITCd (for dorsal) and ITCv (for ventral); the lateral amygdala excites ITCd, which in turn inhibits ITCv, and ITCv suppresses fear responses at the central nucleus. The BA excites both ITCv and the central nucleus, which is always the location of CS learning and has properties similar to the striatum (including a predominance of GABAergic inhibitory cells).

A noteworthy feature of the John et al. model is that cortical inputs can influence the relative strength of activation of the two sets of intercalated cells and thereby influence the balance between tendencies toward fear learning and fear extinction. In an extreme ("cautious") case, if the activation of the ITCd side is above some threshold, the network actually becomes incapable of extinction of a fear response to a CS once that response has been learned. In the opposite case there is a tendency toward relatively fast extinction when contingencies change.

6.4. Multiple Levels: Model-Free and Model-Based Learning

In the first half of the twentieth century, ideas about animal learning were split between behaviorist and cognitive notions. The behaviorist outlook (Thorndike, 1933) treated animals, including humans, as relatively passive learners who rely on strengthening responses via associations with rewards. By contrast, the cognitive outlook (Tolman, 1948) treated animals as active goal seekers who form mental cognitive maps of where rewards could be found.

The consensus among current psychologists is that the brain is involved in both passive and active forms of learning. Active learning is often called *model-based* because it relies on constructing a mental model of the environment (Daw, Niv, & Dayan, 2005; Doya, 1999; Doya, Samejima, Katagiri, & Kawato, 2002; Sutton & Barto, 1998). In the same vein, passive learning is called *model-free*.

Daw et al. (2005) noted that the TD model mainly reproduces model-free phenomena. The same is true of some of the other well-established conditioning models discussed in this chapter, such as READ. The MOTIVATOR and PVLV models discussed in the next two subsections address a mixture of model-based and model-free processes. Notably, the neural substrates of those two models include parts of the prefrontal cortex in addition to subcortical areas.

Daw et al. (2005) proposed that the prefrontal cortex and dorsolateral striatum are control loci for model-based and model-free systems respectively. In later versions (e.g., Dayan, 2007, 2009; Dayan & Berridge, 2014), the model-based controller also includes dorsomedial parts of the striatum. These researchers developed a formal model of these two processes, without assigning parts of the network explicitly to brain regions. Mannella, Gurney, and Baldassarre (2013) developed a more explicit neural formulation of the interaction between two processes, with the nucleus accumbens (ventral striatum) playing a key role as the arbiter of value.

Model-based learning uses a *tree search* that goes through and compares the consequences of different, often multistep, courses of action. Model-free learning, on the other hand, uses what the authors call *caching*, which involves a learned value of an action regardless of the source of that learning. Often the early learning of a new task is model-based but when the task is well learned it becomes model-free: a response is made out of habit. Figure 6.16 shows a schematic interaction between the tree search and caching subsystems as

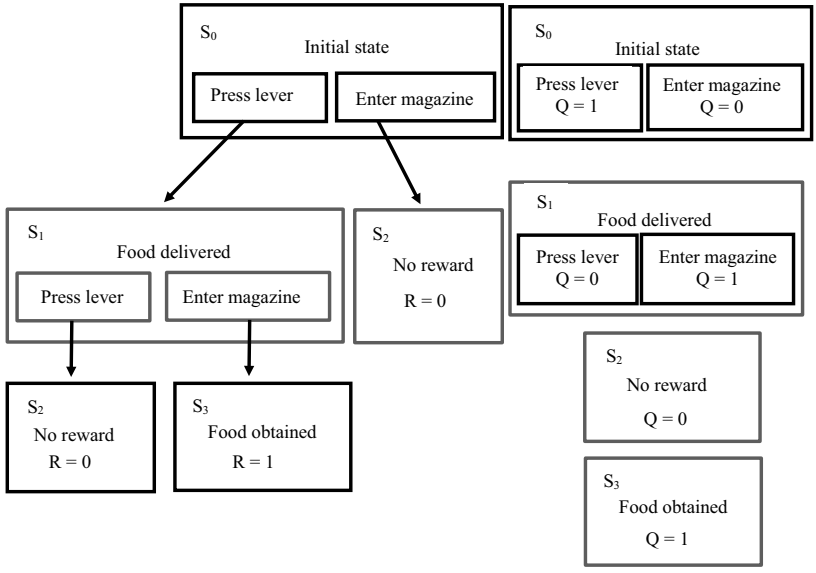


FIGURE 6.16 Representation of a two-stage instrumental conditioning task by (a) a tree search system (model-based) and (b) a cache system (model-free). S_0, S_1, S_2, and S_3 are the four possible states within the task. In (a), R represents whether the reward was obtained. In (b), Q represents the future value of each action, and therefore the likelihood of performing it, regardless of its future consequences. Reward devaluation would suppress the action for (a) but not for (b).

Source: Reproduced from Daw et al., 2005, with the permission of Nature Journals, Inc.

illustrated by a simple discrete-trial, discrete-choice instrumental learning task, whereby a hungry rat is trained to perform a lever press followed by entry to a food magazine to receive a food pellet.

The conditioning paradigm that Daw et al. use to distinguish between the two types of learning is *reinforcer devaluation* (see, e.g., Dickinson & Balleine, 2002). In a typical devaluation study, an animal first learns to perform a particular behavior in order to receive food pellets. After the animal has learned the response, the food pellets are devalued either by satiation or by pairing with illness, and the response becomes less prevalent. Daw et al. note that devaluation requires a model-based process because it depends on updating the value of the response and overriding a habit. Damage to orbitofrontal cortex or amygdala tends to interfere with devaluation.

Subsequent fMRI tests of the roles of prefrontal cortex and striatum partially verified the existence of dissociable model-free and model-based systems. Gläscher, Daw, Dayan, and O'Doherty (2010) designed a sequential probabilistic decision task with choices in two internal states (signified by fractal images) followed by a reward, and the participants learned the transition probabilities between states at successive levels before experiencing rewards. They found that activity in the ventral striatum correlated with reward prediction error, as in many previous studies by other investigators. Yet there was another error signal that they called the *state prediction error* (SPE), which occurred when a state at the first level was followed by the less probable state at the second level. The neural correlates of the SPE were activity in the lateral prefrontal cortex and intraparietal sulcus.

Yet, striatal involvement was seen when Simon and Daw (2011) studied a more complex cognitive map that changed over time. Simon and Daw gave participants a maze where they had a choice of rooms to go to, with doors closed in some directions and open in others that changed randomly, with the ultimate goal being a monetary reward. They found that activations in several areas of the caudate and putamen were selective for the projected positive value of the current state. Significantly, those activations correlated more strongly with the value predicted by a model-based than by a model-free process.

Daw, Dayan, and their associates noted the similarity of their model-free/model-based system dichotomy to various other two-process psychological theories that posit an automatic, reactive system and a controlled, deliberative system (e.g., Kahneman, 2011). Two-process theories are considered in Section 9.5 in the context of decision-making models. In a similar vein, many of the models of executive function tasks like the Wisconsin card sorting and Stroop tasks discussed in Section 9.4 are based on competition between habits and current rewards or task demands. Typically, prefrontal damage interferes with performance on those tasks by weakening the ability of current task contingencies to override habitual responses.

While Daw et al.'s (2005) model does not explicitly include nodes corresponding to brain regions, there are other models with explicit brain regions that capture effects such as reinforcer devaluation (Deco & Rolls, 2005a; Dranias et al., 2008; Frank & Claus, 2006).

6.4.1. Reinforcer Revaluation Models

Dranias et al. (2008) and Grossberg et al. (2008) embedded the dopaminergic reward prediction error in a larger conditioning model called *MOTIVATOR* (an acronym for *matching objects to internal values triggers option revaluations*). These authors discussed which parts of the conditioning process require dopamine and which other parts (mainly involving amygdala, hypothalamus, and OFC) do not require dopamine. The conclusions from their model are roughly consistent with the division between dopamine "wanting"

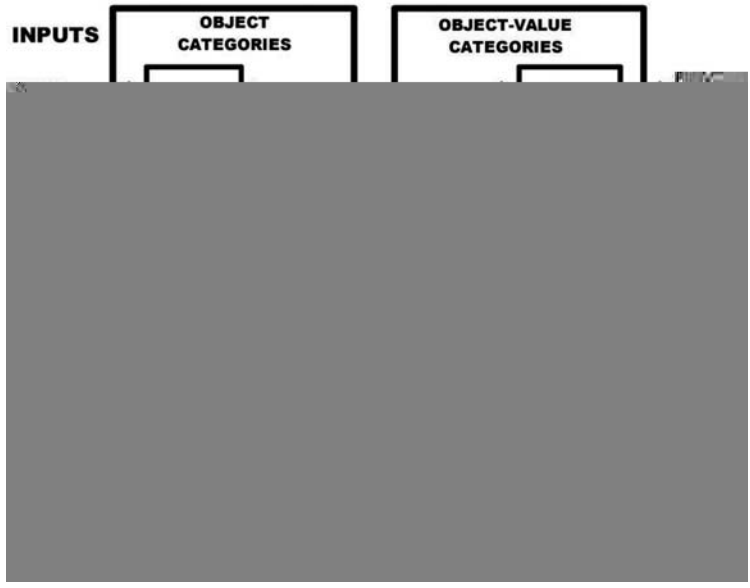


FIGURE 6.17 Overview of MOTIVATOR model. Object categories represent visual or gustatory inputs, in anterior inferotemporal (ITA) and rhinal (RHIN) cortices. Value categories represent the value of anticipated outcomes on the basis of hunger and satiety inputs, in amygdala (AMYG) and lateral hypothalamus (LH). Object-value categories resolve the value of competing perceptual stimuli in medial (MORB) and lateral (ORB) orbitofrontal cortex. The reward expectation filter detects reward or its absence using a circuit of ventral striatum (VS), ventral pallidum (VP), striosomal delay (SD) cells of the striatum, the pedunculopontine nucleus (PPTN), and midbrain dopaminergic neurons of the SNc/VTA (substantia nigra pars compacta/ventral tegmental area).

Source: Reprinted from Dranias et al., 2008, with the permission of Elsevier Science, Inc.

and amygdala/OFC "liking" (Berridge & Robinson, 1998; Berridge, 2007) discussed in Section 5.2. The dopamine–basal ganglia system is involved with calculating reward prediction errors and the amygdala and OFC with setting affective values for stimuli and events.

A schematic of the MOTIVATOR model is shown in Figure 6.17. The model network is divided into four large regions that represent categories of objects (in the applications in these two articles, mainly visual and gustatory objects); values of anticipated outcomes; object-value categories that compare the values of different stimuli; and a reward expectation filter. The first three of these regions together represent an updated version of the CogEM theory discussed in Sections 6.3.3 and 6.3.4, and the fourth is an elaboration of the reinforcement learning subnetwork discussed in Section 6.3.5.

The MOTIVATOR model accomplishes revaluation, including devaluation when a reinforcer has been satiated or paired with aversive consequences, via its object-value categories. The presence of two layers of orbitofrontal cortex allows for contextual updating of positive and negative emotional value calculations that are slower to change in the amygdala (Rolls, 2004).

The models of Frank and Claus (2006) and Deco and Rolls (2005a) similarly explain revaluation and devaluation using the unique connectivity properties of orbitofrontal cortex. The Frank–Claus model also distinguishes lateral OFC, which favors negative emotion, from medial OFC, which favors positive emotion. The Deco–Rolls model employs five functional classes of OFC cells (sensory, reward-related, sensory-intermediate, error-related, and rule-related). Dranias et al. (p. 273) discusses differences between the three models in terms of numbers of training trials required for the revaluations.

6.4.2. The Primary Value/Learned Value Model

Another conditioning model framework that involves basal ganglia, amygdala, and ventromedial prefrontal cortex along with midbrain dopamine nuclei is called *primary value/learned value* (Hazy, Frank, & O'Reilly, 2010; O'Reilly, Frank, Hazy, & Watz, 2007). The model derives its name from the two systems that compose it: the primary value system that is engaged by the US, which learns to inhibit the US-generated dopamine burst, and the learned value system that drives the CS-generated dopamine burst once the US has become reliably associated with reward.

The authors argue that their model is more biologically based and better able to simulate variable time delays between CS and US than the TD model. This is due in part to PVLV being based on associations and not on temporal chaining. O'Reilly et al. (2007) also compare PVLV with the model of Brown et al. (1999), which they call *BBG* (an initialism both of the authors' names and of "bursting in the basal ganglia"). O'Reilly et al. note that PVLV has a similar decomposition to the BBG model but differs from Brown et al. in

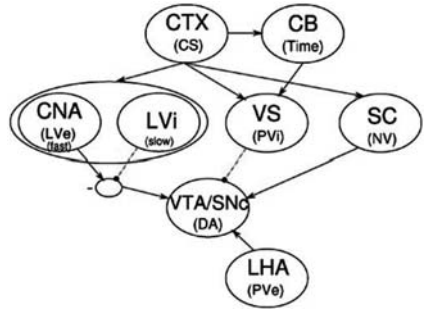


FIGURE 6.18 Biological interpretation of PVLV including novelty value (NV) component. Novelty value (for visual stimuli) is assumed to be driven by known direct projections from the superior colliculus (SC) to midbrain dopaminergic cells (VTA/SNc). CAN = central nucleus of amygdala; VTA = ventral tegmental area; SNc = substantia nigra pars compacta; LHA = lateral hypothalamus; VS = patch-like neurons of ventral striatum; SC = superior colliculus; CB = cerebellum; CTX = sensory neocortex.

Source: Reprinted from Hazy et al., 2010, with the permission of Elsevier Science, Inc.

several key anatomical and functional details. Notably, PVLV locates the excitatory part of the learned value system in the central nucleus of the amygdala, whereas BBG locates it in the nonstriosomal part of the ventral striatum. Also, in PVLV and not in BBG the subsystem for canceling reward bursts can be activated by signals other than time since CS onset. Figure 6.18 shows the updated biological interpretation of the PVLV model by Hazy et al. (2010), which also incorporates data showing that dopamine nuclei respond to novel stimuli.

PVLV and BBG (along with its later version in Tan and Bullock, 2008a, 2008b) are like each other, and, unlike TD, in that they both naturally interface with separate models of psychological processes other than conditioning, such as working memory. TD, by contrast, tries without complete success to incorporate working memory within its reward prediction error model. The larger connections of BBG are seen in the MOTIVATOR model shown in Figure 6.17. Likewise, PVLV draws on connections with a working memory model that includes functions of prefrontal cortex (Frank, Loughry, & O'Reilly, 2001; O'Reilly & Frank, 2006), which is discussed further in Chapter 9. The working memory connection allows the model to capture better than TD the distinctions between delay and trace conditioning, such as data showing that trace conditioning, but not delay conditioning, is impaired by prefrontal and hippocampal damage.

Equations for Networks in Chapter 6

Detailed Description: Sutton–Barto Equations

The key variable in the Sutton–Barto conditioning model, defined by Equations (3.22), is the variable y denoting the current amount of positive reinforcement in the environment – whether from a US or from one or more conditioned CSs. Consider first the typical Pavlovian paradigm called *delay conditioning*, in which the CS onset precedes the US onset but CS and US presentations overlap. First consider the trials before substantial conditioning has occurred. Then the variable y is 0. On the first trial where a US occurs, $\overline{y}$, the variable representing a weighted average of recent past values of y, is still 0, but $y = w_0 x_0$ is positive because w_0, the weight from the US, is set to a high value (in the Sutton and Barto, 1981, simulations, .6) and x_0 represents US intensity. So the difference $y - \overline{y} > 0$. Now let the given CS have the index 1, so that w_1 is the weight from the CS. Since the change in w_1 is proportional to $y - \overline{y}$ times the eligibility trace $\overline{x}_1$ of that CS, this weight will increase if the eligibility trace is high enough. That happens over a range of time delays from the onset of the CS, the same range illustrated in Figure 6.9. Over this range of delays, conditioning occurs to the CS, with varying effectiveness.

When the CS has become conditioned, its outgoing weight w_1 can become as large as the weight w_0 from the US. If that is the case and CS amplitude is the same (in whatever units it's measured) as that of the US, then the arrival of the US causes no change in the value of y. This is because y is constrained to be no larger than its asymptotic value, and that value has already been reached by the CS.

Klopf's Drive–reinforcement Model

The drive–reinforcement model of Klopf (1988), applied particularly to Pavlovian conditioning, was interpreted by its author as a single-neuron model. Hence, the $x_i(t)$ shown in Figure 6.5 can be interpreted either as presynaptic firing frequencies or as CS activity levels. The $w_i(t)$ in that figure can either be synaptic efficacies or associative weights. The single value $y(t)$ can be interpreted either as a postsynaptic firing frequency or as a level of reinforcement.

Klopf's neuronal input–output relationship is similar to the one devised by Sutton and Barto (1981) and shown in Chapter 3, with the addition of a threshold. His equation is

$$y(t) = \left[\sum_{i=1}^{n} w_i(t)x_i(t) - \theta\right]^{+} \tag{6.3}$$

where θ is the postsynaptic firing threshold, and the symbol "+" denotes replacing the expression by 0 if it is negative. After the value $y(t)$ is calculated using the linear weighting function in (6.3), if it is larger than some prespecified maximum firing rate, called $y^{\max}(t)$, it is replaced by $y^{\max}(t)$.

The learning mechanism is based on the changes in synaptic efficacies w_i as a result of changes in both presynaptic signals x_i and postsynaptic signal y. To account for causality in conditioning, the presynaptic signal changes are averaged over a time interval, called τ, which is the longest interstimulus interval at which conditioning is effective. The contributions of presynaptic signals at preceding times are weighted by the strength of pre-to-post connections at those times and by interval-dependent learning rate constants, called c_j for time interval j. Hence

$$\Delta w_i(t) = \Delta y(t-1)\sum_{j=1}^{\tau} c_j \left| w_i(t-j)\right| \Delta x_i(t-j-1) \tag{6.4}$$

where, as always, $\Delta y(t)$ denotes $y(t+1) - y(t)$ (hence, $\Delta y(t-1)$ denotes $y(t) - y(t-1)$), and analogous notation is used for changes in other variables. In addition, the Δx_j are not allowed to be negative, but are set to be 0 whenever they become negative.

The Temporal Difference Model

The temporal difference model of conditioning by Sutton and Barto (1990) includes two sets of equations, for the eligibility traces $\overline{x}_i(t)$ of the various CSs, and for $w_i(t)$, the strengths of association between those CSs and a single US. If $x_i(t)$ denotes the actual ith CS input signal (1 when the stimulus is present and 0 when it is absent), then the ith eligibility trace obeys the equation

$$\overline{x}(t+1) = (1-\delta)\overline{x}_i(t) + x_i(t) \tag{6.5}$$

with δ a constant between 0 and 1. If $y(t)$ denotes the US input signal (again 0 or 1), then the associative strength $w_i(t)$ obeys the learning equation

$$w_i(t+1) = w_i(t) + az_i(t+1)\overline{x}_i(t+1) \tag{6.6}$$

where α is another constant between 0 and 1, and $z_i(t+1)$ is the temporal difference factor (prediction error, i.e., change in total positive primary and secondary reinforcement). The temporal difference factor in (6.6) is in turn given by

$$z_i(t+1) = \left[y(t+1) + \gamma \sum_j w_j(t) x_j(t+1) \right] - \sum_j w_j(t) x_j(t) \tag{6.7}$$

In (6.7), the expression in brackets denotes actual reward experienced (from both the US and the CSs) and the expression to the right of the brackets denotes predicted reward (from the CSs). γ, which is typically slightly less than 1 in Sutton and Barto's simulations, is a factor by which future reinforcement is discounted as compared to present reinforcement.

In Sutton and Barto (1998) and Ludvig, Sutton, and Kehoe (2012), the coefficient 1-δ in Equation (6.5) was interpreted as the product of the time discounting factor γ times a value λ between 0 and 1 that represents the time window over which presentation of the CS can influence learning. The resulting model was labeled TD(λ).

The READ Circuit of Grossberg, Schmajuk, and Levine

Recall from earlier discussion that the recurrent associative dipole (READ) circuit was designed by Grossberg and Schmajuk (1987). As its name implies, this circuit, shown in Figure 6.15, adds to a gated dipole some internal feedback pathways. This feedback creates the possibility of second-order conditioning, both appetitive and aversive. The network also joins the gated dipole to a mechanism for associative learning.

If an unconditioned stimulus (US) is applied to the "on" channel of Figure 6.15 (the side with odd-numbered subscripts), and nonspecific arousal to both the "on" and "off" channels, the following equations arise (see Grossberg, Levine, and Schmajuk, 1992, for discussion):

Arousal + US + Feedback On-Activation

$$\frac{dx_1}{dt} = -ax_1 + I + J + f(x_7) \tag{6.8}$$

where I denotes the tonic arousal input to both channels and J is the specific input to the "on" channel.

Arousal + Feedback Off-Activation

$$\frac{dx_2}{dt} = -ax_2 + I + f(x_8) \tag{6.9}$$

On- and Off-Transmitters (Depletable)

$$\frac{dw_1}{dt} = b(1 - w_1) - cg(x_1)w_1 \tag{6.10}$$

$$\frac{dw_2}{dt} = b(1 - w_2) - cg(x_2)w_2 \tag{6.11}$$

Gated On- and Off-Activations:

$$\frac{dx_3}{dt} = -ax_3 + eg(x_1)w_1 \tag{6.12}$$

$$\frac{dx_4}{dt} = -ax_4 + eg(x_2)w_2 \tag{6.13}$$

Normalized Opponent On- and Off-Activations

$$\frac{dx_5}{dt} = -ax_5 + (h - x_5)x_3 - (x_5 + k)x_4 \tag{6.14}$$

$$\frac{dx_6}{dt} = -ax_6 + (h - x_6)x_4 - (x_6 + k)x_3 \tag{6.15}$$

Total On- and Off-Activations

$$\frac{dx_7}{dt} = -ax_7 + m[x_5]^+ + p\sum_{k=1}^{n} S_k w_{k7} \tag{6.16}$$

$$\frac{dx_8}{dt} = -ax_7 + m[x_6]^+ + p\sum_{k=1}^{n} S_k w_{k8} \tag{6.17}$$

where S_k denotes the amplitude of the kth conditioned stimulus (CS) and x^+, for any real number x, denotes max (x, 0).

On-Conditioned Reinforcer and Off-Conditioned Reinforcer Associations

$$\frac{dx_{k7}}{dt} = S_k\left(-qw_{k7} + r[x_5]^+\right) \tag{6.18}$$

$$\frac{dx_{k8}}{dt} = S_k\left(-qw_{k8} + r[x_6]^+\right) \tag{6.19}$$

The functions f and g in Equations (6.8)–(6.19) can either be sigmoids or *ramp* functions, that is, linear above some threshold and 0 below the threshold. All symbols not discussed so far denote positive constants.

The equations used by Grossberg and Levine (1987) to model attentional effects, such as blocking, are not shown here. They are similar to Equations (6.8)–(6.19) except that they include multiple, variable CS representations and do not include opponent pairs of on- and off-channels. Grossberg and Levine also added some terms to the CS and US short-term memory equations to allow for competition between stimuli and decay in the presence of random background noise.

Equations of Brown, Bullock, and Grossberg (1999)

Ventral striatal activity S is excited by primary reward inputs IR and by CS inputs I_i that are gated by adaptive weights w_iS:

$$\frac{1}{\tau_S}\frac{dS}{dt} = -A_S S + (1-S)\left[\sum_i I_i w_{iS} + I_R w_{RS}\right] \tag{6.20}$$

CS-to-striatal weights w_iS change only when S is positive. They are potentiated by a positively reinforcing dopamine burst N^+ and depressed by a "negatively reinforcing" dopamine depression N^-, described below:

$$\frac{1}{\tau_{w_S}}\frac{dw_{iS}}{dt} = S\left[N^+(I_i w_S^{\max} - w_{iS}) - \beta_{w_S} N^- w_{i_S}\right] \tag{6.21}$$

PPTN activity P is excited by striatal inputs S and primary reward inputs IR:

$$\frac{1}{\tau_P}\frac{dP}{dt} = -\left[1 + U_P w_{UR}\right]P + (1-P)\left[S w_{SP} + I_R w_{RP}\right] \tag{6.22}$$

where U_P is a habituation term that satisfies

$$\frac{1}{\tau_{U_P}}\frac{dU_P}{dt} = -U_P + (1-U_P)P. \tag{6.23}$$

Dopamine cell activity D is excited by the rectified PPTN activity $[P - \Gamma_P]^+$, where Γ_P is a signal threshold, and a tonic arousal signal I_D. The dopamine cell is inhibited in an adaptively timed fashion by the summed spectrum of signals from the striosomes, $\Sigma_{i,j}[G_{ij}Y_{ij} - \Gamma_S]^+ w_{ij}$, thus

$$\frac{1}{\tau_D}\frac{dD}{dt} = -D + (1-D)\left[(P-\Gamma_P)^+ w_{PD} + I_D\right] \qquad (6.24)$$
$$-(D+h_D)\sum_{i,j}\left[G_{ij}Y_{ij} - \Gamma_S\right]^+ w_{ij}$$

D over a period of time is averaged into a tonic signal $\bar{D}$ that satisfies the equation

$$\frac{1}{\tau_{\bar{D}}}\frac{d\bar{D}}{dt} = D - \bar{D}$$

which in turn leads to the calculation of the positive and negative reinforcement signals of Equation (6.21):

$$N^+ = \left[D - \bar{D} - \Gamma_N\right]^+ \qquad (6.25)$$

$$N^- = \left[\bar{D} - D - \Gamma_N\right]^+. \qquad (6.26)$$

If the index i is used for different CSs and j for striosomal cell populations, the activity x_{ij} of the jth striosomal population in response to the ith CS is

$$\frac{dx_{ij}}{dt} = r_j\left[-x_{ij} + (1-x_{ij})I_j\right] \qquad (6.27)$$

where r_j is a variable parameter (a hyperbolic function of j) that determines the adaptive timing of Ca^{++} spikes. In a simplification of the detailed biochemistry of Fiala et al. (1996), each Ca^{++} spike is represented as a product $G_{ij}Y_{ij,}$ with G_{ij} a signal activated by striosomal firing and Y_{ij} the amount of available Ca^{++}. Those two variables satisfy equations of the form

$$\frac{dG_{ij}}{dt} = \alpha_G\left(B_G - G_{ij}\right)f_G\left(x_{ij} - \Gamma_G\right) - B_G G_{ij} \qquad (6.28)$$

$$\frac{dY_{ij}}{dt} = \alpha_Y(1-Y_{ij}) - \beta_Y\left[G_{ij}Y_{ij} - \Gamma_Y\right]^+ \qquad (6.29)$$

In Equation (6.28), $f_G(x)$ is the unit step function (1 for $x > 0$ and 0 for $x < 0$).

The equations for the CS-to-striosome weights are listed in Brown et al. (1999) as

$$\frac{dw_{ij}}{dt} = \alpha_z\left[G_{ij}Y_{ij} - \Gamma_S\right]^+\left(-w_{ij} + \gamma_S(N^+ + N^-)\right) \qquad (6.30)$$

But Tan and Bullock (2008b) noted that (6.30) was a misprint in the 1999 article, which had led O'Reilly et al. (2007) to conclude that the model could lead to endless weight modification. The correct form of the equation for CS-to-striosome weights was

$$\frac{dw_{ij}}{dt} = \alpha_z \left[G_{ij} Y_{ij} - \Gamma_S \right]^+ \left((A_z - w_{ij}) N^+ - w_{ij} N^- \right) \tag{6.30a}$$

The Leabra Equations and Their Use in PVLV

Leabra (O'Reilly & Munakata, 2000; O'Reilly & Frank, 2006) simplifies the geometry of neurons to single points. The equations are difference equations for updating neuron activations at the ends of trials. If V_m is the membrane potential across a neuron, it is updated by three different conductances (excitatory, leak, and inhibitory) g_c and reversal (driving) potentials E_c using the equation

$$\Delta V_m(t) = \tau \sum_c g_c(t) \bar{g}_c \left(E_c - V_m(t) \right) \tag{6.31}$$

The conductances g_c and potentials E_c have three different subscripts (e, l, and i). The constants $\bar{g}_c$ denote the relative influence of the three different conductances.

The excitatory conductance $g_e(t)$ for the jth neuron is based on activations of those other neurons $i = 1, 2, \ldots, n$ that send excitation to the jth neuron times the weights of connections from those neurons:

$$\eta_j = g_e(t) = \frac{1}{n} \sum_i x_i w_{ij} \tag{6.32}$$

The leak conductance g_l is a constant. The inhibitory conductance g_i is computed by a k-winners-take-all function to be described below.

If the voltage V_m at time t is greater than a threshold value θ, then y_j, the activation communicated to other neurons is

$$y_j(t) = \frac{1}{1 + \frac{1}{\gamma \left(V_m(t) - \theta \right)}} \tag{6.33}$$

and if $V_m(t) < \theta$, $y_j(t) = 0$. Before communication to other cells the function of (6.33) is smoothed out via convolution with a Gaussian function:

$$y_j^*(x) \int_{-\infty}^{\infty} \frac{1/\sqrt{2\pi\sigma}\, e^{-z^2}}{2\sigma^2 y_j} (z - x)\, dx$$

The k-winners-take-all inhibition algorithm is implemented by means of a uniform inhibitory current g_i whose value is set such that the $(k+1)$st most excited unit within a layer is below its firing threshold but the kth most excited is above threshold. For each index i, it follows from (6.31) that the level of inhibition necessary to keep the ith unit right at the threshold voltage θ is

$$g_i^{\theta} = \frac{g_e \overline{g}_e (E_e - \theta) + g_l \overline{g}_l (E_l - \theta)}{\theta - E_l} \tag{6.34}$$

From the values defined in Equation (6.34), g_i is calculated to be a value between g_k^{θ} and g_{k+1}^{θ}; namely, $g_i = g_k^{\theta} + q(g_{k+1}^{\theta} - g_k^{\theta})$, with q a parameter between 0 and 1 whose default value is .25.

Learning in Leabra is a combination of Hebbian associative learning and error-driven learning, with the error-driven part actually similar to differential Hebbian learning (see Section 6.1.5). For weights w_{ij} between layers, let the superscript $^+$ above any number x denote max $(x, 0)$, and the superscript $^-$ above x denote min $(x, 0)$. Then the Hebbian component of weight change is

$$\Delta_{\text{hebb}} w_{ij} = y_j^+ (x_i^+ - w_{ij})$$

The error-driven component of learning is first calculated as

$$\Delta_{\text{err}} w_{ij} = x_i^+ y_j^+ - x_i^- y_j^-$$

to which a weight-dependent algorithm is then applied to bound it between 0 and 1:

$$\Delta_{sb\text{err}} w_{ij} = [\Delta_{\text{err}}]^+ (1 - w_{ij}) + [\Delta_{\text{err}}]^- w_{ij}$$

The actual weight change is a normalized linear combination of the Hebbian and error-dependent terms:

$$\Delta w_{ij} = \sum [k_{\text{hebb}} \Delta_{\text{hebb}} + (1 - k_{\text{hebb}}) \Delta_{sb\text{err}}]$$

Exercises for Chapter 6

○1. How might the Rescorla–Wagner model of blocking be extended to account for the phenomenon of unblocking, whereby if $CS_1 + CS_2$ predicts a different level of US stimulation than does CS_1 alone, then CS_2 conditions normally to the US? Does the Sutton–Barto model explain that phenomenon better?

*2. (a) Run a simulation of the Sutton–Barto equations (3.22) where the interstimulus interval is varied. Let the number of trials be 40 and let

each trial be 50 time units long. Let the CS duration be three time steps, the US duration 30 time steps. Use the parameters $\alpha = .9$, $c = .2$, $\beta = 0$, and the initial US associative strength $w_0 = .6$. Plot the asymptotic value of the CS associative strength as a function of interstimulus interval. Let CS and US both have amplitude 1.

(b) Run a simulation of a blocking paradigm using the Sutton–Barto equations and two conditioned stimuli, CS_1 and CS_2. Let CS durations be five time steps and US duration ten time steps. For the first ten trials, CS_1 alone is presented and followed by US, starting immediately after CS_1 termination. For the next ten trials, CS_1 and CS_2 are presented simultaneously, followed immediately by the US. Use the parameters $\beta = 0$, $c = .5$, $\alpha = .6$, $w_0 = .6$.

*3. Simulate excitatory conditioning and extinction in the READ circuit, with the same parameters used by Grossberg and Schmajuk (1987), shown in Equations (6.8)–(6.19), with just a single CS. On each trial, present the CS for a duration of 40 time units, of which the US is also present for the last 15. Let the total trial length be 200 time units. The CS is paired with the US for ten trials, and then presented in the absence of the US for the next ten trials. The parameters in the differential equations (using the terminology of this book, not of Grossberg and Schmajuk) are $a = 1$, $b = .005$, $c = .00125$, $e = 20$, $h = 20$, $k = 20$, $m = .5$, $p = 20$, $q = .005$, $r = .025$. The strength of tonic arousal is ten, while the CS and US have amplitudes 1 and 200 respectively. For the functions $f(x)$ and $g(x)$, use $.05x$ and x respectively if x is positive, and 0 if x is negative. Hint: this has tended to work better in simulations if on each time step the equations for node activities are solved in the steady state, that is, with time derivatives equal to 0. Then the equations for connection strengths are integrated, using a differential equation solving routine, with the new values for node activities substituted in.

*4. Simulate the temporal difference model of anticipatory neural activity with the same parameters used by Suri and Schultz (2001). The equations used are a slight modification of the TD Equations (6.5)–(6.7), namely:

$$p_j(t+1) = \sum_{i=1}^{N \times L} w_{ji}(t) x_i(t+1)$$

$$z_j(t) = u_j(t) + \gamma p_j(t+1) - p_j(t)$$

$$w_{ji}(t+1) = w_{ji}(t) + \alpha z_j(t) \bar{x}_i(t)$$

$$\bar{x}_i(t+1) = \delta \bar{x}_i(t) + (1-\delta) x_i(t+1) \text{ and } \bar{x}_i(t \le 0) = 0$$

Each trial is seven seconds with the time unit in the simulation being 0.1 second, i.e. the total trial time is 420 time units. The first event in a trial

(usually a stimulus) serves as the information to learn prediction signals for the second event (usually the reward). On each trial, present the events $j = 1$ and $j = 2$ for a duration of 60 time units, as shown in Figure 6.19. The events are each present for 20 trials. Each event is represented with 70 phasic representation components $\overline{x}_i(t)$. The prediction error factor $z_j(t)$ increases phasically when the event is presented, and the prediction signal $p_j(t)$ is tonically increased during the intratrial interval. The elements of the weight matrix $w_{ji}(t)$ are incrementally adapted according to the product of the prediction error factor $z_j(t)$ with eligibility traces $\overline{x}_i(t)$ of the temporal event representation $x_i(t)$. The temporal representation $x_i(t)$ is shown in the figure, each representation is activated for one time unit after the previous one, and the first one is activated after one time unit of the event for both

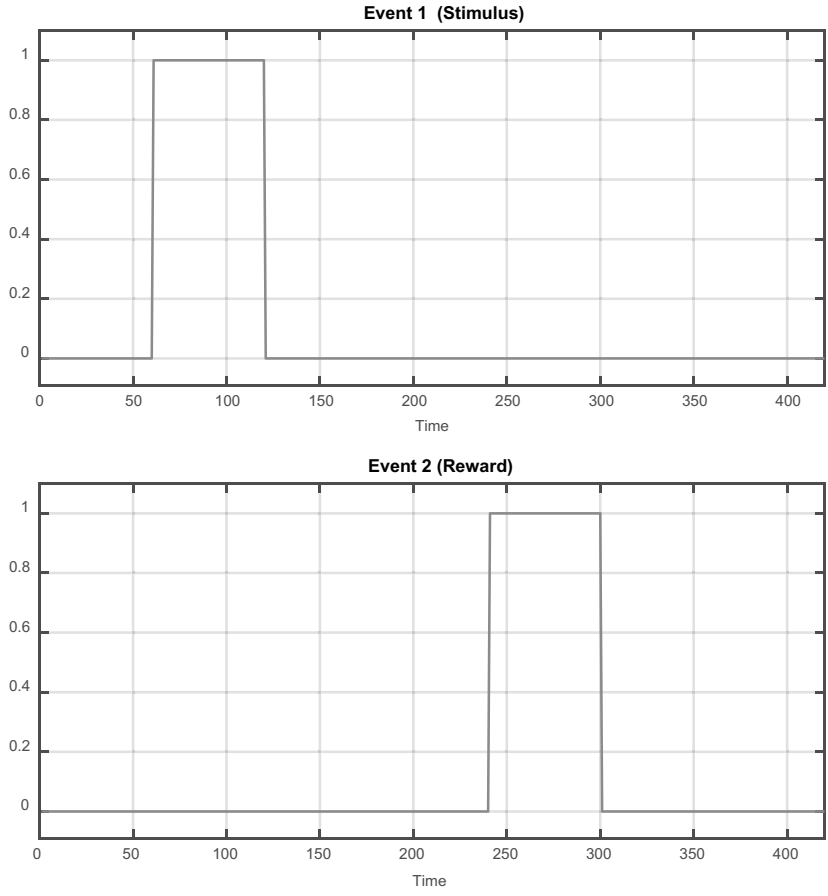


FIGURE 6.19 Time course on a single trial of the association of stimulus and reward in the model of Suri and Schultz (2001).

events. Hint: we have shown two temporal representations for the simulation so we need an $N \times L$ matrix. The parameters in the equations (using the terminology of this book) are $N = 70$, $L = 2$. For simplicity, let the temporal discount factor $\gamma = 0.9$, the learning rate $\alpha = 50$, and the eligibility traces discount factor $\delta = 0.9$.

Some Additional Sources

Modeling of Classical Conditioning

Commons, Grossberg, and Staddon (1991); Dawson (2008); de Pinho, Mazza, and Roque (2006); Gluck, Myers, and Meeter (2005); Mannella, Koene, & Baldassarre, 2009; Meeter, Myers, and Gluck (2005); Myers et al. (1996); Myers, Ermita, Hasselmo, and Gluck (1998); Schmajuk (1997); Schmajuk and DiCarlo (1992).

Modeling of Operant Conditioning

Dragoi (1997); Miller and Shettleworth (2008); Raymond, Baxter, Buonomano, and Byrne (1992); Simen and Cohen (2009); Staddon and Zhang (1991); Suppes, de Barros, and Oas (2012).

Connections with Invertebrate Neurophysiology

Canavier, Baxter, Clark, and Byrne (1993, 1994); Santos, Porto, Romero, Albó, and Pazos (2007).

Temporal Difference Models

Bertin, Schweighofer, and Doya (2007); Dayan, Niv, Seymour, and Daw (2006); Kim, Hwang, Seo, and Lee (2009); Rivest, Kalaska, and Bengio (2014); Singh and Sutton (1996); Zhang (2009).

Other Reinforcement Learning Models

Ashby and Crossley (2011); Izhikevich (2007).

Model-Based and Model-Free Systems

Penner and Mizumori (2011).

Modeling of Fear Conditioning

Schmajuk, Larrauri, and LaBar (2007).

Notes

1. While the term "interstimulus interval" is commonly used in the animal learning literature, the interval is not properly "between stimuli" but rather between their onsets. For this reason, some psychologists prefer the term "stimulus onset asynchrony," or SOA. In fact, Grossberg and Rudd (1989), in discussing light flashes, define SOA as the time between the onsets of two flashes, and ISI as the time between the offset of the first flash and the onset of the second.
2. Houk et al. (1995) included the lateral hypothalamic input on theoretical grounds but did not include anatomical evidence for its existence. But findings from various laboratories have shown that a main source of excitation to dopamine cells in the substantia nigra pars compacta is the pedunculopontine tegmental nucleus (PPTN) of the midbrain and that the main source of excitation to the PPTN is the lateral hypothalamus.

7

MODELS OF CODING, CATEGORIZATION, AND UNSUPERVISED LEARNING

> The Nameless is the origin of heaven and earth; the Named is the mother of all things.
>
> Lao Tzu

> I hate definitions.
>
> Benjamin Disraeli (*Vivian Gray*)

No two sensory events are exactly the same. We never see the same moon twice; even our wife's or husband's face exhibits subtle changes in expression from one viewing to another. Yet, most of us manage to ascribe to the sensory world a fair degree of regularity. How do we decide when two sensory objects are similar enough to be listed in the same category and when they are not?

Of course, the rules for "similarity" vary tremendously with context. An apple and a radish are listed together if foods are classified by color, but not if foods are classified by taste (sweet, bitter, sour, or salt). How to model the capacity for multiple categorizations, in which the context determines which categorization is used, is a problem at the forefront of contemporary neural network theory. But much progress has been made on the simpler problem of constructing neural networks that learn to encode a single categorization of sensory patterns, regardless of context.

Some neural network categorization models involve *supervised* learning; that is, certain output nodes are trained to respond to certain patterns, and the changes in connection weights due to learning cause those same nodes to respond to more general classes of patterns. Other models involve *unsupervised* learning; that is, input patterns are presented in some sequence and the network discovers through self-organization a "natural" categorization of the sensory

world. The distinction between supervised and unsupervised has been made in the engineering literature for over 40 years; see Duda and Hart (1973). Sometimes, in engineering applications, supervised categorization is called *classification*, whereas unsupervised categorization is called *clustering*. This distinction is not rigid because some "unsupervised" networks are actually supervised by an *internal* error signal (for discussion, see Dawes, 1992), but it is a useful means of classifying models.

Neural network categorization algorithms have been quite diverse, but many of them (e.g., Carpenter & Grossberg, 1987a, 1987b; Edelman, 1987; Rumelhart et al., 1986; Rumelhart & Zipser, 1985) have some general points in common. All these networks include modifiable connections between one layer of nodes encoding features of the sensory environment, and another layer of nodes encoding categories of sensory patterns composed of those features (see the generic Figure 1.2 from Chapter 1). Typically, category nodes are initially random or neutral in their responses, but learn specific pattern categories by experience.

To prepare the way for understanding neural categorization, we first consider the slightly simpler issue of how a node in a neural network can learn to respond to particular patterns of activity at other groups of nodes. These patterns of activity, in turn, could represent combinations of sensory features. The next section deals with coding in that sense, not in the sense of how the primary representation of a sensory stimulus is actually formed in the nervous system. Network mechanisms for such higher level coding, which are also discussed in Section 8.2 of Levine (1983), have possible implications for biological organisms during development.

7.1. Interactions between Short-and Long-Term Memory in Code Development

7.1.1. Malsburg's Model With Synaptic Conservation

The study of neural networks for code development essentially began with a seminal article by Malsburg (1973) on the development and tuning of orientation-sensitive cells in the visual cortex. This model is discussed in some detail because its basic structure anticipates many of the more current models of coding and categorization.

Malsburg's (1973) model is based on unmodifiable recurrent excitation and inhibition between "cortical" nodes, combined with modifiable synapses to the "cortex" from an input ("retinal") layer of nodes. He was motivated to develop this model by a body of experimental results on the mammalian visual system. These results suggested that:

> The task of the cortex for the processing of visual information is different from that of the peripheral optical system. Whereas eye, retina and lateral geniculate body (LGB) transform the images in a "photographic" way, i.e., preserving essentially the spatial arrangement of the retinal image, the cortex transforms this geometry into a space of concepts.
>
> (p. 85)

In particular, Malsburg's model, and several subsequent models discussed in this chapter, drew their inspiration from physiological results on single-cell responses to line orientations. These models can explain findings that neurons in the cat or monkey visual cortex respond preferentially to lines of a particular orientation, and that cells responding to similar orientations are grouped close together anatomically, in columns (Hubel & Wiesel, 1962, 1963, 1965, 1968). These models also explain findings that preferred orientations of neurons are influenced by early visual experience (e.g., Blakemore & Cooper, 1970; Hirsch & Spinelli, 1970). (Later results on the various influences on orientation specificity, including the interplay between genetic and developmental factors, are summarized in Ferster & Koch, 1987, and Malsburg & Cowan, 1982.)

Some early models (e.g., Bienenstock, Cooper, & Munro, 1982, p. 32; Grossberg, 1976a, p. 152; Grossberg, 1976b, p. 131; Perez, Glass, & Shlaer, 1974) also address evidence that there is a *critical period* in development of orientation detectors. That is, for a short period of time (in cats, age 23 days to four months; in humans, six months to two years), cortical orientation tuning is much more modifiable than it is either earlier or later.

Malsburg's simulated cortex is organized into two separate populations, excitatory and inhibitory nodes (he called them *cells*). That is, Malsburg's model is *unlumped* in the terminology of Section 4.2. The variation of connection strengths with distance endows the simulated cortex with a crude form of the lateral inhibitory architecture shown in Figure 4.2: narrow-range excitation and broad-range inhibition. Malsburg's laws for lateral interaction between nodes are *additive* rather than *shunting* in the terminology of Sections 4.1 and 4.2.

The arrangement of excitatory-to-excitatory connections p_{ik}, excitatory-to-inhibitory connections r_{ik}, and inhibitory-to-excitatory connections q_{ik} in Malsburg's simulated cortex is shown in Figure 7.1. As that figure shows, excitatory and inhibitory nodes are organized into two parallel planes, each with a hexagonal arrangement of nodes. Excitatory nodes excite neighboring nodes, both excitatory and inhibitory ones, whereas inhibitory nodes inhibit excitatory nodes that are a distance of two away. The signal transmitted by each node is equal to the amount of the signaling node's activity that is above some threshold value. Equations relating all these variables, and the connection weights s_{ik} from simulated retinal afferents, are given at the end of this chapter.

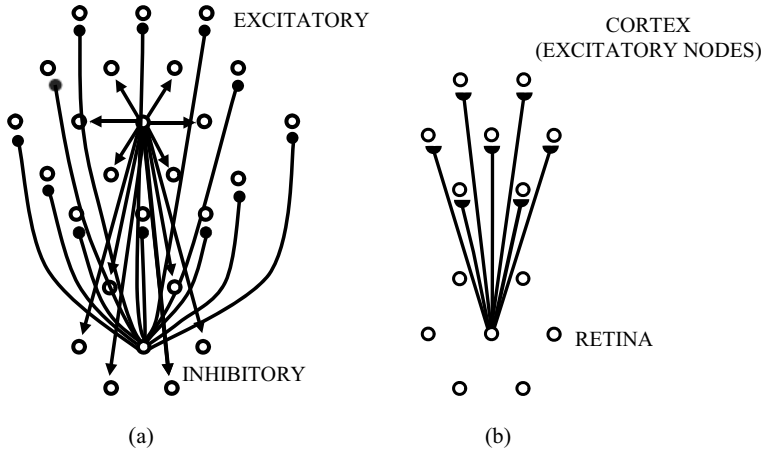


FIGURE 7.1 (a) A small part of the simulated cortex of Malsburg (1973), showing the arrangement of connections between excitatory and inhibitory nodes. (b) Each node of Malsburg's retina has modifiable connections to all excitatory nodes of the cortex.

Of the connections in Malsburg's model, only the connections from retinal afferents to cortical nodes have modifiable weights. The rule for changing these weights combines an associative learning law with a synaptic conservation rule similar to the gamma-system learning law of Rosenblatt (1962; see Section 2.1.3). Synaptic conservation was imposed to prevent the unbounded growth of synaptic strengths that would otherwise result from associative learning. This combination of laws is described as follows (Malsburg, 1973, p. 88):

> . . . if there is a coincidence of activity in an afferent fibre I and a cortical E-cell k, then s_{ik}, the strength of connection between the two is increased to $s_{ik} + \Delta s$, Δs being proportional to the signal on the afferent fibre I and to the output signal of the E-cell k. Then all the s_{jk} leading to the same cortical cell k are renormalized to keep the sum $\Sigma_j s_{jk}$ constant.

Figure 7.2 shows the standard set of stimuli used on Malsburg's model retina. These stimuli correspond to bars of light at different orientations. As shown in Figure 7.3, orientation detectors, such as were found by Hubel and Wiesel (1962, 1963, 1968), develop spontaneously among Malsburg's simulated cortical cells. After 100 learning steps, the lateral excitatory and inhibitory interactions lead to self-organization of cortical nodes, whereby most nodes have preferred orientations and nodes of similar preferred orientations tend to be grouped together. Figure 7.4 shows a similar grouping of biological orientation detectors found by Hubel and Wiesel (1968) in the monkey striate, or primary visual, cortex.

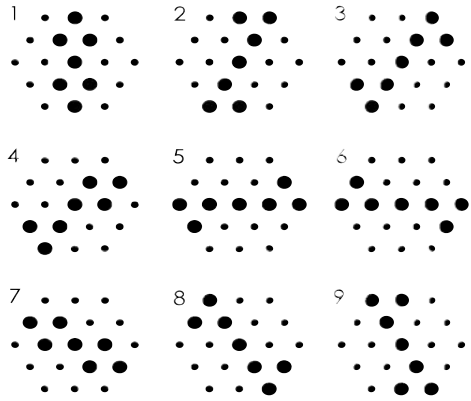


FIGURE 7.2 Standard set of stimuli used on the simulated retina. Larger dots denote locations of activated nodes.

Source: Reprinted from Malsburg, 1973, with permission of Springer-Verlag.

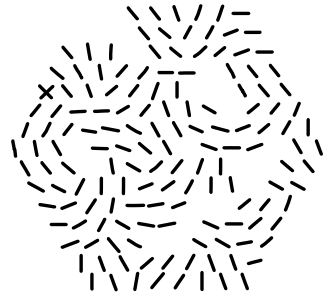


FIGURE 7.3 Simulated cortex after 100 time steps of learning. Each bar indicates the orientation to which the excitatory node at that location is most responsive. Blank spaces represent locations of nodes that never learn to react to any of the standard stimuli.

Source: Adapted from Malsburg, 1973, with permission of Springer-Verlag.

The idea of synaptic conservation is intuitively based on the notion that some chemical substance, whether a transmitter or second messenger (see Section 3.1), is present in a fixed amount at postsynaptic sites and distributed in variable fashion across impinging synapses. This mechanism is necessary for the effects in Malsburg (1973), and in two related models of the visual cortex by Perez et al. (1974) and Wilson (1975). Some more recent categorization models (e.g., Carpenter & Grossberg, 1987a; Rumelhart & Zipser, 1985; Sirosh & Miikkulainen, 1997) also use learning laws whereby strengthening of some synapses weakens other synapses. Such laws are reminiscent of the learning scheme of Rescorla and Wagner (1972; see Chapters 3 and 6), which includes an upper bound on the total associative strength of all stimuli with a given reinforcer.

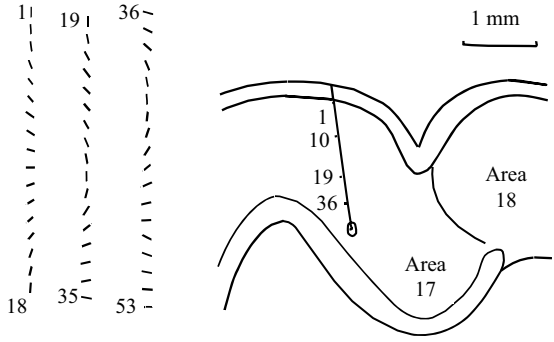


FIGURE 7.4 Reconstruction of an electrode penetration through an area of a monkey's primary visual cortex. Lines to the left indicate preferred orientations of the cells traversed. Areas numbered 17 and 18 are standard terms for subdivisions of the cortex.
Source: Adapted from Hubel and Wiesel, 1968, with permission of Cambridge University Press.

Malsburg's model was extended by Willshaw and Malsburg (1976, 1979) to a model of the development of *topographic maps*, that is, mappings between network layers whereby areas of the upper layer that code for nearby receptive fields are close together. Willshaw and Malsburg applied these maps to models of interactions between the retina and optic tectum in nonmammalian vertebrates. Some later modelers (Kohonen, 1982; Linsker, 1986a, 1986b, 1986c; Sirosh & Miikkulainen, 1994, 1997) developed models of topographic maps that, unlike those of Willshaw and Malsburg, did not include any topographic order in the original inputs. These later models are discussed in Section 7.1.4.

7.1.2. Grossberg's Model with Pattern Normalization

Grossberg (1976a, 1976b) developed a model that has many principles in common with Malsburg's but does not use a synaptic conservation law for learning. He argued that such a conservation law is incompatible with classical conditioning. Moreover, while Malsburg used this law in order to keep synaptic strengths, and therefore total network activity, bounded, it is also possible to achieve boundedness by replacing additive lateral interactions with shunting inhibition and excitation. These arguments are now reviewed.

The argument that synaptic conservation is incompatible with conditioning was given in Grossberg (1976a, p. 149). Suppose a sensory cue S_1 elicits a response pattern R, and then another cue S_2 is paired with S_1 (see Figure 7.5). The pairing leads to a strengthening of the connection from the S_2 node to those nodes whose activation is strongest in the R response. But if, as in Malsburg's model, total strength of synapses impinging on a given node from the previous level is kept constant, strengthening of the connection S_2-to-R would force weakening of the connection S_1-to-R. In reality, secondary conditioning occurs

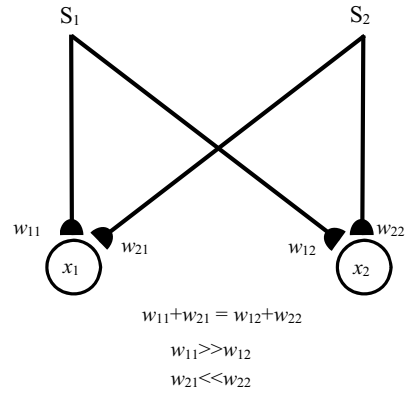


FIGURE 7.5 Argument against conservation of synaptic strength. If w_{11} is much larger than w_{12}, the network shown here learns to respond to stimulus S_1 with a response pattern R, whereby x_1 is much more activated than x_2. Since the two sums are equal, w_{21} is then less than w_{22}, so the network cannot also learn to perform R in response to S_2. If S_2 is paired with S_1, as in secondary conditioning, this is contradictory.

Source: Adapted from Grossberg, 1976a, with permission of Springer-Verlag.

(see Section 6.2), so that S_1 and S_2 can simultaneously be strongly connected to R. The argument that synaptic conservation is unnecessary for boundedness was based on the mathematical theory of shunting on-center off-surround networks (Ellias & Grossberg, 1975; Grossberg, 1973; Grossberg & Levine, 1975; see Section 4.2). This argument uses the nonrecurrent shunting Equations (4.2) (recall the discussion of recurrent versus nonrecurrent inhibition in Section 4.1). If x_i is the activity of the ith node which receives input I_i, it can be shown from those equations (see Exercise 1 of Chapter 4) that the steady-state value of x_i (what it converges to after long times) is

$$x_i = \theta_i \left(\frac{BI}{A+1} \right) \tag{7.1}$$

Since the relative pattern weights θ_i add up to 1, (7.1) says that the sum of all the steady-state activities equals $BI/(A + I)$, which is no larger than B regardless of the total input intensity I and the number of nodes in the network. Hence, total network activity never exceeds B.

Grossberg's argument is based on the factorization of spatial patterns into a product of relative pattern weights θ_i and total intensity I. This concept, sometimes called *factorization of pattern and energy*, was previously discussed in relation to outstar networks in Section 3.2.1. Factorization of pattern and energy also plays a role in models of motor control (Bullock & Grossberg, 1988) and of word recognition (Grossberg & Stone, 1986).

A model for development and tuning of feature detectors, combining lateral inhibition for short-term memory with associative synaptic modification for long-term memory, is discussed in Grossberg (1976b). Figure 7.6 shows the minimal network of that article. This network, like that of Malsburg (1973), includes unidirectional modifiable synapses from an input layer F_1 to a cortical layer F_2, leading to coding of input patterns by cortical nodes. Grossberg (1976c) extended this model to include modifiable feedback from F_2 to F_1. To describe the mutually excitatory dynamics that emerge in a modifiable network with top-down feedback, he coined the term *adaptive resonance*. This work ultimately led to the well-known adaptive resonance theory (ART) of Carpenter and Grossberg (1987a, 1987b; see Section 7.2).

In the network of Figure 7.6, the input-receiving nodes x_{1i} are endowed with a nonrecurrent (feedforward) on-center off-surround anatomy, and the pattern-coding nodes x_{2i} with a recurrent on-center off-surround anatomy (see Figure 4.1). F_1 and F_2 represent successive layers in a hierarchical network. Grossberg suggested that variations on the same hierarchy could be repeated in different brain regions. In Malsburg (1973) and Perez et al. (1974), F_1 was interpreted as either retina or thalamus, and F_2 as visual cortex (see Appendix 2). But F_1 might also be identified with a composite of early processing areas in the retina (*receptors*, *horizontal cells*, and *bipolar cells*) and F_2 with retinal areas closer to the optic nerve (*amacrine* and *ganglion* cells; Grossberg, 1976b). Also, since the visual cortex itself contains several processing stages, identified with cell groups known as *simple*, *complex*, and *hypercomplex* cells (Hubel & Wiesel, 1962, 1963, 1965, 1968; see Section 5.4.1), F_1 and F_2 might be interpreted as different parts of cortex. Based on these cortical cell groups, Grossberg (1976b,

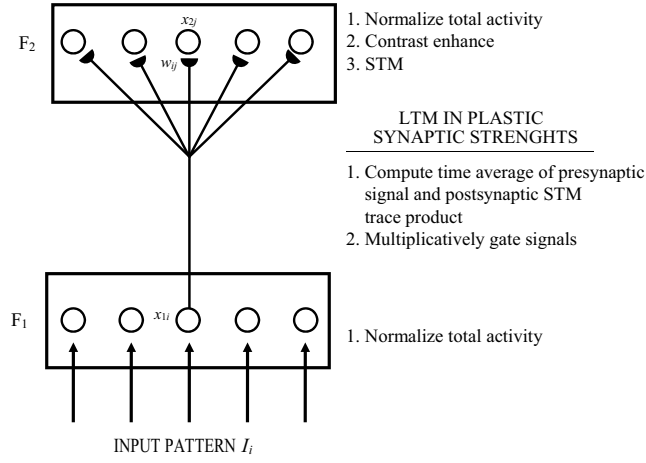


FIGURE 7.6 Minimal model of development and tuning of feature detectors using short-term memory (STM) and long-term memory (LTM) mechanisms.

Source: Adapted from Grossberg, 1976b, with permission of Springer-Verlag.

1976c) proposed architectures that are more complex, adding to Figure 7.6 another layer F_3 whose nodes code activity patterns across F_2. Nor are these architectures restricted to vision: Grossberg (1976b) described yet another interpretation, whereby F_1 is the olfactory bulb and F_2 is olfactory cortex.

In Grossberg (1976b), the input signals I_{2j} to the cortical nodes x_{2j} are linear combinations of the activities at the retinal nodes x_{1i}, weighted by the strengths of retinocortical synapses. The patterns viewed at the retina are assumed to be normalized so that their values θ_i add up to 1. (Recall the discussion in Section 4.1 of pattern normalization in networks with lateral inhibition.) For simplification, it is assumed that the x_{1i} activities represent the input pattern to the retina. Hence, the total signal at time t to a given cortical node x_{2j} due to the retinal pattern $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is

$$S_i(t) = \sum_{k=1}^{n} \theta_k w_{kj}(t) \tag{7.2}$$

where $w_{kj}(t)$ denotes the strength of the synapse from retinal node k to cortical node j. The linear combination in (7.2) also is incorporated in Equation (7.5) at the end of this chapter, from Malsburg (1973).

Grossberg (1976b) discussed several possible long-term memory laws for the synaptic strengths w_{kj}. In one of these laws (shown in Equation (7.10) at the end of this chapter), weights of connections to a node change only when short-term memory at that node is active.

7.1.3. Mathematical Results of Grossberg and Amari

Recall from Section 4.2 that a recurrent competitive network, such as F_2 in Figure 7.6, can either have exactly one or more than one node with positive asymptotic activity as time increases. Hence, sometimes, but not always, competition is *winner-take-all*, that is, exactly one node at the competitive level gets its incoming signals stored in short-term memory (STM). From the viewpoint of categorization, the winner-take-all case is especially interesting. In particular, the "winning" node in the competition for short-term storage could be the one whose incoming signal S_j, as defined by (7.2), is the largest.

In Grossberg (1976b), the criterion of choosing the F_2 node with the largest incoming signal leads to a primitive scheme for categorizing patterns, each pattern defined as a vector of retinal node activities. This largest-linear-signal criterion is also the basis for categorization in Amari and Takeuchi (1978).

Grossberg (1976b) discovered, however, that such a categorization algorithm can lead to miscodings if the network is subjected to many different spatial patterns over time. For suppose that cortical node x_{21} is the most active in response to a particular retinal pattern, say $\boldsymbol{\theta}^1$. Then the vector $\mathbf{w}^1$ of synaptic

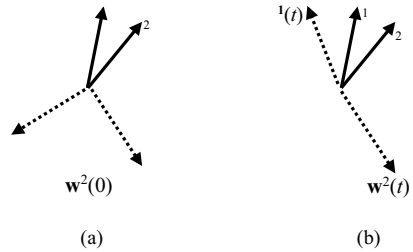


FIGURE 7.7 (a) Vectors of weights from F_1 to two different F_2 nodes at time 0, as compared with spatial patterns $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$. (b) As node x_{21} of F_2 learns $\boldsymbol{\theta}^1$, its bottom-up weight vector gets closer to both $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ than the vector $\mathbf{w}^2$ of bottom-up weights to node x_{22}.

Source: Adapted from Grossberg, 1976b, with permission of Springer-Verlag.

weights from all the retinal nodes to that cortical node becomes closer to $\boldsymbol{\theta}^1$ as time increases (as is shown by Equation (7.9) at the end of the chapter). But, as Figure 7.7 shows, bringing these weights closer to $\boldsymbol{\theta}^1$ could also bring them closer to a different spatial pattern $\boldsymbol{\theta}^2$. Thus if $\boldsymbol{\theta}^2$ is presented, it will now be coded by the node x_{21} and the weight vector $\mathbf{w}^1$ will be attracted to $\boldsymbol{\theta}^2$. Alternate presentation of $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ to the network can lead to an oscillation, or the node x_{21} can become unable to recognize the original pattern $\boldsymbol{\theta}^1$. This forces $\boldsymbol{\theta}^1$ to be recoded by a different cortical node.

To prevent presentation of one pattern from recoding other patterns, Grossberg (1976b) proposed adding to the network of Figure 7.6 some feedback connections from F_2 to F_1. Such feedback allows the network to detect when $\boldsymbol{\theta}^2$ is sufficiently different from $\boldsymbol{\theta}^1$ to be categorized separately from $\boldsymbol{\theta}^1$. In that case, whatever F_1-to-F_2 signals would otherwise be generated by $\boldsymbol{\theta}^2$, a mismatch signal is generated that transiently removes node x_{21} from the set eligible to code $\boldsymbol{\theta}^2$. This mismatch-detecting mechanism is at the heart of adaptive resonance theory, which is discussed more fully in Section 7.2.2.

Grossberg (1976b, 1976c) showed that every solution of the equations for the network of Figure 7.6 approaches some equilibrium point corresponding to a learned category. Amari (1977a, 1977b, 1980) proved an analogous result for a coding network similar to Grossberg's but with additive rather than shunting interactions. Amari and Takeuchi (1978) extended this result to a coding system which includes both excitatory and inhibitory modifiable interlevel synapses.

7.1.4. Feature Detection Models With Random Elements

Bienenstock et al. (1982) constructed a model of the development of orientation detectors in the visual cortex. Their model has much in common with Grossberg's but also includes random elements. Bienenstock et al. added

nonlinear interactions to the previous linear probabilistic model of Nass and Cooper (1975). (Recall from Section 3.2 that Nass and Cooper's model was in turn based on addition of decay terms to the associative model of Anderson, 1973.) In addition to orientation preferences, units in the network of Bienenstock et al. can exhibit preferences for one or another eye; such ocularity preferences are influenced by the opening or closing of either eye during development (see Table 4.1).

In the model of Bienenstock et al. (1982), retinocortical connections are subject to a learning rule that includes the possibility of both synaptic increase (in the manner of Hebb, 1949) and decrease. If I_j is the jth retinal input, and w_j the strength of the synapse to a given cortical node with activity x, then the expression for the rate of change of w_j is of the form

$$-\varepsilon_{w_i} + \Phi(x(t))I_i(t) \tag{7.3}$$

where ε is a decay rate and the function Φ can either be positive or negative. In fact, Φ is positive for x above a certain threshold value, and negative for x below that threshold. That is, in current terminology, the learning law alternates between Hebbian with decay and *anti-Hebbian* with decay. (Another version of this type of learning law appears in Bear et al., 1987.) This law embodies, in the learning equation itself, a form of contrast enhancement of significant inputs, or suppression of random noise, such as is often obtained with lateral inhibition (see Chapter 4). The network also, however, includes some actual lateral inhibition.

The inputs I_j of (7.3) are assumed to be random, with a probability distribution reflecting the distribution of patterns in a given sensory environment. Through computer simulation and mathematical analysis, the connection weights w_j were shown to converge to a steady state that is *selective* with respect to the distribution of the random inputs. That is, the response of the given cortical neuron, which equals the sum $\Sigma_j w_j(t) I_j(t)$, reaches its maximum possible value over a relatively small subset of the set of possible inputs. Bienenstock et al. suggested that this kind of selectivity is analogous to the orientation selectivity of actual cortical neurons.

A series of articles by Linsker (1986a, 1986b, 1986c) carried the idea of multilayer structure further with a network of three (or more) layers with adaptable connection strengths. These layers were intended to be analogs of both the retina and the visual cortex, without mimicking the detailed anatomy of either area. From random inputs at the lowest layer, combined with Hebb-like learning at interlayer connections, emerge at higher layers some *spatial-opponent cells* (that is, nodes with on-center off-surround or off-center on-surround responses, like those of the retinal ganglion cells shown in Figure 4.2). At still higher layers, nodes emerge that are responsive to given

orientations. Finally, the orientation nodes organize themselves into columns in much the same manner as occurs in the models of Malsburg, Grossberg, and Bienenstock et al.

Linsker's model differs from the other models discussed in this section in that prescribed orientations are *not* part of the pattern of its inputs. Rather, the orientation specificity and topographic organization emerge from mathematical properties of the Hebb-like learning laws at interlevel synapses. Orientation specificity also emerges from learning in the models of Sirosh and Miikkulainen (1994, 1997). Sirosh and Miikkulainen found that activity-dependent topographic organization depends on receptive fields of individual nodes being large enough compared with their initial topographic scatter. Kohonen (1982) suggested that the theory of development of this kind of organization is not restricted to topographic maps but could be applied to a feature or attribute space at any level of abstraction.

7.1.5. From Feature Coding to Categorization

The visual orientation detection networks discussed above perform a primitive form of pattern categorization. Hence, this type of network provides one of the bases for the more sophisticated categorizations (e.g., the shape of a dog, the sound of a recorder) that we perform in daily life. As categorizations become more subtle, they depend heavily on the ability to notice common features in disparate, and sometimes noisy, inputs. For example, we classify both a cocker spaniel and a dachshund, with or without tail damage, as dogs.

Recall from the start of this chapter the two general types of networks for pattern categorization: supervised and unsupervised learning. Both of these types of networks combine the coding of input patterns by internal layers of nodes with other functions. In supervised learning, the responses of an output layer to certain given patterns are compared with desired responses. Hence, in addition to a coding system, these networks typically require at some stage an error-correcting or delta learning rule (see Sections 2.1.4, 3.3, and 6.3). In unsupervised learning, the input pattern is usually compared with an internally generated prototype pattern or with some of the previous stimuli. Hence, these networks typically require some architecture for measuring how similar a pattern is to previously encountered ones – in other words, for detecting familiarity or novelty. Computer scientists and engineers sometimes refer to unsupervised learning as *clustering*; that is, finding categorical order in patterns that are presented. Supervised learning they refer to as *classification*, that is, fitting presented patterns into existing categories.

Both unsupervised and supervised learning models are important in the theory of how biological organisms categorize. We make many quick decisions to place items, both sensory stimuli and abstract entities, into either existing

or novel categories, too quickly to depend on supervision. The networks to be discussed in the remainder of this chapter, many of which build on simpler structures like those described in Chapters 3 and 4, model *self-organization*; that is, making sense out of a large amount of incoming data. Yet, over the course of a person's or other animal's lifetime, there is a need to learn from other organisms or the environment some categorizations that self-organization alone will not discover. Hence, Chapter 8 is devoted to supervised categorization models, many of which rely on error correction.

The supervised/unsupervised split between chapters is made for ease of exposition but many widely used classes of algorithms, such as adaptive resonance, back propagation, deep learning, and brain-state-in-a-box, are applied to both supervised and unsupervised learning. Hence, some of the narrative in both chapters transitions between the supervised and unsupervised modes.

7.2. Self-Organization and Unsupervised Categorization Models

7.2.1. Competitive Learning and Self-organizing Maps

The term *competitive learning* is in general usage for multilevel networks that combine associative and competitive principles, including most of the networks discussed in this section and the previous one (see Grossberg, 1987a, for a discussion). The same term is sometimes used more specifically for a subclass of this type of network developed by Rumelhart and Zipser (1985). These researchers studied a simple system capable of detecting pattern regularity and illustrating some basic competitive learning principles.

The Rumelhart–Zipser model is based on a multilayer architecture. The lower level consists of input units (feature detectors), and inputs are treated as binary patterns activating some of these nodes. Nodes in succeeding layers group into "clusters"[1] and there is winner-take-all competition within a cluster. Winning nodes then send signals up to the next layer. There is no feedback from higher to lower layers.

The weights of connections from lower to higher layers change according to a rule similar to the synaptic conservation rule of Malsburg (1973). Only connections to winning units in clusters are modified. The connection weights w_{ij} to a given unit j, from units at the next layer below, add up to a fixed amount, and a proportion of the weight shifts from inactive to active pathways; the equations for this process are given at the end of this chapter. Each cluster of nodes classifies the stimulus set into as many groups as there are units in the cluster. If the arriving inputs fall into "natural clusters," the clusters of network nodes tend to find them. If there are no natural clusters, responses of those

nodes can at times become oscillatory or chaotic rather than converging to an equilibrium.

Rumelhart and Zipser (1985) applied the competitive learning algorithm to a variety of binary patterns – letters, horizontal and vertical lines, and "dipoles," which activated exactly two neighboring input units.[2] In one experiment, the stimuli are letter pairs drawn from the set {AA, AB, BA, BB}. Owing to randomness in the initial weights, the units in higher layers develop activation patterns that correspond either to A or B in the first serial position, or else to A or B in the second position. Moreover, on any given run, only one of the two positions is preferentially detected.

Another of Rumelhart and Zipser's experiments draws letter pairs from the set {AA, BA, SB, EB}. In this case, all higher level units become detectors of the letter in the second position; that is, they learn to respond to either the class {AA, BA} or the class {SB, EB}. This is striking because "A" and "E" are similar in their dot patterns as represented, as are "B" and "S." But the network completely ignores these similarities in favor of identity in the second position. In other experiments where there is no repetition of letters at either position, the responses of higher level units do not reach an equilibrium.

The family of competitive learning models also includes Kohonen's *self-organizing map* (*SOM*), also known as the *self-organizing feature map* (*SOFM*) (Kaski & Kohonen, 1994; Kohonen, 1982, 1984/1995, 1993, 1997). The SOFM in general is a device for teaching a network to represent the properties of the inputs it repeatedly receives, in the manner of Malsburg (1973). Nodes in this type of network start out representing random weight vectors, and an updating algorithm moves those weight vectors progressively closer to the inputs it receives. A good description of how it works is given in Mercado, Myers, and Gluck (2001):

> (1) Choose an input vector at random from the set of all input vectors; (2) measure the similarity between this input vector and all the weight vectors; (3) find the node with the most similar weight vector; (4) update the weights of this node and nearby nodes so that they are more similar to the input vector; and (5) continue this process until all the inputs have been presented to the map.
>
> (p. 40)

The SOFM has been widely used in a variety of computer science applications including optimization problems. It has also, more sporadically, provided the basis for some models of topographic organization in the visual system (Sirosh & Miikkulainen, 1994, 1997) and the auditory system (Mercado, Myers, & Gluck, 2000, 2001).

7.2.2. Adaptive Resonance Theory

Most of the topographic map models discussed in Section 7.1, and the models of Kohonen (1982) and Rumelhart and Zipser (1985), are pure feedforward, with primary receptor layers projecting to higher processing layers but not the reverse. Yet, recall from Section 7.1.3 the argument of Grossberg (1976b) that pure feedforward coding or categorization could be unstable. This led to the idea in Grossberg (1976c) of adaptive resonant feedback between two layers of nodes, corresponding to the extensive feedback connections in the visual system (cortex to lateral geniculate and V4 to V1 and V2; see, e.g., Felleman & Van Essen, 1991).

The adaptive resonance theory (ART) first developed by Grossberg (1976c) is best introduced in the article of Carpenter and Grossberg (1987a), which describes the *ART 1* model for classifying binary (0 or 1 to each node) inputs. The most important modifications of this algorithm for classifying analog inputs (running over a range, typically between 0 and 1) are *ART 2* (Carpenter & Grossberg, 1987b) and *fuzzy ART* (Carpenter, Grossberg, & Rosen, 1991a, 1991b).

Figure 7.8 illustrates the basic structures of ART 1. The F_1 layer is assumed to consist of nodes that respond to input features, analogous to cell groups in

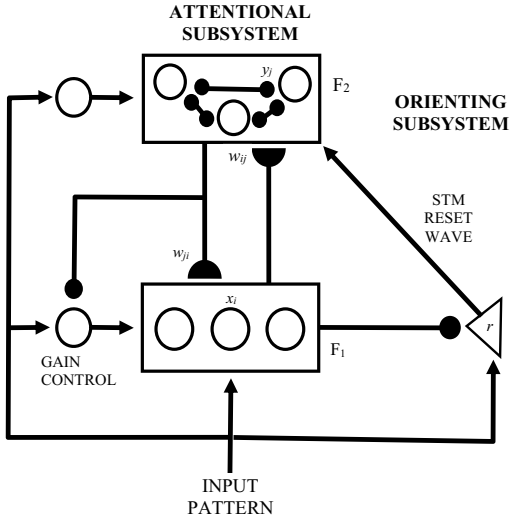


FIGURE 7.8 ART 1 architecture. Short-term memory at the feature level F_1 and category level F_2, and bottom-up and top-down interlevel long-term memory traces, are modulated by other nodes. The orienting system generates a reset wave to F_2 when bottom-up and top-down patterns mismatch at F_1, that is, when the ratio of F_1 activity to input activity is less than a *vigilance levelr*. This wave tends to inhibit recently active F_2 nodes. Functions of gain control nodes are described in the text.

Source: Adapted from Carpenter & Grossberg, 1987a, with permission of Academic Press.

a sensory area of the cerebral cortex. (The detailed structure of how inputs are processed at F_1 is disregarded in this network but is considered in ART 2.) The F_2 layer is assumed to consist of nodes that respond to categories of F_1 node activity patterns. Synaptic connections between the two fields are modifiable in both directions, according to two different learning laws.

The F_1 nodes do not directly interact with each other, but the F_2 nodes are connected in a recurrent competitive on-center off-surround network (see Figure 4.3). As discussed in Chapter 4, such competition is a common device in neural networks, inspired by visual neurophysiology, for making choices in short-term memory. In this version, the simplest form of choice (winner-take-all) is made: only the F_2 node receiving the largest signal from F_1 becomes active. To compute the signal received by a given F_2 node, the activity of each F_1 node in response to the input pattern is weighted by the strength of the bottom-up synapses from that F_1 node to the given F_2 node, and all these weighted activities are added.

Inhibition from the F_2 field to the F_1 field (via the gain control node) serves two related purposes. First, it counteracts excitatory F_2-to-F_1 connections that might otherwise lead to a spurious percept (hallucination) when one thinks about a category. Second, it shuts off most neural activity at F_1 if there is mismatch between the input pattern and the active category's prototype. Only with a sufficiently large match are enough of the same F_1 nodes excited by both the input and the active F_2 category node, which is needed to overcome nonspecific inhibition from F_2.

If match occurs, then F_1 activity is large because many nodes are simultaneously excited by input and prototype. Then F_1 inhibits the activity of the node A representing the orienting subsystem. This stabilizes the categorization of the given input pattern in the given F_2 node. If mismatch occurs, by contrast, F_1 activity is not sufficient to inhibit A, which thereby becomes active. The A node activity leads to F_2 reset which shuts off the active category node as long as the current input is present. The F_2 node receiving the next largest F_1 signal is then tested, and the process repeated. The exact criterion for mismatch is that the ratio of F_1 activity to total input intensity be less than some prescribed parameter. That is, if $[\mathbf{I}]$ is the number of active pixels in the (binary) input pattern, and $[\mathbf{X}]$ the number of F_1 nodes active after combined input and prototype presentation, then mismatch is said to occur if $[\mathbf{X}]/[\mathbf{I}] < r$ for some positive constant r, which is called the *vigilance* of the network. $\mathbf{X}$ can also be thought of as the number of 1s that are common at the same locations between the input pattern $\mathbf{I}$ and the pattern $\mathbf{w}$ of top-down weights from the active category.

The short-term memory (STM) and long-term memory (LTM) equations for this network, shown at the end of this chapter, incorporate all these effects along with two additional rules. The *2/3 rule* says that an F_1 node must be activated by at least two out of three signal sources if it is to generate suprathreshold

output signals. These sources are outside inputs, gain control (activated by inputs but inhibited by F_2), and the top-down signal from the active category node. The *Weber Law rule* says that LTM size should vary inversely with input pattern scale. This rule is designed to prevent a category node that has learned to code a particular binary pattern (1s in particular locations) from also coding every superset pattern (a pattern that has 1s in those same locations and some others). Figure 7.9 shows an example of coding by ART 1; note the importance of the 2/3 rule for code stability.

One of the controversies among cognitive psychologists studying categorization is whether categorization decisions are primarily based on prototypes or on *exemplars* (e.g., Kruschke, 1992; Nosofsky & Zaki, 2002; Smith & Minda, 1998). An exemplar of a known category is defined as an instance of a category member that has been previously experienced; hence, exemplar theories posit that a new input is compared with previously experienced exemplars of a category to decide how closely it fits that category. A prototype of a known category is defined as some kind of weighted average of previously experienced exemplars; hence, prototype theories posit that a new input is compared with this category average to see how closely it fits. Grossberg, Carpenter, and Ersoy (2005) note that exemplar models have the advantage when it comes to capturing familiar patterns, whereas prototypes have the advantage of capturing novel variations of familiar patterns. Using some of the same data previously simulated by exemplar and prototype models, these researchers argued that a version of ART combines the advantages of both types of models, though their model builds prototypes in a different manner (using top-down synaptic weights) than typical prototype models such as that of Smith and Minda (1998). We return to consideration of exemplars and prototypes in Section 7.3.

The categorization shown in Figure 7.9 is an example of the ability of the ART network to learn novel patterns without forgetting old ones. The issue of learning the new while retaining the old is a common concern of cognitive psychologists studying memory. This issue has been given various names: *stability-plasticity dilemma* (Carpenter & Grossberg, 1987a); *catastrophic interference* (McCloskey & Cohen, 1989); and *catastrophic forgetting* (French, 1999).

Neurophysiological support for the basic theoretical constructs of ART has been obtained from a series of experiments over the years by Robert Desimone and his colleagues on visual attention and responses of cells in different parts of the visual cortex (see Section 5.3.2), which led Desimone and Duncan (1995) to develop the notion of biased competition. In particular, Reynolds, Nicholas, Chelazzi, and Desimone (1995) and Reynolds, Chelazzi, and Desimone (1999) showed how spatial attention can protect macaque V2 and V4 cells from the influence of unattended stimuli.

The roles of mismatch and reset in ART receive some support from event-related potential (ERP) studies reviewed by Banquet and Grossberg (1987). These ERP data deal with an auditory paradigm with alternate presentations of two tones that differ in frequency, one tone more probable than the other. Hence, a top-down expectation develops for the more commonly presented tone and the rarer tone is treated as an oddball (see Section 5.3). Various ERP components (the processing negativity, positivity at 120 ms, part of the negativity at 200 ms, and part of the positivity at 300 ms) are higher in amplitude when the oddball tone is presented. Banquet and Grossberg conjectured that those components could reflect the function of the orienting system in an ART network such as in Figure 7.8. The orienting system has been tentatively identified with the hippocampus or some related area (Grossberg, 1984), based on data showing that hippocampectomized animals do not orient to novel stimuli (O'Keefe & Nadel, 1978, p. 250).

7.2.3. Continuous Versions of Adaptive Resonance: ART 2 and Fuzzy ART

The binary, all-or-none nature of the patterns classified by ART 1 is a crude description of the types of patterns that are input to actual brains, both at sensory and higher cognitive levels. Yet the fundamental ideas of adaptive resonance theory – match, mismatch, vigilances, reset, resonance – generalize to more realistic analog (i.e., continuous-valued or grayscale) patterns. Two of the adaptive resonance architectures for processing and classifying analog patterns are known as *ART 2* and *fuzzy ART*.

The ART 2 network of Carpenter and Grossberg (1987b) builds on the ideas of ART 1 with two layers and modifiable synapses in both directions, but adds several sets of processing nodes at the F_1 layer. These extra nodes are designed to contrast-enhance significant parts of the pattern and suppress noise, according to general principles developed in Grossberg (1973).

Although the preprocessing is more complex in ART 2 than in ART 1, the learning laws are simpler. Since supersets are not an issue with analog patterns, the Weber Law rule is dispensed with and the top-down and bottom-up LTM equations are the same (although top-down weights are initially 0, while bottom-up weights are initially random and positive). The matching criterion for ART 2 is also different, since $[X]$ and $[I]$ no longer make sense. In ART 2, the quantity to be compared with the vigilance value is the cosine of the angle between vectors that represent input and prototype patterns.

A simpler way to handle analog patterns in an ART network is *fuzzy ART*, introduced by Carpenter, Grossberg, and Rosen (1991a) and used in many current engineering and computing applications. Fuzzy ART is designed to generalize the mathematical operations of ART 1 from the binary to the analog case, including the criteria for category selection, reset, and resonance. Hence,

the binary selection algorithm used in ART 1 comes out as a special case of fuzzy ART in which the patterns consist solely of 0s and 1s.

Since the inputs to the F_1 layer of fuzzy ART can range between 0 and 1, comparison of input and prototype patterns is done at both the low and high ends of the range. This is done by a technique called *complement coding*, that is, feeding to the layer an actual input pattern followed by its "complementary" pattern whose values are 1 minus the original values (e.g., following the pattern [.8 .4 1] by [.2 .6 0]). In terms of sensory processing, Carpenter et al. noted, this is analogous to including the operations of receptors that are turned off by specific stimuli as well as those turned on by these stimuli; biological visual systems contain both ON and OFF cells (Kuffler, 1953).

The combined operation of complement coding and fuzzy sets works as follows. A complement-coded pattern **I** at F_1 is compared with a top-down weight pattern **w**. Intersecting of **I** with **w** as in ART 1 is generalized to a *fuzzy intersection* which takes the minimum of corresponding components of the two vectors. The value of [**X**] used in the matching criteria is then the sum of these fuzzy intersections over all components, and this is divided by [**I**] (now interpreted as the sum of all the input components) and compared with the vigilance value *r*.

The family of ART networks has several advantages over competing neural networks for classification. First, ART networks exhibit considerable stability. Theorems proved in Carpenter and Grossberg (1987a) show that (in ART 1 at least) after a given collection of input patterns has been learned, the categorization of those patterns is not perturbed by an arbitrary barrage of new inputs. Second, the category prototypes against which an input is tested change over time to reflect the type of patterns that are most frequently observed in the environment. Third, the model allows for influences on the feature and category layers from other subsystems external to these layers, such as the attentional and orienting systems. Hence, it is part of an interconnected set of theories of many other cognitive processes (see, e.g., Grossberg, 1982, 1987b, 1988; Levine, 1983). The self-organizing map (Kohonen, 1984/1995, 1997), used in unsupervised categorization, and the back propagation algorithm (Rumelhart, Hinton, & Williams, 1986), used in supervised categorization (see Section 8.1), are fairly successful in reproducing cognitive data but are not founded in principles that embed categorization in a network for attention, learning, and memory.

Recent network extensions of ARTMAP, the supervised version of ART (see Section 8.4), have also made the distinction between ambiguous and unambiguous classifications, with "I don't know" being an allowable response at certain stages of the algorithm. In particular, Carpenter and Ross (1995) developed a network called ART-EMAP, where ARTMAP was synthesized with a process of spatial and temporal evidence accumulation. At one of the network's stages, a decision criterion is added which delays identification of

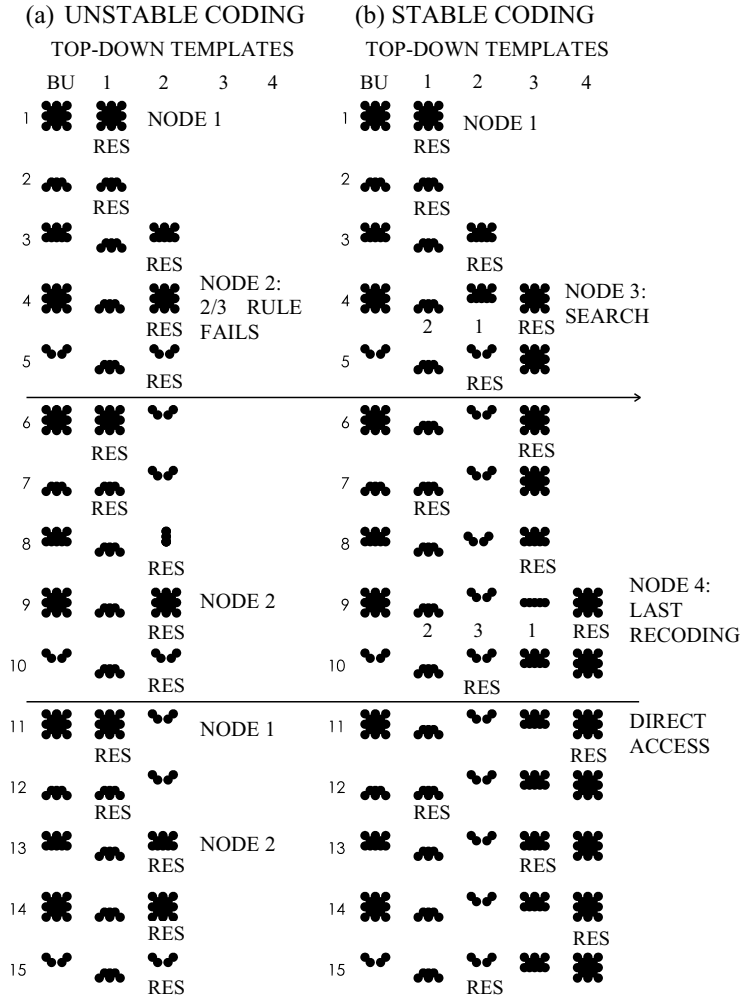


FIGURE 7.9 Categorization of binary patterns by ART 1. The same input sequence, four patterns *A*, *B*, *C*, and *D* in the order *ABCAD*, is presented repeatedly in both (a) and (b). In (a), top-down inhibitory gain control (see Figure 7.8) is weak, and the 2/3 rule is violated. This leads to ceaseless recoding of A. In (b), after some initial recoding, all patterns resonate (as shown by the symbol "RES") in distinct stable categories.

Source: Reprinted from Carpenter and Grossberg, 1987a, with permission of Academic Press.

ambiguous objects when predictions are made with sufficiently low confidence. This network was particularly applied to three-dimensional object recognition, and multiple views eventually resolved the ambiguity. Another ART-based network that constructs classifications of 3-D objects from multiple 2-D views is the VIEWNET of Bradski and Grossberg (1995).

ARTMAP deals successfully with nonuniformity in the knowledge base. Other modelers have added to the ART framework an explicit mechanism for selective attention (see Chapters 4 and 6 of this book) to some of the features in the input space. Weingard (1990) developed a "self-organizing analog field" model that adds attentional modulation to an ART network, modulation that allows for dynamic setting of the vigilance parameter. Attentional modulation of ART also appears in models of frontal lobe executive function (Levine & Prueitt, 1989) and multiattribute decision-making (Leven & Levine, 1996; Levine, 2012), both discussed in Chapter 9.

Carpenter and Grossberg (1989) extended the ART architecture to include multiple levels of nodes, denoting increasing levels of abstraction in the network's categorizations. Also, that article includes an explicit mechanism for category search and reset, which is absent in their 1987 simulations. The reset mechanism is based on some known qualitative properties of chemical transmitter storage, utilization, and release by neurons. This work is further extended in the *ART 3* model of Carpenter and Grossberg (1990). Future extensions could relate these levels to the increasing levels of abstraction as one moves further forward in the prefrontal cortex (Christoff & Gabrieli, 2000; Christoff et al., 2009).

Carpenter and Tan (1995) and Kant (1995, 1996) both added to ART networks mechanisms for *interpreting* the categorizations that the system made. That is, placing a set of patterns into the same category is described in terms of a set of rules for what constitutes membership in that category. Carpenter and Tan applied this rule making to medical diagnosis, and Kant applied it to understanding decisions made by both customers and experts regarding bank savings schemes (see Section 9.5 for other decision-making models). These recent articles illustrate that the influences on feedback between ART's F_1 and F_2 layers can be widely varied to reflect the action of other important neural subsystems.

More detailed connections of ART with brain processes appear in later models that integrate the top-down attentional processes of ART with other sensory processes, in particular the models called *LAMINART* (Grossberg, 1999; Raizada & Grossberg, 2001) and *SMART* (Grossberg & Versace, 2008). LAMINART integrates the ART layering with the known laminar structure of the visual cortex, partly in order to solve a paradox between the requirements of visual attention and of preattentive grouping; that paradox and its solution are discussed under visual models in Section 9.2.

SMART (an acronym for "synchronous matching adaptive resonance theory") builds on that same laminar structure and adds several model features linking adaptive resonance processes to interactions between the thalamus and cortex. These features include the following:

- Match can occur at several points in the visual system. For example, at the lateral geniculate there can be a match between bottom-up retinal inputs and top-down modulatory expectations from the primary visual cortex (V1). At the pulvinar there can be a match between bottom-up inputs from V1 modulatory inputs from a higher visual level, V2.
- The resonant state caused by matching at either level is related to spike synchronization in the gamma frequency range (20–70 Hz). This frequency allows for spike timing–dependent plasticity (STDP).
- In case of mismatch, by contrast, there are slower beta frequency (4–20 Hz) oscillations, and STDP is disabled at this lower frequency.

7.2.4. Edelman and Neural Darwinism

Edelman (1987), using a theoretical development based partly on analogies with immunology, proposed that the nervous system performs a selection between pattern encodings in which the "fittest" encodings survive. (This is the meaning of his book's title, *Neural Darwinism*.) Reeke and Edelman (1987) describe simulations of a categorization network incorporating this selection idea.

Although Edelman's pattern selection idea was sometimes regarded as a new view of the brain (Rosenfield, 1988), his philosophical approach is not essentially different from the approaches of many other neural modelers (Levine, 1988). Darwinian selection among encodings is a striking metaphor for the much older idea of competition between neurons and neuron groups. Moreover, modifiability of visual or somatosensory maps, which forms the biological basis of Edelman's arguments, is also the basis for modifiable interlayer synapses proposed by Malsburg, Grossberg, Bienenstock et al., Linsker, and others (see Section 7.1).

Edelman and his colleagues may have made a greater contribution in proposing a theory about which functional groups of neurons in living animals correspond to "nodes" in a neural network. The basic idea of this theory (see Edelman, 1987, and Levine, 1988, for details) is that chemical markers, called *cell adhesion molecules* (CAMs), determine boundaries between groups of neurons. Sensory inputs during development alter the distribution of CAMs, in a way that has not been fully described. In adult life, there is less shifting of cell group boundaries, and the main mechanism for change, in this theory, is synaptic modification via a non-Hebbian associative learning law (see Edelman & Reeke, 1982; Finkel & Edelman, 1985). In this law, modification of a synapse depends not only on activities of the two neurons connected by that synapse but on activities of all neurons in a group.

An example of the categorization network based on Edelman's scheme includes two subnetworks, called "Darwin" and "Wallace," which perform

different processing stages. The Darwin part of the network includes "recognizers" or feature detectors. On the Wallace side is an abstracting network that responds to patterns of activity but is insensitive to translation or rotation. This network is discussed further in Section 7.3, as it bears on the problem of translation invariance.

7.3. Translation and Scale Invariance

Most of the categorizations of visual patterns we have discussed so far involve patterns of input activities at defined locations. But, in reality, we see the letter A, for example, as an A regardless of its location in the visual field. Minsky and Papert (1969) and many others have noted that this problem of *translation invariance* adds considerable difficulty to modeling pattern categorization.

Some categorization algorithms can be modified in ways that partly solve the problem of translation invariance (and, in some cases, rotation and scale invariance as well). One example is the averaging of weight changes that enables the back propagation network to distinguish a T from a C regardless of position or rotation. Another example is the mechanism of Reilly, Cooper, and Elbaum (1982) whereby multiple prototypes can map into one classification node. Carpenter, Grossberg, and Mehanian (1989) developed an invariant recognition mechanism that adds to an ART 2 network a log-polar Fourier filter (Casasent & Psaltis, 1976; Szu, 1986) for achieving translation and rotation invariance. This mechanism assumed that the pattern to be classified had previously been segmented, using a variant of the boundary contour system of Grossberg and Mingolla (1985b).

The *Neocognitron* of Fukushima (1980) and Fukushima and Miyake (1982) also achieves translation invariance by taking a previously developed multilevel hierarchical classifier that is position-dependent (the *Cognitron* of Fukushima, 1975) and adding further levels. Fukushima drew his inspiration from the physiological findings of Hubel and Wiesel (1962, 1965), discussed in Section 7.1, that the cat or monkey visual cortex contains a hierarchy of cell types ranging from *simple* to *complex* to *hypercomplex*. In real animals, as one ascends this hierarchy, the highest layers tend to respond most selectively to complicated pattern features but least selectively to location.

Fukushima's model mimics the hierarchy developed by Hubel and Wiesel, and extends it beyond hypercomplex cells to cells in association areas of the cortex. As shown in Figure 7.10, this model includes modifiable connections between successive hierarchical layers, as do most of the models of this chapter. The cells in each layer are organized so as to have the same receptive field structure but at different positions.

The Neocognitron has been most extensively used in computer vision, although it remains a useful model of visual cortical interactions. Some of the recent advances in Neocognitron architecture are reviewed in Fukushima

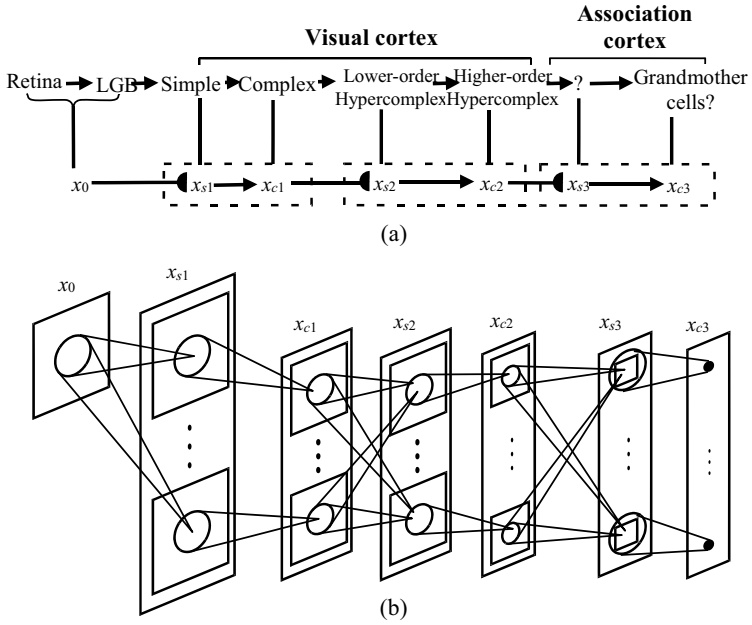


FIGURE 7.10 Neocognitron architecture. (a) Schematic of layers in the network's hierarchy and their rough correspondence to the actual biological hierarchy of Hubel and Wiesel. LGB stands for the lateral geniculate body of the thalamus. Grandmother cell is a colloquial term for a neuron that fires only in response to a very specific pattern. (b) More detailed diagram of the network's interconnections and their locations within layers.

Source: Adapted from Fukushima, 1980, with permission of Springer-Verlag.

(2013). Many of the advances involve adding top-down connections to achieve such visual functions as selective attention, recognition and completion of partly occluded patterns, and restoring occluded contours.

Other details of visual cortical anatomy (see Section 5.4.2) inspired the design of the *What-and-Where filter* by Carpenter, Grossberg, and Lesher (1998). This network is based on the findings that the visual cortex contains two parallel pathways for computing what an object is and where it is located (Ungerleider & Mishkin, 1982). In the Carpenter et al. network, there is a Where filter that includes an array of orientation detectors followed by competitive interactions between position, orientation, and size scales. These are used to determine a maximum likelihood position, orientation, and size for a given object in a scene. Then these spatial parameters are filtered out to yield a What image that is independent of the three Where variables. This invariant image can thereby be used to place the object in the same category as other objects related to it by translation, rotation, and dilation.

A different approach to translation and rotation invariance (and scale invariance, as well) is seen in the "Wallace" part of the network of Reeke and Edelman (1987). Reeke and Edelman (1984) describe this approach as probabilistic matching (p. 188). More specifically:

> Wallace begins with a tracing mechanism designed to scan the input array, detecting object contours and tracing along them to give *correlations of features* that . . . respond to some of their characteristics, such as junctions of various types between lines.
>
> (p. 189, my italics)

Nigrin (1993, Chapters 6 and 7), like Fukushima, achieved translation invariance through a hierarchy of levels, but his architecture requires a somewhat smaller number of nodes than does the Neocognitron. In Nigrin's network, successive levels of the hierarchy contain fewer and fewer nodes, and therefore the location of the input object becomes progressively centered. This is achieved through cooperation and competition not between nodes, as in most of the networks for vision in Chapter 4 of this book, but between interlevel *connections*.[3] This type of centering can lead to identical classifications of two objects that are in different parts of the visual field but consist of the same patterns relative to their centers. Nigrin (1993, Chapter 7) also shows how a variant of his architecture can accomplish scale invariance.

Equations for Networks in Chapter 7

Detailed Description: Adaptive Resonance Theory

Adaptive resonance theory (ART) networks are a family including a wide range of precise architectures. The most important classifications within ART are unsupervised (self-organizing) versus supervised and analog (fuzzy) versus binary. ART 1 of Carpenter and Grossberg (1987a), the unsupervised version which classifies binary patterns, captures much of the essence of this family of networks.

The computations performed by ART 1 on each input pattern can be subdivided into distinct stages: (1) input and bottom-up filtering; (2) resonant category choice; (3) updating of top-down weights; (4) reset for arrival of the next input. First, the pattern arrives at F_1 (the feature level in Figure 7.8) and is filtered through the bottom-up synaptic weights w_{ij}. As in the coding schemes of Amari and Takeuchi (1978), Grossberg (1976a),

Malsburg (1973), and others, the node at the category layer F_2 receiving the largest bottom-up signal is tentatively chosen. Second, the input is compared with the top-down prototype, which is encoded by the set of synaptic weights from the chosen node. If there is a sufficient match (using the vigilance criterion), the choice is made more permanent; this is the *resonance*, and it is called *adaptive* because the prototype resonating with the input reflects the learning of previous inputs by the node at F_2.

If the match is insufficient, either another node with a prototype pattern is found to produce resonance, or an "uncommitted" F_2 node – that is, one from which the synaptic weights are all 0s so far – is dedicated to the new pattern. Either way, the F_2 node finally chosen now influences both F_1 node activities and F_2-to-F_1 synaptic weights, via the function $f(x_j)$ for that node (see Equations (7.22) and (7.24) below). Finally, the input is shut off and the F_1 activities all decay to close to 0 to receive the next input pattern.

Carpenter and Grossberg (1987a, pp. 60–63) describe this sequence of computational steps as an example of a more general type of ART activation sequence

$$I \rightarrow X \rightarrow S \rightarrow T \rightarrow T \rightarrow Y \rightarrow U \rightarrow V \rightarrow X^* \tag{7.4}$$

The letters in (7.5) are explained as follows. I is the input pattern, which is 1 or 0 at each F_1 node. In general, this pattern is transformed into a pattern X of activation of the nodes; in the case of typical ART 1 computations, X quickly becomes the same vector as I. S is the set of output signals from F_1. They are filtered through the bottom-up weights, and T is the ensuing pattern of inputs to the F_2 layer.

Contrast enhancement within the on-center off-surround field (see Chapter 4) of F_2 nodes transforms T into a different ("sharper") activation patterns among the nodes at F_2. In the case of a typical ART 1 network, the contrast enhancement is of the winner-take-all variety. That is, only the one F_2 node representing the (old or new) category in which the choice is made to classify the input remains activated. U is the output of Y, and the $U \rightarrow V$ transformation is the reverse of the $S \rightarrow T$ transformation; that is, U is filtered through the top-down weights from the chosen F_2 node (which reflects an average of previously perceived members of the appropriate category) to produce a pattern V at F_1.

The pattern V is called a *top-down template* or *learned expectation*: it is heavily influenced by learning of previous input patterns. The combination of input pattern I and template pattern V leads to a combined activation pattern X^* at F_1. It is the number of 1s in this combined pattern X^* that needs to be at least a certain fraction (the vigilance) of the number of 1s in the input-based pattern X for resonance to be said to occur.

Of the four computational stages listed earlier, the LTM weight updating (Stage 3) takes far more time steps than the other three steps (bottom-up filtering, category choice, and preparation for the next input).

Malsburg's and Grossberg's Development of Feature Detectors

The equations of Malsburg (1973) for excitatory and inhibitory node activities $x_{E,k}$ and $x_{\mathrm{I},k}$ are of the form

$$\frac{dx_{E,k}}{dt} = -a_k x_{E,k}(t) + \sum_i p_{ik} x^*_{E,i}(t) + \sum_i s_{ik} A^*_i(t) - \sum_i q_{ik} x^*_{I,i}(t) \tag{7.5}$$

$$\frac{dx_{I,k}}{dt} = -a_k x_{I,k}(t) + \sum_i r_{ik} x^*_{E,i}(t)$$

where the *s denote signals from other nodes (excitatory nodes in the case of E_i*, inhibitory nodes in the case of I_{i}*, and retinal afferents in the case of A_{i}*). These equations had previously been used, for a different purpose, by Grossberg (1972d).

Malsburg's associative law is

$$s_{ik}(t+1) = s_{ik}(t) + hA^*_i E^*_k \tag{7.6}$$

that is, retinal-to-cortical connection weights grow with the cross-correlation of presynaptic retinal afferent activity and postsynaptic excitatory cortical node activity. The other connection weights, p_{ik}, q_{ik}, and r_{ik}, do not change over time.

In the model of Grossberg (1976b), the activities x_{1i} of the V_i obey the nonrecurrent STM equations

$$\frac{dx_{1i}}{dt} = -Ax_{1i} + \left(B - x_{1i}\right) I_i - x_{1i} \sum_{k \neq i} I_k \tag{7.7}$$

where I_i are the inputs to the "retinal" nodes F_{1i}. Note that Equations (7.7) are identical to (4.10) with x_{i} replaced by x_{1i}. To store patterns of retinal node activity in cortical STM, the activities x_{2j} of the F_{2j} are governed by the recurrent equations

$$\frac{dx_{2j}}{dt} = -Ax_{2j} + \left(B - x_{2j}\right)\left[f\left(x_{2j}\right) + I_{2j} \right] - x_{2j} \sum_{k \neq j} f\left(x_{2k}\right) \tag{7.8}$$

where f is a signal function (typically sigmoid) and I_{2j} represents the excitatory input to F_{2j} from the F_1 level. Equations (7.8) are a subcase of Equations (4.1), which was developed in Grossberg (1973).

Grossberg (1976b) listed many possible laws for long-term memory (LTM) and for how the LTM traces affect the interlevel signals I_{2j} of (7.9). In general,

I_{2j} is a weighted sum of the form $\Sigma_i \theta_i w_{ij}$, where the θ_i are the *normalized* activities at the F_1 level and the w_{ij} are the weights of the corresponding connections from F_1 to F_2. The w_{ij} in turn obey equations such as

$$\frac{dw_{ij}}{dt} = \left(-w_{ij} + \theta_i\right) x_{2j} \tag{7.9}$$

Equation (7.9) says that the weights of connections to cortical node j change only when short-term memory at that node is active, so that x_{2j} is nonzero. This equation bears a close resemblance to the outstar learning law (3.14), except that in the outstar weights of connections *from* a given node change only when short-term memory at the node is active. Indeed, the subnetwork of Figure 7.6 consisting only of a single cortical node and its connections behaves like a reverse outstar, and is frequently called an *instar*.

If a constant spatial pattern is presented through time to the network defined by Equations (7.8) and (7.9), a theorem in Grossberg (1976b) shows that this pattern is learned by the vector of connection weights to some F_2 node from the F_1 field. More precisely, recall from Section 7.1 that the total signal at time t to a given cortical node x_{2j} due to the retinal pattern $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ is

$$S_j(t) = \sum_{k=1}^{n} \theta_k w_{kj}(t) \tag{7.2}$$

Grossberg showed that if j is such that the value of S as defined by (7.2) is the largest, then the angle between the vector $\mathbf{w}_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$ of weights to the jth node and the input vector $\boldsymbol{\theta}$ is decreasing for all time and approaches 0. The result of Amari and Takeuchi (1978) also states that asymptotic weights to an F_2 node have a "winning dot product" with the original pattern. (Recall from Section 3.2 that the dot product of two vectors, of the same number of components, is found by multiplying the two vectors component by component and then summing the products.)

Rumelhart and Zipser's Competitive Learning Equations

The competitive learning algorithm of Rumelhart and Zipser (1985) is defined by changes in the interlevel weights w_{ij}. At each of Layers 2 and 3, node j receives at each time step a signal equal to $\Sigma_i x_i w_{ij}$, the sum taken over nodes i in the next lower layer, where x_i is 1 if node I is active and 0 otherwise. (The algorithm can be extended to arbitrarily many layers.) If the lower layer is Layer 1 (input units), activity or inactivity is determined by which nodes are excited by the input pattern. If the lower layer is Layer 2, activity or inactivity is determined by whether the ith node won or lost its intracluster competition at the last time step. Within each cluster of the top two layers in turn, the node receiving the largest (linearly weighted) signal wins and the others lose.

The w_{ij} in turn change at each time step according to a rule similar to the gamma system rule of Rosenblatt (1962). No change occurs in connections to losing nodes; hence $\Delta w_{ij} = 0$ if unit j loses. For connections to winning nodes, weight is shifted from inactive to active pathways. Each of the input pathways to a winning node gives up some proportion k (between 0 and 1) of its weight, and that weight is then distributed equally among the active input pathways (cf. Malsburg, 1973). Hence

$$\Delta w_{ij} = k\left[\frac{c_i}{n} - w_{ij}\right] \text{ if unit } j \text{ wins} \tag{7.10}$$

where n is the total number of active units at the next lower layer in the current pattern, and c_i is 1 if unit i at the lower layer is active and 0 otherwise. Note that

$$n = \sum_i c_i \tag{7.11}$$

Equations (7.10) and (7.11) together imply that $\Sigma_i w_i$ remains constant for a given j. In the simulations done by the authors, that sum is kept equal to 1.

Adaptive Resonance Equations

In the ART 1 network, the STM activity of the ith F_1 node is denoted by x_i and the STM activity of the jth F_2 node by x_j. The convention that the subscript i relates to F_1 and j to F_2 is observed throughout. Hence, w_{ij} values represent bottom-up LTM strengths (synaptic weights) and w_{ji} values represent top-down LTM strengths.

The STM traces at F_1 are assumed to change quickly under the influence of shunting (multiplicative) excitation from outside inputs and from top-down signals, and shunting inhibition from F_2 (mediated by the gain control node). Hence

$$\varepsilon \frac{dx_i}{dt} = -x_i + \left(1 - A_1 x_i\right)\left(I_i + D_1 \sum_j f\left(x_j\right) w_{ji}\right) - \left(B_1 + C_1 x_i\right) \sum_j f\left(x_j\right) \tag{7.12}$$

where ε is small (.05 or .1). The function f of Equation (7.12) is defined by f = 1 if node j of F_2 is active and 0 if node j is inactive. Only one node of F_2 is active at a time. After an input comes in, the jth F_2 node receives a bottom-up signal equal to $T_j = D_2 \Sigma_i h(x_i) w_{ij}$, where D_2 is a positive constant and h is the Heaviside (unit step) function. The category chosen to be active (that is, to be tested for match or mismatch between bottom-up and top-down patterns) is the one for which T_j is the largest.

The activities x_j of the F_2 nodes are in turn governed by the equations

$$\varepsilon \frac{dx_j}{dt} = -x_j + \left(1 - A_2 x_j\right)\left(g(x_j) + T_j\right) - \left(B_2 + C_2 x_j\right)\sum_{k \neq j} g(x_k) \tag{7.13}$$

where g is a sigmoid function and T_j is as defined earlier. The summation in Equation (7.13) indicates lateral inhibition in an on-center off-surround field.

The LTM trace of the top-down pathway from w_j to w_i obeys the learning equation

$$\frac{dw_{ji}}{dt} = f(x_j)\left[-w_{ji} + h(x_i)\right] \tag{7.14}$$

where h is again the unit step function. The bottom-up pathway obeys an equation similar to (7.14) except that the synaptic decay term embodies the Weber Law rule. As discussed in Section 7.2, this rule is designed to prevent access to a category by supersets of the category prototype; this is achieved by selectively decreasing bottom-up signals from input patterns that activate large numbers of F_1 nodes. The design is achieved through competition between LTM traces, resulting in equations of the form

$$\frac{dw_{ij}}{dt} = Kf\left(x_j\right)\left[-E_{ij} w_{ij} + h\left(x_i\right)\right] \tag{7.15}$$

where

$$E_{ij} = h(x_i) + \left(\frac{1}{L}\right)\sum_{k \neq 1} h(x_k)$$

with K a constant, $L > 1$, and the sum taken over all F_1 indices k not equal to i. It can be shown that the other rule described above, the 2/3 rule, is satisfied by choosing the parameters of (7.13) such that max $(1, D_1) > B_1 > D_1$. Equations (7.14) and (7.15) guarantee that learning only occurs at synapses to or from active category nodes.

The ART 2 equations, which are not given here, also involve shunting excitation and inhibition in the manner of Equations (7.12) and (7.13). The shunting terms in the F_1 STM equations reflect influences from six extra sets of input processing nodes shown in Figure 7.10, designed to suppress noise that is characteristic of analog patterns. The LTM laws are simpler than in ART 1, because superset encoding is irrelevant when patterns consist of continuous values rather than just 1s and 0s. Hence the Weber Law rule is not used and top-down and bottom-up LTM equations are nearly the same.

The fuzzy ART algorithm is based on inputs **I**, which are m-dimensional vectors $(I_1, \ldots, I_M)$, with each component Ii being in the interval [0, 1]. The weight vector associated with category j (and with its corresponding node at F_2) is

$\mathbf{w}_j = (w_{j1}, \ldots, w_{jM})$. It is assumed that *both* top-down weights from and bottom-up weights to the jth category node equal that vector. At the start of the run, all those weights are set to 1, implying that the category node is uncommitted.

The three parameters that determine fuzzy ART dynamics are a choice parameter $\alpha > 0$; a learning parameter β between 0 and 1; and a vigilance parameter ρ between 0 and 1. After an input arrives, it is initially placed in the category that has the maximum value of the *choice function*

$$T_j = \frac{|\mathbf{I} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|} \tag{7.16}$$

where the fuzzy intersection operator $\wedge$ is defined by $(\mathbf{x} \wedge \mathbf{y})_i = \min(x_i, y_i)$, and the norm $|\cdot|$ by

$$|x| = \sum_{i=1}^{M} x_1$$

In the case of binary patterns, fuzzy intersection reduces to set intersection and norm to number of 1s in the pattern, and (7.28) reduces to the choice function for ART 1. The category with largest T_j is called J; in the case where more than one are tied for largest, the node with the smallest index is chosen.

The vigilance criterion for resonance is

$$\frac{|I \wedge w_j|}{|I|} \geq \rho \tag{7.17}$$

As in ART 1, mismatch reset occurs and a new category node is chosen if the vigilance criterion (7.18) is not met.

Finally, the weight vector $\mathbf{w}_j$ is updated according to the equation

$$\mathbf{w}_i^{(\text{new})} = \beta\left(\mathbf{I} \wedge \mathbf{w}_j^{(\text{old})} + (1-\beta)\mathbf{w}_j^{(\text{old})}\right)$$

Fast learning corresponds to $\beta = 1$ in the above equation.

Exercises for Chapter 7

**1. Consider the competitive learning model of Grossberg (1976b) defined by Figure 7.7. Let the activities x_{1i} of the F_1 layer encode the relative input intensities, i.e.,

$$x_{1i} = \frac{I_i}{\sum_j I_j} = \theta_i$$

Let $x_{2j} = 1$ if $S_j > \max(\varepsilon, S_k: k \neq j)$,
0 if $S_j < \max(\varepsilon, S_k: k \neq j)$,

where for $j = 1, 2, 3$, $S_j = \Sigma_i \theta_i w_{ij}$, $i = 1, 2$, and $\varepsilon = .05$. For each j, let the vector of synaptic weights to node j of F_2 be $\mathbf{w}_j = (w_{1j}, w_{2j})$, let $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and let $\mathbf{w}_j$ obey the equation

$$\frac{dw_j}{dt} = x_{2j}\left(-\mathbf{w}_j + \boldsymbol{\theta}\right)$$

with initial conditions $\mathbf{w}_1(0) = (1, 0)$, $\mathbf{w}_2(0) = (.4, .4)$, and $\mathbf{w}_3(0) = (1, 0)$. The current pattern is considered as being encoded in category j where j satisfies $S_j > \max(\varepsilon, S_k: k \neq j)$.

Choose a spatial pattern $\boldsymbol{\theta} = (y, 1-y)$ for some positive y not equal to .2 or .8.

Extra credit: Alternate presentations of 2 or more spatial patterns over time and see if stable coding remains true. That may need to be done by computer.

*2. Based on Kohonen self-organizing maps, use the algorithm given in Angéniol et al. (1988) for the classical Euclidean traveling salesman problem. Given a set of cities M defined by their positions in the plane, each city has coordinate (x_1, x_2). An approximate tour in our approach is given by a set of nodes N joined together in a one-dimensional ring. The algorithm has two steps, iteration step, and processing step. Each node j on the ring is characterized by the two coordinates (c_1, c_2) of the associated point in the plane. Each node is also related to its two neighbors in the ring; nodes $(j - 1) \bmod N$ and $(j + 1) \bmod N$. At the beginning of the process, only one node exists which is located at point (0, 0) in the plane, The number of nodes, N, grows subsequently according to a node creation process at each iteration. Surveying the city i comprises the following steps:

Step 1: Find the node j_c which is closest to city i: for each node j, compute its potential: $V_j = (x_1^i - c_1^j)^2 + (x_2^i - c_2^j)^2$ and determine the winning node j_c by competition: $V_j = \min(V_j)$. Now, node j_c is associated with city i.

Step 2: Move node j_c and its neighbors on the ring toward city i. The distance each node will move is determined by a function $f(G, n)$, where G is a gain parameter, and n is the distance measured along the ring between nodes j and j_c:

$$n = \min((j - j_c) \bmod N, (j_c - j) \bmod N)$$

The new position c_{new} of node j is defined by

$$c_{\text{new}} = c_{\text{old}} + f(G,n)\exp\left(\frac{-n^2}{G^2}\right)$$

where the gain G is discounted by a factor α at each iteration, that is, $G_{new} = (1 - \alpha)G_{old}$.

Each node j can be chosen by one city i in each iteration. If a node j was not a winner for any city during three complete surveys, that node is deleted. The evolving population of nodes, which sometimes has duplication, iteratively organizes toward a solution that may not be optimal (in the sense of shortest total distance) but is close to optimal.

Perform a simulation: create $N = 12$ cities whose coordinates are obtained from random variables uniformly distributed between –10 and 10. Let $M = 10$, $\alpha = 2$, and the initial value of $G = 5$. How many iterations are needed to complete the tour, that is, have a path that traverses all 12 of the cities? (Owing to randomness the answer will not always be the same.)

○3. How could the ART 1 network of Carpenter and Grossberg be modified to include a statement of the degree of certainty in the choice made? In particular, if a pattern just barely passes the vigilance criterion for more than one category at a time, how would "ambiguity detection" be built in? Note: *Recording* ambiguity is not the same as *resolving* ambiguity.

*4. Run a simulation of the ART 1 network, using Equations (7.24) to (7.27). For these equations, choose any parameter settings that obey the following inequalities from Table 1 of Carpenter and Grossberg (1987a):

$A_1 \geq 0$, $C_1 \geq 0$, $\max(1, D_1) < B_1 < 1 + D_1$, $0 < \varepsilon << 1$ (where "<<" is an imprecise term meaning "much smaller than"), K is close to 1, $L > 1$, $0 < z_{ij}(0) < L/(L - 1 + M)$, $1 \geq z_{ji}(0) > (B_1 - 1)/D_1$.

Let F_1 consist of a 5-by-5 "pixel" array, and let the patterns A, B, C, and D be the ones shown in Figure 7.11, in succession.

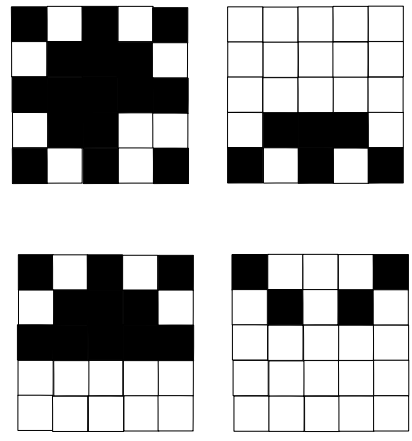


FIGURE 7.11 Input patterns used in the ART 1 simulation of Exercise 4.

(a) Following the simulations shown in Figure 6 of Carpenter and Grossberg (1987a), present these patterns repeatedly in the order *ABCAD*, with vigilance level $r = .8$. Show that after several repetitions, the patterns *A*, *B*, *C*, and *D* are eventually coded in separate, stable categories at F_2.

(b) Lower the value B_1 to a value below max (1, D_1), which means that the 2/3 rule is violated. Show that periodic recoding of pattern A can occur.

(c) Progressively lower the vigilance level *r* by increments of .1 until two or more of the patterns are coded in the same category, then until all four are coded in the same category.

Some Additional Sources

Coding in the Visual Cortex

Adorján, Barna, Érdi, and Obermayer (1999); Adorján, Levitt, Lund, and Obermayer, 1999; Song and Abbott (2001); Swindale (1980).

Adaptive Resonance Theory Networks

Exact Implementation via Differential Equations

Raijmakers and Molenaar (1997); Raijmakers, van der Maas, and Molenaar (1996).

Mathematical Theory

Georgiopoulos, Heileman, and Huang (1991, 1992); Heileman, Georgiopoulos, and Abdallah (1995).

Other Neural Networks for Unsupervised Categorization and Classification

Lo (2010).

Notes

1. Rumelhart and Zipser use the term "cluster" for groups of nodes, as opposed to the usage of that word described elsewhere in this chapter for groups of similar patterns.
2. Rumelhart and Zipser's "dipole" is unrelated to the gated dipole of Section 3.3.
3. Competition between connections appears to be Nigrin's invention. It may, however, provide a basis for synaptic conservation laws such as that of Malsburg (1973), and also possibly the maximum associability in the nonneural theory of Rescorla and Wagner (1972) discussed in Chapters 3 and 6. A somewhat similar mechanism is found in ART 3 (Carpenter & Grossberg, 1990).

8

MODELS OF SUPERVISED PATTERN AND CATEGORY LEARNING

> Thinking in its lower grades is comparable to paper money, and in its higher forms it is a kind of poetry.
>
> Havelock Ellis (*The Dance of Life*)

> To rule is easy, to govern difficult.
>
> Johann Wolfgang von Goethe

The use of neural networks for supervised learning of predetermined classifications dates back to the early work of Rosenblatt (see Chapter 2 of this book). His perceptron learning theorem (Rosenblatt, 1962) states that:

> Given an α-perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists; let all stimuli in W occur in any sequence, provided that each stimulus must reoccur in finite time; then beginning from an arbitrary initial state, an error correction procedure will always yield a solution to $C(W)$ in finite time.
>
> (p. 596)

Since Rosenblatt wrote, many other researchers have employed variants of the error correction or delta learning rule (e.g., Anderson & Murphy, 1986; Barto & Anandan, 1985; LeCun, 1985; McClelland & Rumelhart, 1985; Parker, 1985; Sutton & Barto, 1981, 1998; Werbos, 1974; Widrow, 1962; Widrow & Hoff, 1960). Rumelhart et al. (1986) showed how a multilayer network incorporating this rule can be applied to pattern classification.

8.1. The Back Propagation Network and PDP Networks

Recall the derivation at the end of Chapter 3 of the back propagation learning algorithm, whereby the delta rule for changing connection weights to output units generalizes to a rule for changing weights to hidden units. Within such a feedforward scheme, this method provides, in a sense, the most efficient weight changes for encoding the particular input–output mappings desired. (Indeed, before it was used for categorization the back propagation algorithm was employed in optimization problems; see Werbos, 1974, 1988.) At each iteration of this learning algorithm, the weights are set for a network of input, hidden, and output units, with the generic architecture shown in Figure 8.1.

The types of nonlinear input–output relationships that can be learned by back propagation (BP) networks are essentially arbitrary. This is one of the main reasons for the broad appeal of such networks. Some relationships that can be taught to BP cannot be taught to networks without hidden units, because these relationships map dissimilar inputs into similar outputs or vice versa. Neither can they be taught to networks whose activation functions are all linear, in which case hidden units provide no advantage. The best known example of such a relationship is the *exclusive OR* (XOR) mapping of binary variables (see Table 8.1). (It is to be noted that XOR is also difficult for some nonlinear unsupervised categorization networks such as adaptive resonance networks; see Stork, 1989a.)

Mathematically, this type of input–output relationship can be considered as an arbitrary mapping from an n-dimensional space of vectors to an m-dimensional space, where n is the number of input nodes and m the number of output nodes. (See Appendix 1 for discussion of n-dimensional vectors. The ARTMAP network to be discussed in Section 8.4 also can be treated as a device to map one multidimensional space to another.) Several analytical results (e.g., Hornik, Stinchcombe, & White, 1989) have shown that a back propagation network can essentially learn any such mapping if the numerical values of the output activations do not get arbitrarily large. If these mappings are interpreted as discriminations between perceptual classes, Sontag and Sussmann (1989) proved that back propagation networks can at least learn all the types of discriminations that can be learned by the perceptrons of Rosenblatt (1962).

The universality of the back propagation method has led to its use in a wide variety of computing and engineering applications. Perhaps the first of its applications to become widely known was reading aloud. The NETtalk algorithm of Sejnowski and Rosenberg (1986) and Rosenberg and Sejnowski (1986) employs a BP network to associate written language to spoken sounds. Many researchers have also applied BP to character recognition. Early BP simulations often involved learning of a single letter discrimination, such as the discrimination between the letters "T" and "C" (regardless of rotation) that was taught to this network by Rumelhart et al. (1986; see Exercise 3 of this

chapter). This kind of simulation has been extended, using multiple output units, to the supervised learning of entire categorizations, such as teaching the network to discriminate between the ten possible digits in hand-printed zip codes (e.g., Weideman, Manry, Yau, & Gong, 1995).

But back propagation network algorithms typically require a large number of iterations, with no universal guarantee that the weights will converge to the desired state. This state is mathematically defined as the global minimum of an error function, and the system can be trapped in a local minimum that is different from the global minimum (see Figure 4.12). Conditions for convergence of the BP algorithm have been the object of much mathematical investigation. White (1987) placed the back propagation method in the context of other nonlinear regression methods. He gave some conditions that guarantee convergence to the desired global minimum, noting that his conditions (such as nonexistence of an alternative local minimum) are frequently not met in applied problems.

Also, the network varies enormously in how many steps it requires to converge to the mapping it is supposed to learn. Typically, the convergence rate is strongly dependent on the number of hidden units, and that number must be decided separately for each application. Criteria for selecting this number are also suggested by Theorem 3 of White (1987), but in most cases the optimal number of hidden units needs to be determined empirically.

A variety of authors have developed algorithms for finding the best number of hidden units. One of the best of these algorithms is that of Hirose, Yamashita, and Hijiya (1991), which involves adding a hidden unit when the network becomes trapped in a local minimum. If the network converges to the global minimum to start with, hidden units are removed until it no longer converges, and then one is put back.

The setting of the learning rate is considered in Rumelhart et al. (1986) and McClelland and Rumelhart (1988, Ch. 6); the 1988 book chapter also discusses other BP implementation issues. There is a tradeoff in this setting: Too small a learning rate can make convergence excessively slow, but too large a learning rate can cause oscillations that make convergence impossible. A method has been found for preventing such oscillations, thereby allowing a larger learning rate. This method involves adding to the weight change equations a *momentum* term which biases change in the same direction that the last previous change was made (see the end of this chapter for mathematical details). This reduces the likelihood of rapid oscillation between weight increases and decreases. Many back propagation networks in applications include a momentum term.

Back propagation is not generally believed to be a mechanism actually used in the brain. This is because the feedback in the network is not of electrical signals but of synaptic weights, and no mechanism is known for such weight transport in real nervous systems (see Figure 8.2). Yet some researchers have suggested different possible biological bases for back propagation.

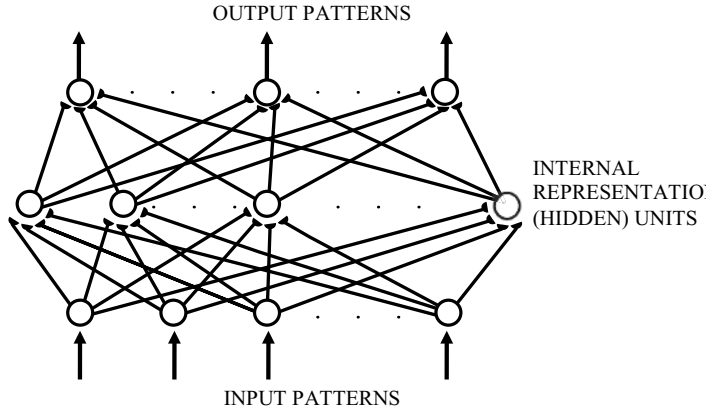


FIGURE 8.1 Generic architecture for a three-layer back propagation network. Error signals from output nodes, if their response to the input pattern is not the desired one, propagate backwards from hidden-to-output weights to input-to-hidden weights. In the process, the hidden units learn internal representations, that is, learn to encode certain classes of input patterns.

Source: Adapted from Rumelhart et al., 1986, with permission of MIT Press.

Input	*Output*
00	0
01	1
10	1
11	0

TABLE 8.1 The logical exclusive OR (XOR) relationship

Variables are assumed to be binary: "1" corresponds to "Yes" or "True," "0" to "No" or "False." Exclusive OR of two propositions is true whenever one of the propositions is true but *not both at once*.

Stork (1989b) constructed a minimal collection of neurons and synapses implementing back propagation but not requiring actual weight transport. His architecture is intricate and based on some unnatural assumptions, such as symmetry of connection weights (see Section 4.2) and lack of dependence of synaptic modifications on postsynaptic activity (see Sections 3.1 and 3.2). Dayhoff, Hameroff, Swenberg, and Lahoz-Beltra (1993) and Werbos (1992a) related back propagation to backward flows in *microtubules*, which are part of the structural support system of all biological cells including neurons. Levine (1996) proposed that weight transport (not necessarily "backward") might arise from suitable networks that include nodes responsive to the combination of

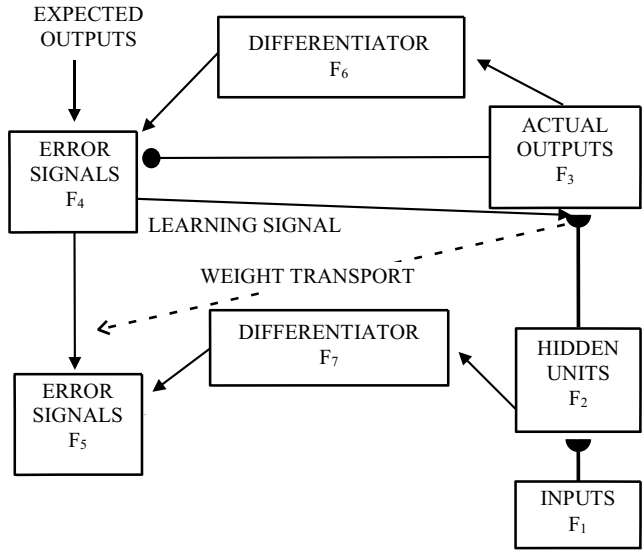


FIGURE 8.2 Possible circuit diagram of the back propagation model. Postulated interactions among node levels F_1, F_2, and F_3 suggest additional levels F_4, F_5, F_6, and F_7. Transport of weights from F_2-to-F_3 to F_4-to-F_5 pathways makes this circuit biologically implausible.

Source: Reprinted from *Cognitive Science*, *11*, S. Grossberg, From interactive activation to adaptive resonance, 23–63, copyright Cognitive Science Society, Incorporated, used by permission.

activities of pairs of other nodes, an idea borrowed from a model of cortical processing by Guigon, Dorizzi, Burnod, and Schultz (1995). Surprising examples of backward flowing signals at synapses or neuronal dendritic trees were reported by Stuart, Spruston, Sakmann, and Hauser (1997) and reviewed in Waters, Schaefer, and Sakmann (2005).

Whether a back propagation process can or does occur in biological nervous systems remains an open question. Yet there are other architectures with qualitatively similar properties that are arguably more biologically realistic because they do not involve transport of weights. One of these was discussed by O'Reilly (1996a), based on earlier work of Hinton and McClelland (1988). O'Reilly showed that the quantity calculated in the back propagation algorithm for the impact of a given hidden unit on the error can be obtained in a different manner without weight transport. Specifically, this quantity can be expressed as a difference of two activation quantities, one related to the target pattern, the other to the current output pattern. He suggested how these two activations could be generated in the cerebral cortex by suitable combinations of long-term potentiation, long-term depression, and the actions of calcium ions. This form of error correction is one of the bases for the Leabra model (O'Reilly, 1996b), which combines Hebbian associative learning and error-correcting learning at

the same nodes. Leabra in turn is the basis for the PVLV conditioning algorithm discussed in Section 6.4.2.

Another possible method for making the back propagation algorithm more biologically realistic was introduced by Lillicrap et al. (2016) and involves using a random matrix rather than the precise matrix of downstream weights to modify upstream synapses. This method is applicable to deep learning networks with more than three layers so will be discussed in Section 8.3.

To illustrate use of BP in a specific problem domain, let us return to the case of teaching a network to discriminate between a "T" and a "C" regardless of position or orientation (Rumelhart et al., 1986). Figure 8.3 illustrates the different rotations of the T and C. The network for solving the T-versus-C problem is shown schematically in Figure 8.4. Each rotation has to be taught

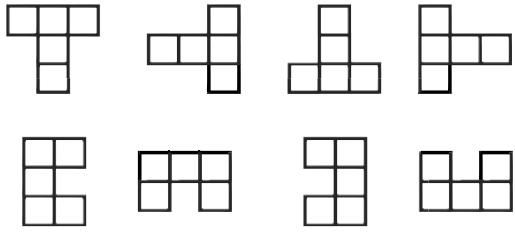


FIGURE 8.3 Stimulus set for the T-versus-C problem. The set consists of a block T and a block C in each of four orientations. One of the eight patterns is presented on each trial.

Source: Reprinted from Rumelhart et al., 1986, with permission of MIT Press.

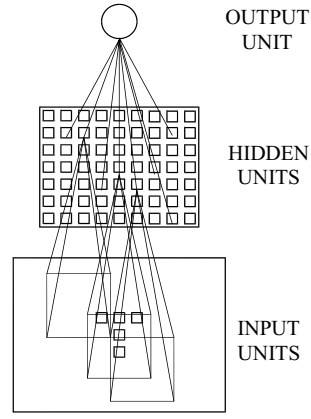


FIGURE 8.4 Schematic diagram of a network for solving the T-versus-C problem. Hidden units are organized into a two-dimensional grid with each unit receiving input from a square 3 × 3 region. The output unit is trained to take on the value 1 if the input is a T (at any position or orientation) and 0 if the input is a C.

Source: Reprinted from Rumelhart et al., 1986, with permission of MIT Press.

to the network separately. But translation invariance is achieved by adding an additional transformation to the rule for learning input-to-hidden-unit connections. To make the learning of a pattern independent of its location in the visual field, all hidden units are constrained to learn exactly the same pattern of weights. This is accomplished by adding together the weight changes dictated by the delta rule for each unit and then changing all weights by averages of those amounts.

Averaging weight changes over the entire visual field is a *nonlocal* transformation, that is, a transformation in which a region is directly affected by an arbitrarily large region around it. Since locality is usually assumed in biological models, it might be desirable to achieve the same effect by averaging, at each connection, over some smaller diameter of the visual field. As is discussed in Section 7.3, the problem of translation- and rotation-invariance still poses a challenge for theories of neural categorization.

The back propagation models used to categorize patterns typically have connections that are exclusively feedforward, except for the transport of weights in the reverse direction. Various modelers, among them Jordan (1986b), Pineda (1987, 1989), and Elman (1990), have added recurrent or feedback connections (see Chapter 4) to a basic back propagation network. A common use of recurrent back propagation networks has been to learn and reproduce time sequences, such as occur in speech or in motor behavior. A more recent model in this school (Botvinick & Plaut, 2006) is discussed in Section 9.3 along with other models of sequence learning and production.

Classification algorithms that are tightly supervised have some distinct advantages in speed and reliability for applications in which particular, known outputs are desired. By contrast, living organisms are internally supervised by reinforcement and drive systems that function as *critics* (see Grossberg, 1971; Barto et al., 1983; Werbos, 1993, 2003; Section 6.3.2 of this book) but do not dominate classification decisions. Hence, living neural systems seem to include learning modules that find without direct supervision the naturally recurring input classes in the environment, but then are subject to attentional control. This attentional control causes finer distinctions to be made among those inputs that are most important to the system's goals.

Yet the proponents of back propagation networks argue that, even if a network structure bears little resemblance to brain structure, the insights that emerge from designing the network guide our understanding of brain mechanisms relevant for certain cognitive processes, including effects of lesions to particular regions. These insights typically emerge from the response patterns that develop in groups of hidden units after a large number (hundreds or thousands, usually) of iterated applications of the back propagation algorithm. These unit response patterns are referred to as *internal representations* (Rumelhart, Hinton, & Williams, 1986).

Back propagation based models of neural and psychological processes typically fit into a larger class called *parallel distributed processing* (*PDP*) models. This term is defined as follows (Rumelhart & McClelland, 1986a):

> These models assume that information processing takes place through the interactions of a large number of simple processing elements called units, each sending excitatory and inhibitory signals to other units. In some cases, the units stand for possible hypotheses. . . . In these cases, the activations stand roughly for the strengths associated with the different possible hypotheses, and the interconnections among the units stand for the constraints the system knows to exist between the hypotheses. In other cases, the units stand for possible goals and actions . . . and the connections relate goals to subgoals, subgoals to actions, and actions to muscle movements. In still other cases, units stand not for particular hypotheses or goals, but for aspects of these things.
>
> (Vol. I, p. 10)

For understanding the PDP modeling approach, the word "distributed" is key. PDP models tend to start with units that represent broad classes of entities. Their interconnections start out at time 0 with little defined structure, and the internal representations emerge through extensive training of the weights. The internal representations are typically interpreted as not residing at single units but distributed across multiple units.

In contrast to the PDP models, some other prominent neural network models are called *localist* because single nodes or units play a prominent role in representing concepts. Localist models notably include many models of Grossberg and his colleagues. The controversy between distributed and localist neural network models, discussed in Exercise 7 of Chapter 2, is very much alive today (see, in particular, Bowers, 2017, and Thomas & French, 2017). On the localist side, Bowers (2017) argued that there is considerable data showing that some neurons respond selectively to specific faces or other high-level information (the so-called "grandmother cells"). He also noted that, even in some PDP models, sometimes units emerge that are highly stimulus-selective in their responses. On the distributed side, Thomas and French (2017) argued that grandmother cell responses exist but may not be causally related to the animal's ability to recognize a category. They cited modeling results showing that in a self-organizing map designed to model monkey data on categorization performance, removal of category-specific neurons from the simulation had no effect on performance, and the absence of neuroscience data showing loss of recognition of single objects due to localized brain damage.

In reality, human and animal nervous systems probably use a mixture of localist and distributed codes. In evolution, the design of functional architectures is opportunistic, and "one size fits all" is rarely the rule.

8.2. "Semantics without Categorization" in PDP Networks

The PDP approach tends to favor learned over innate mechanisms for cognitive processes. Yet, modelers who apply it to specific processes, most notably in the domain of language, often add to the generic structure of Figure 8.1 some groups or units representing concepts that could be either innate or learned. For example, the PDP model of Rumelhart and McClelland (1986b) for learning the past tense of English verbs includes phonological representations of root forms (as input units); phonological representations of past tenses (as output units); and feature representations, based on combinations of phonemes, of both the root form and past tense (as hidden units) that have modifiable associative connections from root to past tense representations. The PDP model of word recognition and naming by Seidenberg and McClelland (1989) includes representations of orthography, phonology, meaning, and context.

An influential PDP approach to semantic cognition leads to learning of category-related properties of objects and concepts without building explicit categories (Rogers & McClelland, 2004, 2011; Rumelhart & Todd, 1993). Rogers and McClelland (2011) noted that many categorization models assume for simplicity the existence of a discrete set of categories such that each object under consideration is placed in one and only one category. By contrast, they noted that the same object can be classified in many ways depending both on context and on the required degree of specificity: "Lassie, for example, belongs to the categories *collie*, *dog*, *pet*, *animal*, and *movie star*" (Rogers & McClelland, 2011, p. 89). Other theorists have also noted the one-to-many aspect of categorization and dealt with it, for example, by varying vigilance levels (e.g., Carpenter, Grossberg, & Reynolds, 1991) or including attribute-selective attention (Kruschke, 2011). Yet, Rogers and McClelland started from that point to eschew categorization altogether and simply model the process of inferring properties of concepts from what is known about related concepts. They adapted a model by Rumelhart (1990) and Rumelhart and Todd (1993) that was in turn a network representation of a classic theoretical model in cognitive science, the spreading activation model of Collins and Quillian (1969).

The back propagation network of Rogers and McClelland (2011) includes the typical input, hidden, and output layers, with inputs being concepts and outputs being their potential attributes. In addition, the hidden layer is influenced by an extra layer of relations, of which there are four: "is a," "is," "has," and "can."[1] The resulting network simulates a range of data on generalization of properties, such as how quickly one can conclude that a canary can sing or has skin from knowing that most birds can sing and all animals have skin. These researchers mapped the concepts into clusters in concept space that have similar weights in multidimensional space (e.g., "sunfish" is closer

to "salmon" than it is to "robin" or "oak"), with the distances between concepts being different for different contexts determined by relations.

The long period of training in back propagation networks and the fact that many of their internal representations are distributed over the same nodes make them particularly vulnerable to catastrophic forgetting (French, 1999). A way to overcome catastrophic forgetting in such networks, and in the later, more biologically plausible Leabra networks, was suggested by McClelland, McNaughton, and O'Reilly (1995) based on the different learning properties of the neocortex and hippocampus. McClelland et al. proposed that new items are learned quickly in the hippocampus and this protects them from interference by other memories stored in the cortex. In their schema, later reinstatement of memories of those new items in the cortex proceeds slowly and gradually by interleaving those memories with others. Hence, the theory relates to those of other investigators who have posited that memories are temporarily stored in the hippocampus before being transferred to the cortex for long-term storage.

The theory of McClelland et al. (1995) provides one account of the difference between episodic and semantic memories. In their theory, updated in O'Reilly and Rudy (2000), episodic memories are based in the hippocampus or, more broadly, the hippocampal system which also includes perirhinal, parahippocampal, and entorhinal cortices (see Section 5.3), and accurately represents events as they have occurred. Semantic memories, by contrast, are based in the cortex and represent a kind of statistical average of many episodic memories that may be different in detail but closely related to one another (which can be considered a form of categorization of past events).

Grossberg and Merrill (1996) and Franklin and Grossberg (2017) objected to the mechanism of McClelland et al. (1995) on the grounds that the hippocampus does not contain perceptual representations, such as those in the cortex and thalamus, that are specialized enough to encode all the relevant or salient aspects of a novel event. They added that it is not clear from the available physiological evidence that synapses in the cortex in fact learn slower than those in the hippocampus. Grossberg and Merrill attribute to the hippocampus a role that does not include actual memory storage but a modulatory influence on cortical processing, via spectral timing (see Section 6.3.4) and the orienting system of ART (see Section 7.2.2).

Yet, O'Reilly and Rudy's (2000) theory may be more plausible if it is restricted to parahippocampal cortex and not the actual hippocampus. Otto and Eichenbaum (1992) review evidence for some memory storage in parahippocampal cortex but not in the hippocampus proper. Recall from Section 5.3 that cortical inputs about features of objects ("what") converge in the perirhinal and lateral entorhinal areas, whereas details about their location ("where") converge in the parahippocampal and medial entorhinal areas (Eichenbaum et al., 2007). Hence, one stream seems to process events whereas the other stream processes the context of those events. The two streams meet

in the hippocampus proper, which also has neurons that register the time of events (MacDonald, Lepage, Eden, & Eichenbaum, 2011).

Reasoning about category properties rather than explicit categorization was also emphasized in the cognitive architecture of Sun and Zhang (2006). Sun and Zhang's work is based not on back propagation but on a two-layer connectionist-symbolic architecture called CLARION, which is described in the next chapter (Section 9.6.2).

Through the 1990s and early 2000s, a number of developments in machine learning applications led researchers to add more hidden layers to back propagation networks. This development enabled the network to simultaneously encode internal representations at multiple levels of abstraction. Such expanded back propagation networks are the major, though not the only, focus of the recent modeling trend known as *deep learning*.

8.3. Deep Learning

Hinton (2007) described the neuro-behavioral aspect of the fundamental problem motivating the transition from three-layer back propagation to deep learning that includes more hidden layers:

> To enable the perceptual system to make the fine distinctions that are required to control behavior, sensory cortex needs an efficient way of adapting the synaptic weights of multiple layers of feature-detecting neurons.
>
> (p. 428)

Hinton (2007) went on to note that standard back propagation requires training the network with data that are labeled as members of an existing category or class of objects. He discussed some examples of recurrent networks whereby the training data are generated using top-down connections rather than labeled as category members, and learning consists of adjusting top-down weights to maximize the likelihood that those data would be generated. Bottom-up connections are still used to determine activations of the features of those data.

The primary payoff of models such as Hinton's is in computing applications; specifically, in Hinton (2007), recognition of poorly written digits. Yet he discusses possible applications to cortical modeling:

> In particular, is the initial perception of sensory input closely followed by a reconstruction that uses top-down connections? . . . All that is required is that there are two phases that differ in the relative balance of bottom-up and top-down influences, with synaptic potentiation in one phase and synaptic depression in the other.
>
> (p. 433)

This proposed interplay of top-down and bottom-up processes is reminiscent of other models with closer ties to neuroscience, such as ARTMAP (see Section 8.4) and COVIS and SUSTAIN (see Section 8.7).

These considerations led a variety of modelers to develop networks that build on the back propagation structure described in Section 8.1 but include training at multiple levels (see Bengio & Lee, 2015, and LeCun et al., 2015, for reviews). Many of these models have been inspired both by the hierarchical structure of the visual cortex and by the Neocognitron model (Fukushima, 1980; Fukushima & Miyake, 1982; see Section 7.3) that captures this visual hierarchy.

One of the earliest networks to incorporate an approximation of the visual hierarchy was designed by LeCun et al. (1989, 1990) for recognizing badly handwritten digits. LeCun and his colleagues set out to build enough internal structure into the back propagation network to minimize the need for preprocessing inputs. The network at various points performs *convolution*, a mathematical operation using integral calculus that bears some similarity to a correlation, of incoming signals with feature maps representing features that become more abstract at deeper layers of the network. Such a *convolutional neural network* or *ConvNet* is one of the commonest types of deep learning neural networks (see LeCun et al., 2015, p. 439). Several ConvNet-based systems for handwriting recognition or optical character recognition have been employed by Microsoft. ConvNets have also been applied successfully to speech recognition, document reading, face recognition, and object detection in natural images. Thus far there have been few applications of ConvNets and other deep learning neural networks to modeling actual brain systems or behavior. Deep learning has also been combined with reinforcement learning (see Chapter 6) to build machines that have been successful in playing Atari games (Mnih et al., 2015).

There is still potential for using deep learning algorithms in models of brain systems with hierarchies of abstraction; this includes not only the visual and auditory sensory systems but also the prefrontal cortex executive function system (Badre & D'Esposito, 2007; Christoff & Gabrieli, 2000; Christoff et al., 2009; Koechlin & Hyafil, 2007). Some biologically based models of the visual system with different architectures are described in Section 9.2 and models of the executive system in Section 9.4. Many of the models in Chapter 9 build on models of simpler cognitive processes described in Chapters 3 and 4.

The majority of deep learning applications described in this section, like the majority of back propagation applications, have involved supervised learning. An application of a convolutional deep network to unsupervised learning appears in Lee, Grosse, Ranganath, and Ng (2011). Lee and colleagues looked for high-level structures in unlabeled visual scenes using a convolutional

example of a *deep belief network*. A deep belief network (Hinton, Osindero, & Teh, 2006) is a network where each layer encodes statistical dependencies among the nodes in the layer immediately below and can be trained to approximately maximize the likelihood of the training data. In the network of Lee et al. (2011), the first layer learns edge detectors, the second layer object parts, and the third layer complete objects. These representations enable the network to infer hidden object parts from high-level object information.

Typical deep learning networks rely on error-driven weight adjustments by the same method used in three-layer back propagation networks, namely updating of weights at lower levels based on feedback to synapses at higher levels. Hence, deep learning networks inherit from back propagation the doubts about biological plausibility. Lillicrap et al. (2016) noted that the Leabra model introduced by O'Reilly (1996a, 1996b) had not completely overcome these doubts because O'Reilly's method made some symmetry assumptions that may not be realistic in the brain. Lillicrap and his colleagues showed mathematically that approximations comparable to those from back propagation networks could be obtained by subjecting all the weights in network to influence by a randomly chosen matrix, rather than a matrix of downstream synaptic weights as in typical back propagation. This innovation, like Leabra, provides the benefits of back propagation without the nonlocality that makes back propagation biologically dubious.

Yet, none of the networks described in this section has been closely connected with known neural pathways involved in specific cognitive functions. The next section, and Section 8.7, provide examples of networks for supervised categorization that have made more contact with cognitive neuroscience data.

8.4. ARTMAP: A Family of Supervised Adaptive Resonance Networks

The early successes of ART in unsupervised categorization led Carpenter, Grossberg, and Reynolds (1991) to apply the same architectural principles used in ART 1 to supervised categorization. These researchers observed that pure self-organization is insufficient to predict the consequences of categorization decisions under many conditions. For example, self-organization tends to place two stimuli in the same category if they have similar sensory or cognitive features, even when the effects produced by those two stimuli are radically different. One example occurs in a benchmark machine learning problem that these authors simulated, that of distinguishing between poisonous and edible mushrooms, which often look quite similar. A less spectacular example occurs in the letter classification dilemma shown in Figure 8.5.

The basic operation of a supervised ART network is to learn an arbitrary nonlinear mapping from patterns in an n-dimensional space to other patterns

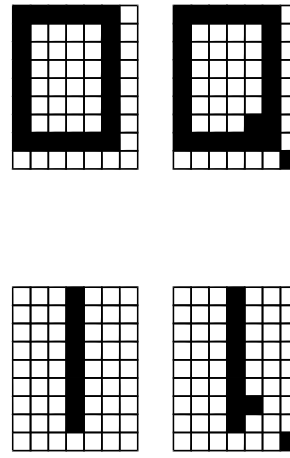


FIGURE 8.5 The same two extra pixels which change an "O" into a "Q" in (a) change an "I" into a noisy "I" in (b).

Source: Reprinted from Stork, *Journal of Neural Network Computing*, *1*, 26–42 (New York: Auerbach Publishers). Copyright 1989 Warren, Gorham, & Lamont, Inc. Used with permission.

in an *m*-dimensional space; this is why the authors coined the term *ARTMAP* for their network. (Note: such a mapping need not be a function in the mathematical sense of having a unique value, as the network can learn one-to-many as well as one-to-one and many-to-one mappings.) In particular, if inputs to such a network are represented as arrays of numerical values of features, the network can learn categories with irregular boundaries. These mappings are learned by connecting a pair of ART 1 modules known as ART_a and ART_b via associative memory links, as shown in Figure 8.6.

Learning in ARTMAP is unidirectional, from ART_a to ART_b. During the training phase, both ART modules receive streams of input patterns: The patterns at ART_b are interpreted as consequences of those at ART_a (e.g., poisoning as a consequence of ingesting a specific mushroom). Then associative learning takes place between the F_2, or category, layers of the two ARTs. As a result of this learning, in the testing phase, as each new input comes in to ART_a not only is it classified in a category there but a category is also predicted at ART_b. However, initial predictions (in either the training or testing phase) may not be correct; consequently, an orienting system and a vigilance parameter (see the earlier description of ART 1) are required to reset the system if predictions are incorrect. This orienting system is called a *map field*, as shown in Figure 8.7.

The essentials of the ARTMAP predictive architecture are carried over in *fuzzy ARTMAP* (Carpenter et al., 1992). In fuzzy ARTMAP, two fuzzy ART modules are combined with inter-ART associative learning and a map field to

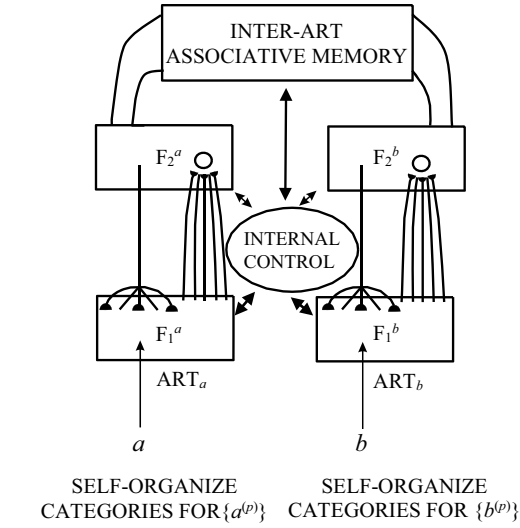


FIGURE 8.6 An ARTMAP system for supervised learning includes two ART modules linked by an inter-ART associative memory. Internal control structures actively regulate learning and information flow.

Source: Adapted from *Neural Networks*, *4*, G. A. Carpenter, S. Grossberg, & J. H. Reynolds, ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, 565–588, Copyright 1991, with permission from Elsevier Science.

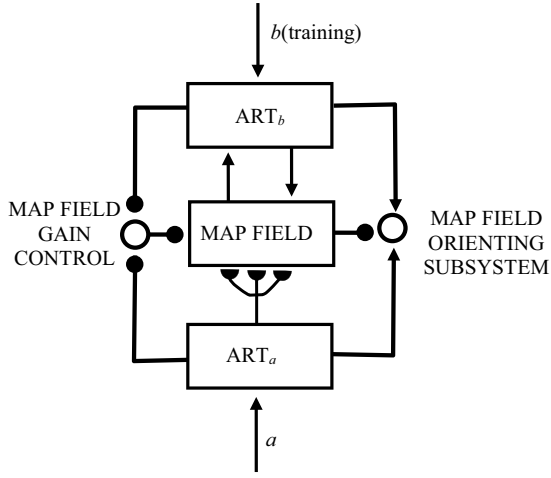


FIGURE 8.7 Block diagram of ARTMAP. The map field along with its orienting subsystem and gain control connect the two ART modules, ART_a and ART_b, and control resonance and reset of predictions.

Source: Adapted from *Neural Networks*, *4*, G. A. Carpenter, S. Grossberg, & J. H. Reynolds, ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, 565–588, Copyright 1991, with permission from Elsevier Science.

learn arbitrary mappings between multidimensional analog patterns. As in the case of fuzzy ART (Carpenter, Grossberg, & Rosen, 1991a, 1991b), complement coding (see Section 7.2.3 for a definition of this term) is typically used for fuzzy ARTMAP inputs both to indicate those nodes, or features, that are activated strongly and those that are not activated.

8.5. Exemplars, Prototypes and Rules in Models of Category Learning

Section 7.2.2 noted that categorization relies both on exemplars, which are specific instances of input patterns that belong to a category, and prototypes, which are averages across many exemplars of that category. Sometimes categorization also involves explicit rules for what features or criteria make a pattern a member or nonmember of a category.

There are many nonneural cognitive models that classify a new input based on comparison of the input with stored exemplars, and others that classify based on comparison of the input with a prototype. Grossberg, Carpenter, and Ersoy (2005) pointed out the different advantages of both types of models. Prototype models are better at capturing abstraction of key properties that psychologically define a category. The importance of prototypes is supported by some human behavioral data. Reaction times to exemplars of a category are shortest for exemplars that are close to the prototype (e.g., Mervis & Rosch, 1981); for example, a sparrow is recognized as a "bird" sooner than an ostrich is. Moreover, if a subject is taught exemplars that are random variations on a general pattern of dots, the prototype formed is close to the average of these variations. The subject then learns the prototype faster than any of the actual exemplars even without having seen the prototype (e.g., Posner & Keele, 1970). On the other hand, exemplar models are better at capturing episodic memories and specific important stimuli that are frequently encountered, such as the face of a family member. Some cognitive modelers of categorization have attempted to combine the strengths of both types of models in *rule-plus-exceptions* models (e.g., Palmeri, Nosofsky, & McKinley, 1994), whereby categories are defined mostly by nearness to prototypes except for a few exemplars that are distant from category centers.

A series of models by Kruschke and his colleagues, starting with *ALCOVE* (*attention learning covering map*; Kruschke, 1992), embed the exemplar theory in a connectionist framework, without reference to specific brain regions. ALCOVE is derived from a widely used nonconnectionist exemplar model called the *generalized context model* (Nosofsky, 1986). The model is based on a feedforward network whose structure bears some similarity to back propagation networks, with input nodes coding novel inputs, hidden nodes coding stored exemplars, and output nodes coding categories to which inputs

might be assigned. Yet the representations are distributed only over a few nodes, a design that avoids the catastrophic interference common in PDP models, most of which have extensive overlap of nodes representing different concepts.

As in other category models discussed in this chapter, inputs to be placed in categories of ALCOVE are treated as patterns (i.e., vectors) of numerical activations of different attributes or features. An error-driven learning rule adjusts the relative attention paid to different attributes, and attributes whose appearance is correlated across the input set become selectively enhanced. The reallocation of attention is a main feature of a later variation on the model known as *RASHNL* (*rapid attention shifts and learning*: Kruschke & Johansen, 1999). Other variations on the ALCOVE design have added context-specificity of categorization (e.g., Denton & Kruschke, 2006) and pooling of "expert" classifiers (Denton, Kruschke, & Erickson, 2008).

The Kruschke family of exemplar-based models has been able to simulate a wide range of behavioral data on categorization. This includes a form of base rate neglect (Gluck & Bower, 1988b), whereby participants given the symptom profiles of patients with two fictitious diseases tend to overestimate the probability of the rarer disease given symptoms that are highly compatible with that condition. Also, they simulated some data showing that *linearly separable* categories are not necessarily easier to learn than categories that are not linearly separable, contrary to a prediction of some prototype models. Linear separability means that if each input is described numerically by an array of node activation quantities, the criterion for membership in a category is that some linear function of those quantities be larger than some set value. Kruschke (2011) noted that one limitation of these models is a difficulty in generating categorization rules, though they have been able to simulate "rules plus exceptions" such as the learning of past tenses of English verbs, which was also simulated by Rumelhart and McClelland (1986b).

Also, ALCOVE and related models are not embedded in real-time networks for other cognitive processes such as memory, perception, and conditioning. Neither are categorization models based on prototypes, such as that of Knapp and Anderson (1984), which is derived from the earlier linear models of Anderson and his colleagues (see Section 3.2.2). The next subsection discusses a different type of categorization model, by Anderson et al., one that comes from repeatedly applying the same linear transformation to a pattern and then classifying the original pattern in a cluster determined by what it converges to for large time.

The use of explicit rules in computational models of categorization has been mainly in models that include explicit connections with brain regions. Two classes of models of that type – COVIS and SUSTAIN – are discussed in Section 8.7.

8.6. Brain-State-in-a-Box Models

A class of categorization models based on linear associative learning, widely known by its nickname of *brain-state-in-a-box*, has been studied since 1977 by Anderson and his colleagues. In early versions of these models (Anderson et al., 1977; Anderson & Mozer, 1981), pattern classifications are unsupervised. In later versions (Anderson & Murphy, 1986), an error-correcting learning rule allows learning of particular desired classifications.

The brain-state-in-a-box (BSB) model was first derived by Anderson et al. (1977) as an offshoot of the linear model with saturation by Nass and Cooper (1975; see Section 3.2). Recall that saturation was imposed on a linear network with associative memory and positive feedback, in order to prevent activities of the nodes in the network from becoming unbounded. The "box" from which the model derives its name is shown in Figure 8.8. It is an abstract square, cube, or, more generally, hypercube of possible *n*-dimensional activity vectors which constitute the numerical bounds for possible node activities.

The BSB model associates vector patterns of activities at a set of nodes with other patterns at the same nodes. The matrix consisting of the connection weights between nodes provides feedback that transforms the pattern, as is developed shortly. The network then converges to one of the characteristic system states corresponding to corners of the "box" in Figure 8.8. Categorization of the original input pattern is based on which one of these corners is reached. (A good, accessible introduction to matrices and vectors as they relate to neural networks is found in Jordan, 1986a. Some of that material is summarized in Appendix 1 of this book.).

Recall from Section 3.4.1 the distinction between *autoassociative* and *heteroassociative* encoding (Kohonen, 1977, 1984). The connectivity matrix is designed to associate a pattern (input) **x** to another pattern (response) **y**; this is called autoassociative if the two patterns are equal, and heteroassociative otherwise. The BSB model is applicable to both of these two types of encoding.

A brief description of this algorithm follows; more mathematical detail appears at the end of this chapter. As in Anderson's earlier models discussed in Chapter 3, the input pattern is interpreted as the initial state of a vector of node activities $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))$. These activities x_i vary between preassigned maximum and minimum values, which for mathematical simplicity are taken to be 1 and –1. The vector $\mathbf{x}(0)$ at $t = 0$ represents the input pattern to be classified; values of $\mathbf{x}(t)$ at subsequent times are given by the rule

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) \tag{8.1}$$

where **A** is the connectivity matrix of the system, or rectangular array of its connection weights.

Equation (8.1) represents positive feedback as it might occur in the brain, due to the past operation of a Hebbian associative learning law. This feedback

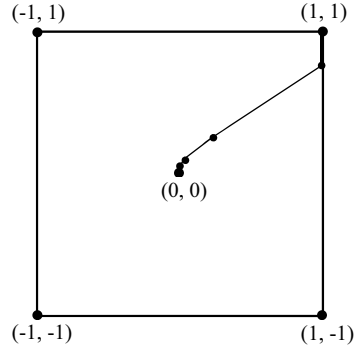


FIGURE 8.8 A two-dimensional example of the "box" from which the brain-state-in-a-box model (Anderson et al., 1977) derives its name. The system state at any time is given by a vector of (in this case, two) numbers between the bounds for node activity (in this case, –1 and 1). Points along the curve drawn inside the box denote the system state at successive times. In this case, the points that are drawn arise if the initial state is (0, 0.05) and the connectivity matrix (see Equation (8.1)) is $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$.

has the desirable property of enhancing significant activities or stimuli, but often has an additional property that is undesirable. Repeated application of (8.1) to a pattern vector will drive the state of the system outside the box of Figure 8.8, that is, cause values of some or all of the x_i to become larger than 1 or smaller than –1. This is particularly true if the input pattern is one of the significant states of the system, which correspond to the vectors known as *eigenvectors* of the matrix **A** (see the equations at the end of this chapter for a definition of this term).

To prevent activities from leaving the box, Anderson et al. (1977) imposed an additional rule whereby, if any of the activities becomes greater than 1 as a result of (8.1), it is replaced by 1, and if that activity becomes less than –1 it is replaced by –1. The effect is to make the system converge to one of the corners of the box, which represent states in which all activities are 1 or –1. Hence, the BSB model, like all neural network models, includes a method for keeping activities within bounds, corresponding to the limits on possible firing frequencies of actual neurons. Grossberg (1978c) argued that this particular method (linearity plus "hard" saturation) is not biologically realistic and creates some inevitable distortions of pattern weights because distinctions among values close to 1 or –1 are lost. He argued that an inherently nonlinear mechanism is needed to avoid such distortion. (For the gist of his reasoning, see Exercise 1 of Chapter 4). Anderson and Silverstein (1978) responded that linear transformations do occur in some regions of real nervous systems (as in the response of the horseshoe crab eye to visual stimuli). They added that a simple linear model illustrates some properties that are likely to be retained in more complex nonlinear models.

Anderson and Murphy (1986) combined BSB with the delta or error-correction learning rule (in a form used previously by McClelland & Rumelhart, 1985) to deal with the problem of noisy encoding. Previous BSB networks succeeded in distinguishing mutually orthogonal (perpendicular) input vectors, but did not always succeed in distinguishing inputs whose vector directions were relatively close together. For this reason and others, the connectivity matrix learned from previous pattern associations sometimes failed to yield the desired encoding if the data were somewhat noisy.

In the Anderson–Murphy simulations, the connectivity matrix **A** is initially chosen at random. Then the system is fed a desired association of a vector **x** to another vector **y**. On each time step, the matrix **A** is changed by a correction term based on the difference between **y** and the product **Ax** (see the end of the chapter for details). Once the desired matrix has been found, the system goes through the saturating-linear BSB algorithm as in Anderson et al. (1977).

Anderson and Murphy's supervised version of the BSB model has been applied to processing linguistic inputs that are converted to vectors of 1s and –1s by means of ASCII codes. This model has reproduced the disambiguation by context of words with more than one meaning; for example, the word "bat" by itself could mean a flying animal, "ball" could mean a dance, and "diamond" could mean a jewel, but all three words together must refer to baseball.

Recently, BSB has been used more in computing and industrial applications than in modeling of psychological or cognitive neuroscience data. Perhaps the most interesting application of this type of modeling to psychologists is the work of Abdi, Valentin, and O'Toole (1997), who applied an autoassociative network based not on BSB but on a variant of Anderson (1972) to classifying human faces by gender. The face classification is of particular interest since it employs a modified linear autoassociator designed to allow for the possibility of selective attention to different parts of the feature space (Abdi, Valentin, Edelman, & O'Toole, 1996).

8.7. Some Brain-Based Models: COVIS and SUSTAIN

Ashby and his colleagues (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Ennis, & Spiering, 2007) have developed a neural theory of perceptual categorization based on a combination of declarative and procedural learning. They call their model *COVIS*, an acronym for *competition between verbal and implicit systems*. In Ashby et al. (2007), the procedural part of COVIS was elaborated and updated into a model of categorization by experts called *SPEED* (*subcortical pathways enable expertise development*).

The starting point for the inquiry that led to COVIS was a set of behavioral results on categorization of multiattribute geometric patterns, some of which could be categorized explicitly by easily stated rules and others for which the rules could not be easily expressed (Maddox & Ashby, 1993). An example of

an explicit rule given in Ashby et al. (1998) deals with classifying rectangles, whereby the rule is "Respond A if the stimulus rectangle is taller than it is wide; respond B if the rectangle is wider than it is tall" (p. 444). An example of a categorization that does not yield to an explicit rule is classifying stimuli consisting of circles with oriented lines inside them, because orientation of the line and radius of the circle cannot be directly compared. Categorization in the latter case requires what the authors called information integration, which in turn requires learning of actual connections between representations of specific stimuli and categories. The "competition" in the model's name arises because the participants need to discover which type of rule is more appropriate for a particular categorization problem, so both the verbal and procedural systems are active in the early stages of the task.

The original version of COVIS is shown in Figure 8.9. While both the verbal (rule-based) and implicit (procedural) systems involve loops between the cortex, basal ganglia, and thalamus, rule-based category learning is centered in the prefrontal and cingulate cortices and head of the caudate nucleus, and procedural learning in the tail of the caudate nucleus. The procedural learning system works somewhat analogously to the learning by exemplars in the Kruschke and Nosofsky models.

Both model systems require dopamine to learn. The architecture led to predictions, which the data generally supported, that different lesions should affect different types of categorization. The model predicts that prefrontal damage or decreased dopamine input to the anterior cingulate should impair only the rule-based type of categorization, whereas Parkinson's disease (which can affect both the head and tail of the caudate) and caudate damage should impair both types.

The SPEED model of Ashby et al. (2007) maps the development of expertise or automaticity (two words that they openly conflate) in categorization. In simulations using that model, striatal pathways and dopamine are active in the early stages of task but later activity is confined to the sensory, associative, and motor cortex. They note an analogy between the temporary activity of the striatum in their model and the temporary activity of the hippocampus in the memory model of McClelland et al. (1995) (see Section 8.2). Some of the equations for the SPEED model are listed at the end of this chapter.

Whenever a model relies on two separate systems, the question arises as to how the two systems are integrated – or, in this case, how a decision is made as to which one to activate. Ashby et al. (1998) discuss this issue without coming to a conclusion, hinting that lateral inhibition somewhere in the striatum might be a mechanism for such a decision.

The dual-process idea also occurs in the discussion by Ashby et al. (1998) of data such as Knowlton and Squire's (1993) showing that amnesics who have difficulty remembering recently presented category exemplars are normal at categorization, at least in the early stages. Yet, Nosofsky and Zaki (1998) were

able to simulate Knowlton and Squire's data using a one-process exemplar model by manipulating the value of a sensitivity parameter. Grossberg et al. (2005) suggested that the sensitivity in that model was analogous to the vigilance in their own ART model. The data of Nosofsky et al. (2012) discussed in Section 5.3 also point to the plausibility of such a parameter manipulation. This discussion suggests that even when experimental results point to dissociations between two or more processes, models can play a unifying role in suggesting how the system can simultaneously incorporate all these processes.

Knowlton and Squire's data were also simulated in the *SUSTAIN* ("supervised and unsupervised stratified adaptive incremental network") model developed in Love, Medin, and Gureckis (2004) and given a neural basis in Love and Gureckis (2007). The authors of these articles acknowledge the kinship of their work both to exemplar models and to ART (see McDonnell & Gureckis, 2011, for the related idea of "adaptive clustering"). SUSTAIN covers a different database than COVIS, and so the two models may not be in contradiction. Unlike COVIS, SUSTAIN includes unsupervised as well as supervised categorization, is applied primarily to conceptual rather than perceptual inputs, and does not include an implicit learning module.

SUSTAIN attempts to synthesize exemplar-based, prototype-based, and rule-based models by positing that category representation is based on what the authors call clusters, which is subtly different from the definition of clustering given earlier in the chapter. Clusters as they use the term are not necessarily entire categories, but may instead be subcategories of items that are commonly thought of together, or commonalities across many categories. Notably, they define a cluster as "a bundle of features that captures conjunctive relationships across features (e.g., *wings*, *flies*, *and has feathers* tend to co-occur") (Love & Gureckis, 2007, p. 91).

In training episodes, SUSTAIN starts with one cluster based on the first training item. More clusters are added when there are surprising events, that is, novel stimuli in the case of unsupervised learning or prediction errors in the case of supervised learning. If there is no surprise the new stimulus is placed in the dominant cluster and that cluster's prototype moves in the direction of that stimulus (a mechanism very close to what occurs in adaptive resonance networks). Love and Gureckis (2007) propose that the familiarity and novelty processing in this network involves prefrontal cortex (subregion not specified), perirhinal cortex, and hippocampus. PFC compares the current stimulus with perirhinal representations of familiar stimuli, and this circuit influences encoding of surprising events by the hippocampus (see Ranganath & Rainer, 2003; Ranganath et al., 2004). Love and Gureckis summarize that

> . . . the hippocampus constructs codes, the perirhinal cortex generates a familiarity or fit signal, and the PFC monitors and directs encoding and retrieval processes. In terms of SUSTAIN, cluster activations relate to

the fit signal generated by the perirhinal cortex, with cluster evaluation processes carried out by the PFC. When an event is deemed surprising by the PFC, the hippocampus attempts to construct a new cluster.

(pp. 92–93)

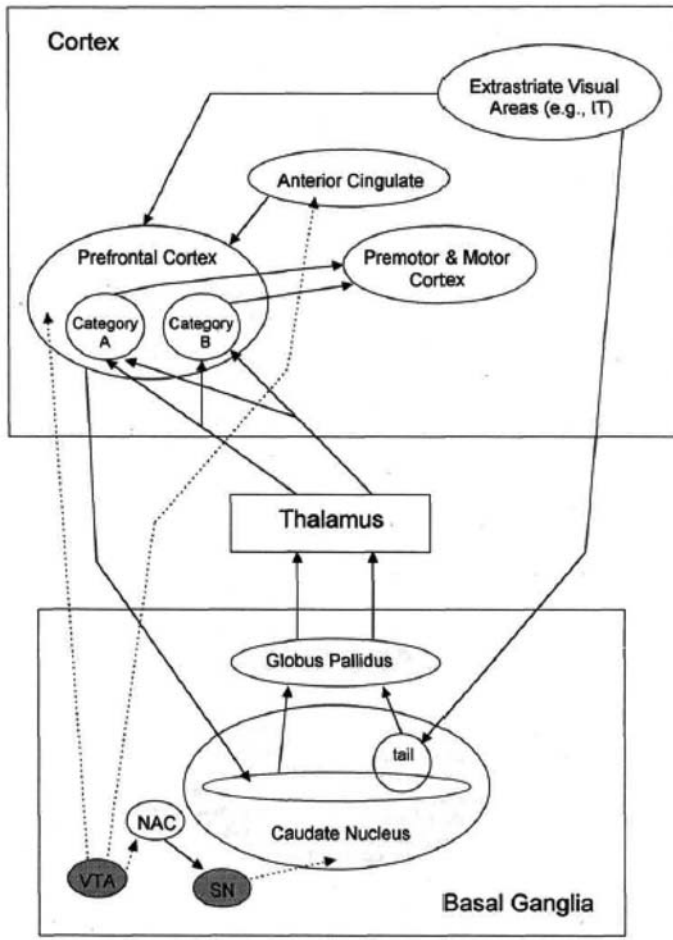


FIGURE 8.9 The fundamental structure of the COVIS model based on competition between verbal and implicit systems. Both systems include loops through basal ganglia, thalamus, and cortex. The verbal system is centered on prefrontal and cingulate executive regions, and the implicit system on the tail of the caudate which receives inputs from high-order visual cortex. Dotted lines denote dopamine pathways. VTA = ventral tegmental area, SN = substantia nigra, NAC = nucleus accumbens, IT = inferotemporal cortex.

Source: Ashby, Alfonso-Reese, Turken, & Waldron, *Psychological Review*, *105*, 442–481, 1998. Copyright 1998 by the American Psychological Association. Reprinted by permission.

Equations for Networks in Chapter 8

Some Implementation Issues for Back Propagation Equations

Recall from Sections 3.3 and 3.5 that the back propagation network of Rumelhart et al. (1986) uses nonlinear (typically sigmoid) activation or input-output functions *f*. The output of unit *j* (output or hidden) is

$$y_{pj} = f(\text{net}_{pj}) = f\left(\sum_j w_{ij} y_{pi}\right) \tag{3.24}$$

with the sum taken over units *i* from the previous layer.

In most of the simulations done by Rumelhart et al., a "bias term" θ_j is added to net_{pj}, the total signal received by unit *j*. This bias, which can be positive or negative, is interpreted as the spontaneous activation level of the *j*th unit, regardless of what inputs it is or is not receiving. Also, these simulations use a specific form of the activation function, namely, the *logistic function*, $f(x) = 1/(1 + \exp(-x))$. Hence, the special case of (3.24) used therein is

$$y_{pj} = \frac{1}{1 + \exp\left(-\left(\sum_i w_{ij} y_{pi} + \theta_j\right)\right)} \tag{8.2}$$

The logistic function has the convenient property that its derivative equals

$$f'(x) = f(x)\big(1 - f(x)\big) \tag{8.3}$$

By (8.3), the change in *f* is largest when *f* is 1/2, and 0 when *f* is 0 or 1. In other words, output units that have not "made up their mind" whether to respond (with a 1) or not respond (with a 0) are subject to the greatest weight changes. In applied problems, as Rumelhart et al. (1986, p. 329) point out, values above .9 or below .1 are usually taken to be "decisions."

The logistic function also simplifies the equations given in Section 3.5 for changes of weights to both output and hidden units. Recall that the expression for the error signal at an output unit, derived from the chain rule for derivatives, is

$$\delta_{pj} = f'\left(\text{net}_{pj}\right)\left(t_{pj} - y_{pj}\right) \tag{3.10a}$$

where y_{pj} is the actual value of the *j*th output unit's activity and t_{pj} its desired value. Combining (3.26) with (8.2) and $y_{pj} = f(\text{net}_{pj})$, we obtain

$$\delta_{pj} = \left(t_{pj} - y_{pj}\right) y_{pj} \left(1 - y_{pj}\right) \tag{8.4}$$

From Section 3.5, error signals at hidden units are related to error signals at output units by

$$\delta_{pj} = f_j'\left(\text{net}_{pj}\right)\sum_k \delta_{pk} w_{jk} \tag{3.10b}$$

where the summation is over all output units k that receive inputs from the jth hidden unit. Equation (3.10b), combined with (3.26), (8.3), and $y_{pj} = f(\text{net}_{pj})$, yields

$$\delta_{pj} = y_{pj}\left(1 - y_{pj}\right)\sum_k \delta_{pk} w_{jk} \tag{8.5}$$

Rumelhart et al. originally made the change in the weight w_{ij} proportional to the product of activation level y_{pi} and error term δ_{pj}, the latter defined by (8.3) or (8.4). But it was found that, with such a rule, too low a constant of proportionality (learning rate) makes learning much too slow, whereas too large a learning rate can lead to wild oscillations. They prevented oscillations at high learning rates by including a *momentum* term that makes the direction of present weight changes partly dependent on the direction of recent past changes. Hence, the rule they actually used is

$$\Delta w_{ij}\left(t+1\right) = \beta\left(\delta_{pj} y_{pi}\right) + \alpha \Delta w_{ij}\left(t\right) \tag{8.6}$$

where β (the learning rate) and α (the momentum) are two separate positive constants.

ARTMAP Equations

ARTMAP consists of two ART 1 modules, labeled "*a*" and "*b*," and a map field that connects them. The effects of the algorithm described by (7.12) through (7.15) and the constraints on gain control, 2/3 rule matching, and so forth, can be simplified in a "fast learning" case as follows (see Carpenter, Grossberg, & Reynolds, 1991):

Let the feature layer F_1 have M nodes, and let the category layer F_2 have N nodes. Let $\mathbf{I}$ denote the current input vector. In this case of binary inputs, the norm or absolute value symbol $|\mathbf{I}|$ denotes the number of components of the vector $\mathbf{I}$ that are equal to 1.

All F_2 nodes are initially uncommitted; hence, at time 0 the weights to a given F_2 node, labeled Z_{ij}, are equal from all the F_1 nodes, that is,

$$Z_{ij}(0) = \alpha_j \tag{8.7}$$

where the small positive numbers α_j are such that $\alpha_1 > \alpha_2 > \ldots > \alpha_n$ and each α_j is less that the value $1/(\beta + |\mathbf{I}|)$, with β also being a small positive number. If z_{ji} denote the top-down weights from F_2 to F_1, then

$$z_{ij}(0) = 1 \tag{8.8}$$

As particular F_2 nodes become active (by processes to be shown below), their vectors of z_{ji} values, denoted as $\mathbf{z}_j$, become mixtures of 1s and 0s representing prototype patterns. The binary F_1 output vector $\mathbf{x} = (x_1, \ldots, x_M)$ then becomes

$$\begin{aligned} \mathbf{x} &= \mathbf{I} \quad \text{if } F_2 \text{ is inactive} \\ &\mathbf{I} \cap \mathbf{z}_J \text{ if the } J\text{th } F_2 \text{ node is active} \end{aligned} \tag{8.9}$$

where the intersection (() denotes a binary vector that has 1s in the components where both vectors on either side of the symbol have 1s. This leads to an input from the F_1 field to each F_2 node; the input T_j to the jth F_2 node is

$$\begin{aligned} T_j &= |\mathbf{I}|\,\alpha_j \quad \text{if } j \text{ is the index of an uncommitted node} \\ &\frac{|\mathbf{I} \cap \mathbf{z}_J|}{\beta + |\mathbf{z}_J|} \quad \text{if } j \text{ is the index of a committed node} \end{aligned} \tag{8.10}$$

The initial choice of F_2 node is the one with the largest value of T_j. Then the F_2 output vector is denoted by $\mathbf{y} = (y_1, \ldots, y_N)$.

Resonance occurs if for the chosen node J (the one with the largest T_j value),

$$|\mathbf{I} \cap \mathbf{z}_J| \geq \rho |\mathbf{I}| \tag{8.11}$$

where ρ is the vigilance parameter (remembering that in ARTMAP, unlike ART 1 standing alone, the value of the vigilance can change with learning). If (8.11) is satisfied and the input is placed in the category coded by node J, then both the top-down weight vector $\mathbf{z}_J$ from that node and the bottom-up weight vector $\mathbf{Z}_J$ to that node are updated. The new values of these vectors are

$$\begin{aligned} \mathbf{z}_J &= \mathbf{I} \cap \mathbf{z}_J^{(\text{old})} \\ \mathbf{Z}_J &= \frac{\mathbf{I} \cap \mathbf{z}_J^{(\text{old})}}{\beta + \left| \mathbf{I} \cap \mathbf{z}_J^{(\text{old})} \right|} \end{aligned} \tag{8.12}$$

The ARTMAP algorithm links together two fast-learn ART 1 modules, called ART_a and ART_b, each obeying Equations (8.7) through (8.12), with the superscripts *a* and *b* denoting which one a given variable refers to. The algorithm also incorporates an inter-ART module with a map field, via the rules to be described.

In addition to the variable vigilance (see below), another change from previous versions of ART is that the map field F^{ab} (which has the same number of nodes as F_2^b, the category field of ART_b) can *prime* ART_b. That is, if F^{ab} sends nonuniform input to F_2^b in the absence of an input to ART_b, F_2^b remains inactive but its subsequent activity is influenced by the input from F^{ab}. Specifically, as soon as the input to F_2^b from F^{ab} arrives, F_2^b chooses the node K receiving the largest input from the map field. This node in turn sends to F_1^b the top-down input $\mathbf{z}_K^b$

In the full ARTMAP network, the output vector of x_1^a is denoted by $\mathbf{x}^a = (x_1^a, \ldots, x_{Ma}^a)$ and the output vector of x_2^a by $\mathbf{y}^a = (y_1^a, \ldots, y_{Na}^a)$, with analogous superscripts for the outputs of the ART_b module. Denote by $\mathbf{x}$ the output of the map field F_2^b. Between input presentations, the five vectors $\mathbf{x}^a$, $\mathbf{y}^a$, $\mathbf{x}^b$, $\mathbf{y}^b$, and $\mathbf{x}$ are all set to 0.

Let $\mathbf{w}_j = (w_{j1}, \ldots, w_{jNb})$ be the vector of weights from the jth F_2^a node to F^{ab}. The vector $\mathbf{w}_j$ starts with every component equal to 1, but converges during resonance with the jth category active to the current value of the map field output vector $\mathbf{x}$. The vector $\mathbf{x}$ in turn obeys

$$\mathbf{x} = \begin{cases} y^b \cap \mathbf{w}_J & \text{if the } J\text{th } F_2^a \text{ node is active and } F_2^b \text{ is active} \\ w_J & \text{if the } J\text{th } F_2^a \text{ node is active and } F_2^b \text{ is inactive} \\ y^b & \text{if the } F_2^a \text{ node is inactive and } F_2^b \text{ is active} \\ 0 & \text{if the } F_2^a \text{ node is inactive and } F_2^b \text{ is inactive} \end{cases} \tag{8.13}$$

Let ρ_a be the ART_a vigilance and ρ the map field vigilance. If there is mismatch between the current prediction of an ART_b category by ART_a and current ART_b activity, that is (by the first line of (8.13))

$$|\mathbf{x}| = |y^b \cap \mathbf{w}_J| < \rho |\mathbf{y}^b| \tag{8.14}$$

then if $\mathbf{a}$ denotes the vector of inputs to ART_a, ρ_a is increased until it is slightly larger than $|\mathbf{a}|\mathbf{z}_J^a|/|\mathbf{a}|$. By (8.13), this means that

$$|\mathbf{x}^a| = |\mathbf{a} \cap \mathbf{z}_J^a| < \rho_a |\mathbf{a}| \tag{8.15}$$

where J is the index of the active F_2^a node. That leads to F_2^a reset: either activation of a new F_2^a node for which neither mismatch condition (8.13) or (8.14) holds, or if no such node exists, shutting down of F_2^a as long as the current input is on.

The learning of weights in ARTMAP can be broken down into nine cases: inputs $\mathbf{a}$ and $\mathbf{b}$ to the two ART 1 modules could appear alone, or one before the other; $\mathbf{a}$ could make a prediction of ART_b activity based on prior learning, or not; if $\mathbf{a}$ makes a prediction, $\mathbf{b}$ could confirm or disconfirm the prediction. If it is assumed that all $|\mathbf{a}|$ are the same constant, the changes in the top-down and map field weight vectors $\mathbf{z}_J^a$, $\mathbf{z}_K^b$, and $\mathbf{w}_K$ are listed, and these lead to corresponding changes in the bottom-up weights $\mathbf{Z}_j^a$ and $\mathbf{Z}_k^b$ by (8.11):

a *only*, *no prediction*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$

a *only*, *with prediction*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$

b *only*: $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$

a *then* **b**, *no prediction*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$, $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$, and $\mathbf{w}_J \rightarrow \mathbf{y}^b$, the effect of which is that category J of ART_a learns to predict category K of ART_b

a *then* **b**, *prediction confirmed*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$, $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$

a *then* **b**, *prediction not confirmed*: this leads to selection of a new ART_b category K, then increase of the ART_a vigilance ρ_a as described earlier. If an ART_a category is found for which resonance occurs both within ART_a and between the two ARTs (i.e., (8.13) and (8.14) are both false), then $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$ and $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$, and if node J is uncommitted, $\mathbf{w}_J \rightarrow \mathbf{y}^b$. If no such node is found, then only $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$.

b *then* **a**, *no prediction*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$, $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$, and $\mathbf{w}_J \rightarrow \mathbf{y}^b$, the effect of which is that category J of ART_a learns to predict category K of ART_b

b *then* **a**, *prediction confirmed*: $\mathbf{z}_J^a \rightarrow \mathbf{z}_J^{a(\text{old})} \cap \mathbf{a}$, $\mathbf{z}_K^b \rightarrow \mathbf{z}_K^{b(\text{old})} \cap \mathbf{b}$

b *then* **a**, *prediction not confirmed*: like the case of ***a** then **b**, prediction not confirmed*.

The supervised classification of analog vectors using fuzzy ARTMAP (Carpenter et al., 1992) largely follows the lines of ARTMAP and Equations (8.6) through (8.14) with the same changes as were applied in going from ART to fuzzy ART. That is, the intersection operator "$\cap$" is replaced by the fuzzy AND operator "$\wedge$," which takes the minimum of corresponding components of the two vectors (see the earlier section on fuzzy ART equations), and the norm operator "|" is interpreted as the sum of all components.

Inputs to both ART modules in fuzzy ARTMAP, like those to fuzzy ART, are complement-coded. That is, an input is followed by its "complementary" pattern whose values are 1 minus the original values (e.g., following .8, .4, 1 by .2, .6, 0), creating a vector with constant norm equal to twice the number of nodes in the appropriate layer.

Brain-State-in-a-Box Equations

Section 8.7 listed the equation of Anderson et al. (1977) for linear transformation of the node activity vector (initially an input vector) over time. This equation is reproduced here:

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) \tag{8.1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the vector of node activities at any given time, and $\mathbf{A}$ is the connectivity matrix of the system. But it was noted that Equation (8.1) can create instability through positive feedback; that is, it can drive the state vector

of the system outside the hypercube (box) of Figure 8.8. To prevent such instability, Anderson et al. added a rule whereby any node activity driven above 1 by (8.0) is replaced by 1, and symmetrically, any activity driven below –1 is replaced by –1. Hence, if the subscript $_i$ below any vector denotes its ith component, i between 1 and n, (8.1) is replaced by the equation

$$x(t+1)_i = \max\left(-1, \min\left(1, \left[(I + A)x(t)\right]\right)\right), \quad i = 1, \ldots, n \tag{8.16}$$

where **I** denotes the identity matrix (1 along the main diagonal, 0 elsewhere).

Equation (8.16) is the one that Anderson et al. (1977) actually used in their simulations. The matrix **A** is constrained to be *symmetric*; that is, for each i and j, the entry a_{ij} in the ith row and jth column is identical to a_{ji}. Symmetry simplifies a matrix's mathematical properties. For example, a symmetric n-by-n matrix always has n mutually orthogonal *eigenvectors*, that is, n-dimensional vectors $\mathbf{x}_i$ such that $\mathbf{A}\mathbf{x}_i$ is a constant multiple of $\mathbf{x}_i$. The constant that is multiplied is called an *eigenvalue* of the matrix **A** (see Jordan, 1986a, for discussion). In order to simulate the positive feedback process involved in classification, Anderson et al. used a matrix for which all eigenvalues are positive, and all eigenvectors are corners of the box in Figure 8.8 (e.g., see Exercise 2 of this chapter).

Anderson and Murphy (1986) used a variant of (8.16) that incorporates the initial state (i.e., input) vector $\mathbf{x}(0)$, in order to keep input information present. Their equation for the change in the state vector **x** over time is

$$\left(\mathbf{x}(t+1)\right)_i = \max\left(-1, \min\left(1, \left[\alpha \mathbf{A}\mathbf{x}(t) + \Gamma \mathbf{x}(t) + \delta \mathbf{x}(0)\right]_i\right)\right), \quad i = 1, \ldots, n$$

where α, Γ, and δ are different positive constants.

In addition, Anderson and Murphy (1986) include a scheme for learning the correct matrix **A**. The optimal **A** for forming a heteroassociative connection **x**-to-**y**, with **x** and **y** both vectors, is one for which

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{8.17}$$

Anderson and Murphy's error-correcting scheme for incrementing the matrix, based on the rule of Widrow and Hoff (1960), was designed to move **A** closer to satisfying (8.17). Their equation for the change in **A** is

$$\Delta \boldsymbol{A} = \beta\left[\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right) \cdot \boldsymbol{x}\right] \tag{8.18}$$

The learning rate β in (8.18) can be either fixed or adjustable. The matrix learning scheme defined by (8.18) can also, the authors pointed out, be used in auto-associative encoding by letting **y** = **x**. In that case, the learning scheme becomes an algorithm for making **x** an eigenvector of **A** with an eigenvalue equal to 1.

Equations for the SPEED Model of Ashby et al. (2007)

In the SPEED model of the automatic part of categorization (Ashby et al., 2007), sensory inputs are treated as gestalts and not broken down by attributes or features. When a stimulus is presented, the activation of sensory cortical unit K at time t is

$$I_k(t) = \frac{1}{\alpha}\exp\left(-\frac{d\left(K,\text{stimulus}\right)^2}{2\alpha^2}\right) \tag{8.19}$$

where α is a constant, and $d(K, \text{stimulus})$ is the distance (in stimulus space) between the stimulus preferred by unit K and the presented stimulus. Equation (8.19) is an example of a *radial basis function* (e.g., Buhmann, 2003).

Striatal (medium spiny) unit activations $S_j(t)$ are determined by weighted sums of activations in all visual cortical units projecting to a given striatal unit and by lateral inhibition from other striatal units, thus:

$$\begin{aligned}\frac{dS_j(t)}{dt} = &\left[\sum_K W_{K,j}(n) I_k(t)\right]\left[1 - S_j(t)\right] \\ &- \beta_S S_M(t) - \gamma_S\left[S_j(t) - S_{\text{base}}\right] + \sigma_S \varepsilon(t) S_j(t)\left[1 - S_j(t)\right]\end{aligned} \tag{8.20}$$

where β_S, γ_S, S_{base}, and σ_S are constants, $M \neq J$, $I_K(t)$ is as in Equation (8.19), $w_{K,J}(n)$ is the connection weight from cortical unit K and striatal unit J on trial n, and $\varepsilon(t)$ is white noise.

Activation in the jth unit of the globus pallidus at time t, denoted by $GJ(t)$, is described by

$$\frac{dG_j(t)}{dt} = -\alpha_G S_j(t) G_j(t) - \beta_G\left[G_j(t) - G_{\text{base}}\right] \tag{8.21}$$

where α_G, β_G, and G_{base} are constants. This indicates inhibition from the corresponding striatal unit and decay to baseline.

Activation in the jth unit of the thalamus obeys an equation similar to (8.21):

$$\frac{dT_j(t)}{dt} = -\alpha_T G_j(t) T_j(t) - \beta_T\left[T_j(t) - T_{\text{base}}\right] \tag{8.22}$$

with one major difference. T_{base} is set to a value higher than the expected spontaneous firing rate of thalamic neurons to also include excitatory inputs to the thalamus, most notably inputs from the PFC.

To each of the units in the striatum, globus pallidus, and thalamus there corresponds a unit in the premotor cortex. The activations of those premotor nodes are controlled by the equations

$$\frac{dE_j(t)}{dt} = \left[\alpha_E T_j(t) + \sum_K V_{K,J}(t) I_k(t)\right]\left[1 - E_j(t)\right] - \beta_E E_k(t) - \gamma_E \left[E_j(t) - E_{\text{base}}\right] + \delta_E \varepsilon(t) E_j(t)\left[1 - E_j(t)\right]$$

where $v_{K,j}(n)$ is the connection weight from visual cortical unit K and premotor unit j on the nth trial, and other symbols are defined analogously to those in previous equations.

There are as many premotor units as there are possible responses. In tasks with only one possible response, it is assumed that a response is initiated when the integral over time of the activation in the premotor unit first exceeds a threshold τ. In tasks with two possible responses, A and B, it is assumed that response A is given when the integral over time of $E_A(t) - E_B(t)$ first exceeds τ, and response B when that same integral is first less than $-\tau$.

As for the weights, the two critical synapses are between sensory association and motor cortex units and between sensory association cortex and spines of striatal units. The sensory-motor weights are modified by *two-factor*, that is, Hebbian associative learning, and the cortico-striatal weights are modified by *three-factor* learning that also incorporates the effects of dopamine. It is assumed that weights, unlike node activities, change only at the end of trials, so while node activities are defined by differential equations, weights are defined by difference equations.

If $v_{K,J}(n)$ denotes the strength of the synapse on trial n between sensory cortical unit k and premotor unit J, then the two-factor learning equation for that synaptic weight is the difference equation

$$v_{k,J}(n+1) = v_{k,J}(n) + \alpha_v \sum_t I_k(t)\left[\sum_t E_J(t) - \theta_{\text{NMDA}}\right]^+ \times \left[1 - v_{k,J}(n)\right] - \beta_v \sum_t I_k(t)\left[\theta_{\text{NMDA}} - \sum_t E_J(t)\right]^+ v_{k,J}(n) \tag{8.23}$$

where the sum over t denotes total activation of the appropriate unit over the course of the trial; α_v and β_v are constants; and θ_{NMDA} is the activation threshold of the NMDA receptor. The term in (8.22) starting with α_v describes the conditions under with LTP occurs, and the term starting with β_v describes the conditions under with LTD occurs.

As for three-factor learning, if $w_{K,J}(n)$ denotes the strength of the synapse on trial n between sensory cortical unit k and striatal unit J, then

$$
\begin{aligned}
w_{k,J}(n+1) = w_{k,J}(n) \qquad & (8.24)\\
&+ \alpha_w \sum_t I_k(t) \left[\sum_t S_J(t) - \theta_{\text{NMDA}} \right]^+ \left[D(n) - D_{\text{base}} \right]^+ \left[1 - w_{k,J}(n) \right]\\
&- \beta_w \sum_t I_k(t) \left[\sum_t S_J(t) - \theta_{\text{NMDA}} \right]^+ \left[D_{\text{base}} - D(n) \right]^+ w_{k,J}(n)\\
&- \gamma_w \sum_t I_k(t) \left[\theta_{\text{NMDA}} - \sum_t S_J(t) \right]^+ w_{k,J}(n)\\
&- \phi_w \left(1 - \frac{1 - \left[D(n) - D_{\text{base}} \right]^+}{1 - D_{\text{base}}} \right) w_{k,J}(n)
\end{aligned}
$$

where D_{base} is the baseline firing rate of dopamine cells and $D(n)$ is the amount of dopamine released following feedback on trial n. The second line of Equation (8.24) describes the conditions that produce LTP (striatal activation above the threshold for NMDA receptor activation and dopamine above baseline), and lines three and four describe conditions that produce LTD (striatal activation above the NMDA threshold but dopamine below baseline, or striatal activation below NMDA threshold). The last line models a slow decay in synaptic strength that occurs when dopamine stays at baseline levels over a long period of time. In all simulations, the initial cortical-striatal weights $w_{i,J}(0)$ are set to be strong enough to cause one or more striatal units to fire when a novel stimulus is first presented.

The dynamics of the dopamine release variable $D(n)$ are simpler than in some other models (e.g., Brown et al., 1999), and different depending on whether the response on trial n is correct or not. If $P(C)$ is the probability of a correct response, then a correct response yields a value $D(n) = D_{\text{base}} + [1 - P(C)](1 - D_{\text{base}})$, whereas an incorrect response yields $D(n) = D_{\text{base}} - P(C)D_{\text{base}}$.

Exercises for Chapter 8

*1. Simulate a back propagation network with 0 biases. Let the input units form a 5(5 grid, and each hidden unit respond to a 3(3 subgrid, as described in Rumelhart et al. (1986) and shown in Figure 8.10.

Use initial weights of .5 for all connections. Teach the network the "T–C" discrimination, so that it should respond with a 1 to T, 0 to C, in any rotation but centered at (3,3) in the visual field, as shown in Figure 8.11.

Use the sigmoid function $1/(1 + \exp(-x))$. Do not correct for edge effects. Use a positive momentum term (you can experiment with that). Look at the final response patterns of the hidden units after several thousand iterations.

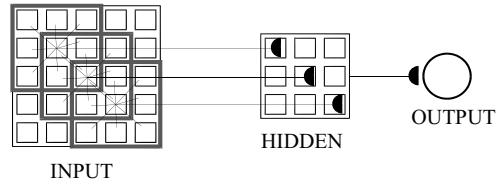


FIGURE 8.10 Example of the three-layer back propagation network used in the simulation of Exercise 1.

○2. Refer to the discussion of "semantics without categorization" in Section 8.2 of this chapter. Rogers and McClelland (2011) argued that it was possible to learn category-related properties of objects and concepts without explicitly putting them into categories. Comment on whether you believe that explicit categorization is necessary for appropriate concept learning. You might decide that categorization is important in some cognitive contexts and not in some others.

*3. Simulate the brain-state-in-a-box equations of Anderson et al. (1977):

$$x(t+1)_i = \max\left(-1, \min\left(1, \left[(I + A)x(t)\right]\right)\right), \qquad (8.16)$$
$$i = 1, \ldots, n$$

(a) Choose $n = 4$, and choose the matrix **A** to be

$$\begin{bmatrix} 3/2 & 0 & 0 & 1/2 \\ 0 & 3/2 & 1/2 & 0 \\ 0 & 1/2 & 3/2 & 0 \\ 1/2 & 0 & 0 & 3/2 \end{bmatrix}.$$

This matrix has the two four-dimensional hypercube corners **P** = (1, 1, –1, –1) and **Q** = (1, –1, 1, –1) as eigenvectors, both with eigenvalue 1. Define the 16 starting points $\mathbf{Q}_i$, $i = 0$ to 15, by $\mathbf{Q}_0 = \mathbf{P}$, $\mathbf{Q}_{15} = \mathbf{Q}$, and $\mathbf{Q}_i = (1, r_i, s_i, -1)$, where $r_i = \cos\theta_i + \sin\theta_i$, $s_i = -\cos\theta_i + \sin\theta_i$, $\theta_i = 12i$ degrees. Thus the $\mathbf{Q}_i$ have their first and last components fixed at 1 and – 1, their middle two at equal spaces along the arc of the circle between (1, –1) and (–1, 1).

Test the "categorizations" made by the network based on the $\mathbf{Q}_i$ as starting positions, that is, whether the final state of the network is **P** or **Q**. With no noise added to the $\mathbf{Q}_i$, the final state should always be **P** for $i = 0$ to 7, **Q** for $i = 8$ to 15. Then test the categorizations with Gaussian noise of various standard deviations[2] added to each component of each $\mathbf{Q}_i$.

(b) Do the same as in part (a) except with the matrix **A** equal to

$$\begin{bmatrix} 3/2 & 0 & 1/2 & 0 \\ 0 & 3/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 3/2 \\ 1/2 & 0 & 3/2 & 0 \end{bmatrix}.$$

This matrix has **P** and **Q** as eigenvectors with eigenvalues 1 and 2, respectively, giving the network a bias in favor of going to the corner **Q**.

*4. "Circle-in-the-square" problem (Carpenter et al., 1992): Use fuzzy ARTMAP to identify which points of a square lie inside or outside a circle whose area equals half that of the square. Using complement coding, if the input **a** is a point on the square with coordinates (a_1, a_2), $0 < a_1, a_2 < 1$, the input is coded in the F_1 level of ART_a as $(a_1, a_2, 1-a_1, 1-a_2)$. ART_b has two categories, inside and outside, and in complement coding form the input to F_1^b is (1, 0) for inside, (0, 1) for outside. Use one training epoch presenting a set of points, then on each simulation test responses to those same points.

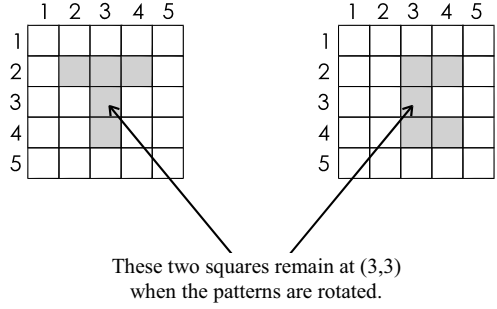


FIGURE 8.11 Basic "T" and "C"input patterns used by Rumelhart et al. (1986). The network of Figure 8.10 can be trained to classify all of the four rotations of the "T" in one category and all of the four rotations of the "C" in another category.

Some hints on the circle-in-the-square:

(1) In the simulations by Carpenter et al. (1992), as the training set increased from 100 to 100,000 exemplars (points on the square), the rate of correct predictions increased from 88.6% to 98.0%. At the same time, the number of ART_a categories created by the algorithm increased from 12 to 121.

(2) In each simulation, as a consequence of the fuzzy ART algorithm, the categories established in ART_a will be represented geometrically as rectangles, with the corners having in each coordinate the minimum

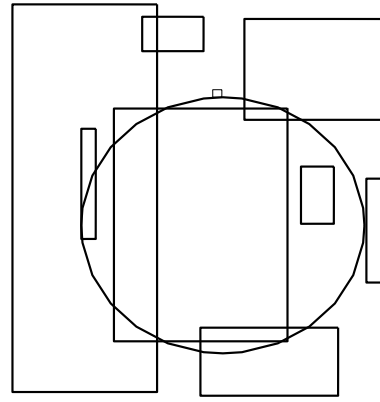


FIGURE 8.12 Fuzzy ARTMAP category rectangles for circle-in-the-square simulations with 100 exemplars and 12 ARTa categories.

Source: Adapted from Carpenter et al., copyright 1992 IEEE, with the permission of the publisher.

and maximum of the abscissa or ordinate for current members of the category. Keep track of where the rectangles are located. The ones near the circle should become smaller as there are more exemplars and more categories. An example of such rectangles is shown in Figure 8.12.

(3) Performance can be improved by "voting," that is, having more than one simulation (say, about five simulations) of the same set of exemplars but presented in different orders, and taking a "majority vote" of the decisions as to whether a given point is inside or outside the circle.

Some Additional Sources

Back Propagation Networks for Categorization

Apfelbaum and McMurray (2015); Tijsseling and Gluck (2002).

Deep Learning Networks

Kim, Choi, and Lee (2015); Tissera and McDonnell (2016); Zorzi, Testolin, and Stoianov (2013).

ARTMAP Networks

Carpenter (1997); Carpenter and Gaddam (2010); Carpenter and Markuzon (1998); Carpenter, Milenova, & Noeske (1998); Carpenter & Olivera (2012); Marriott and Harrison (1995); Vigdor and Lerner (2007); Williamson (1996).

BSB Networks

Mathematical Theory

Hui and Zak (1992); Lillo, Miller, Hui, and Zak (1994); Perfetti (1995).

Further Extensions of the General Algorithm

Anderson (1993); Anderson and Sutton (1995, 1997); Sutton and Breiter (1994).

Application to Learning Arithmetic

Anderson (1998); Anderson, Spoehr, and Bennett (1994).

Other Networks for Supervised Classification

Ashby and Crossley (2011); Gluck and Bower (1988a, 1988b); Hélie, Paul, and Ashby (2012).

Notes

1. Relations such as "is a" and "has" also appear in the model of analogy making by Jani and Levine (2000) discussed in Section 9.6.1.
2. Some computer systems have access to a package that generates Gaussian noise, that is, generates a random variable with a normal ("bell-shaped") distribution. If you do not have access to such a program, the following procedure (Box & Muller, 1958) can be used for this purpose. First, let u_1 and u_2 be two random variables *uniformly* distributed between 0 and 1, each obtained, for example, by the algorithm in Chapter 2, Exercise 3. Then $x_1 = [-2 \ln u_1]^{1/2} \sin (2\pi u_2)$, $x_2 = [-2 \ln u_1]^{1/2} \cos (2\pi u_2)$ are two independent Gaussian variables of mean 0 and standard deviation 1; to get a variable of standard deviation s, multiply x_1 or x_2 by s.

9

MODELS OF COMPLEX MENTAL FUNCTIONS

> I suppose an entire cabinet of shells would be an expression of the whole human mind; a Flora of the whole globe would be so likewise, or a history of beasts; or a painting of all the aspects of the clouds. Everything is significant.
>
> Ralph Waldo Emerson

> We may affirm absolutely that nothing great in the world has been accomplished without passion.
>
> Georg Hegel (Philosophy of History)

9.1. How Do We Model Complex Brain Functions?

As recent brain imaging results supplement animal, lesion, and behavioral studies to provide more pieces of the cognitive neuroscience puzzle, the goal of a self-consistent computational model of the entire brain seems inviting and possibly attainable. How can we approach the search for a brain model that is fairly comprehensive as well as illuminating, and making predictions about, important cognitive functions?

Some theories proceed bottom-up from the neuroanatomy and neurophysiology, looking at the entire brain and mapping out its connectivity patterns (e.g., Sporns, 2010, 2012). Others proceed mainly top-down from overarching functions such as optimizing some utility or quality-of-life function over time, which Werbos (e.g., 1992b, 2009) sees as a major organizing principle (though not as an explanation for everything the brain does). Still others, combining bottom-up and top-down approaches, focus on specific processes (vision, conditioning, decision-making, etc.) and design models to bridge the paradoxes those processes entail (e.g., Grossberg, 2000).

This chapter is organized in terms of models for different processes. It partly parallels Chapter 5, which is organized in terms of cognitive neuroscience results for some of these same processes. The reader seeking to compare different models of a particular process should ideally keep in mind some of the criteria of Meeter et al. (2007) discussed in Chapter 1. Those authors proposed that, in addition to fitting known data and making predictions about novel data, the model should build on assumptions that make biological and behavioral sense. They also proposed that models of different levels of the same process should be interconnected and mutually consistent. In thinking of the brain as interconnected, we can extend the latter prescription to models of different processes that influence each other. Hence, the reader should try to discover unstated connections between models discussed in different sections of this chapter, and notice how some models discussed in this chapter build on architectures developed for simpler functions in earlier chapters.

Researchers who develop models of complex brain function vary widely in their relative emphasis on neuroscientific versus behavioral or cognitive details. Yet this book has an overall bias toward models that are initially motivated by efforts to incorporate the requirements of behavioral and/or cognitive tasks, with the details of the models further constrained by specific neural as well as behavioral data.

9.2. Models of Vision and Visual Attention

Chapter 4 of this book presents early models of illusory contours and other illusory percepts that are by-products of the cortical system that compensates for imperfections in the retinal uptake of visual stimuli. These illusions typically arise preattentively as part of the process of grouping objects in the visual environment. Chapter 7 presents models of top-down attentional influences on the same cortical visual system.

Yet, as noted in Chapter 5, there is a paradox between the separate cognitive requirements of the grouping and attentional operations. Top-down attentional feedback requires inhibiting the perception of some objects seen by lower brain levels. Also it is important that top-down attentional connections not be able to activate primary visual levels above threshold (*supraliminally*), in order that we not have hallucinations of seeing illusory objects just by thinking about them. Yet the grouping process requires perceptual filling-in parts of contours that are not physically present in the environment, and in fact neurons at the visual area V2 respond to lines and curves in those illusory contours as they respond to actual lines and curves of the same orientation (Peterhans & von der Heydt, 1989; von der Heydt & Peterhans, 1989).

9.2.1. Modeling Grouping and Attention

Grossberg (1999) ruled out the possibility that attention and preattentive grouping engage different cortical areas, noting that visual areas V1 and V2 are involved in both sets of processes. He then developed a theory that synthesizes the attentional matching of ART (Carpenter & Grossberg, 1987a; Chapter 7 of this book) with the preattentive grouping performed by the boundary contour system (BCS) (Grossberg & Mingolla, 1985b; Chapter 4 of this book). The synthesis developed in that article turns out to engage the laminar architecture of the visual cortex.

The starting point for Grossberg (1999) was the observation that different layers of the cortex have characteristically different structures of competitive and cooperative interactions. Perceptual grouping starts in Layers 2 and 3, where complex pyramidal cells excite each other using monosynaptic long-range horizontal connections, and inhibit each other using short-range disynaptic inhibitory connections (see Hirsch & Gilbert, 1991; McGuire, Gilbert, Rivlin, & Wiesel, 1991). These are the bipole interactions discussed in Chapter 4 that support illusory contours. If there are two collinear line segments that are separated, the neurons responding to those lines excite one another, leading to an inward perceptual grouping (Peterhans & von der Heydt, 1989). Yet, because of the short-range inhibition there is no outward grouping from a single line (Hirsch & Gilbert, 1991).

Direct inputs from the retina via the lateral geniculate nucleus (LGN) go to other cortical layers, mainly Layers 4 and 6. Layer 4 activates Layers 2 and 3. Layers 2 and 3 send excitatory signals to Layer 6, which in turn connects back to Layer 4 via a nonrecurrent on-center off-surround network (Grieve & Sillito, 1995). This *folded feedback* (the term used in Grossberg, 1999) enables different possible grouping to compete with one another to select those groupings that make the most "sense" perceptually.

Based on the on-center off-surround structure of attentional influences of V2 on V1 (Bullier et al., 1996) and of V1 on LGN (Sillito et al., 1995), the model of Grossberg (1999) extends the structure of these interlayer connections to connections between layers in these different regions, as shown in Figure 9.1. For both V2-to-V1 and V1-to-LGN connections, the attentional feedback in the model is modulatory, that is, enough to enhance excitation caused by external stimulation but too weak to cause neurons to fire in the absence of such stimulation. The modulatory nature of attention is in keeping with the ART matching rule (see Chapter 7) designed to avoid hallucinations, and has been supported by a variety of data, notably that Layer 4 EPSPs elicited by Layer 6 stimulation are much smaller than EPSPs caused by stimulation of LGN axons or of neighboring Layer 4 sites (Stratford, Tarczy-Hornoch, Martin, Bannister, & Jack, 1996).

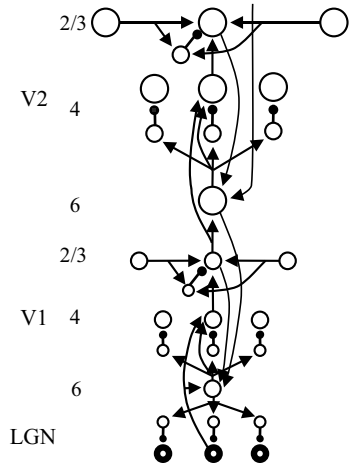


FIGURE 9.1 Connections between different layers of lateral geniculate and visual areas V1 and V2 in the model of Grossberg and Raizada. V2 has the same laminar pattern as V1 but ata larger spatial scale. V1 Layer 2/3 projects to V2 Layers 6 and 4, while LGN projects to layers 6 and 4 of V1. Higher cortical areas send feedback to Layer 6 of V2, and V2 sends feedback to Layer 6 of V1.

Source: Adapted with permission from Grossberg & Raizada, 2000.

Variants of a network model based on Figure 9.1 were simulated by Grossberg and Raizada (2000) and Raizada and Grossberg (2001, 2003); the equations from Raizada and Grossberg (2001) are given at the end of this chapter. This model reproduces a range of visual data involving interactions between attention and grouping, including the greater effect of attention on low-contrast than on high-contrast stimuli (DeWeerd, Peralta, Desimone, & Ungerleider, 1999); attention flows along real and illusory contours (Roelfsema et al., 1998); and a difference in effects of flanker stimuli on responses of V1 neurons to a high-contrast versus a low-contrast Gabor stimulus[1] (Polat, Mizobe, Pettet, Kasamatsu, & Norcia, 1998). Raizada and Grossberg (2001) called their model *LAMINART* because it combines the laminar organization of the visual cortex with the multilayer architecture of adaptive resonance theory or ART (see Sections 7.2 and 8.4).

How the connections of Figure 9.1 are built up in the course of development was the topic of the article by Grossberg and Williamson (2001), which extended some of the coding ideas discussed in Section 7.1. The Grossberg–Williamson model constrains the development of connections both within and between cortical layers by regulating balance between excitatory and inhibitory connections within Layer 2/3 and between Layers 4 and 6. The course of growth of connections is governed by two developmental rules. One of the rules is that axons are attracted to cell targets when the source and target cells are

both active. The second rule is that target cells compete for axons from the same source cell, so that connections to target cells that only receive weakly correlated signals are removed. The first rule embodies a form of Hebb-type associative learning and the second helps to limit the growth of associative strengths.

Associative learning and competition among synapses are also main features of the visual cortex model called *LISSOM* (*laterally interconnected synergetically self-organizing map*), best described in the book by Miikkulainen, Bednar, Choe, and Sirosh (2005). LISSOM has been used to model such data as tilt aftereffects (Bednar & Miikkulainen, 2000) and illusory contour formation (Choe & Miikkulainen, 2004).

The LISSOM network is designed to model the same general dataset on self-organizing or orientation detectors, perceptual grouping, and illusory contours as the LAMINART models of Grossberg, Raizada, and Williamson. Yet there are several significant differences between the two sets of models. First, LISSOM relies on plasticity of lateral connections within the visual cortex as well as the afferent connections with the retina, whereas LAMINART has plasticity only in the afferent connections. Second, LAMINART relies on shunting excitation and inhibition (see Section 4.2), whereas all the excitation and inhibition in the different versions of LISSOM is additive or subtractive. Equations for both the basic LISSOM and the spiking version developed by Choe and Miikkulainen (2004), called PGLISSOM, are given at the end of this chapter; spiking was added in order to incorporate a role for neural signal synchronization in perceptual grouping. Third, the LAMINART models include retina, LGN, and the layers of both V1 and V2, whereas LISSOM includes only retina and two layers of V1, corresponding to Layers 2/3 and 4. The greater number of layers in LAMINART allows it to capture top-down attentional influences on vision, as described earlier in this section. Yet, one aspect of vision that is included in the most recent versions of LISSOM, and not yet in LAMINART, is differences between parts of the retinal visual field; specifically, between periphery and fovea and between upper and lower hemifields.

The use of spikes for synchronization of responses to inputs to model perceptual grouping has been a feature of many other visual cortex models (e.g., Borisyuk, Kazanovich, Chik, Tikhanoff, & Cangelosi, 2009; Terman & Wang, 1995; Wang, 2000). Yet synchronization has been achieved in other visual models without explicit spiking (e.g., Grossberg & Somers, 1991; Grossberg & Grunewald, 1997).

Another model of visual attention and search (but not grouping) that includes top-down influences is by Usher and Niebur (1996). Their model, based on integrate-and-fire dynamics (see Appendix 1), includes three layers that are labeled input, sensory memory, and working memory and correspond respectively to V1, inferotemporal (IT), and prefrontal cortex. Usher and Niebur's model reproduces monkey data on IT cell responses to targets and

distractors (Chelazzi, Miller, Duncan, & Desimone, 1993; Motter, 1994). These monkey data indicated that attentional search is parallel rather than serial, with cells responding to both targets and distractors at first and then responding only to targets at the end of trials.

9.2.2. Bayesian Approaches to Visual Perception

In this century there has been considerable interest in applications of *Bayesian modeling* to neural and cognitive systems. That type of modeling lacks a precise definition, but it is based on the claim that the brain is nearly optimal in performing specific tasks. Hence, such models emphasize the statistical properties of the task environment more than the internal constraints of the organism performing the task. The models are called Bayesian because they update probabilities based on feedback using Bayes's rule for calculating posterior (after-the-fact) probabilities from prior (before-the-fact) probabilities.

Bayesian models have been applied to a wide range of cognitive processes including reasoning (Koechlin, 2014; see Section 9.6), reinforcement learning (Dayan & Daw, 2008), and many others. Yet, perhaps their commonest use has been in modeling visual perception. The work of Geisler and his colleagues, based on the notion of an *ideal observer*, is particularly noteworthy (e.g., Geisler, 2011; Geisler & Diehl, 2003; Geisler, Perry, Super, & Gallogly, 2001).

Geisler is what Bowers and Davis (2012) call a *methodological Bayesian* rather than a *theoretical Bayesian*. A theoretical Bayesian is one for whom optimality is central to his or her theory of the mind (e.g., Oaksford & Chater, 2007, 2009). A methodological Bayesian is not committed to such a theory but finds optimality a useful tool for constructing models that can account for behavior that is not quite optimal but close to it. Indeed, Geisler and Diehl (2003) state that evolutionary constraints limit optimality in visual processes:

> While ideal observer theory provides an appropriate benchmark for evaluating perceptual and cognitive systems, it will not, in general, accurately predict the design and performance of real systems, which are limited by a number of factors . . . Here, we mention only one of these factors: Evolution through natural selection is an incremental process where each change must produce an increase in fitness; thus the real observer may correspond to a local maximum in the space of possible solutions, whereas the ideal observer corresponds to the global maximum in the space of possible solutions.
>
> (p. 381)

Geisler (2011) reviews models inspired by the ideal observer in several visual domains: pattern detection and discrimination; perceptual grouping;

shape, depth, and motion perception; and visual attention. Unlike the models described in the last section, these models tend to start with an entire visual scene, then extract and draw inferences from the scene's statistical properties. An example is the model by Geisler et al. (2001) of perceptual grouping. The grouping model starts with extracting edges in a scene by means of detecting local contrasts, then classifying pairs of edge elements using three parameters: distance between element centers, orientation difference between elements, and angle between the first element and the line between their centers. Based on correspondence in these parameters, the model then generated the likelihood of two given edge elements belonging to the same contour.

The Bayesian models of Geisler and his colleagues rely on formal probability calculations and are not embedded in a connectionist network. These models do not describe interactions between retina, thalamus, and cortex; in fact, there is some suggestion that the ideal observer is an "ideal retina" and the cortex is a source of extra noise for retinal perception.

Bowers and Davis (2012) critique the whole Bayesian approach to modeling cognitive processes. They cite examples showing that prior probabilities, likelihoods, and utility functions can be chosen to explain just about any behavior as optimal. Bowers and Davis also argue that the evidence for rational probability calculation is weak in both experimental psychology and neuroscience. They also cite the very evolutionary constraints that Geisler and Diehl (2003) discussed to illustrate that the influences on behavior are not limited to the current task, in contrast to the Bayesian method of focusing on the current environment. In the case of perception, similarly, Ramachandran (1990) has characterized the process as a "bag of tricks."

Yet, Bowers and Davis acknowledge that Bayesian theorists have contributed to understanding of how the brain updates perceptions, concepts, and behaviors based on feedback from the environment. These types of updating can be accommodated by non-Bayesian theories that do not assume optimality, or assume it only on small constrained tasks.

9.3. Models of Sequence Learning and Performance

A variety of models have been propounded for learning and performance of sequences. Sequence performance models have been applied to movement sequences such as typing, speech, and music performance. Sequence learning models have been applied to short-term recall in the correct order of serially presented items such as words or numbers; some of those models have also been extended to encompass data on free recall of items. Many of the same principles have been utilized in both sequence learning and sequence performing models.

Some of the models are descendants of earlier models of *spatiotemporal pattern learning*, that is, learning of a time sequence of spatial patterns. Many

of the earlier models build on architectures such as back propagation and adaptive resonance that were previously used, among other things, to categorize spatial patterns that do not include a time element (see Chapters 7 and 8).

9.3.1. Early Models of Spatiotemporal Pattern Learning

Several investigators added recurrent interactions to the basic form of the supervised back propagation network in order to train a network to produce a specified time sequence of outputs. The first of these was Jordan (1986b), who developed what he called the *sequential network*. The sequential network is a standard back propagation network with the addition of some feedback and some *plan units* activated by external stimuli. The net effect is to have a decaying memory of past events blended with current plans. More specific sequence learning problems along these lines were simulated by many investigators, most notably Elman (1990). Elman's network, which was largely based on learning associations between successive items, can learn several linguistic tasks, among them discovering the notion of word and discovering lexical classes from word order.

Nigrin (1993) pioneered the modeling of spatiotemporal pattern processing using adaptive resonance theory (ART). In networks of this type, if a sequence consists of several items in order, they are converted into a spatial pattern by converting order information into relative strength of activation. In order to achieve presentation of the sequence in the correct order, it was shown that first items need to be activated to a greater degree, and later items to a lesser degree. Also, these networks incorporate an *invariance principle of memory*, formulated by Grossberg (1978b): as new items are presented, the activation of old items may change but the relative proportion of their activations should not. These insights incorporate the argument of Lashley (1951) that the characteristic errors in sequence learning (reproducing the correct elements but in the wrong order) argue against such learning being based on chains of associations between successive elements of the sequence. Conversion of order into relative activation is also the basis of several models discussed in the next subsection that incorporate specific neurophysiological data (Grossberg & Pearson, 2008; Rhodes, 2000; Rhodes & Bullock, 2002; Rhodes, Bullock, Verwey, Averbeck, & Page, 2004).

Nigrin introduced into his *SONNET* (self-organizing neural net) network some mechanisms for learning the asymmetric inhibitory connections among lists of varying lengths, a spin-off of the *masking field* introduced by Cohen and Grossberg (1986, 1987). By such a mechanism, his network can independently learn both a list and various sublists, and control contextually which chunk of the list is activated. Also, SONNET can learn lists which include repetitions.

A few of the early models included analogs of brain regions involved in sequence learning, notably prefrontal cortex and basal ganglia. Another early ART-based sequence model was developed by Bapi and Levine (1994, 1997) to simulate data on frontal lobe involvement in sequence learning and classification. Bapi and Levine's networks include the ability to learn many sequences composed of rearrangements of the same elements by encoding them at sequence detector nodes. These networks include analogs of some regions of prefrontal cortex and basal ganglia, yet are based in part on the associative chaining between successive elements that Lashley's (1951) arguments refuted. Such sequence detection nodes and prefrontal and basal ganglia representations also appear in the model of Dominey, Arbib, and Joseph (1995) for learning a sequence of eye movements based on associations between visual cues and target positions. The network of Dominey et al. also shares with that of Jordan (1986b) a use of reward and punishment signals to change the weights between context elements and generators of sequence elements. Beiser and Houk (1998) developed a network whereby prefrontal and basal ganglia cells learn sequence representations, but their network does not model the actual reproduction of the sequences. Brown, Bullock, and Grossberg (2004) developed an elaborate biologically realistic network involving both prefrontal and striatal areas in control of saccadic eye movements. The Brown et al. network, called TELOS, is notable for its ability to simulate the difference between reactive and planned movements.

9.3.2. Models of Learning and Performance of Sequences of Movements

One of the more ambitious biologically based models of sequence learning and performance is found in Rhodes and Bullock (2002) and Rhodes (2000).[2] Rhodes and Bullock, like Nigrin (1993), based their neural architecture on an understanding of sequence learning as a parallel more than a serial process. As discussed in Section 5.5 of this book, the parallel representation of sequence elements also received support from studies of neurons in the monkey prefrontal cortex by Averbeck, Chafee et al. (2002, 2003) and Averbeck, Crowe et al. (2003), showing that neural correlates of all the sequence elements were active to different degrees during performance of an entire motor sequence.

Rhodes and Bullock (2002) and Rhodes (2000) were further motivated by behavioral data about learning to produce sequences of length 6 or less, particularly sequences of key presses or sounds (Klapp, 1995; Sternberg et al., 1978/1980; Verwey, 1996). These data include in particular: (1) *Sequence length effect on latency*. If the GO stimulus is defined as the stimulus that triggers the need to start sequence production, latency is defined as the time interval between the GO stimulus and the onset of the first response. That

latency increases with the number of items in the sequence. (2) *Sequence length effect on production rate*. Mean inter-response interval (IRI) also increases with sequence length. (3) *Serial position effect on IRIs*. In the early stages of learning, mean IRI varies nonmonotonically with position in the sequence. Even after practice the IRI between the responses to the last two items remains shorter than the preceding ones. (4) *Ratio effect*. The latency described in (1), otherwise known as the sequence start time, is longer than the mean response time between performance of successive sequence items.

These combined neural and behavioral data motivated a model called N-STREAMS based on the idea of *competitive queuing*. The term competitive queuing was introduced by Houghton (1990) in a model of learning monosyllabic words, and its foundational ideas were outlined in Grossberg (1978b). A definition of competitive queuing (CQ) models is as follows (Bohland, Bullock, & Guenther, 2010):

> . . . items and their serial order are stored via a primacy gradient utilizing the simultaneous parallel activation of a set of nodes, where relative

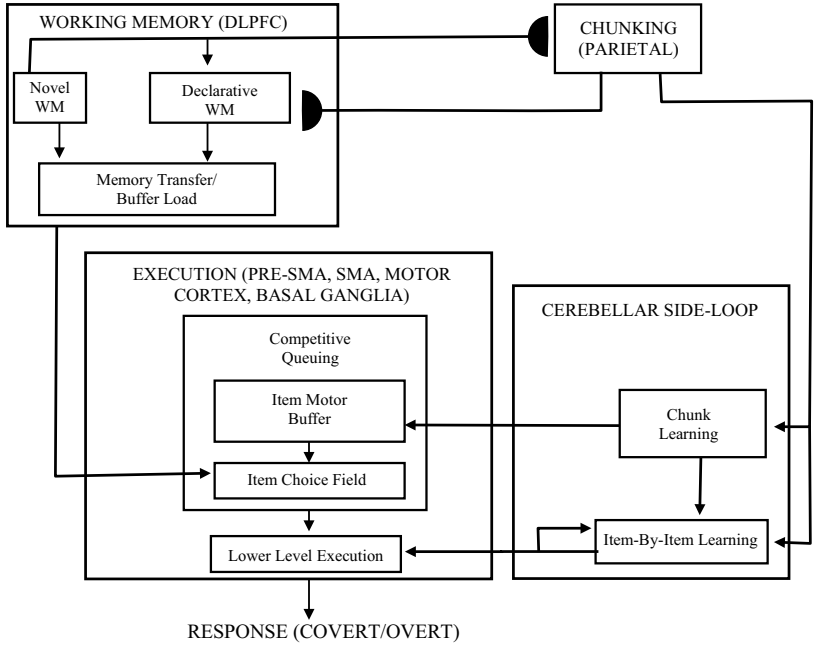


FIGURE 9.2 Global architecture of the N-STREAMS model, with tentative brain region assignments. Chunking refers to the grouping together of subsequences that are presented repeatedly.

Source: Adapted from Rhodes & Bullock, 2002, with the permission of Daniel Bullock.

> activation levels of the content-addressable nodes code their relative order in the sequence. This parallel working memory plan, which can be characterized as a *spatial pattern* in a neuronal map, can be converted to serial performance through an iterative competitive choice process in which i) the item with the highest activation is chosen for performance, ii) the chosen item's activation is then suppressed, and iii) the process is repeated until the sequence reaches completion.
>
> (p. 1505)

Figure 9.2 shows the large-scale architecture of the N-STREAMS model, with tentative assignments of modules to brain areas. Rhodes (2000) and Rhodes and Bullock (2002) get into considerable detail about excitatory and inhibitory connections within the execution module. The lower level execution part of the execution module is borrowed from a model of voluntary movement generation and control called *VITE* (an acronym for *Vector Integration to Endpoint*) developed in a series of articles starting with Bullock and Grossberg (1988).

9.3.3. Models of Serial Recall

The same competitive queuing principles have informed many models of immediate recall of item sequences presented in working memory. Among these are Page and Norris (1998), Botvinick and Plaut (2006), and Grossberg and Pearson (2008).

The model of Page and Norris (1998) was a cognitive model without contact with neuroscience, but set the tone for later models by positing a coding of presented items in a list by relative activation; thus the authors called it a "primacy model." Their primacy model excluded not only chained associations between successively presented items but explicit representations of the order of presentation of items. Order representations do appear, though, in the later models of Botvinick and Plaut (2006) and Grossberg and Pearson (2008). The Botvinick–Plaut model is a three-layer PDP-type model with internal representations at recurrently connected hidden units. The Grossberg–Pearson model integrates working memory with the laminar structure of the cerebral cortex.

Botvinick and Plaut (2006) noted a paradox about interitem associations. The paradox is that while the data previously described refute chaining models, other data involving letter pairs (*bigrams*) points to the importance of learned sequential associations. Specifically, words containing bigrams commonly presented in sequence (such as CK) are easier to remember than words containing less common bigrams (such as KC). Botvinick and Plaut reproduced both datasets using the PDP model of Figure 9.3. The emergent hidden unit

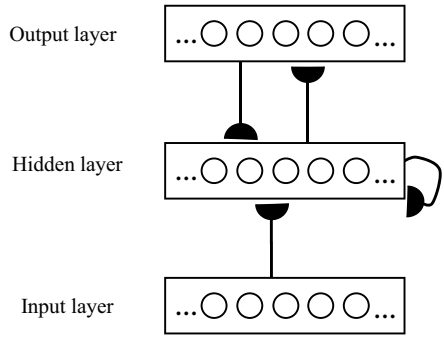


FIGURE 9.3 Architecture of the network used by Botvinick and Plaut (2006) to simulate immediate serial recall data.

Source: Reproduced from Bowers et al., 2009, with the permission of the American Psychological Association.

internal representations code both list items and orders within lists. Notably, conjunctions of item and order (such as "A in list position 2") are coded not by individual hidden units but by vectors of hidden unit activations.

Botvinick and Plaut's model, like many PDP models, relies on an extensive (up to a million cycles) training episode. It is trained on some sequences and then tested on other sequences which may or may not be the same, but usually have some of the same elements, as the training sequences. Thus, the network has considerable capability to generalize from what it has been taught to similar but different sequences. Yet, Bowers, Damian, and Davis (2009) argued that the conjunctive coding (simultaneous coding for item and order) aspect of the Botvinick–Plaut model limits its generalization capability. Bowers et al. presented simulations suggesting that if the network sees a particular letter in different positions but never sees that letter in one specific position, it will have trouble learning future sequences with the letter in the untaught position.

Bowers et al. (2009) segued from their criticism of the serial recall model to a criticism of the context-dependent aspect of PDP models in general. Botvinick and Plaut (2009) answered this criticism by saying that the context dependence is not intrinsic to PDP models but emerges from the internal representations, which in turn emerge from training. The difference in viewpoints arises in part from earlier debates regarding the importance of symbols in cognition. Several authors with a cognitive science background, starting with Fodor and Pylyshyn (1988), criticized PDP models on the grounds they do not include the type of context-independent representations that constitute the elements of human thought. Yet, as Bowers (2002) noted, there is already a rich history of neural network models that include context-independent representations. This includes the serial recall models as discussed above whereby

order information is not coded separately from item information but simply arises from relative activation levels.

Botvinick and Plaut (2006) noted that their model is compatible with a variety of neurophysiological data whereby responses of prefrontal neurons change over several steps of sequence encoding. Yet, the back propagation learning and the requirement of massive amounts of training argue against their model's plausibility as a neural representation of the immediacy of the process of serial recall.

9.3.4. Model of Both Serial and Free Recall

The LIST PARSE network of Grossberg and Pearson (2008) models immediate serial recall by extending the laminar cortical architecture discussed in Section 9.2 to the prefrontal cortex. In so doing, LIST PARSE also can simulate human data on *free recall*, that is, recalling as many items as possible from a list without regard to the order of their presentation. The same network can simulate the aforementioned monkey data on prefrontal neuron responses during learning of a motor sequence (Averbeck et al., 2002; Averbeck, Chafee et al., 2003; Averbeck, Crowe et al., 2003).

The LIST PARSE network consists of a cognitive working memory, a motor working memory, and a trajectory generator. The cognitive working memory is assumed to be located in the ventrolateral prefrontal cortex and the principal sulcus, and the motor working memory in the dorsolateral prefrontal cortex. The laminar structure of the cognitive working memory plays a major role in the model. Layers 4 through 6 of that part of the network are assumed to be involved in filtering and temporary storage of incoming items on which they perform normalization and contrast enhancement (see Chapter 4). The more superficial layers, 2 and 3, are assumed to group these items, as they do in the laminar visual models discussed in Section 9.2. In this case the groupings are based on sequential order, and the groups form chunks of items that are remembered as units.

Grossberg and Pearson (2008) noted that their model includes a computational implementation of the influential psychological model of working memory by Baddeley (1986). Baddeley described working memory as a set of interactions between a central executive controller and two subsidiary systems called the *phonological loop* and *visuospatial sketchpad*. The LIST PARSE model contains a computational analog of the phonological loop that includes pathways involving the inferior parietal lobule, the ventrolateral PFC (cognitive working memory), dorsolateral PFC (motor working memory), and covert rehearsal (inferior parietal and anterior insula). The pathways involving some of those same areas and the dorsal and ventral visual streams (see Sections 5.2 and 9.2) are a computational analog of the visuospatial sketchpad. The control

of these subsystems by volitional functions such as gating and gain control are analogous to Baddeley's executive controller.

Like the competitive queuing models of Rhodes and Bullock, LIST PARSE does not rely either on associative chaining or explicit representations of numerical order. Sequential order of items is represented by relative activation of those items within the deeper layers of the cognitive working memory. Yet, toward the end of their article Grossberg and Pearson (2008) note some behavioral data indicating that we do develop explicit representations of positions of items within sequences. For example, when participants are recalling one list, intrusions from other recently presented lists tend to go into the same or a close serial position as they occupied on their correct lists. Also, as Botvinick and Plaut (2006) noted as well, repeated presentation of an item in the same serial position in lists across many trials tends to make that item easier to learn. Based on these data, Grossberg and Pearson suggested future extensions of their network to include conjunctive coding of item, order, and position.

Prefrontal cortex layering also plays a role in the models of prefrontal–basal ganglia interactions in working memory by Frank, Loughry, and O'Reilly (2001) and O'Reilly and Frank (2006). The article of Frank et al. (2001) considered two paradoxical requirements of working memory: the need for robust maintenance of items and the need for rapid task-relevant updating of memory representations. Based on the capacity of the prefrontal cortex to sustain representations of sensory events over delays (e.g., Fuster, 1997), these authors assigned the robust maintenance to the frontal lobes. The updating function they assigned to the basal ganglia based on this region's importance for disinhibiting motor movements whose details are organized elsewhere (e.g., Chevalier & Deniau, 1990). Stimuli input to this network are represented twice in prefrontal cortex, once in a "maintenance" layer (assumed to be in cortical layers 2–3 or 5–6) and once in a "gating" layer (assumed to be in cortical layer 4) influenced by striatum through its disinhibition of thalamus. Those inputs recognized as relevant to task performance also have corresponding representations in the striatum. Frank et al. applied this network to working memory tasks involving sequences such as the *1–2–AX task*, whereby the participant is asked to respond to a sequence A–X if they saw a 1 more recently than a 2 and to a different sequence B–Y if they saw a 2 more recently than a 1.

The subsequent article by O'Reilly and Frank (2006) built on Frank et al. (2001) but noted that the earlier article had not answered the question of how the striatum learns which stimuli are relevant. They answered that question in the 2006 article using the reward structure of the Primary Value Learned Value algorithm, with actor–critic structure, discussed in Section 6.4.2. The expanded model combining sequence working memory with PVLV was extended to

simulate two other tasks. One of these tasks is called *store–ignore–recall* (*SIR*). The task is to store a particular stimulus called S, maintain and ignore S over a sequence of other stimuli, and then recall S when another stimulus called R appears. The other is an analog of Baddeley's (1986) phonological loop, a task that requires encoding and reproducing a sequence of phoneme inputs.

Another sequence model, by Taylor and Taylor (2000), includes basal ganglia–thalamocortical loops (Alexander, DeLong, & Strick, 1986) but particularly emphasizes roles for premotor, supplementary motor, and presupplementary motor areas of cortex. In particular Taylor and Taylor reproduced many of the monkey data of Tanji and Shima (1994) discussed in Section 5.5, and related data by Halsband, Matsuzaka, and Tanji (1994), on sequences whose elements were a push, pull, and turn.

The Taylor–Taylor sequence storage and generation model is a variant of a more general model of the functions of basal ganglia–thalamocortical loops known as *ACTION* (e.g., Taylor & Alavi, 1996). The sequence data simulations are in two parts, one dealing with motor sequences guided by visual cues and the other dealing with sequences guided by internal cues, corresponding to the visual guided and memory conditions in the Tanji and Shima (1994) task (see Section 5.5). The activity patterns of various network nodes reproduce some of the experimentally found cell types. These cell types include initiators that respond to the auditory tone used to signal that the movement would be visually guided; memory cells that indicate which movement is to be made in response to an internal GO signal; premovement cells that are active in response to a GO signal if a specific movement is called for; movement cells that are active while the movement is being made; cells responsive to a specific order of movements; and cells active in the delay between two movements.

9.3.5. Models of Word Recognition and Reading

The oldest significant neural network model of letter and word recognition is the *interactive activation* model of McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982). These researchers sought to explain the fact that recognition of letters is heavily dependent on context. For example, letters can be identified more accurately when they appear in words than when they appear in random sequences. This advantage of words partially extends to nonwords that are pronounceable by an English speaker (such as MAVE or REET). McClelland and Rumelhart concluded that modeling of these context effects has to involve both top-down and bottom-up processes, that is, feedback between "letter level" and "word level" nodes.

Grossberg and Stone (1986) also included top-down and bottom-up processes in their model of word recognition. But they stated that McClelland

and Rumelhart's letter–word distinction poses a problem in the case of words of one letter: It has been shown (Wheeler, 1970) that those letters that are also words in English (A and I) are no easier to recall than other letters. If A and I were represented on both the letter and word levels, they would have a selective advantage over other letters, which is false. Hence, the letter A and the word A need to have separate representations.

These considerations led Grossberg and Stone to replace the concepts of letter and word nodes with the more abstract concepts of *item* and *list* nodes. Their architecture is also used in speech recognition (Cohen & Grossberg, 1987; Cohen et al., 1987). There the lists or "chunks" (auditory, in this case) need not be words, and there can also be competition between words when one is a "sublist" of the other (e.g., "SELF" and "MYSELF").

The Grossberg–Stone theory of word recognition also relies on the distinction between *pattern* (what is processed) and *energy* (how strongly is it processed). In the word recognition domain, the pattern–energy distinction takes the form of a distinction between *attentional priming* and *attentional gain control*. The priming stimulus ("pattern") is encoded at the F_2-to-F_1 synaptic weights, and gain control ("energy") from other nodes determines the relative amount of attention paid to that prime. In the network equations, these two factors are multiplied, an example of the factorization of pattern and energy discussed in Chapter 7.

Grossberg and Stone explained a number of experimental results on word recognition. These results include reaction times to words versus nonwords after different primes (related words, unrelated words, or neutral stimuli like strings of Xs); "word superiority" effects that can cause a tendency to misclassify nonwords that differ from words only by one letter; and word frequency effects.

Seidenberg and McClelland (1989) developed a model of word recognition and naming based on back propagation. This model is pointedly different from McClelland's own interactive activation model, because it assumes distributed representations of words whereas interactive activation assumes localist representations (explicit word nodes). In particular, Seidenberg and McClelland's network makes *lexical decisions* (decisions about whether or not a presented string of letters is an English word) without explicitly including representations of words. Their network consists of 400 orthographic (spelling) units, 450 phonological (sound) units, and 100 to 200 hidden units in between. After learning spelling–sound associations, the network produces a sound in response to a string of letters that may or may not be a word. There is feedback between orthographic and hidden units but the connections from hidden to phonological units are unidirectional.

On the database that Seidenberg and McClelland (1989) used, the network mispronounced only 77 out of 2897 words. The learning is faster for words

that are more frequent, and when words are of lower frequency it is also faster for words whose pronunciation is more regular based on the spelling (e.g., "mint" versus "pint"). High-frequency words, such as "have," do not show such a regularity effect. The learning of pronouncing a particular letter string is influenced by the pronunciation of strings with which it shares some of the letter sequence, so there is also some slowing down in the learning of what the authors called *regular inconsistent* words. These are words whose own pronunciation is regular but have irregular "neighbors": for example, "gave" is close to "have." The slowest learning is in what the authors called *strange* words that have irregular pronunciations combined with spelling parts not shared with any other words, such as "fugue" and "yacht."

Lexical decisions are modeled in the Seidenberg–McClelland network by means of similarity of the incoming string with stored patterns of orthographic activation. Their one simulation that explicitly considers latency of lexical decision involves *pseudohomophones*, that is, nonwords that if pronounced in a regular manner would sound like words (e.g., "brane" which sounds like "brain"). The model simulates data showing that when pseudohomophones are compared with control nonwords that do not have the same sound as words (e.g., "brone"), the latency of reading aloud is shorter but the latency of lexical decision is longer, because of confusion with the actual word.

Since the Grossberg–Stone and Seidenberg–McClelland models there have been several other models of reading and/or lexical decision, all involving orthographic, phonologic, and in some cases semantic nodes. These models have been both localist (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Davis, 2010; Perry, Ziegler, & Zorzi, 2007) and distributed (e.g., Harm & Seidenberg, 1999, 2004). Some of the localist models are similar to or inspired by the sequence models of Page, Grossberg, and others described earlier in Section 9.3. While there are substantial variations within both model types, in general the localist models posit that reading words engages a "lexical" pathway and reading pseudowords engages a "nonlexical" pathway. By contrast, the distributed models lack word representations and so rely on a "division of labor" between orthographic and phonologic processing for reading.

Taylor, Rastle, and Davis (2013) performed a meta-analysis of 36 brain imaging studies of word and pseudoword reading. They found some areas of cortex that responded more to words than to pseudowords and some did the reverse. Also, some areas responded more to regular words (in terms of correspondence between spelling and pronunciation) and some the reverse. As with other imaging studies, there was a kind of inverted-U relationship between activation of a brain region on a task and involvement of that region in the processes in the task. That is, a region does not become active in tasks that do not engage the function of the region, but also becomes less active when the processing involved takes little effort. Yet the results of their meta-analysis did

not succeed in helping to decide between the different types of cognitive models. So far there has not to my knowledge been a model of the processes of reading and lexical decision that explicitly includes brain regions.

The meta-analysis of Taylor et al. (2013) also illuminated likely effects of two kinds of dyslexia. There is *phonological dyslexia*, whereby pseudoword reading is impaired but real and irregular words can be read normally. This would be expected to disrupt the normal functioning of brain regions involved in the nonlexical part of reading. There is also *surface dyslexia*, whereby pseudoword reading is intact but there are difficulties in reading irregular words. This would be expected to disrupt brain regions involved in retrieving lexical representations. The most severe form of dyslexia is *deep dyslexia*, whereby both irregular word and pseudoword reading are impaired. Deep dyslexia has been modeled by the back propagation network of Plaut and Shallice (1993).

9.4. Models of Executive Function and Cognitive Control

Many aspects of executive function are captured by the working memory models discussed in the last section. Other models capture aspects of cognitive control. Specifically, there have been many models of cognitive control tasks affected by prefrontal lesions, such as the Wisconsin card sorting task, delayed response, and Stroop task.

9.4.1. Models of the Wisconsin Card Sorting Test

Recall from Section 5.6 that the Wisconsin card sorting test (WCST) is a standard test used by clinical neuropsychologists to test the executive capacity of cognitive flexibility. The WCST task is to classify cards which differ by three criteria: color, shape, or number of the designs on the face of the cards (Figure 5.6), with the experimenter changing the criterion used after the participant makes ten consecutive correct classifications. Damage to the DLPFC particularly impairs WCST performance, leading typically to perseveration on the first classification (e.g., Milner, 1964).

Since the late 1980s there have been several computational models that have simulated typical performance of DLPFC-damaged and undamaged participants on the WCST. In all of these models there is a node or set of nodes such that "lesions" of those nodes lead to perseverative responses. The earliest WCST models (Dehaene & Changeux, 1991; Kimberg & Farah, 1993; Leven & Levine, 1987; Levine & Prueitt, 1989) were built mainly on abstract neurocognitive capacities with loose assignment of modules to brain regions. More recent WCST models (e.g., Monchi & Taylor, 1999; Amos, 2000) include explicit analogs of the loops connecting prefrontal cortex, basal ganglia, and thalamus.

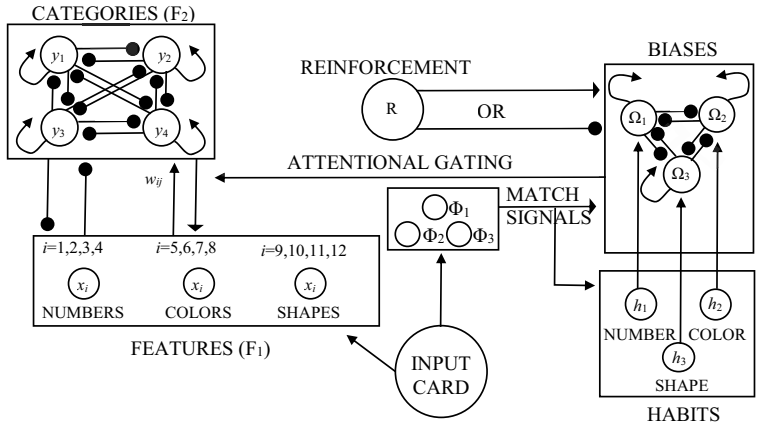


FIGURE 9.4 Network used to simulate Milner's Wisconsin card sorting data. Habits and biases are explained in the text.

Source: Modified from *Proceedings of the First International Conference on Neural Networks* (p. II-270) by S. J. Leven and D. S. Levine, 1987, San Diego: IEEE/ICNN. Copyright 1987 IEEE. Adapted by permission.

Leven and Levine (1987) simulated the card sorting data using a network (also presented in Levine & Prueitt, 1989) based on adaptive resonance theory (see Figure 7.8). In their WCST network (Figure 9.4), the nodes in the feature field F_1 code numbers, colors, and shapes, whereas the nodes in the category field F_2 code template cards. F_1 divides naturally into three subfields (number, color, and shape); corresponding to each subfield is a "habit node" and a " bias node." The habit nodes code how often classifications have been made, rightly or wrongly, on the basis of each feature. The bias nodes additively combine habit node activities with reinforcement signals (the experimenter's "Right" or "Wrong"), then gate the excitatory signals from F_1 to F_2. A network parameter measuring the gain of reinforcement signals to bias nodes was varied. The network with high reinforcement gain acted like Milner's normal subjects. The network with low reinforcement gain acts like Milner`s frontal patients, learning the first classification as quickly as the normal network but remaining stuck in that classification for the remaining trials. Hence the network treats habit and reinforcement as separate, and sometimes competing, influences on response selection.

Leven and Levine noted that perseveration due to frontal damage can be overridden by attraction to novelty, as in the monkey data of Pribram (1961). Pribram placed a peanut under a junk object several times, unobserved by a monkey. Each time this was done, he added a new object to the scene and waited for the monkey to choose which object to lift for food. On the first trial with a novel object present, normal monkeys tended to choose another object

that had previously been rewarded, whereas monkeys with lesions of the ventral frontal cortex chose the novel object immediately. Levine and Prueitt (1989) simulated the novelty data using an example of a *dipole field* (Grossberg, 1980). In a dipole field, each of several sensory stimuli has an "on" and an "off" channel structured like the two competing channels of a gated dipole (see Figure 3.7), and each channel becomes transiently active when an input to the opposing channel is turned off. Hence, the grated dipole provides a network instantiation of opponent processing which can be used to model counterfactual comparisons. Recall from Section 3.3.4 that the gated dipole operates by means of transmitter habituation; hence, if the outputs of dipoles corresponding to different stimuli compete, novel stimuli that have not been habituated have an advantage over familiar stimuli.

Another WCST simulation was developed by Dehaene and Changeux (1991). Dehaene and Changeux's model was intended to represent more general cognitive and inferential capabilities than are manifested by the WCST, which is a simple test for mental flexibility in the face of changing context. Their model includes representations of the input features, color, shape, and number; rule-coding clusters that represent the three different possible card classification rules; nodes entitled "current intention" that represent dynamic tendencies to follow each of those rules; a reward node; and an error cluster that became active when the network receives feedback that its classification is incorrect.

Dehaene and Changeux's model proceeds from a different fundamental mode of organization than does Levine and Prueitt's, being based on primary neurophysiological considerations and not on previously established neural network structures such as ART. Their memory nodes in particular were found to have activity patterns somewhat like those of DLPFC neurons that remain active during delay tasks (Fuster, 1973), and their model had previously been used to model delayed response deficits with prefrontal lesions (Dehaene & Changeux, 1989). For explaining responses to unexpected lack of a reward they include a mechanism analogous to the transmitter depletion in a gated dipole but possibly more biologically realistic, stating that "fast synaptic depression may result from the desensitization of receptor molecules, mediated by their allosteric[3] transitions in the postsynaptic membrane" (p. 75).

Yet most parts of Dehaene and Changeux's network can be mapped fairly closely either into Levine and Prueitt's WCST model or their novelty preference model (see Levine, Parks, & Prueitt, 1993, for discussion). For example, Dehaene and Changeux's memory and intention nodes are closely analogous to Levine and Prueitt's feature and category fields. Also, their rule-coding clusters are similar structurally and functionally to the bias nodes of Levine and Prueitt (1989). Dehaene and Changeux added to their model a feature they called "episodic memory," though it differs somewhat from the

common usage of that term by psychologists (see Tulving, 1972). Their version of episodic memory keeps track of rules that had been previously tried and not led to reinforcement, and selectively reduced the activation of nodes representing such rules. This is analogous to the opponent processing mechanism (via the gated dipole network) used by Levine and Prueitt (1989) to selectively enhance representations of novel inputs.

Kimberg and Farah (1993) simulated the WCST using a computational model that is based not on a neural network but on the heuristic programming concept of *production system*. A production system combines procedural knowledge in the form of productions, which are instructions of the form "If condition X holds, THEN perform action Y" (Kimberg & Farah, 1993, p. 415) with declarative knowledge in the form of working memory representations. These researchers showed that, if prefrontal damage is interpreted as weakening of working memory associations, their production system model can account for results on the WCST, Stroop Test, a motor sequencing task, and a context memory task.

Monchi and Taylor (1999) and Monchi, Taylor, and Dagher (2000) also emphasized the working memory aspects of prefrontal function on tasks like the WCST. These researchers also emphasized that tasks activate the basal ganglia and thalamus in addition to prefrontal cortex. They found further evidence of basal ganglia involvement in the deficits of Parkinson's disease patients on the WCST and other working memory tasks (e.g., Owen et al., 1992). They modeled these tasks using analogs of basal ganglia–thalamocortical loops (Alexander, DeLong, & Strick, 1986), in a variant of the ACTION network (Taylor & Alavi, 1996) also used in models of sequence learning (e.g., Taylor & Taylor, 2000; see Section 9.3.3).

The Monchi–Taylor models rely extensively on inhibition and task-selective disinhibition of sensory and attribute representations. This is based on the division of the basal ganglia into two pathways, called the *direct* and *indirect pathways* (Alexander & Crutcher, 1990; Alexander, DeLong, & Strick, 1986). The internal segment of the globus pallidus (GPi) sends inhibitory projections to the thalamus (mediodorsal or ventrolateral), which in turn projects to frontal (including both prefrontal and motor) cortex. The direct pathway involves inhibitory projections from the striatum (i.e., caudate and putamen) to the inhibitory GPi, so its net effect on cortical processing is excitatory. The indirect pathway, by contrast, involves inhibition from striatum to external globus pallidus, which then inhibits the subthalamic nucleus, which then excites GPi, which in turn inhibits the thalamus. So the net effect of the indirect pathway on cortical processing is inhibitory. Variations on the model were also used to simulate the classical delayed response task and a delayed visual matching task.

Hence, in performing tasks such as the WCST and delayed response, the direct and indirect pathways allow for selective disinhibition of those

representations that are relevant for the current task. Monchi and his colleagues also modeled how such selection is disrupted in conditions such as Parkinson's disease and schizophrenia. Both of those conditions tend to involve poor performance on these working memory tasks: Parkinson's disease involves basal ganglia abnormalities (in particular, weakening of the direct pathway) and schizophrenia involves reduced prefrontal activity.

Another WCST model that has been applied to Parkinson's disease and schizophrenia (and also to Huntington's disease, which is somewhat opposite to Parkinson's in its effects on the basal ganglia) is that of Amos (2000). Amos's model is clinically focused and anatomically simplified in comparison with Monchi et al.'s: it includes prefrontal cortex, basal ganglia, and thalamus but not the connections from thalamus back to prefrontal cortex. However, Amos's model makes more accurate predictions about types of error that different patient groups will show on the WCST. Specifically, it predicts that schizophrenics, like those who are damaged in the DLPFC, will show perseverative errors, that is, tend to classify cards on the basis of rules that are previously correct. Parkinson's and Huntington's patients, on the other hand, will show more random errors because the basal ganglia damage will prevent them from selecting responses on the basis of reward.

The clinical concerns were also paramount in the WCST model of Bishara et al. (2010), which is a cognitive process model whose nodes do not correspond to brain regions. Free parameters are kept to a minimum because the goal is not a detailed neural explanation but predictability at the level of individuals as well as groups; the four parameters used in the model represent decision consistency, attentional focus, sensitivity to reward, and sensitivity to punishment. These researchers also looked at the differences on the WCST between participants who are and are not substance dependent, most often on either alcohol or stimulants. They found that the substance dependent individuals could be modeled either by lowered decision consistency or lowered sensitivity to punishment.

Kaplan, Şengör, Gürvit, Genç, and Güzeliş (2006) developed another model of the WCST that was designed to capture the two possible effects of prefrontal damage that could impede task performance. One effect was perseveration and mainly associated with loss of executive function due to DLPFC damage. The other effect was distractibility and mainly associated with loss of response inhibition due to OFC damage. These prefrontal regions are not explicitly included in the model network but partially represented by a "Hopfield network" (e.g., Hopfield, 1982; see Section 4.2.4) and a "Hamming network."

All of these models of the WCST have different emphases: some are designed mainly to capture different functional properties, others focus on realistic anatomy, and still others are primarily intended to reproduce clinical lesion data. Such diversity indicates that none of these models is yet definitive

or generally accepted. While some of the models also, with variations, reproduce data other than the WCST on cognitive effects of prefrontal damage, none is yet embedded in a comprehensive model of prefrontal executive function.

9.4.2. Models of the Stroop Test

Cohen and Servan-Schreiber (1992) did some network simulations of three cognitive tasks that require attention to the current context. One of these was the Stroop test – previously modeled in Cohen, Dunbar, and McClelland (1990) without reference to frontal involvement – whereby the subject sees the word for a color printed in ink of either the same or a different color and must state the *color of the ink*. The primary response is to actually read the word, so reaction time is slower if the ink color and word do not match; for example, if the word "red" is written in green ink. People with dorsolateral frontal damage, and many schizophrenics, have an even slower reaction time than other people under these incongruent conditions. Cohen and Servan-Schreiber also simulated a *continuous performance* task, whereby subjects were instructed to respond to a target pattern while receiving a steady stream of other stimuli, and a lexical disambiguation task. All these tasks required the subject to perform a nondominant but contextually appropriate response.

Cohen and Servan-Schreiber, using the back propagation network shown in Figure 9.5, simulated the deficits of schizophrenics on all three tasks, which they attributed to a deficit of dopamine inputs to the dorsolateral prefrontal cortex. Their network includes a node that selectively influences gains in two competing neural pathways (e.g., in the Stroop test, pathways coding words and colors), and that is assumed to be decreased in activity in the case of dorsolateral frontal damage or schizophrenia. Although Cohen and Servan-Schreiber used a back propagation learning algorithm, which so far appears to be anatomically unrealistic, they captured some qualitative functional relationships that are important for a wide class of tasks that involve prefrontal executive function.

The model of Kaplan, Şengör, Gürvit, and Güzeliş (2007) builds on the models of Cohen and his colleagues but integrates further knowledge about functions of specific brain regions, such as the roles of DLPFC in inhibitory control of attentional selection, OFC in impulse control, ACC in conflict monitoring, and basal ganglia in performing automatic or habitual actions. In a similar manner to the WCST model by the same group, Kaplan et al. did not include realistic anatomy of these brain regions but built their model around functional modules corresponding to these various roles: sensory and motor networks along with modules for word reading; color naming; habitual responses; directing of attention; inhibition; and error detection. The sensory, motor, and inhibitory modules were structured as Hopfield networks.

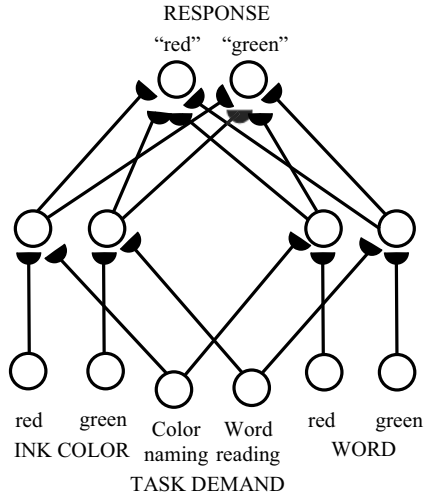


FIGURE 9.5 Network architecture for performing the Stroop task.

Source: Reproduced from Cohen, Dunbar, & McClelland, 1990, by permission of the American Psychological Association.

A more biologically realistic extension of Cohen et al. (1990) is the model of Herd, Banich, and O'Reilly (2006). This model includes abstract representations of colors and words in addition to the task and rule representations of the previous model. The refined model was able to explain some previously puzzling fMRI data, such as increased activity in regions processing information that the task requires the participant to ignore.

9.4.3. Models of General Cognitive Control

A few researchers have started to build models that link working memory, attention, and other cognitive control functions. These models show how brain networks are involved in monitoring of the environment and of task requirements.

Rougier, Noelle, Braver, Cohen, and O'Reilly (2005) noted that some previous models (including those of Cohen et al., 1990, and Dehaene & Changeux, 1991) had posited prefrontal (PFC) rule representations that influence processing in the posterior cortex but had not explained how these rule representations might develop over time. Rougier et al. explained rule formation using a network that includes PFC modulation of posterior cortex. The PFC area incorporates the two complementary functions of maintaining neural activity patterns over time and rapidly updating new representations (see Frank et al., 2001). Updating is implemented by an adaptive gating mechanism

that mimics functions of the basal ganglia and midbrain dopaminergic nuclei involved in reinforcement learning (Montague, Dayan, & Sejnowski, 1996). The network including PFC, but not the posterior part alone, can extract from task-based training any rule based on selection of one of the features of a stimulus as the most important. Examples of the network were applied to both the Stroop task and WCST.

Related models (Brown, Reynolds, & Braver, 2007; Herd et al., 2014) have simulated the process of switching between tasks. The model of Brown et al. focused on cognitive data involving multiple tasks with several dimensions of stimuli. The data dealt with costs in reaction time due to several sources: switching tasks, changing response instructions on the same task, and presenting incongruent stimuli. One example of incongruency occurs in task-switching paradigms when "a feature of the target stimulus is associated with an incompatible response according to the currently irrelevant task" (Brown et al., 2007, p. 40). The network designed to simulate these data includes a task-switching subnetwork and a supervisory control system. The network regions were not explicitly assigned to brain regions but the network's design was inspired by known regional functions such as the role of the ACC in conflict monitoring.

The model of task-switching by Herd et al. (2014) is based on earlier models of PFC–basal ganglia interactions in working memory (Frank et al., 2001; O'Reilly & Frank, 2006) and includes individual differences in task-switching ability. Herd et al. reviewed evidence that task-switching ability is separate from, and sometimes even negatively correlated with, other executive functions such as updating and response inhibition. Their model includes a PFC module influencing parietal cortex, and strength of executive function relates to strengths of both PFC influence on parietal areas and internal recurrent connections in PFC. Yet these same recurrent connections were sometimes found to lead to "stickiness" that interferes with ability to switch tasks. Stickiness can also come from overly long persistence of basal ganglia "go" signals from previous tasks.

More sophisticated extensions of these task control models are found in Collins and Frank (2013) and Ranti, Chatham, and Badre (2015), which extend prefrontal–basal ganglia interactions to multiple hierarchical levels. These authors base their theory on results showing that the PFC is arranged hierarchically, with cognitive processing tending to become more abstract as one moves further forward in the frontal lobe (Badre & D'Esposito, 2007; Christoff & Gabrieli, 2000; Koechlin, Ody, & Kouneiher, 2003; Koechlin & Hyafil, 2007). The PFC region at each level of abstraction is part of a loop that includes a corresponding region of striatum. The Collins–Frank model is designed to select abstract task sets in response to arbitrary cues and then select actions in response to stimuli once the rule is in effect.

In the Collins–Frank model, task sets (abstract states) are encoded in the PFC. Candidate actions are encoded in "stripes" within premotor cortex, which compete via lateral inhibition. Sensory stimuli are encoded in parietal cortex. As in previous working memory models (e.g., O'Reilly & Frank, 2006), the basal ganglia perform the gating function of selectively inhibiting or disinhibiting action stripes and task sets. Sensory projections to basal ganglia are plastic, with dopamine involved in reinforcement learning. Characteristic patterns of errors on complex discrimination tasks are obtained by weakening PFC-to-basal-ganglia connections relative to the strength of parietal-to-basal-ganglia connections.

Ranti et al. (2015) made minor changes in the Collins–Frank model and applied it to learning a hierarchy of rules. They also did experiments involving visual stimuli that varied along seven dimensions at different levels of abstraction. Their data were reproduced by their network, which exerts parallel cognitive control at all levels by means of multiple interconnected prefrontal–basal ganglia loops. In the model, the highest (most abstract) level is more likely than the others to be engaged first, but only slightly more: the parallel control is paramount.

Alexander and Brown (2015) also included levels of abstraction in their model that captured the interplay between anterior cingulate (ACC) and dorsolateral prefrontal cortex (DLPFC) in cognitive control (see Miller & Cohen, 2001). Their model, called HER (hierarchical error representation), is built on a previous model called PRO (predicted response–outcome) of the ACC (Alexander & Brown, 2011). PRO posits that the ACC learns and predicts the likely outcomes of actions (whether good, bad, or indifferent) and compares actual with predicted outcomes, using both TD and back propagation at different loci.

In the HER model of Alexander and Brown (2015), ACC error signals are conveyed to DLPFC, which then calculates the contributions of task-relevant stimuli to the error. The DLPFC error representations in turn update the ACC's predictions of the likely consequences of actions. The addition DLPFC allows the error signal at the ACC to be sensitive to past stimuli: in PRO it was sensitive only to current stimuli. The ACC and DLPFC representations are repeated at different levels of abstraction of the relevant stimuli and error signals at neighboring levels influence one another, which is the "hierarchical" part of the network's name.

The primary simulations in Alexander and Brown (2015) deal with the 1–AX task previously modeled by Frank et al. (2001) and O'Reilly and Frank (2006). Recall from earlier in this chapter that this is a hierarchical continuous performance task that involves one search nested inside another. Specifically, the participant must respond to the sequence AX if s/he has seen a 1 more recently than a 2, but to the sequence BY is s/he has seen a 2 more recently

than a 1. Alexander and Brown note that the nodes in their 1–AX simulations are task-specific and do not generalize readily to other cognitive control tasks. Still, their model is perhaps the first significant plausible computational model of the interaction between two key cognitive control regions (ACC and DLPFC).

9.5. Models of Decision-Making

The dominant paradigms in the study of human preference decisions deal with choices between gambles, that is, between options that involve different probabilities of gains or losses. The gains or losses most often have involved money because that is the easiest thing to quantify, but sometimes instead have involved human or animal lives. Researchers in the 1960s and 1970s largely treated probabilistic choices among different commodities as equivalent, but in this century some points of emotion-related nonequivalence between commodities have been studied and modeled.

For brevity this section does not include models of action selection based on perceptual inputs. There have been several models of that type of decision-making, most of them based on interplay between the cortex and the basal ganglia, with its gating and reinforcement learning functions discussed in Sections 6.3 and 9.3. A few influential perceptual decision models are listed at the end of this chapter.

The types of decisions considered here involve value judgments between options that are usually presented in words (except in the case of animal foraging). These options typically involve quantifiable entities like money or lives and mention abstractions such as probabilities. Before the late 1960s the dominant quantitative theory of preference decisions was based on calculating the "utility" or value of each possible amount of money or lives and multiplying those utilities by the probability of occurrence of a gain or loss of that amount, then summing over all possible outcomes (with losses counted as the negative of equal gains). Then it was assumed that decision-makers are rational and choose the gamble with the highest "expected utility" so obtained.

The results of Tversky and Kahneman (1974, 1981) discussed in Section 5.7 directly contradicted the predictions of that rational theory. So did a theoretical example by Allais (1953), verified by later experimenters, whereby a preference between two probabilistic monetary alternatives changes when an element common to both alternatives is removed. In particular, the effects of how alternatives were framed and the tendency to be risk seeking with losses but risk averse with gains demanded a new quantitative model. The dominant descriptive quantitative model of decision-making became *prospect theory* (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). The main difference between prospect theory and the earlier expected utility theory is that

probabilities in prospect theory are nonlinearly weighted. The nonlinear weighting magnifies psychological differences between low nonzero probabilities and impossibility or between high nonunity probabilities and certainty.

Prospect theory has an impressive record of reproducing data, but does not have a natural mechanistic basis that allows it to be mapped easily into a neural network. Moreover, recent results (e.g., Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004) suggest that the overweighting of low and underweighting of high probabilities applies only to *decisions from description*, that is, quick real-time choices between two gambles whose probabilities are presented explicitly. In fact, the weighting sometimes reverses when the same choices are made between gambles whose probabilities are learned by feedback over repeated trials. Hence, many researchers since the 1980s have developed other cognitive and neural theories that explain the same results as prospect theory using plausible network architectures.

9.5.1. Early Connectionist Models of Decision Processes

Grossberg and Gutowski (1987) set out to model framing effects such as are found in the Asian disease problem (see Section 5.7). They noted that the choices between gambles depend on the reference point with which the gambles are being psychologically compared. For example, in the loss frame of the Asian disease problem the two plans are compared with the reference point of no lives lost, whereas in the gain frame the same two plans are compared with the reference point of no lives saved. Since such counterfactual comparison is a form of opponent processing, Grossberg and Gutowski applied to the decision data a version of the gated dipole (Grossberg, 1972a, 1972b; see Section 3.3.4), which is designed for just such comparisons.

The Grossberg–Gutowski theory, known as *affective balance theory*, does not include analogs of specific brain regions, but incorporates qualitative properties that have shaped more recent brain-based models. The Grossberg–Gutowski version of the gated dipole has also been applied to connectionist modeling of multiattribute consumer preference (Leven & Levine, 1996) and of animal foraging under predation risk (Coleman, Brown, Levine, & Mellgren, 2005). More recent decision models (e.g., Levine, 2012) have used variants of this gated dipole structure that involve explicit analogs of both the direct and indirect striatal pathways and amygdalar positive and negative valence loci.

A different approach to decision modeling that has also shaped later neural models is the *decision field theory* of Busemeyer and Townsend (1993). Those authors used expected utility (EU) theory as their starting point but set out to model key effects that EU cannot explain, notably the variability of individual preferences and the effects of deliberation time on preferences. Busemeyer and

Townsend accounted for these effects by making their model dynamic and stochastic, and also including approach and avoidance gradients previously developed in theories of classical conditioning. They posited that decisions typically occur when an activation of one of the options reaches a particular threshold.

Roe, Busemeyer, and Townsend (2001) mapped decision field theory into a connectionist network. They extended the theory of Busemeyer and Townsend (1993) to include decisions between two or more options with multiple attributes. Their multiattribute model relies on time-varying attribute-selective attention weights, and on lateral inhibition between option representations whose strength depends on the psychological distance between options. This network provides explanations for three effects in multiattribute decision-making that violate EU theory predictions and that had not previously been explained by the same model. All three effects involve starting with two options A and B that are dissimilar (e.g., A is a car that is high in power and low in economy and B is a car that is high in economy and low in power) and adding a third alternative C. The *similarity effect* means that if C is similar in attribute values to A, it competes with A more than with B, thereby increasing the probability of choosing B. The *attraction effect* means that if C is worse than A on all attributes it increases the probability of choosing A. The compromise effect means that if C is intermediate between A and B (e.g., has medium values of both economy and power) it tends to be preferred to either A or B. The decision field theory network of Roe et al. (2001) did not include explicit brain regions, but Busemeyer, Jessup, Johnson, and Townsend (2006) discussed potential neural realizations of that network. Busemeyer et al. drew parallels between the lateral inhibition of Roe et al. (2001) and processes in the basal ganglia–thalamocortical loops. Also, those authors reviewed data on monkey motor decisions that are compatible with the threshold effects in decision field theory.

Usher and McClelland (2004) developed a different multiattribute decision model to account for the compromise, similarity, and attraction effects, one based in part on their previous model of perceptual choice (Usher & Niebur, 1996). Usher and McClelland included in their model a value function that was steeper for losses than for gains. This was included to account for one of the key generic findings of Tversky and Kahneman: *loss aversion*, the tendency for losses of a given amount to have more influence on decisions than gains of the same amount. Usher and McClelland also included distance-independent lateral inhibition and typical nonlinear activation functions. They noted that Roe et al. (2001), because they did not include such an asymmetric value function and nonlinear activation functions, needed to include distance-dependent lateral inhibition and propagation of negative node activations. In particular, neural models by others (including previous models from their own

group and Grossberg's) have tended to avoid propagation of negative activations on grounds of biological realism. They discussed various differences in the two models in terms of what predictions they made with numerical perturbations of the available alternatives.

Starting about 2004, there arose other models of decision data that incorporated knowledge about the roles of brain regions. Some of these models, to be discussed in the next subsection, were specifically designed to simulate the Iowa gambling task, the most popular clinical test of decision-making competence (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; see Section 5.4.2). The Iowa gambling task involves deciding on the good or bad qualities of decks of cards based on feedback from sampling the deck, and researchers starting with Barron and Erev (2003) have made the distinction between decisions from feedback and decisions from description. Other neural network models, to be discussed in the subsection after the next, have been primarily targeted to simulate data on decisions from description, such as the early findings of Tversky, Kahneman, and their colleagues. These models have suggested mechanistic explanations for many of the data explained mathematically by prospect theory.

9.5.2. Models of the Iowa Gambling Task

Recall from Section 5.4.2 that in the Iowa gambling task (IGT) the participant on each trial must draw a card from one of four decks of cards shown on a computer screen, and each deck provides a different probabilistic distribution of gains and losses of play money. Two of these decks have higher short-term payoffs than the other two, but over time the decks with high immediate payoffs are disadvantageous in the long run (i.e., the expected value of earnings from those decks is negative), whereas the decks with low immediate payoffs are advantageous. Hence the IGT is a test of ability to learn from feedback and to inhibit the impulse to pursue short-term gain.

Participants with damage to either OFC or amygdala cannot learn the advantageous strategy effectively. Also, several investigators have studied IGT difficulties in participants who suffer from many conditions including Parkinson's disease, Huntington's disease, and drug abuse. A series of cognitive models based on decision field theory (e.g., Busemeyer & Stout, 2002; Yechiam, Stout, Busemeyer, Rock, and Finn, 2005) have had some success at reproducing these clinical patterns through IGT model parameter variations.

There are at least three IGT models in the literature that include brain regions. These models vary in emphasis but all incorporate a role for OFC or, more broadly, ventromedial prefrontal cortex (VMPFC) in long-term evaluation of the goodness or badness of options.

Wagar and Thagard (2004) were particularly interested in reproducing the physiological effects of prefrontal and amygdalar lesions. They modeled the

influence on IGT choices of covert emotional reactions, as indicated by what Bechara et al. (1994) termed *somatic markers*. Somatic markers are the bodily representations that are gradually built up to stimuli that take on positive or negative emotional significance. Bechara et al. posited that somatic markers precede conscious evaluation of which decks are good or bad. Wagar and Thagard's GAGE model includes amygdalar (bodily state), VMPFC (emotional evaluation), and hippocampal (context) influence on throughput of stimulus representations that is governed by the nucleus accumbens. The nucleus accumbens in turn feeds back to VMPFC (presumably via the thalamus, which the model does not explicitly include). Their model includes spiking and spike timing–dependent plasticity. It also includes training episodes of several thousand time steps before the actual running of the deck choices.

While the original IGT experiment included two good and two bad decks, the Wagar-Thagard model simplifies it to one good and one bad deck. The model of Levine, Mills, and Estrada (2005) includes all four decks, with one of the bad decks leading to infrequent large losses and the other to more frequent small losses. Between the two bad decks, the network of Levine et al. (2005) exhibits more avoidance of the bad deck that is punished more frequently, a result confirmed in human participants by another laboratory (Yechiam & Busemeyer, 2005). This network, like GAGE, includes amygdala, (two layers of) OFC, and (direct and indirect pathways of) striatum, but also includes anterior cingulate for plan generation and conflict resolution. This model updates deck evaluations but unlike GAGE it does not emphasize somatic markers. Somatic markers were left out because results of Maia and McClelland (2004) cast doubt on Bechara et al.'s idea that unconscious emotional reactions precede explicit knowledge of which decks were good and bad. The model of Levine et al. is based on shunting nonlinear equations with plasticity at two loci and does not require training.[4]

The IGT model of Frank and Claus (2006) is part of a general model of connections between OFC and striatum. To a previous model of striatum that was sensitive to frequency but not magnitude of reward and punishment, Frank and Claus added top-down influences from an OFC that was also sensitive to magnitude. Their combined OFC-striatal model also reproduces various conditioning data dealing with reversal and devaluation.

9.5.3. Models of Other Decision Data and Phenomena

Perhaps the most comprehensive brain-based model of decision data to date is the ANDREA (affective neuroscience of decision through reward-based evaluation of alternatives) model of Litt, Eliasmith, and Thagard (2008; see also Litt, Eliasmith, and Thagard, 2006, and Eliasmith, 2005). The model of Litt et al. reproduces general phenomena accounted for by prospect theory,

including loss aversion and the effects of frames and reference points. The network includes spiking neurons in the amygdala and OFC, representing valuation, and also DLPFC, ACC, striatum, and dopamine and serotonin nuclei.

ANDREA was designed more to reproduce general qualitative properties in a neurocomputational setting than it was to simulate specific data. It does, however, simulate data of Mellers, Schwartz, Ho, and Ritov (1997) showing that desirability of a specific dollar outcome is strongly influenced by comparison with what is expected and by how surprising it is (a low-probability gain being more valued that a high-probability one). Mellers et al. (1997) previously explained these data through a computational theory called *decision affect theory*, which is in many ways similar to the affective balance theory of Grossberg and Gutowski (1987). Litt et al. explained the data of Mellers et al. by means of temporal difference (TD) calculations (see Section 6.3.2), with dopamine mediating positive prediction errors and serotonin mediating negative prediction errors (see Daw et al., 2002). The model does not distinguish between decisions from experience and decisions from description.

Another neural-based approach to "prospect theory" phenomena is the DECIDER model of Levine (2012) and AlQaudi, Levine, and Lewis (2015). This model involves shunting nonlinear equations based on a combination of adaptive resonance theory (see Chapter 7), gated dipole theory (see Section 3.3.4), and a psychological account of decision-making and memory known as *fuzzy trace theory* (e.g., Reyna & Brainerd, 2008). Fuzzy trace theory posits that we store events simultaneously in two separate memory traces: *verbatim* and *gist* traces. The verbatim trace stores a stimulus exactly as it is presented, such as numerical values (e.g., of dollars won or lost or lives saved or lost) and probabilities of those values occurring, whereas the gist trace stores what the person regards as the essential meaning of the stimulus. Levine and colleagues treated gist as a category of possible options with some of the attributes selectively enhanced, as in the Wisconsin card sorting model of Levine and Prueitt (1989). This led to a network model based on an adaptive resonance module with attributes interpreted as being in the amygdala or a superficial layer of OFC, categories in another layer of OFC, reset in the anterior cingulate, and behavioral decisions filtered through the striatal direct and indirect pathways.

The networks of Levine and his colleagues incorporate the fuzzy trace theory explanation for framing effects and prospect theory probability weights, which is based on gists that ignore detailed numerical probabilities in favor of the simpler choice of "some versus none" (some chance versus no chance of gaining or losing money, saving or losing lives, etc.). That explanation was supported by the results of Reyna and Brainerd (1991) showing that framing

effects in the Asian disease problem are enhanced by making the possibility of no lives lost or saved more salient and reduced or eliminated by not explicitly presenting that possibility. AlQaudi et al. (2015) simulated that result, and Levine (2012) simulated results of Rottenstreich and Hsee (2001) suggesting that the probability weight distortion was larger for more emotional commodities (a kiss versus a moderate amount of money). The results described above involve decisions from description; the DECIDER network has recently been extended to decisions from experience by addition of a rough analog of dopamine-dependent plasticity in corticostriatal synapses (Levine, Chen, & AlQaudi, 2017).

Fuzzy trace theory (FTT) falls under the rubric of dual-process theories. According to its proponents, there is a gradual shift from predominantly verbatim to predominantly gist processing as we develop from childhood to adolescence to adulthood. Gist processing allows us to ignore irrelevant details and see the commonalities that unite many of our experiences, and so Reyna and Brainerd regard it as a more advanced form of processing than verbatim. Yet, gist processing also makes us prone to some types of errors, including false memories and misleading decision heuristics. A different dual-process theory of decision-making is the *cognitive-experiential theory* (e.g., Epstein, 1994; see Kahneman, 2011, for a recent variation). The two processes are described by Mukherjee (2010) as:

> an associative affect-based mode of decision making (System A) and a deliberative rule-based mode of decision making (System D). Processing in System A is intimately influenced by mood and emotional states of mind and involves how one feels about a particular prospect. On the other hand, processing in System D is analytical in nature and can involve computational operations. Hence, the affective system is driven by pre-conscious, less effortful, experiential considerations, and the deliberative system is driven by conscious, more effortful, numerical and logical considerations.
>
> (p. 243)

Epstein and other proponents of this theory regard System D as more advanced than System A, in contrast to the claims of FTT. Mukherjee (2010) developed a nonneural cognitive model of decisions that integrate the processing done by the two systems into decisions that synthesize the two modes. This led to a theory of decisions that combine the two modes to varying degrees in different individuals and different contexts. Mukherjee's model accounts for the effects of emotion in Rottenstreich and Hsee (2001) and the effects of removing common alternatives in the paradox of Allais (1953).

9.5.4. Models of Neuroeconomics

The knowledge gained from both imaging and behavioral experiments on people's decisions involving money and goods has led to the growth of a new field called *neuroeconomics* (see for example Camerer, Loewenstein, & Prelec, 2005; Glimcher, Dorris, & Bayer, 2005; Glimcher & Rustichini, 2004). For humans, economic entities are learned so well that they become equivalent in many ways to biological reinforcers. For example, Montague and Berns (2002) have found common sites in the basal ganglia and dopamine system for encoding monetary and food rewards. On the negative side, Sanfey, Rilling, Aronson, Nystrom, and Cohen (2003) found that the insular cortex, which encodes pain and disgust, is activated by the receipt of an unfair reward in an ultimatum game.

Thus far, few neural network models have directly addressed the growing findings in neuroeconomics. Yet, a few researchers have applied network theories developed to model other phenomena, notably classical conditioning and reinforcement learning (see Chapter 6), to economic decision-making. Some of the models deal explicitly with economic choices, whereas others deal with more abstract rewards. Yet, results of various fMRI and animal studies reviewed by Montague and Berns (2002) show that monetary rewards activate some of the same brain systems that respond to primary biological rewards; in other words, the brain computes a "common currency" between different kinds of rewards.

Mengov, Egbert, Pulov, and Georgiev (2008) and Mengov (2014) applied the READ model of psychological opponent processing (Grossberg & Schmajuk, 1987) to economic decisions made by their experimental participants. Their work was also influenced by the opponent processing based decision models of Grossberg and Gutowski (1987) and Leven and Levine (1996). The 2008 article modeled a binary choice and the 2014 article generalized the earlier model to a choice among four suppliers of the same product. The model of the later article reproduced data from participants who learn over several trials the likely amount of the product they will receive from each supplier and then react emotionally to actually receiving more or less of it than expected. The READ model only reproduces emotionally based decisions and not decisions based on deliberative strategies. The final decision rule in the model is based on maximizing across options a linear function of three factors. The authors' labels for those factors are "STM and dynamic neural balances in response to economic options"; "emotional long-term memories and economic reputations"; and "remembered particular consumer satisfaction."

Application of the class of TD models (Sutton & Barto, 1998) to economic decision-making can be seen in the three interrelated articles of Egelman, Person, and Montague (1998), Montague and Berns (2002), and Bogacz,

McClure, Li, Cohen, and Montague (2007), all of which combine modeling with economic experiments. Egelman et al. (1998) devised a game in which participants have to decide on each of many trials between two stimuli, A and B, on the basis of the reward they expect to receive from each. The reward on a particular trial depends both on the current choice and on the percentage of choices allocated to each stimulus. There is a medium range of allocations at which the rewards to A and B are equal, and participants tend to be attracted to that range. Attraction to this medium value is an example of *event matching*, or allocating one/s choices in proportion to expected payoffs, a common phenomenon in animal as well as human learning.

In one version of Egelman et al.'s (1998) game, the reward values were set to make the matching strategy optimal. In another version, the matching strategy was suboptimal and the best long-term strategy was to choose stimulus A every time, which was hard to discover because it was not advantageous in the short term. In the second case, 14 out of the 26 participants still chose the matching strategy, but a minority who were more risk seeking chose the optimal strategy instead. Montague and Berns (2002) showed that a TD learning model can account for the matching behavior in either case, but not for the optimal behavior of some participants in the second case. Bogacz et al. (2007) explained the optimal behavior by an extension of the TD model to include eligibility traces. The influence of eligibility makes reinforcement learning depend not solely on the last choice but on several preceding choices (time-weighted) as well.

Rustichini and Padoa-Schioppa (2015) modeled monkey data on OFC cell activities in response to choices between two different juices. Their model is based on several thousand spiking neurons, including shunting and other types of interactions, and is an adaptation of a previous model of perceptual choice (e.g., Wang, 2002). The model approximated the behavior of three types of OFC neurons. One type of neuron encodes the values of individual juices. A second type encodes the outcome of the binary decision between the two juices. A third type encodes the value of the chosen juice.

9.6. Models of Thinking and Problem Solving

A few modelers have ventured into the largely unknown territory of thinking, reasoning, and problem solving, territory that has been traditionally part of symbolic artificial intelligence. The models so far arrived at for such functions as analogy learning and concept formation cannot properly be called "computational cognitive neuroscience" because they include only sketchy analogies with actual brain processes. Still, these models combined with extensions of models like those discussed in the last two sections are likely to be a step on the way to genuine brain-based models for the same cognitive functions.

9.6.1. Models of Analogy Making

There have been relatively few neural network or connectionist models of the processes of learning, making, and reasoning from analogies. Some of these models use connections among very abstract nodes to explain analogies between complex sentences and to reproduce results of creativity tests from cognitive psychology laboratories. These models notably include different versions of LISA (Hummel & Holyoak, 1997, 2003). Other models have used more brainlike processes to model inference in more limited domains, such as simple proportional analogy queries of the form "A is to B as C is to what?" These models notably include the work of Jani and Levine (2000) and Choe (2004).

Hummel and Holyoak (1997) used LISA to model the processes of access to analogies and mapping from one domain to another. Hummel and Holyoak (2003) extended their earlier work to model the process of reasoning from the mapping thus formed to properties of the second domain. Theirs is a connectionist model that includes nodes representing propositions (e.g., "John loves Mary"), subpropositions (e.g., John as the lover, Mary as the beloved), roles (e.g., lover and beloved), fillers (e.g., John and Mary), and properties (e.g., adult, human, male, female, has-emotion). The later versions of the model rely extensively on what the authors call *dynamic binding*: temporary and quickly learned synchronization between roles and the fillers for the roles (e.g., in the example described above, between "Mary" and "beloved").

Hummel and Holyoak (2003) describe LISA as a "symbolic-connectionist theory," meaning that it is strongly influenced by symbolic artificial intelligence despite having node activities and connection weights. Yet, Knowlton, Morrison, Hummel, and Holyoak (2012) found some general correspondences between processes in LISA and some processes in prefrontal cortex and related brain regions. The dynamic binding between roles and fillers could be implemented by synchronized neural oscillations in the gamma (>30 Hertz) range, perhaps involving both prefrontal and posterior cortex. The various levels of representation in LISA could be mirrored by the increasing abstraction of representations as one goes forward within prefrontal cortex (e.g., Badre & D'Esposito, 2007).

The analogies described by LISA involve parallels between a situation within a "driver" domain and within a "recipient" domain. These parallels involve different people or different objects playing, and therefore temporarily binding to, the same roles, so that further properties of the recipient domain can be inferred from corresponding properties in the driver domain. The model of Jani and Levine (2000) is designed to simulate a different sort of analogies in which there is a change from a construct in the driver domain to another construct the recipient domain. These are proportional analogies, common on college entrance examinations, in which one needs to fill in

the last element, such as "red square : red circle :: yellow square : ?" or "apple : red :: banana : ?."

Jani and Levine's (2000) model starts with the adaptive resonance (ART) model of categorization (Carpenter & Grossberg, 1987a), which is based on a feature layer F_1 and a category layer F_2 (see Chapter 7). Three more layers are added to ART. First there is the abstract category layer F_3, consisting of generalizations such as color, shape, taste, fruit, and word. The authors noted that the relationship between layers $F_2$and F_3 are of the form "IS-A" (e.g., "apple is a fruit") in contrast to the "HAS-A" relationship between $F_2$and F_1 (e.g., "apple has color red"). Second is the relation layer F_4, which includes "IS-A" and "HAS-A" and other key relationships for analogical mapping such as "activate," "suppress," "maintain," and "change." As the authors describe, "in the transition from apple to banana, 'yellow' is activated, 'red' is suppressed, 'fruit' and 'color' are maintained, and 'red' is changed to 'yellow') (Jani & Levine, 2000, p. 154). Each of the activation, suppression, maintenance, and change operations carries with it a different learning law. Finally there is the modulator layer F_5, which includes nodes representing the transitions between analogy items (1-to-2, 2-to-3, 1-to-3, and 3-to-4) and a node representing a form of weight transport. Weight transport sometimes occurs with maintenance and suppression weights; for example, in the analogy "apple is to red as banana is to yellow," the "maintain red" weight going from item 1 to item 2 can be "transported" to "maintain yellow" going from item 3 to item 4.

The weight transport and the unconventional learning laws could be criticized on grounds of biological plausibility, but the authors relate them to the work of other theorists who have proposed various laws whereby a neuron multiplicatively gates signals from two other neurons (Dehaene, Changeux, & Nadal, 1987; Guigon, Dorizzi, Burnod, & Schultz, 1995). The Jani–Levine model is complementary to, not opposed to, the Hummel and Holyoak (1997, 2003) model, so is likely to be also compatible with the prefrontal data discussed by Knowlton et al. (2012).

A different approach to modeling proportional analogies, tied to known interactions between the cerebral cortex and thalamus, is described in Choe (2004). In Choe's model each item in a proportional analogy is represented at three network levels corresponding to cortex, thalamus, and thalamic reticular nucleus (TRN). The TRN is inhibitory, surrounds the thalamus, and serves as a filter that selectively inhibits some of the competing nodes (see the extended ART model of Grossberg & Versace, 2008). Sequential presentation of the first three items in a proportional analogy leads to activation of cortical nodes both for those items and for another item which solves the proportional analogy. If the network's leaky integrator equations (see Appendix 1) are chosen to fit data on relative transmission speeds between these brain regions, TRN activity

filters out those thalamic representations driven by sensory inputs in favor of those thalamic representations driven only by cortical feedback. The result is that the last (solution) item of the analogy, which is not one of the inputs, becomes the only one represented by cortical activity.

9.6.2. Models of Creativity and Concept Formation

Like the analogy models, the few neural network models that have dealt with creative problem solving have had loose connections with neuroscience and been partly influenced by symbolic artificial intelligence. Some of these models are designed to reproduce some behavioral data on creative brainstorming (Doboli, Minai, & Brown, 2007; Doboli, Brown, & Minai, 2009; Farajidavar, Levine, Kohn, & Paulus, 2010; Iyer et al., 2009). Other models are designed to recreate the properties of insight in solving problems with unconventional solutions (Hélie & Sun, 2010).

Experimental brainstorming studies typically ask participants to generate ideas about open-ended problems (e.g., "What might be the consequences if everyone from now on is born with an extra thumb?" "What can be done to improve life at your university?" "What would you do for a pleasant vacation?") While many organizations believe that group interactions generate more ideas for solving a given problem than individuals acting alone, many experimental results have found the opposite to be true (see Paulus, Levine, Brown, Minai, & Doboli, 2010, for review). The disadvantage of groups is due to several factors including cognitive interference and evaluation apprehension. Hence, a variety of other experiments have involved manipulations that might overcome the disadvantage of participants in groups. Some of these manipulations have been studied in network models, including priming with hints from an unconventional category (Iyer at al., 2009) and rest or incubation periods (Farajidavar et al., 2010). The major part of the modeling effort, though, has been directed to simulating the underlying processes of generating ideas and evaluating these ideas for usefulness and novelty.

Figure 9.6 shows a generic architecture for idea generation, with qualitative assignments of different parts to brain regions. The feature and concept modules, under the influence of the context, cues, and cognitive control, comprise an *idea generation subsystem* (IGS) that uses stored semantic knowledge to generate ideas as combinations of concepts already stored in the network. The *critic* receives the generated ideas and produces feedback about the ideas' usefulness and novelty based on its domain knowledge about the current context. The network uses attractor dynamics within each semantic level of the IGS, along with interactions between those levels. The version of the IGS actually implemented in Iyer et al. (2009) includes a subnetwork of concept units along with two different subnetworks of feature units, *type*

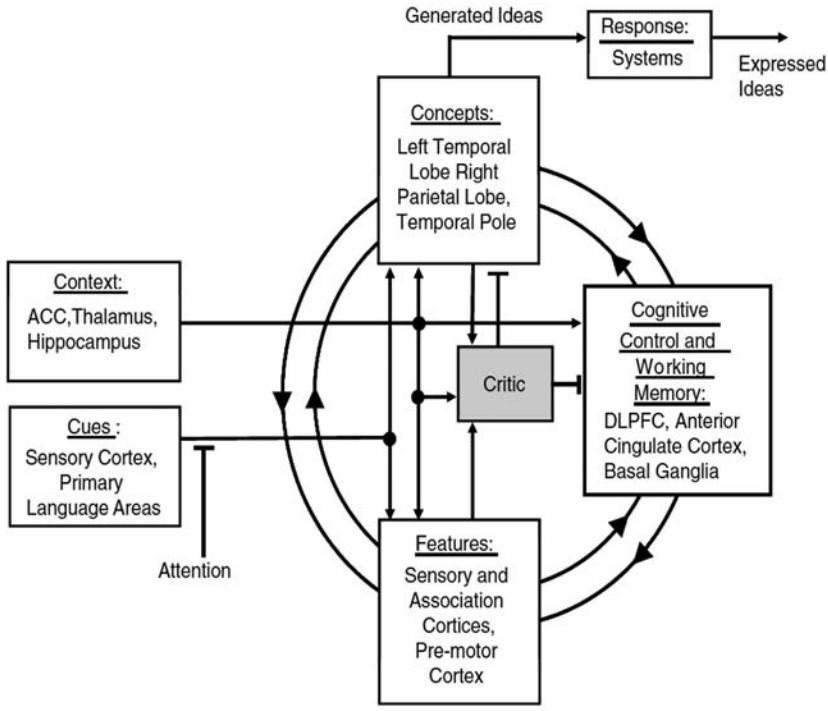


FIGURE 9.6 Basic architecture for an idea generation model, with approximate assignments of modules to brain regions.

Source: Adapted from Iyer et al., 2009, with the permission of Elsevier Science, Inc.

units representing functional types (e.g., "accommodation" or "food") and descriptive units representing attributes (e.g., "expensive" or "sweet-tasting"). There is also a dynamic selector network that groups together concepts that are similar on specific subsets of features, thereby helping to bias the search in favor of concepts relevant to the requirements of the current context.

Hélie and Sun (2010) developed a model of creative problem solving that combines explicit (rule-based) and implicit (associative) processes. Their analysis was based on the popular idea by Wallas (1926) of the four stages in the successful creative process. These four stages are preparation (understanding the problem and acquiring domain knowledge), incubation (removing attention from the problem after an impasse is reached), illumination (or insight, becoming conscious of the solution), and verification. Hélie and Sun modeled the incubation and insight phases. Their model reproduces a wide range of cognitive psychological data about effects of priming and of incubation length.

The Hélie–Sun model, called *EII* (for *explicit–implicit interaction*), is based on the two-layer architecture called *CLARION* (Sun, 2002), which had previously been used to model other cognitive processes including categorical reasoning (Sun & Zhang, 2006; see Chapter 7). The bottom layer deals with implicit knowledge, which is formed by associations and guided by intuition, whereas the top layer deals with explicit, rule-based knowledge. CLARION is further divided into a non-action-oriented part that encodes declarative knowledge (as is used in Hélie & Sun, 2010) and an action-oriented part that encodes procedural knowledge. These articles make no specific assignments of parts of the network to brain regions.

The model's simulations of creative problem solving data rely on the encoding at the top level of EII of a range of potential hypotheses for their solution. For example, one of the problems they model involves explaining the following story:

> A man walks into a bar and asks for a glass of water. The bartender points a shotgun at the man. The man says, "Thank you," and walks out.
>
> (Hélie & Sun, 2010, p. 1011, quoted from Durso, Rea, & Dayton, 1994, p. 95)

The correct solution is that the man had the hiccups, which either water or being startled could remedy. Both the model and experiment participants try to obtain the solution by asking yes/no questions. Those participants and those instances of the model that solved the problem correctly arrived at different knowledge structures for associations between words than did those who failed to solve it. Hypotheses are activated probabilistically according to a Boltzmann distribution with a parameter that measures how broadly the conceptual space is searched for the answer. The model captures how long incubation tends to increase the value of this search parameter and therefore the likelihood of obtaining the correct solution. The model also captures results showing that obtaining the solution is hindered by too much reliance of explicit processing, which tends to restrict the search to conventional associations. This is analogous to results reviewed by Iyer et al. (2009) showing that primes from rarely accessed categories are more helpful in brainstorming than primes from frequently accessed categories.

9.6.3. Reasoning, Bayesian Inference, and the Prefrontal Cortex

A model of human reasoning with closer connections to neuroscience is the PROBE model of Koechlin and his colleagues (Collins & Koechlin, 2012; Donoso, Collins, & Koechlin, 2014; Koechlin, 2014). Their model involves choices between behavioral strategies. It was applied to experimental results by the same authors on different versions of a cognitive task with feedback, a

complex analog of the Wisconsin card sorting test. Probabilities of the decision maker adopting specific behavioral strategies are updated each time the model receives positive or negative feedback, using calculations based on Bayes's rule for deriving postoutcome probabilities from preoutcome probabilities.

PROBE's formal calculations of probabilities are not based on interactions between nodes in a connectionist network. Yet, the model can simulate the authors' behavioral results, including context effects and individual differences between decision-makers. The individual differences deal in part with tendencies toward exploitation (repeated use of strategies that proved to be successful) versus exploration (trying out novel strategies).

Based on years of work on the functions of different prefrontal areas, these authors developed hypotheses about the roles of different prefrontal regions in the performance of this task, and in the model. Donoso et al. (2014) verified those hypotheses qualitatively through fMRI studies. They found that activity in the anterior (including dorsolateral) PFC correlated with reliability of the current strategy. The OFC and ventral striatum were sensitive to the outcome of the current strategy. The frontopolar cortex was sensitive to the reliability of alternative strategies not currently in use, so was assumed to play a role in exploration.

Koechlin and his colleagues describe the model as "Bayesian" because it updates probabilities based on reinforcement learning. Various Bayesian theorists (see in particular Oaksford & Chater, 2007, 2009) claim that much human behavior which appears to be irrational (such as the types of decisions discussed in Section 9.5) is actually rooted in probabilistic inference.

Recall from Section 9.2.2 that Bowers and Davis (2012) critique the Bayesian approach on the grounds that there is little proof, either behavioral or neuroscientific, that the brain performs tasks in an optimal or nearly optimal manner. Their critique applies to multiple domains: reasoning, perception, and motor control. Bowers and Davis argue further that brain processes are constrained by biological and evolutionary considerations that often have little relevance to the current task. Rather than being optimal, the brain and mind make use of the neural structures that are available to solve cognitive and behavioral problems as best they can. Despite the use of the term "Bayesian," the model of Collins and Koechlin (2012) is not inconsistent with the Bowers–Davis critique. Indeed, those authors state that "actor selection is "based on a '*satisficing*' criterion based on task set reliability" (p. 8, authors' italics). Satisficing is a term coined by Simon (1956) to denote arriving at solutions that are "good enough" but not necessarily optimal. The degree to which human or animal behavior is optimal remains one of the most lively and important controversies not only in neural network theory but in cognitive and behavioral neuroscience.

Space and time considerations prevented the author from adding detailed expositions of the recent state of neural network modeling in several other

important domains. Among these domains are motor control, episodic memory, language learning, perceptual decision-making, and consciousness. A few key sources in each of these areas (some of them cited in other chapters) are listed at the end of this chapter.

Equations for Networks in Chapter 9

Wisconsin Card Sorting Test Model

In the Wisconsin card sorting test (WCST), the experimenter goes twice through a special deck of cards each with a number (one, two, three, or four) of a design (triangle, star, cross, or circle) of a given color (red, green, yellow, or blue), yielding a total of $2 \times 4 \times 4 \times 4 = 128$ cards. The subject must classify each card as being "like" one of four templates as shown in Figure 5.6: one red triangle, two green stars, three yellow crosses, or four blue circles.

Now let us go through what happens on a trial with a single card input in the neural network of Figure 9.4 by Leven and Levine (1987) and Levine and Prueitt (1989). For definiteness, let us say that the input card only has one feature at most in common with each template card, say, one green circle. Aside: in the network (and probably in real life as well), presentation of cards that have two or more features in common with a template can make set switching on the WCST more difficult. For example, if it is at the point on the test where the subject has learned the color criterion but the experiment is teaching him or her to switch to a shape criterion, the subject presented with the "two blue circles" card will typically classify it as like the "four blue circles" template. The experimenter will say "Right," since the subject has made the correct match on the shape criterion. But, since the choice also matches the template on the color criterion, unlearning of the color rule will be slowed down. For this reason, Nelson (1976) did a modification of the WCST in which ambiguous cards like the two blue circles card were eliminated, reducing the number of card presentations from 128 to 48.

Return to the trial of the network on the "one green circle" input. The "one" feature detector at the F_1 layer of the network is activated, and in turn activates the "one red triangle" node at the F_2 layer. Also "green" at F_1 activates "two green stars" at F_2, and "circle" at F_1 activates "four blue circles" at F_2. So depending on whether she or he is classifying by number, color, or shape, the subject could choose any of three possible templates;

the template nodes are in competition and the one most activated is chosen. (A human subject could decide to classify the input with the "three yellow crosses" template out of perversity, just because NONE of the features match. But such a sense of humor has not yet been built into the neural network!)

What regulates the competition between the card category nodes at F_2? The activations of the three nodes (one, green, and circle) at F_1 are equal, so whichever F_2 node has the strongest bottom-up connection weight from F_1 wins the competition. But these bottom-up connections are gated by the strength of the bias nodes for the corresponding features (respectively, number, color, and shape). So essentially the choice the subject makes will be determined by which bias node activity Ω_k is the largest.

The relative activities of the number, color, and shape bias nodes depend on the history of choices made on other input cards. Specifically, every time the network has classified an input as being like a template it matches on a feature, the bias node activity has increased if the experimenter has rewarded that choice, and decreased if the experimenter has punished the choice. In the "frontally damaged" version of the network, the amount of that increase or decrease in bias node activity is much smaller than it is for the "normal" version of the network. Hence, if the "frontally damaged" network is rewarded on early trials for color choices, its color bias will move way ahead of its number and shape biases, and subsequent negative reinforcement will be too weak in its effects to change that ordering. But, in the "normal" network, negative reinforcement for color-matching choices that do not match on shape eventually reduce the color bias to the point that the shape bias can overtake it if choices that match on shape are rewarded.

Equations of Raizada and Grossberg (2001)

Retina

The model retina has at each position (i, j) both an ON-cell, u_{ij}^{+}, whose receptive field has a narrow on-center and a Gaussian off-surround, and an OFF-cell, u_{ij}^{-}, with a narrow off-center and a Gaussian on-surround. The retinal cell activities in response constant visual inputs I_{ij} have the equilibrium values

$$u_{ij}^{+} = I_{ij} - \sum_{p,q} G_{pq}\left(i, j, \sigma_1\right) I_{pq} \tag{9.1}$$

and

$$u_{ij}^{-} = -I_{ij} + \sum_{p,q} G_{pq}\left(i, j, \sigma_1\right) I_{pq} \tag{9.2}$$

where $G_{pq}(i, j, \sigma)$ is a two-dimensional Gaussian kernel, given by

$$G_{pq}\left(i, j, \sigma\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left(\left(p-i\right)^2 + \left(q-j\right)^2\right)\right) \tag{9.3}$$

Lateral Geniculate

The ON and OFF cells of the LGN, v_{ij}^{+} and v_{ij}^{-}, are excited by the half-wave rectified ON and OFF cells of the retina, respectively. These retinal inputs are also multiplicatively gain-controlled by on-center off-surround feedback from V1 Layer 6. Layer 6 cells, x_{ijk}, at position (i, j) and sensitive to orientation k, send on-center excitation, A_{ij}, to LGN neurons at the same position, and send a two-dimensional Gaussian spread of off-surround inhibition, B_{ij}, to LGN neurons at the same and nearby positions:

$$\frac{1}{\delta_v}\frac{dv_{ij}^{+}}{dt} = -v_{ij}^{+} + \left(1 - v_{ij}^{+}\right)\left[\, u_{ij}^{+}\right]^{+}\left(1 + A_{ij}\right) - \left(1 + v_{ij}^{+}\right) B_{ij} \tag{9.4}$$

$$\frac{1}{\delta_v}\frac{dv_{ij}^{-}}{dt} = -v_{ij}^{-} + \left(1 - v_{ij}^{-}\right)\left[\, u_{ij}^{-}\right]^{+}\left(1 + A_{ij}\right) - \left(1 + v_{ij}^{-}\right) B_{ij} \tag{9.5}$$

where A_{ij} and B_{ij} represent feedback from layer 6 given by

$$A_{ij} = C_1 \sum_{k} x_{ijk} \text{ and } B_{ij} = C_2 \sum_{p,q,k} G_{pj}\left(i, j, \delta_1\right) x_{ijk}$$

with G_{pj} being the Gaussian defined in Equation (9.3).

LGN Inputs to Cortical Simple Cells

At each position, (i, j), and for each orientation, k, the model has a simple cell with two parts: an ON subregion R_{ijk}, which is excited by LGN ON cells beneath it and is inhibited by LGN OFF cells at the same position; and an OFF subregion L_{ijk}, which has the reverse relation to the LGN channels. This physiology is embodied in the equation for the ON subregion by subtracting the half-wave rectified LGN OFF channel,$[v_{pq}^{-}]$, from the rectified ON channel,$[v_{pq}]$, and convolving the result with a rectified difference of Gaussians. The OFF subregion, L_{ijk}, is similarly constructed:

$$R_{ijk} = \sum_{p,q}\left(\left[v_{pq}^{+}\right]^{+} - \left[v_{pq}^{-}\right]^{+}\right)\left[D_{pqij}^{(k)}\right]^{+}$$

$$L_{ijk} = \sum_{p,q} \left(\left[v^{-}_{pq} \right]^{+} - \left[v^{+}_{pq} \right]^{+} \right) \left[-D^{(k)}_{pqij} \right]^{+}$$

where

$$D^{(k)}_{pqij} = G_{pq}\left(i - \delta \cos \vartheta, j - \delta \sin \vartheta, \sigma_2\right) - G_{pq}\left(i + \delta \cos \vartheta, j + \delta \sin \vartheta, \sigma_2\right)$$

with $\delta = \sigma_2/2$ and $\theta = \pi\ (k - 1)/K$, k ranging from 1 to $2K$ with K being the total number of orientations.

Simple cell activity is given by the rectified sum of the activities of each subfield, minus their difference, that is

$$S_{ijk} = \gamma \left[R_{ijk} + L_{ijk} - \left| R_{ijk} - L_{ijk} \right| \right]^{+}$$

Layer 6 Cells

V1 Layer 6 cells, x_{ijk}, receive input from the LGN, which is represented by the contrast-polarity pooled oriented input, C_{ijk}. They also receive two types of folded feedback excitation. The first type is intracortical feedback from above-threshold pyramidal cells in V1 Layer 2/3, z_{ijk}. These are passed through a thresholding signal function, F, given by

$$F\left(z_{ijk}, \Gamma\right) = \max\left(z_{ijk} - \Gamma, 0\right)$$

where G is the threshold value. The second type of folded feedback is intercortical attentional feedback from V2, x_{ijk}^{V2}. In attentional simulations, an additional term, *att*, is added to the excitatory channel, implementing a two-dimensional Gaussian spread of attentional signals, centered on the attended location and exciting all orientations equally. This attentional term is applied both to V1 and to V2. In the nonattentional simulations, *att* = 0. Thus

$$\frac{1}{\delta_c} \frac{dx_{ijk}}{dt} = -x_{ijk} + \left(1 - x_{ijk}\right)\left(\alpha C_{ijk} + \phi F\left(z_{ijk}, \Gamma\right) V_{21} x_{ijk}^{V2} + att\right) \tag{9.6}$$

Equation (9.6) is solved at equilibrium, that is, by setting the left-hand side to be 0. The equations for Layer 6 of V2 are identical to (9.6) except that the V2-to-V1 feedback term $V_{21} x_{ijk}^{V2}$ is absent.

Layer 4 Cells

Spiny stellate cells in Layer 4 receive a pooled oriented input C_{ijk} which is equal to the sum of the two inputs from simple cell inputs sensitive to opposite

orientations, namely $S_{ijk} + S_{ij(k+K)}$, where K is half the number of orientations represented in the network. This is based on studies showing that Layer 4 simple cells that are sensitive to opposite contrast polarities pool their outputs at Layer 2/3 complex cells. These same spiny stellate cells also receive an on-center off-surround input from Layer 6. The off-surround interactions are determined by a two-dimensional kernel, W_{pqrijk}. The combination of these influences leads to the following equation for the spiny stellate cells:

$$\frac{1}{\delta_c}\frac{dy_{ijk}}{dt} = -y_{ijk} + \left(1 - y_{ijk}\right)\left(C_{ijk} + \eta^{+} x_{ijk}\right) - \left(y_{ijk} + 1\right) f\left(\sum_{p,q,r} W^{+}_{pqrijk} m_{pqr}\right) \qquad (9.7)$$

where f is the sigmoid function $f(x) = \mu x^n/(\nu^n + x^n)$ and the m_{pqr} values represent activities of inhibitory interneurons. Equation (9.7), like (9.6), is solved at equilibrium. The equations for the inhibitory interneuron activities are

$$\frac{1}{\delta_m}\frac{dm_{ijk}}{dt} = -m_{ijk} + \eta^{-} x_{ijk} - m_{ijk} f\left(\sum_{p,q,r} W^{+}_{pqrijk} m_{pqr}\right)$$

Layer 2/3 Cells

The pyramidal cells in Layer 2/3, z_{ijk}, receive excitatory input from Layer 4 cells, y_{ijk}, at the same position and orientation and also long-range bipole excitation from the thresholded outputs of other Layer 2/3 pyramidals with collinear, coaxial receptive fields, $F(z_{ijk})$. Inhibitory interneurons in Layer 2/3, s_{ijk}, also synapse onto these pyramidals. As with the inhibitory kernels in Layer 4, W^{+} and W^{-}, the Layer 2/3 cells synapse onto each other through linearly scaled versions of the self-organized kernels grown in the model of Grossberg and Williamson (2001). Layer 2/3 pyramidals also receive short-range inhibition from inhibitory interneurons at the same position and of the same orientation, s_{ijk}. This inhibition operates through a self-organized short-range kernel, T^{+}. Layer 2/3 cells also receive attentional feedback, whose strength is measured by two coefficients called a^{23}_{excit} and a^{23}_{inhib}. All these effects lead to the following equations for Layer 2/3 pyramidal cell activities:

$$\frac{1}{\delta_z}\frac{dz_{ijk}}{dt} = -z_{ijk} + \left(1 - z_{ijk}\right)\left(\lambda[y_{ijk}]^{+} \sum_{p,q,r} H_{pqrijk} F\left(z_{pqr}, \Gamma\right) + a^{23}_{\text{excit}} att\right) - \left(z_{ijk} + \Psi\right)\sum_{r} T^{+}_{rk} S_{ijr}$$

where s_{ijrw} are Layer 2/3 inhibitory interneuron activities that satisfy the equations

$$\frac{1}{\delta_S}\frac{d}{dt}S_{ijk} = -S_{ijk} + \sum_{p,q,r} H_{pqrijk} F\left(z_{pqr}, \Gamma\right) + a_{\text{inhib}}^{23} att - S_{ijk} T_{rk}^{-} S_{ijr}$$

Feedforward Projections from V1 to V2

The thresholded output of V1 Layer 2/3 projects forward to Layers 6 and 4 of V2, with activities x_{ijk}^{V2} and y_{ijk}^{V2}, respectively, following the same pattern as the LGN forward projections to Layers 6 and 4 ofV1. The equations for these projections are

$$\frac{1}{\delta_C}\frac{dx_{ijk}^{V2}}{dt} = -x_{ijk}^{V2} + \left(1 - x_{ijk}^{V2}\right)\left(V_{12}^{6} F\left(z_{ijk}, \Gamma\right) + \Phi F\left(z_{ijk}^{V2}, \Gamma\right) + att\right)$$

where z_{ijk}^{V2} is the activation in Layer 2/3 of V2

and

$$\frac{1}{\delta_C}\frac{dy_{ijk}^{V2}}{dt} = -y_{ijk}^{V2} + \left(1 - y_{ijk}^{V2}\right)\left(V_{12}^{4} F\left(z_{ijk}, \Gamma\right) + \eta^{+} x_{ijk}^{V2}\right) - \left(y_{ijk}^{V2} + 1\right) f\left(\sum_{p,q,r} W_{pqrijk}^{+} m_{pqr}^{V2}\right)$$

Equations for the LISSOM Model

The versions of LISSOM utilized by Bednar and Miikkulainen (2000) to model tilt aftereffect and by Choe and Miikkulainen (2004) to model perceptual grouping are closely related but slightly different, because the latter version includes spikes to model synchronization of inputs. We present both versions in turn.

In Bednar and Miikkulainen (2000), both retinal ganglion cells and cortical neurons are organized in two-dimensional arrays. The response at time 0 of cortical neuron (i, j) is a weighted sum of retinal activations, namely

$$\eta_{ij} = \delta\left(\sum_{ab} \xi_{ab}^{+} \mu_{ij,ab}\right)$$

where ξ_{ab} is the activation of retinal ganglion cell (a, b); μ_{ij}, ab is the afferent weight from (a, b) to (i, j); and σ is a piecewise linear approximation of a sigmoid activation function. Over a short time the response defined by Equation (9.8)

is influenced by lateral connection weights between cortical neurons. If $E_{ij,kl}$ is the excitatory connection weight from cortical neuron (k, l) to (i, j), and $I_{ij,kl}$ is the inhibitory connection weight from (k, l) to (i, j), then at each time step

$$\eta_{ij}(t) = \sigma\left(\sum_{a,b} \xi^{+}_{ab}\mu_{ij,ab} + \gamma_e \sum_{k,l} E_{ij,kl}\eta_{kl}(t-1) - \gamma_i \sum_{k,l} I_{ij,kl}\eta_{kl}(t-1)\right) \tag{9.9}$$

where γ_e and γ_i are scaling factors that determine the relative strengths of excitation and inhibition.

All weights, both afferent and lateral, are modified at smaller time steps according to an associative learning rule followed by presynaptic normalization. If we give the symbol w (subscripted) to all weights (whether μ for afferent, E for excitatory lateral, or I for inhibitory lateral) and the symbol x to all activities (ξ for retinal or η for cortical), associative learning with presynaptic normalization leads to the rule

$$w_{ij,mn}(t+\Lambda t) = \frac{w_{ij,mn}(t) + \alpha n_{ij} x_{mn}}{\sum_{m,n} w_{ij,mn}(t) + \alpha n_{ij} x_{mn}} \tag{9.10}$$

where α is the learning rate.

The network of Choe and Miikkulainen (2004) differs from that of the 2000 article both in including spikes and in having two cortical layers, one of which has broader excitatory interactions which it uses to perform preattentive grouping. Each connection in the network does exponentially decayed summation of incoming spikes. If x is the input spike ($x = 1$ if a spike occurs at a given time and 0 if it doesn't) then the decayed sum of spikes is

$$s(t) = \sum_{n=0}^{t} x(t-n)\mathrm{e}^{-\lambda n} \tag{9.11}$$

where λ is a decay rate (different for excitatory lateral, inhibitory lateral, and between-layer connections). Spikes are based on comparing an input with a threshold, which is partly dependent on the neuron's previous activity. That threshold adds a base term to terms representing absolute and relatively refractory periods:

$$\theta(t) = \theta_{\text{base}} + \theta_{\text{abs}}(t) + \tau\theta_{\text{rel}}(t) \tag{9.12}$$

In (9.12), the absolute term is ∞ if there has been spike over a certain time period. The relative term is a leaky integrator similar to (9.11), namely

$$\theta_{\text{rel}}(t) = y(t) + \theta_{\text{rel}}(t-1)\mathrm{e}^{-\lambda_{rel}}$$

where y is the incoming spike (1 or 0). The input to the spike generator of the cortical neuron (i, j) at each layer is given by

$$\sigma_{ij}(t) = g\left(\gamma_a \sum_{r,s} \xi_{rs}\mu_{ij,rs} + \gamma_c \sum_{p,q} \zeta(t-1)_{pq}\, v_{ij,pq} + \gamma_e \sum_{k,l} \eta_{kl}(t-1) E_{ij,kl} - \gamma_i \sum_{k,l} \eta_{kl}(t-1) I_{ij,kl}\right) \quad (9.13)$$

where the γ's denote relative strengths of afferent, interlayer, and excitatory and inhibitory lateral contributions; ξ_{rs} is the input level of retinal neuron (r, s), with $\mu_{ij,rs}$ the corresponding afferent connection weight; ζ_{pq} is the decayed sum of spikes of the cortical neuron (p, q) of the other layer than where (i, j) is, with $v_{ij,pq}$ the corresponding interlayer connection weight; $\eta_{kl}(t-1)$ is the decayed sum of spikes of the neuron (k, l) of the same layer as (i, j), with $E_{ij,kl}$ and $I_{ij,kl}$ the corresponding excitatory and inhibitory connection weights; and g is a piecewise linear function of the form

$$g(x) = \begin{cases} 0 & \text{if } x < \delta \\ 1 & \text{if } x > \beta \\ \dfrac{x-\delta}{\beta-\delta} & \text{otherwise} \end{cases}$$

As in the 2000 article, all weights (the label w given to μ for afferent, v for interlayer, E for excitatory and I for inhibitory within each layer) are modified with associative learning followed by normalization, this time postsynaptic, so (9.10) is modified to

$$w_{ij,mn}(t+\Lambda t) = \frac{w_{ij,mn}(t) + \alpha n_{ij} x_{mn}}{\sum_{i,j} w_{ij,mn}(t) + \alpha n_{ij} x_{mn}}$$

Equations for the Wisconsin Card Sorting Test Model of Levine and Prueitt

The equations were originally presented in Leven and Levine (1987), but that is a conference proceedings that is not easily available: the same equations can be found in Levine and Prueitt (1989), which is a journal article much more accessible. The network of Figure 9.4 for simulating the card sorting data of Milner (1964) used the naming convention of x_i for feature node activities and y_j for category node activities. As shown in that figure, i = 1, 2, 3, 4 represent those numbers in order; i = 5, 6, 7, 8 represent colors, in the order red, green, yellow,

blue; i = 9, 10, 11, 12 represent shapes, in the order triangle, star, cross, circle. The equation for each x_i reflects shunting excitation from the category nodes weighted by the top-down synaptic weights w_{ji} and from the input card if it includes the appropriate feature. It also reflects shunting inhibition from all category nodes, in order that category node activity should not be perceived as input (following Carpenter & Grossberg, 1987a). Hence,

$$\frac{dx_i}{dt} = -Ax_i + (B - Cx_i)\left(I_i + \sum_{j=1}^{4} f(y_j) w_{ji}^2 \right) - Dx_i \sum_{j=1}^{4} f(y_j) \qquad (9.14)$$
$$i = 1, 2, \ldots, 12$$

where A, B, C, and D are positive constants, and f is a sigmoid function. I_i is a short-duration input signal that has a constant large value on cards that include the feature corresponding to the ith node and a 0 value on other cards.

The category node activities y_j (j = 1, 2, 3, 4 for the categories defined by the template cards in order as shown in Figure 5.6 of Chapter 5) are excited by the x_i, weighted by the appropriate bias nodes, and by the "bottom-up" synaptic weights w_{ij}. The bias nodes Ω_k have the subscript k = 1 for number, k = 2 for color, and k = 3 for shape. For each feature node i, the corresponding bias number is $k = [(i+3)/4]$, where for any number u the symbol $[u]$ represents the greatest integer not exceeding w. The node y_j is inhibited by other category nodes. Hence

$$\frac{dy_j}{dt} = -Ay_j + (B - Cy_j)\left(f(y_j) + \sum_{i=1}^{12} g\left(\Omega_{\left[\frac{i+3}{4}\right]} x_i \right) w_{ij}^1 \right) \qquad (9.15)$$
$$-Dy_j \left(\sum_{r \neq j} f(y_r) + I \right) \quad j = 1, 2, 3, 4$$

where f is the sigmoid function used in the equations for x_i, g is a saturating threshold-linear function, and I is an inhibitory signal, lasting longer than I_i, that occurs with any input. The effect of I is to prevent perseveration of the category choice made to the previous input card (which does not tend to occur, even in frontal patients). For any given input card, it is assumed that the template card chosen is the one whose corresponding, y_j is the largest after a certain number of time steps.

The w_{1ij} and w_{2ji} are not modifiable; effective short-term changes in associative strength are delivered through changes in the bias node activities which gate the w_{1ij} signals. The values of these synaptic strengths are large if node i represents a feature present on template card j, which occurs if $i = j$, j + 4, or j + 8 (e.g., if j = 1 for the "one red triangle" template, and i is either 1 for "one," 5 for "red," or 9 for "triangle"). Hence w_{1ij} = 5 if $i = j$, j + 4, or j + 8, and w_{1ij} = .2 otherwise. $w_{2ji} = w_{1ij}/5$ for all i and j.

The bias node activities Ω_k, k = 1, 2, 3 decay more slowly than feature or category node activities, thus exhibiting a form of intermediate-term memory for the duration of the card sorting test. A bias node is only subject to shunting excitation and inhibition if a "match signal" occurs between the input and the chosen template on the given feature. That signal is computed at the kth match signal generator node by

$$\Phi_k = \sum_{i=4k-3}^{4k} w_{ji}^2 I_i \tag{9.16}$$

where j is the index of the chosen template. For example, suppose the input card shows one blue cross and the template card shows four blue circles. Then for the color bias node, k = 2, so the range of the summation in Equation (9.14) is from 5 to 8. I_i is positive for i = 8 (the color blue), and I_i = 0 for i = 5, 6, 7. Since j = 4, $\Phi_k = w_{48} I_8 = 5I_8$, which is large. The "matching" bias node is excited by itself, by the corresponding habit node h_k (if its activity exceeds a threshold θ_1), and by positive reinforcement from the experimenter. That node is inhibited by other bias nodes and by negative reinforcement. Hence,

$$\frac{d\Omega_k}{dt} = -E\Omega_k + \left\{ \begin{array}{l} \left(F - \Omega_k\right)\left(\left(h_k - \theta_1\right)^+ + \alpha R^+ + g\left(\Omega_k\right)\right) \\ -\Omega_k \left(\left(\alpha R^-\right) + G\sum_{r \neq k} g\left(\Omega_r\right) \right) \end{array} \right\} f\left(\Phi_k\right) \tag{9.17}$$

for positive constants E, F, G, where $u+$ denotes max (u, 0) and u^- denotes max ($-u$, 0); the reinforcement signal R is 1 for a correct choice and –1 for an incorrect choice. The strength of the reinforcement signal, α, is 4 for the undamaged version of the network and 1.5 for the frontal patient version. The functions f and g are the same ones used in Equations (9.14) and (9.15).

Habit nodes h_k are influenced only by the match signals Φ_k, as defined in Equation (9.14). For example, the color habit is strengthened by choices made (whether correctly or incorrectly) on which template and input have the same color, and weakened by choices on which template and input have different colors. Hence,

$$\frac{dh_k}{dt} = -Hh_k + \left\{ \left(J - h_k\right)\left(\Phi_k - \theta_2\right)^+ - \left(\Phi_k - \theta_2\right)^- \right\} \tag{9.18}$$

where θ_2 is a threshold and H and J positive constants.

Exercises for Chapter 9

○1. The Bayesian approach to neural network modeling is mentioned twice in this chapter – in relation to vision in Section 9.2.2 and to reasoning in Section 9.6.3. It has also been applied to reinforcement learning. Recall that this type of modeling is based on the assumption that the brain performs particular tasks in a nearly optimal fashion, and therefore the tasks can be essentially modeled by considering the structure of the environment in which they are performed.

Do you believe that optimality is a useful organizing criterion for modeling how the brain performs tasks? Pick a specific domain of neural or cognitive functioning and argue either for or against (or both!) the utility of a Bayesian approach to modeling that domain.

○2. Choose two to four related experimental findings on working memory. Using any of the neural architecture principles found in this chapter or Chapters 3, 4, 6, 7, or 8, develop and attempt to simulate a model of these findings, incorporating known roles of regions in the prefrontal cortex and basal ganglia.

*3. Run a simulation of the Levine and Prueitt (1989) model of the Wisconsin card sorting test as described by Equations (9.14)–(9.18) and the intervening text. Over time let the input I_i run through a random arrangement of all possible cards, each presented twice, for a total of 4 (number) × 4 (color) × 4 (shape) × 2 = 128 trials. Set initial values of the x_i in (9.14) and the y_j in (9.15) to 0. Set initial values of the Ω_k in (9.17) and the h_k in (9.18) to 1.

For the parameters in the equations use the values $A = 10$, $B = 5$, C = 1, $D = 1$, $E = .01$, $F = 3$, $G = 10$, $H = .5$, $J = 3$, $\theta_1 = 1$, $\theta_2 = .1$, $I = 100$. Set the input intensities I_1 to 5 if feature i is present and 0 otherwise. $w_{1ij} = 5$ if $i = j$, $j + 4$, or $j + 8$, and $w_{1ij} = .2$ otherwise. $w_{2ji} = w_{1ij}/5$ for all i and j. For the sigmoid function use $f(x) = \arctan(x - 1) + \pi/2$, and for the threshold-linear function use $g(x) = 0$, $x < .5$.

$x - .5$, $.5 \leq x \leq 3$

2.5, $x > 3$.

4. Simulate the parallel distributing model of the Stroop effect in Cohen et al. (1990). Study the training length (number of epochs, training phase) and the reaction time (measure of time, testing phase). The network is shown in Figure 9.7, where the inputs are either 0 or 1. The training signal, *target output*, is a 2-element depending on which response is correct as shown in last two columns of Table 9.1.

Use Table 9.1 to train a standard feedforward backpropagation neural network with one hidden layer and sigmoid function as

$$a_j(t) = \frac{1}{1+e^{-\text{net}_k}} \text{ and}$$
$$\text{net}_j(t) = \sum_i x_i(t) w_{ij} + b_j(t), \text{ and}$$
$$i = (1 \ldots .5), \text{ and } j = (1 \ldots 4)$$

Let the output layer train the nodes with the sigmoid

$$y_k(t) = \frac{1}{1+e^{-\text{net}_k}} \text{ and}$$
$$\text{net}_k(t) = \sum_k a_j(t) w_{jk} + b_k(t), \text{ and}$$
$$j = (1 \ldots .4), \text{ and } k = (1 \ldots 2)$$

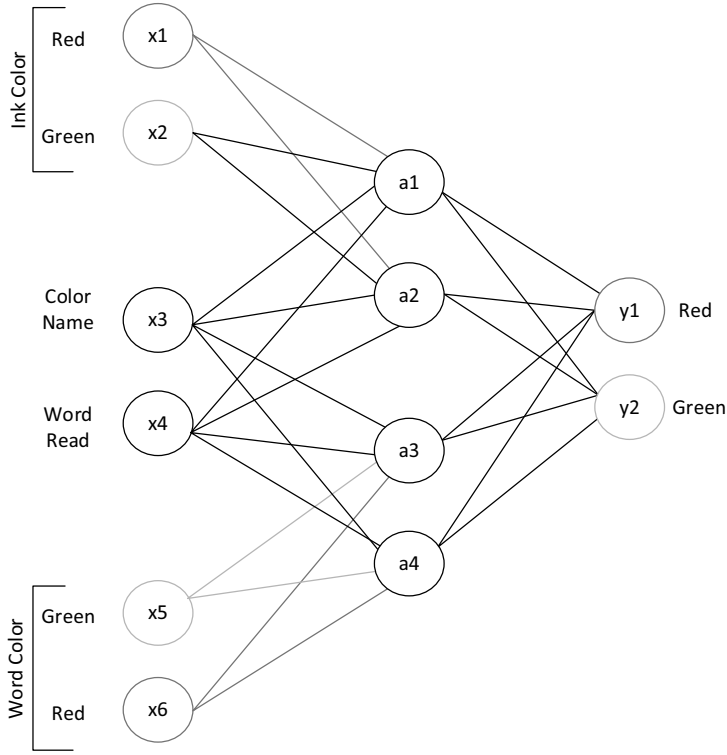


FIGURE 9.7 PDP network for the Stroop test simulation.

Task demand		*Color input*		*Word input*		*Word output*	
Color ($x_i = 1$)	*Word* ($x_i = 2$)	*Green* ($x_i = 3$)	*Red* ($x_i = 4$)	*Green* ($x_I = 5$)	*Red* ($x_i = 6$)	*Green* ($y_k = 1$)	*Red* ($y_k = 2$)
1	0	0	1	0	0	0	1
1	0	1	0	0	0	1	0
0	1	0	0	0	1	0	1
0	1	0	0	1	0	1	0

TABLE 9.1 Training data for the Stroop simulation

Let the hidden and output bias parameters $b_j(t) = b_k(t) = 1$, $\forall$ (j, k) for simplicity.

For the testing phase using Table 9.2 as the eight different stimuli for testing, and use the weight generated in the training phase along with the response and attentional selection mechanism for the hidden layer sigmoid as:

$$\overline{a_J}(t) = \frac{1}{1 + e^{-\overline{net_J}}}$$

$$\overline{\text{net}_J}(t) = \tau\left(\sum_i \left(x_i(t) w_{ij} + b_j(t)\right)\right) + (1 - \tau)\overline{\text{net}_J}(t-1)$$

$$i = (1 \ldots .5), \text{ and } j = (1 \ldots 4)$$

	Task demand		*Color input*		*Word input*	
Case	*Color* ($x_i = 1$)	*Word* ($x_i = 2$)	*Green* ($x_i = 3$)	*Red* ($x_i = 4$)	*Green* ($x_i = 5$)	*Red* ($x_i = 6$)
Task specification	1	0	0	0	0	0
Control	1	0	0	1	0	0
Conflict	1	0	0	1	1	0
Congruent	1	0	0	1	0	1
Task specification	0	1	0	0	0	0
Control	0	1	0	0	0	1
Conflict	0	1	1	0	0	1
Congruent	0	1	0	1	0	1

TABLE 9.2 Test data for the Stroop simulation.

Now for the output layer, test the nodes with the following sigmoid function. Iteratively, increment t and solve the three following equations. Compare the output $y_k(t)$ to the threshold of one and record the time for as the reaction time and the correct response.

$$\overline{a_k}(t) = \frac{1}{1 + e^{-\overline{net_k}}}$$

$$\overline{\text{net}_k}(t) = \tau\left(\sum_k \left(\overline{a_k}(t) w_{jk} + b_k(t)\right)\right) + (1-\tau)\overline{\text{net}_k}(t-1)$$

$$i = (1 \ldots .4), \text{ and } j = (1 \ldots 2)$$

The evidence accumulator equation is as follows:

$$y_k(t) = y_k(t-1) + \alpha\left(\overline{a_k}(t) - \max\left(\overline{a_k}(t)\right)\right) + \sigma v_j(t)$$

Try σ = 0.01 to 0.1, α = 0.01 to 0.1 were $v_j(t)$ Gaussian random variable sampled with N(0,1), zero mean, and one standard deviation. (Hint: the reaction time is ordered faster for the congruent, control, and conflict, respectively.)

Some Additional Sources

Models of Vision and Visual Attention

Corchs and Deco (2002); Deco and Rolls (2005b); Gori, Giora, Yazdanbakhsh, and Mingolla (2011); Grossberg (2001, 2003, 2007); Hoyer and Hyvärinen (2002); Huang and Grossberg (2010); Humphreys and Muller (1993); Portilla, Strela, Wainwright, and Simoncelli (2003); Riesenhuber and Poggio (1999); Schwartz, Sejnowski, and Dayan (2006); Yazdanbakhsh and Grossberg (2004).

Models of Spatiotemporal Pattern Processing

Giles, Kuhn, and Williams (1994); Mannes (1992); Marshall (1990, 1995).

Models of Sequence Learning and Performance

Berns and Sejnowski (1998); Contreras-Vidal and Schultz (1999); Davelaar (2007); Dehaene, Changeux, and Nadal (1987); Durstewitz, Seamans, and Sejnowski (2000a, 2000b); Minai and Levy (1993); Nakahara, Doya, and Hikosaka (2001).

Models of Executive Function and Working Memory

Ashby et al. (2005); Humphries, Stewart, and Gurney (2006).

Models of the Stroop Test

Cohen, Usher, and McClelland (1998); Phaf, Heijden, and Hudson (1990).

Models of Cognitive Control

Botvinick, Niv, and Barto (2009); Botvinick and Cohen (2014).

Models of Decision-Making

Levine and Perlovsky (2008); Sieck and Yates (2001); Teodorescu and Usher (2013); Usher and Zakay (1993).

Models of Neuroeconomics

Kim, Hwang, Seo, and Lee (2009).

Models of Analogy Making

Eliasmith and Thagard (2001).

Models of Motor Control

Brown, Bullock, and Grossberg (2004); Bullock and Grossberg (1988, 1989); Grossberg and Kuperstein (1986/1989); Guenther (1994, 1995); Kawato (1995); Kawato, Furukawa, and Suzuki (1987); Kawato, Isobe, Maeda, and Suzuki (1988); Kawato and Samejima (2007).

Models of Episodic Memory

Franklin and Grossberg (2017); Greve, Donaldson, and van Rossum (2010); Meeter, Myers, & Gluck (2005); Metcalfe (1994); Newman, Gupta, Climer, Monaghan, & Hasselmo (2012); Shastri (2001, 2002); Rolls and Deco (2015); Subagdja and Tan (2015); Werbos (2012).

Models of Language Learning

Garagnani and Pulvermüller (2013); Harley (1996); Harm and Seidenberg (1999, 2004); Kajić, Gosmann, Stewart, Wennekers, and Eliasmith (2017); Penke and

Westermann (2006); Plaut, McClelland, Seidenberg, and Patterson (1996); Seidenberg and McClelland (1989); Thomas, Forrester, and Ronald (2013); Thomas, Purser, Tomlinson, and Mareschal (2012); Westermann and Ruh (2012).

Models of Perceptual Decision-Making

Bogacz and Gurney (2007); Bogacz (2009); Brown, Bullock, and Grossberg (2004); Gurney, Prescott, and Redgrave (2001a, 2001b); Usher and McClelland (2001); Usher and Niebur (1996).

Models of Consciousness

Grossberg (2017); Maia and Cleeremans (2005); Reggia (2013); Taylor (1997, 1999); Taylor and Mueller-Gaertner (1997).

Notes

1. A Gabor stimulus is a sine wave grating seen through a Gaussian window, a popular stimulus in experimental vision laboratories.
2. These two articles are a technical report and a dissertation, so not easily accessible. Yet, copies of both articles and the computer code for the model therein are available from Daniel Bullock at danb@bu.edu. Summaries of their principles can also be found in the published journal articles of Bullock (2004) and Rhodes et al. (2004).
3. The Google definition of "allosteric" is "relating to or denoting the alteration of the activity of a protein through the binding of an effector molecule at a specific site."
4. Neural networks that do not require training typically represent decision processes of mature organisms, which presumably have been "trained" over an organism's lifetime by experience. Levine (2016) sets out some tentative pathways, involving the hippocampus and cortex interacting in episodic memories, for training decision processes.

APPENDIX 1

MATHEMATICAL TECHNIQUES FOR NEURAL NETWORKS

Difference and Differential Equations

The equations for neural networks involve changes over time in two types of variables – node activities and connection strengths. The simplest way to describe such changes is to assume they take place at discrete time intervals – every second, say, or every 250 milliseconds. In that case, time is measured in whole number intervals. Hence the equations derive the value of a particular variable at time $t + 1$, with t being an integer (whole number), if the value of the same variable at time t is known. If the variable (activity or connection weight) is called x, we have the generic equation

$$x(t+1) = x(t) + \Delta x(t) \tag{A1.1}$$

where Δ is a symbol that means "amount of change." Thus, $\Delta x(t)$ represents the total of all changes in x within the given time period. Equation (A1.1) is called a *difference equation* because the term $\Delta x(t)$ represents the difference between a variable at time $t + 1$ and the same variable at time t.

In actual network models, the $\Delta x(t)$ term of (A1.1) is replaced by some algebraic expression involving $x(t)$ itself and other network variables (node activities or connection strengths). This expression reflects influences on the node or connection whose activity is x by excitation, inhibition, and modulation from the same node or connection, or from elsewhere in the network.

Example: The Sutton–Barto Difference Equations

One of the simpler examples of a neural network described by difference equations is the network of Sutton and Barto (1981). Figure 3.5 of Chapter 3

shows the major variables in that network: the conditioned stimulus traces x_i, the connection weights w_i, and the output y. The equations for that network, however, list some additional variables not shown in that figure: the eligibility traces $\bar{x}_i(t)$ and the representation $\bar{y}$ of ongoing reinforcement node activity. Figure A1.1 expands the earlier figure to include these additional variables.

There is no difference equation for the stimulus traces $x_i(t)$ themselves, which simply reflect what is taking place in the sensory environment. But the *i*th eligibility trace obeys a difference equation reflecting the influence from the corresponding (*i*th) stimulus trace. This is

$$\bar{x}_i(t+1) = \alpha \bar{x}_i(t) + x_i(t) \tag{3.22a}$$

where α is some number between 0 and 1. (Note: the labels "a," "b," etc., in parentheses on the right denote parts of a system of equations that are given the same number, in this case, (3.22).)

Let us analyze what (3.22a) says about the change in $\bar{x}_i$ over time. For definiteness, let us choose a specific value for α – say $\alpha = .8$. To find the difference $\Delta\bar{x}_i(t)$, we subtract $\bar{x}_i(t)$ from the expression for $\bar{x}_i(t+1)$. This yields

$$\begin{aligned} \Delta\bar{x}_i(t) &= \bar{x}_i(t+1) - \bar{x}_i(t) = \alpha\bar{x}_i(t) + x_i(t) - \bar{x}_i(t) \\ &= -(1-\alpha)\bar{x}_i(t) + x_i(t) = -.2\bar{x}_i(t) + x_i(t) \end{aligned} \tag{A1.2}$$

Equation (A1.2) says that $\bar{x}_i(t)$ is negatively influenced by its own decay back to a baseline, at a rate .2, and positively influenced by the actual stimulus $x_i(t)$. Some examples of runs with specific values are shown in Figure A1.2.

Now consider the equation for the ongoing reinforcement level $\bar{y}$. That equation is

$$\bar{y}(t+1) = \beta\bar{y}(t) + (1-\beta)y(t) \tag{3.22b}$$

where β is another constant between 0 and 1.

Again, let us put a specific value for β (say .6) into (3.22b) and analyze what that equation says about the change in $\bar{y}$ over time. Subtracting $\bar{y}_i(t)$ from the expression for $\bar{y}_i(t+1)$, we obtain

$$\begin{aligned} \Delta y(t) &= (.6\bar{y}(t) - (1-.6)y(t)) - \bar{y}(t)_i \\ &= (.6-1)\bar{y}(t) + (1-.6)y(t) \\ &= (1-.6)(y(t) - \bar{y}(t)) = .4(y(t) - \bar{y}(t)) \end{aligned} \tag{A1.3}$$

Equation (A1.3) says that the crucial influence on the dynamics of $\bar{y}$ is the *difference* between actual and ongoing (or expected) amount of reinforcement. The factor .4 represents a learning rate, namely the speed at which the expected value is updated.

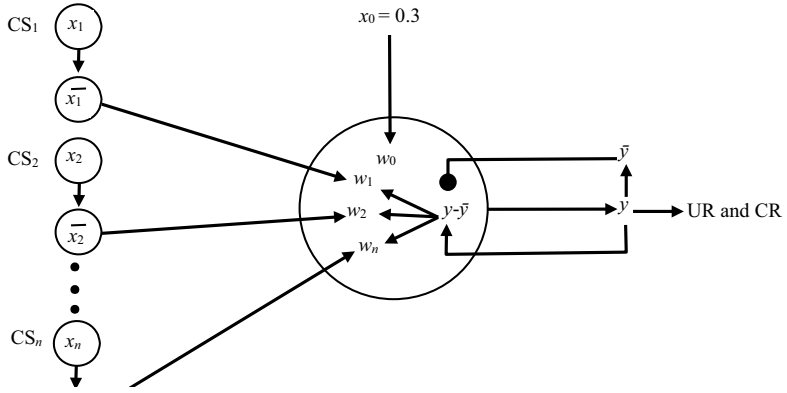
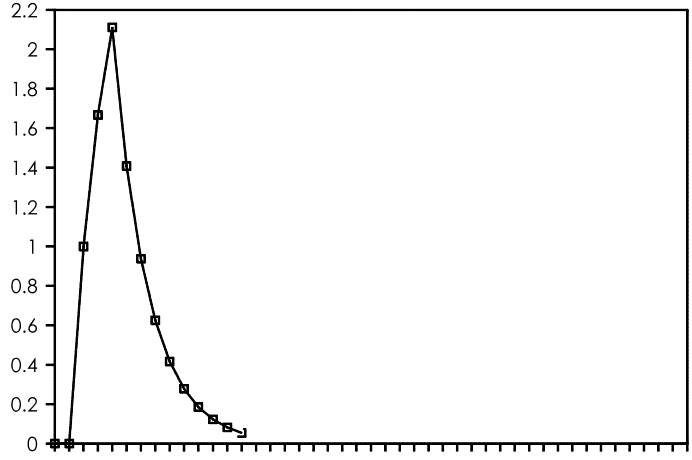


FIGURE A1.1 Extension of Figure 3.5 to include mutual influences among all the variables in Sutton and Barto's Equations (3.22). Additional variables are the eligibility traces $\bar{x}_i$ and the ongoing reinforcement level $\bar{y}$; see text for details.



How is the reinforcement level or output $y(t)$ calculated? Like the conditioned stimulus (CS) traces, this level is calculated instantaneously, based on the levels of all the CSs and of the unconditioned stimulus (US), and the connection weights from the CS representations to the output node.

For the ith CS, denote the stimulus level by $x_i(t)$ and the connection weight by $w_i(t)$. The corresponding stimulus level and weight for the US will be called

$x_0(t)$ and $w_0(t)$. Hence, the combined signal from all the CS nodes is $x_0(t)w_0(t)$ plus the sum of all the products $x_i(t)w_i(t)$, written in the summation or *sigma* notation:

$$\sum_i w_i(t)x_i(t) + w_0(t)x_0(t)$$

where the sum is taken over all i between 1 and n. If, for example, $n = 4$, then

$$\sum_i w_i(t)x_i(t) = w_1(t)x_1(t) + w_2(t)x_2(t) + w_3(t)x_3(t) + w_4(t)x_4(t)$$

This combined CS signal is transformed by a sigmoid activation function f (see Figure 2.7b from Chapter 2). All these terms combine in the equation

$$y(t) = f\left(\sum_i w_i(t)x_i(t) = w_0(t)x_0(t)\right) \tag{3.22d}$$

Finally, we discuss the changes in the n synaptic connection strengths $w_i(t)$. The changes in these variables, denoted $\Delta w_i(t)$, are influenced by a learning rule. Recall from Section 3.3 that this is an associative rule whereby presynaptic (CS) activity is correlated not with absolute postsynaptic (US) activity, but rather with *change* in postsynaptic activity. Hence

$$w_i(t+1) = w_i(t) + c(y(t) - \overline{y}(t))\overline{x}_i(t) \tag{3.22c}$$

where c is a positive constant that denotes the rate of learning.

Differential versus Difference Equations

The assumption made in the last section is that changes in the network take place at discrete time intervals (such as once every second). In biological neural systems, it is probably more realistic to assume that the interacting changes in the network take place continuously. Differential equations involve *derivatives*, or rates of change, of these variables, which are in turn approximations of the average Δf's for very small times, as will be explained below.

As an intuitive example of a derivative, one can look at what is actually measured by the speedometer of a car. As shown in Figure A1.3, speed is measured as distance covered divided by time elapsed. If f indicates the position on the road and t indicates the current time, then the speed of driving in any given time period is measured as change in f divided by change in t, or in the notation introduced above, as $\Delta f/\Delta t$. But over what length of time should speed be taken? At 12:00, you get different results if you measure the speed of travel since 11:50, or since 11:59, or since 11:59 and 50 seconds. It is for

precise description of measurements such as these that Newton and Leibniz (independently) developed the idea of derivative, one of the two basic ideas of calculus, in the late eighteenth century.

Suppose the measured speed of the car (see Figure A1.3) traveling for 10 minutes up to noon is, say, 22 mph, for 1 minute it is 21 mph, for ten seconds it is 20.5 mph, and for five seconds it is 20.2 mph. One can say that as the time gets smaller and smaller, that is closer to an instant (zero length) of time, the speed during that time gets "closer and closer" to 20 mph, which is called the limiting speed. ("Closer and closer" is an intuitive term, related to the more precise mathematical concept of *limit*, which is discussed in Edwards & Penney, 2007/2014; Swokowski, 1988; Thomas & Finney, 1988, or any other calculus textbook.)

For any function that varies with time – such as a moving vehicle's position, or a node activity or a connection weight in a neural network – the *derivative* or *rate of change* of f is defined as the value that the quantity $\Delta f/\Delta t$ gets "closer and closer" to. For complex reasons based on the sociology of mathematics, there are *three* equivalent notations for the derivative of f with respect to time: $df/dt, f'$, and $\dot{f}$. As shown in Figure A1.4, if the function is graphed with respect to time, and the curve is approximated near a given time by a straight line, then the derivative is indicated by the slope of that straight line, that is, how fast that line rises or falls as you move to the right.

Assuming sufficiently short time intervals, the same set of network interactions can be described by *either* a difference or a differential equation formulation. Take, for example, a single one of Sutton and Barto's equations, such as the equation for the eligibility trace $\overline{x}_i(t)$:

$$\begin{aligned}\Delta\overline{x}_i(t) &= \overline{x}_i(t+1) - \overline{x}_i(t) = \alpha\overline{x}_i(t) + x_i(t) - \overline{x}_i(t) \\ &= -(1-\alpha)\overline{x}_i(t) + x_i(t) = -.2\overline{x}_i(t) + x_i(t)\end{aligned} \tag{A1.2}$$

Since the time changes from t to $t + 1$, $\Delta t = 1$, so $\Delta\overline{x}_i/\Delta t$ is the same as $\Delta\overline{x}_i$. As Δt gets small, $\Delta\overline{x}_i/\Delta t$ gets closer and closer to $d\overline{x}_i/dt$, the derivative of $\overline{x}_i$. Hence, the differential equation form of (A1.3) is

$$d\overline{x}_1(t) = -(1-\alpha)\overline{x}_1(t) + x_1(t)$$

Similarly, given a system of differential equations for the interacting variables in a neural network, each differential equation can be approximated by a difference equation, assuming the time steps are "small enough" for a good approximation. This is the basis for the *Euler method* of numerically solving differential equations on a computer.

There are other widely used methods of greater accuracy than the Euler, such as the *fourth-order Runge–Kutta method*. Also, many widely used commercial mathematical software programs such as MATLAB and Mathematica include

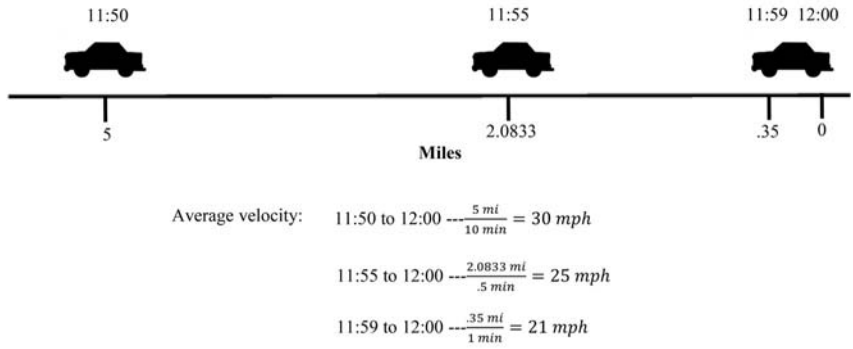


FIGURE A1.3 Schematic of positions at different times, just before 12:00 noon, of an automobile traveling eastward on a straight road, with a speedometer reading at noon of 20 mph.

more accurate ordinary differential equation (ODE) solvers. The Euler method, however, is serviceable for most network applications. Detailed descriptions of both the Euler and Runge–Kutta methods can be found in any introductory junior- or senior-level textbook on differential equations (e.g., Boyce & DiPrima, 2017; Braun, 2012; Rainville, Bedient, & Bedient, 2014) or on numerical analysis (e.g., Burden & Faires, 2005; Greenspan & Casulli, 1993). We give a capsule description of the Euler method in the next section, introducing it by example. The example we use is based on Grossberg's outstar equations, previously introduced in Section 3.2.

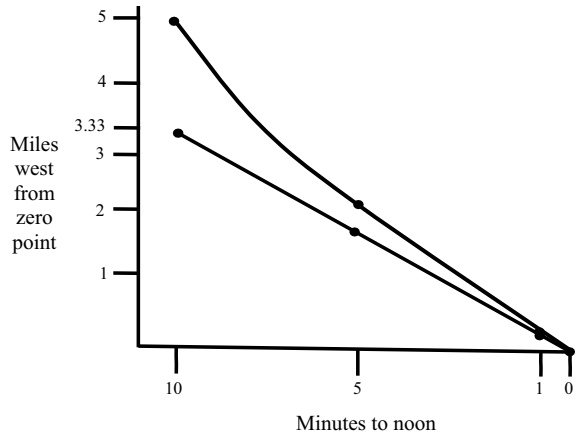


FIGURE A1.4 The curve in this graph represents the position of the car depicted in Figure A1.3, as a function of time. The line represents the linear approximation to that curve at the point (0,0). Its slope is (3.33 miles)/(10 minutes) = 1/3 miles/min = 20 miles/hour.

Outstar Equations: Network Interpretation and Numerical Implementation

The *outstar* (Grossberg, 1968a) is depicted in Figure 3.2. In an outstar, one node (or vertex, or cell population) v_1, called a *source*, projects to other nodes v_2, v_3, . . ., v_n, called *sinks*. (The three dots after v_3 are a generally accepted notation for an indeterminate number of numbers or variables that fit into a general form.)

As discussed in Section 3.2, the source node activity x_1 is affected positively by the source node input I_i, and negatively by exponential decay back to a baseline rate (interpreted as 0). Recalling that the rate of change of x_1 as a function of time is described by its derivative, dx_1/dt, this leads to a differential equation of the form

$$\frac{dx_1}{dt} = -ax_1 + I_1 \tag{3.12}$$

where a is a positive constant (the decay rate). The activities x_i of the v_i, $i = 2, \ldots, n$ obey an equation similar to (3.12), with the addition of an effect of the source node activity. Hence,

$$\frac{dx_1}{dt} = -ax_1(t) + bx_1(t-\tau)w_{1i}(t) + I_i(t) \tag{3.13}$$
$$i = 2, \ldots, n$$

where b is another positive constant (*coupling coefficient*) and τ is a transmission time delay.

The synaptic weights, or long-term memory traces, w_{1i} at the source-to-sink synapses, in one version of the theory, have a passive decay which is counteracted by correlated activities of x_1 (with a time delay) and x_i, thus

$$\frac{dx_{1i}}{dt} = -cw_{1i} + ex_1(t-\tau)x_1 \tag{3.14}$$

But if x_1 is interpreted as encoding the sound A and x_i as encoding the sound B, Equation (3.14) implies that the association between A and B is weakened while the network is not actively hearing A. Hence, Grossberg modified this equation to make the association decay when A is presented without being followed by B, but remain constant when A is not presented at all. This change can be achieved by replacing (3.15) (with the time delay τ set to 0) by

$$\frac{dx_{1i}}{dt} = x_1(-cw_{1i} + ex_1) \tag{3.15}$$

so that w_{1i} remains unchanged while $x_1 = 0$ but decreases while $x_1 > 0$ and $x_i = 0$.

We next go through a simple example of the outstar equations using the simple Euler method. In our example, there are only two sink nodes – x_2 and x_3, with corresponding synaptic weights (from the source node) w_2 and w_3. We also assume there is no decay of memory in the absence of source node stimulation, that is, we use Equations (3.12), (3.13) (for $i = 2$ and 3), and (3.15) (for $i = 2$ and 3) – five equations in all. As for the constants in those equations, set $a = 5$, $b = 1$, $c = .1$, $d = 1$, $\tau = 0$. So the specific forms of the equations become

$$\frac{dx_i}{dt} = -5x_1(t) + I_1(t) \tag{A1.4}$$

$$\frac{dx_2}{dt} = -5x_2(t) + x_1(t)w_{12}(t) + I_2(t)$$

$$\frac{dx_3}{dt} = -5x_3(t) + x_1(t)w_{13}(t) + I_3(t)$$

$$\frac{dx_{12}}{dt} = x_1(t)(-.1w_{12}(t) + x_2(t))$$

$$\frac{dx_{13}}{dt} = x_1(t)(-.1w_{13}(t) + x_3(t))$$

Now to solve Equations (A1.4) numerically, it only remains to set the inputs I_1, I_2, and I_3, and the starting values of x_1, x_2, x_3, w_{12}, and w_{13}. We set up an example in which the source node input arrives at regular intervals, followed by inputs to the sink nodes which remain in a regular proportion. Hence, in the terminology of Section 3.2, these sink node inputs form a *spatial pattern*. In particular, let $I_1 = 2$ on every tenth time step, starting with the first, and 0 on all other time steps. Let pattern $I_i = \theta_i I$, where $\theta_2 = .7$ and $\theta_3 = .3$ on time steps directly *after* those times when $I_1 = 2$, and 0 on other time steps (see Figure A1.5). (Hence $I_2 = 1.4$ and $I_3 = .6$ when they are not zero.) Suppose every time step is of length .1 (which is the largest value typically used in numerical examples; an actual step size of 1 leads to too much inaccuracy).

Consider, for example, Equation (A1.4a) for the source node activity. Suppose this node is inactive, that is has activity 0, at time 0. Then the simple Euler method says that change in x_1 over the time .1 divided by change in time (which is .1) will be represented by the right-hand side of (A1.7a), namely, it will equal $-5x_1 + I_1$, at the previous time. To calculate the new value of x_1, then, we obtain

$$x_1(.1) = x_1(0) + \Delta x_1(0) = x_1(0) + .1(-5x_1(0) + I_1(0)) \tag{A1.5}$$

(Equation (A1.5) is an approximation to the original differential equation (A1.4a), but the computer program treats it as an exact statement). Substituting

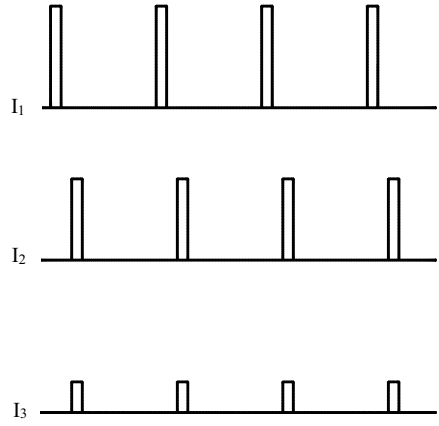


FIGURE A1.5 Examples of inputs to an outstar network with one source node x_1 and two sink nodes x_2 and x_3. The sink node inputs form a spatial pattern (i.e., remain proportional) and uniformly lag behind the source node inputs, both occurring at regular time intervals. This leads to learning of the association between x_1 activation and the given spatial pattern.

0 for every occurrence of $x_1(0)$ in (A1.5), and 2 for I_1 (since the input is on for that time step), we derive

$$x_1(.1) = 0 + .1(-5(0) + 2) = .2$$

The simple Euler method applies this same process repeatedly at later time steps. That is, (A1.5) generalizes to

$$x_1(t+1) = x_1(t) + .1(-5x_1(t) + I_1(t)) \tag{A1.6}$$

for any time t. At the next nine time steps, the input I_1 is 0. So we obtain

$$x_1(.2) = x_1(.1) + .1(-5x_1(.1)) = .2 + .1(-5).2 =$$
$$x_1(.3) = x_1(.2) + .1(-5x_1(.2)) = .1 + .1(-5).1 =$$

etc.

More generally, the rule for the Euler method is for any variable, call it y, and a time step of size Δt,

(y at time $(t + \Delta t)$ = (y at time Δt) × (expression on the right hand side of the differential equation for y) (A1.7)

The last parenthetical expression in (A1.7) is some function of all the network variables, with the values of those variables at time t substituted in. Figure A1.6 shows the values of the five variables in our outstar system over 30 time steps.

To illustrate how differential equations are transformed into difference equations, let us apply (A1.7) to the equations (A1.4). We list C and MATLAB versions of a program segment written to solve these equations numerically. The program must start with initial values for the variables, which can either be obtained from a random number generator over some range or read in arbitrarily. Since the interesting phenomenon in outstars is the convergence of relative x's and relative w's to θ's, the initial values of x_2 and x_3 should *not* be set proportional to θ_2 and θ_3, and the same for the initial values of w_{12} and w_{13}. For simplicity, we set them arbitrarily here. The program calculates these variables over 5000 time steps with step size .1.

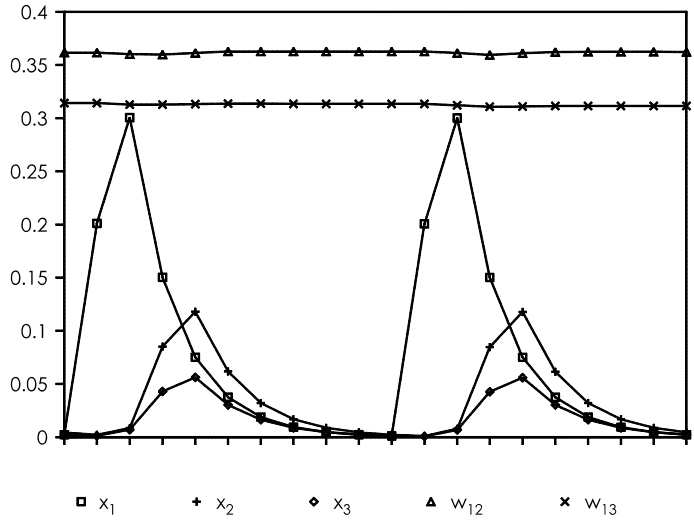


FIGURE A1.6 Graph of the outstar variables over time for the parameters shown in Equations (A1.4) and representative initial values.

C Version

```
     #include <stdio.h>

main ()
     {
          FILE *fp;
          char *arg;
```

```
/* declaration above needed to open a file */
          float x1=0.0, x2=0.2, x3=0.5, w12=0.2, w13=0.5;
          float i1, i2, i3;
          float x1old, x2old, x3old, w12old, w13old;
          int k, i;
          /* open file*/
          arg = "top";
          fp = fopen(arg, "w");
          /*         */
          fprint (fp, "                outstar dataset\n");
               k = 1;
               i = 1;
               while (i < 5000)
               {
                    fprintf(fp, " %8.6f %8.6f ", w12, w13);//
prints values of w12 & w13 with 8 digits to 6 decimal places
                    fprintf(fp, "    %8.6f %8.6f %8.6f \n", x1, x2, x3);
                    x1old = x1;
                    x2old = x2;
                    x3old = x3;
                    w12old = w12;
                    w13old = w13;
                    if (k == 1)
                        i1=2.0;
                        else
                        i1=0.0;
                    if (k == 2)
                        {
                        i2=1.4;
                        i3=0.6;
                        }
                        else
                        i2=i3=0.0;
                    x1=x1+0.1*(-5.0*x1old+i1);
                    x2=x2+0.1*(-5.0*x2old+x1old*w12old+i2);
                    x3=x3+0.1*(-5.0*x3old+x1old*w13old+i3);
                    w12=w12+0.1*x1old*(-.1*w12old+x2old);
                    w13=w13+0.1*x1old*(-.1*w13old+x3old);
                    if (k == 10)
                        k=1;
                        else
                        k++;
                    i++;
```

```
                    }
    }
return 0;
}
```

MATLAB Version

```
function sim_p_407
clear all; close all; clc;
x1 = 0;
x2   = 0.2;
x3   = 0.5;
w12 =0.2;
w13 =0.5;
k = 1;
i = 1;
while (i < 5000)

    x1old = x1;
    x2old = x2;
    x3old = x3;
    w12old = w12;
    w13old = w13;

if (k == 1)
        i1 = 2.0;
else
        i1=0.0;
end

if (k == 2)
        i2 = 1.4;
        i3 = 0.6;
else

        i2=0.0;
        i3=0.0;
end
    x1=x1+0.1*(-5.0*x1old+i1);
    x2=x2+0.1*(-5.0*x2old+x1old*w12old+i2);
    x3=x3+0.1*(-5.0*x3old+x1old*w13old+i3);
    w12=w12+0.1*x1old*(-.1*w12old+x2old);
    w13=w13+0.1*x1old*(-.1*w13old+x3old);
```

```
if (k == 10)
        k = 1;
else
    k=k+1;
    i=i+1;
end

    input(i,:)= [i1,i2,i3];
    states(i,:) = [x1, x2, x3, w12,w13];
end% while
l = 20;
subplot(211); stairs(1:l, input(1:l,:));
ylabel('Inputs'); legend ('i_1','i_2','i_3');
subplot(212); plot(1:l, states(1:l,:)); ylabel('States');
legend ('x_1','x_2','x_3','w_{12}','w_{13}' );

end
```

In the foregoing programs, the "old" values (values at the previous time step) of all the variables are preserved so that the variables can all be updated in succession. Within each programming language, various differential equation solving packages are available that do all the updating simultaneously.

Vectors and Matrices

The word "vector," derived from the Latin for "carrier," was originally used in physics to describe an entity with a direction, such as a velocity or a force. Directions in a plane can be described by an ordered pair of two numbers denoting the magnitude of velocity, force, or whatever in the two cardinal directions. Likewise, directions in space can be described by an ordered triple of three numbers. The notation for a vector can either be horizontal (a *row vector*) or vertical (a *column vector*). Section 3.2 uses the row notation, so that, for example, the vector made from the ordered triple 2, 4, 6 is written (2, 4, 6).

Hence, in mathematics, "vector" came to be abstracted to mean an ordered array of any number of real numbers. In neural networks, this array can mean, for example, the pattern of inputs to a specified collection of nodes, or the pattern of activations of those nodes, or the pattern of weights of connections to or from one of the nodes. This means that a vector is a convenient shorthand for a pattern of inputs, activities, or weights.

The numbers that constitute a vector are called the *components* of the vector. If the vector has two or three components, it can represented by an arrow: for example, if node 1 has activation 2 and node 2 has activation 1.5, the vector (2, 1.5) of node activities can be represented by the arrow shown in

Figure A1.7. If it has more than three components, it is still sometimes useful to visualize the vector either as an arrow or as a point in space whose coordinates are the vector's components.

Vectors can be treated as mathematical objects in themselves, and are represented in this book by boldface variable names. Mathematical operations on vectors are at the heart of programming in MATLAB, in which it is typically more efficient to do these operations on a vector of quantities than sequentially (via a DO loop) on the individual quantities themselves.

Two vectors can be added *if* they have the same number of components. In that case the components of the sum of the vectors are the sums of the corresponding components of the original two vectors. For example, the sum of the two vectors (2, 4, 6) and (3, –1, –2) is (2 + 3, 4 + (–1), 6 + (–2)) = (5, 3, 4). Also, any vector can be multiplied componentwise by any real number, which is called a *scalar* because it "scales" the magnitude of the vector (without changing its direction if it is a positive scalar, or reversing its direction if it is a negative scalar). For example, 2 × (2, 4, 6) = (4, 8, 12).

Dot Products

Vectors are usually not multiplied by each other in the same manner as numbers. Yet there is a scalar-valued product (again, of two vectors with the same number of components), called the *dot product* or *inner product*) that plays key roles in several neural network models. The dot product is mentioned in Section 3.2.2 of Chapter 3 and is used to describe transformations of inputs in linear systems (e.g., Anderson, 1970; Jordan, 1986a; Kohonen, 1977).[1]

If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are two vectors with the same number n of components, the dot product $\mathbf{x}\ \mathbf{y}$ is defined as the sum $\Sigma_{i=1}^{n} x_i y_i$. For example, the dot product (2, 4, 6) (3, –1, –2) = 2(3) + 4(–1) + 6(–2) = 6 – 4 – 12 = –10.

The dot product is a measure of similarity between two vectors. This is because, if the vectors are described geometrically as in Figure A1.7, the dot product can be used to calculate the angle between them. First, the inner product of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ with itself is $\Sigma_{i=1}^{n} x_i^2$. For example, $(2, 3.5) \cdot (2, 3.5) = 2^2 + (3.5)^2 = 16.25$. By the Pythagorean theorem from plane geometry, this is also the square of the length of the hypotenuse of the right triangle shown in Figure A1.7, and the hypotenuse is none other than the vector (2, 3.5). In general, the length of the vector $\mathbf{x}$ can be defined as the square root of $\mathbf{x} \cdot \mathbf{x}$. This square root is known as the *norm* of $\mathbf{x}$ and written $\| \mathbf{x} \|$. From plane geometry and trigonometry it can be shown (the demonstration is omitted here) that for any two vectors $\mathbf{x}$ and $\mathbf{y}$ the angle between $\mathbf{x}$ and $\mathbf{y}$ is the angle θ such that

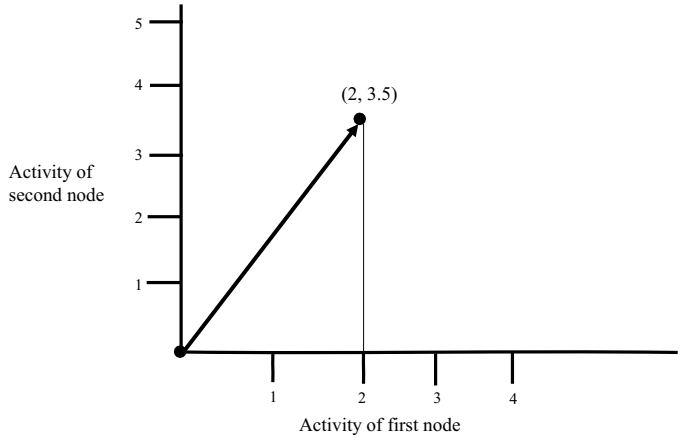


FIGURE A1.7 Vector representation of activities of two nodes

$$\cos\theta = \frac{x \cdot y}{||x||\,||y||} \tag{A1.8}$$

By analogy, (A1.8) is true for any two vectors with n components for any positive integer n.

Recall that the cosine of θ is 1 when θ is zero, zero when θ is 90°, and –1 when θ is 180°. Equation (A1.8) shows that if **x** and **y** point in approximately the same direction (a good "match"), the dot product will be relatively large and positive. If **x** and **y** point in opposite directions, the dot product is large and negative. If **x** and **y** are perpendicular their dot product is 0. The angle cosine criterion for pattern matching is used in many neural models include some adaptive resonance models (Carpenter & Grossberg, 1987b) and some PDP models.

Matrices

Just as a vector is a one-dimensional array of numbers, a matrix is a two-dimensional array of numbers. An example is the matrix of connection weights between all the nodes of a fully connected network, or from all the nodes in one layer of a network to all the nodes in another layer to which it connects.

If the array has m rows and n columns, it is called an $m \times n$ (or *m-by-n*) *matrix*. Matrices are usually denoted by bold capital letters, and the numbers in the matrix are listed inside square brackets. For example,

$$\begin{bmatrix} 2 & 4 & 6 \\ 3 & -1 & -2 \end{bmatrix}$$ is a 2 × 3 matrix, and

$$\begin{bmatrix} 5 & 2 \\ 1 & -4 \end{bmatrix}$$ is a 2 × 2 matrix.

Matrices with the same number of rows as columns are called *square matrices* and turn out to have special significance. The numbers inside the brackets are called the *entries* of the matrix.

Addition and scalar multiplication are defined for matrices in the same way as they are for vectors. An $m \times n$ matrix can be multiplied by a vector with n components, by treating the vector as a column and then taking the dot product of each row of the matrix with the vector. The result is another vector with m components. For example,

$$\text{if } \mathbf{A} = \begin{bmatrix} 2 & 4 & 6 \\ 3 & -1 & -2 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

then the first row of $\mathbf{Av}$ is $(2, 4, 6) \cdot (1, 2, 3) = 2(1) + 4(2) + 6(3) = 2 + 8 + 18 = 28$. The second row of $\mathbf{Av}$ is $(3, –1, –2) \cdot (1, 2, 3) = 3(1) + (–1)(2) + –2)(3) = 3 – 2 – 6 = –5$. Hence, the product $\mathbf{Av}$ is the two-component vector

$$\begin{bmatrix} 28 \\ -5 \end{bmatrix}$$

Multiplying a system vector on the left by a matrix of connection weights therefore generates another system vector with either the same or different numbers of components: this is the basis of either autoassociation or hetero-association in linear systems. Autoassociation can occur if $\mathbf{A}$ is a square matrix, so the number of components will be unchanged after multiplication by $\mathbf{A}$; a good example of this process is the brain-state-in-a-box (BSB) model of Anderson et al. (1977; see Exercise 2 of Chapter 8). In networks such as BSB, it is of particular interest to find vectors whose direction remains unchanged after multiplication by $\mathbf{A}$, that is, find vectors $\mathbf{x}$ such that $\mathbf{Ax} = \lambda\mathbf{x}$ for some real (preferably positive) constant λ. If that equation is satisfied, λ is called an *eigenvalue* of $\mathbf{A}$ and $\mathbf{x}$ an *eigenvector* corresponding to λ. In BSB and many other systems, in neural networks and other applications, the eigenvectors represent steady state of large-time tendencies of the system, with the eigenvectors corresponding to the largest positive eigenvalues being the most significant.

Recall that an $m \times n$ matrix can be multiplied by an n-component column vector, which can be regarded as an $n \times 1$ matrix. By stacking the columns together and repeating the dot product–based multiplication for each column, the same $m \times n$ matrix can be multiplied by an $n \times p$ matrix for any positive integer p, with the result being an $m \times p$ matrix. For example,

$$\text{if } \mathbf{A} = \begin{bmatrix} 2 & 4 & 6 \\ 3 & -1 & -2 \end{bmatrix} \text{ (which is } 2 \times 3\text{) and}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 5 \\ 2 & 1 \\ 3 & 0 \end{bmatrix} \qquad \text{(which is } 3 \times 2\text{),}$$

the first column of the product **AB** is the same as **Av** in the previous example, namely

$$\begin{bmatrix} 28 \\ -5 \end{bmatrix}$$

The second column of **AB** has in its first row the dot product $(2, 4, 6) \cdot (5, 1, 0) = 2(5) + 4(1) + 6(0) = 14$, and in its second row the dot product $(3, -1, -2) \cdot (5, 1, 0) = 3(5) + (-1)(1) + (-2)(0) = 14$. Hence the complete matrix product **AB** is the 2×2 matrix

$$\begin{bmatrix} 28 & 14 \\ -5 & 14 \end{bmatrix}$$

If **A** and **B** are both $n \times n$ square matrices, both products **AB** and **BA** are defined and are also n × n, but in general **AB** and **BA** are not equal. For example,

$$\text{if } \mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 3 & 4 \\ 2 & 5 \end{bmatrix},$$

$$\text{then } \mathbf{AB} = \begin{bmatrix} 8 & 13 \\ 5 & 9 \end{bmatrix} \text{ but } \mathbf{BA} = \begin{bmatrix} 10 & 7 \\ 9 & 7 \end{bmatrix}.$$

That is, multiplication of matrices is not *commutative* in the way that multiplication of real numbers is.

Yet there are two special cases in which matrix multiplication is commutative. For square matrices the *identity matrix* of order n is defined to be the matrix **I** whose entries 1 along the main diagonal and 0 elsewhere; for example, the identity matrix of order 3 is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For the identity matrix **I** of order n, it is easy to show that $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ for any $n \times n$ matrix **A**. Also, for the majority of square matrices **A** (the conditions under which this holds are beyond the scope of this review), there is another matrix $\mathbf{A}^{-1}$, called the *inverse* of **A**, such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

The mathematics of vectors and matrices is called *linear algebra*. The chapter by Jordan (1986a) goes into more details of linear algebra than are covered here, including concepts such as linear independence and bases. Jordan (1986a) also includes applications of linear algebra to PDP networks and approximation of nonlinear systems by linear systems.

The remaining subsections of this appendix are not necessary for the student to perform the simulation exercises in the book, but aid the student in following some of the mathematical discussions elsewhere in the text. The section on the chain rule for derivatives is provided as background for the derivation ofthe back propagation algorithm in Section 3.5. The section on dynamical systems is provided as background for the discussions of equilibrium states and energy functions at various points in Sections 4.2, and 4.5. The section on integrate-and-fire methods is provided as background for many recent biologically realistic neural models, including some found in Chapters 6 and 9 of this book.

The Chain Rule and Back Propagation

The chain rule determines the derivative of a composite function, that is, a function whose argument is itself a function of another variable. Some examples of composite functions in biological applications are given in Gentry (1978, pp. 250–253). In one of his examples, a nerve impulse is translated into a muscular movement. The muscle reaction is a function of the number of acetylcholine ions liberated at neuromuscular junctions by the nerve impulse, and the number of ions liberated is itself a function of the number of millivolts in the impulse.

If f is a function of some variable y, and y is in turn a function of another variable x, then the derivative of *f as a function of x*, written df/dx, is the product of the derivative of f as a function of y with the derivative of y as a function of x. That is

$$\frac{df}{dx} = \left(\frac{df}{dy}\right)\left(\frac{dy}{dx}\right) \tag{A1.9}$$

If f, instead of being a function of a single variable y, is a function of several variables called $y_1, y_2, \ldots, y_n$, each a function of x, the rule (A1.8) generalizes to one involving the *partial derivatives* of f. The partial derivative $\partial f/\partial y_i$, for each i, is defined as the rate of change of f as y_i is varied, with all the other variables kept constant. Then the derivative of the composite function f becomes the sum of contributions from the variables $y_1, y_2, \ldots, y_n$, thus:

$$\frac{df}{dx} = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial y_i} \right) \left(\frac{dy_i}{dx} \right) \tag{A1.10}$$

Both (A1.9) and (A1.10) are frequently used to obtain derivatives of complex expressions in neural network equations. For example, they are used in the derivation, seen at the end of Chapter 3, of the changes of weights to hidden units in the three-layer backpropagation network, given the changes of weights to output units. A detailed justification follows now for some of the steps in the earlier derivation.

First, recall that the jth output unit (on the pth pattern) receives a signal equal to the linear sum of the outputs y_{pi} from the hidden layer weighted by the connections w_{ij}. This signal is called

$$\text{net}_{pj} = \sum_i w_{ij} y_{pi} \tag{3.9}$$

If f is the activation function of unit j, then the output of unit j is

$$y_{pj} = f(\text{net}_{pj}) = f\left(\sum_j w_{ij} y_{pi} \right) \tag{3.24}$$

Recall, also, that the total error in the pth output pattern is measured in terms of deviation of the output pattern vector from a target pattern vector ($t_{p1}, t_{p2}, \ldots, t_{pn}$). Since deviation from the target could be either positive or negative, the differences are squared, leading to a total error signal

$$E_p = \frac{1}{2} \sum_j \left(t_{pj} - y_{pj} \right)^2 \tag{3.25}$$

Then the response change δ_{pj} should be based on how much the jth unit contributes to the incorrectness of the response. This is done by taking the negative derivative of the total error E_p as a function of y_{pj}.

The t_{pj} in (3.25) are constants, and for a given output unit j the only part of Expression (3.24) that changes with the output signal y_{pj} is the part corresponding to that j, namely, $1/2(t_{pj} - y_{pj})^2$. From standard formulas for derivatives

of polynomial functions (e.g., Swokowski, 1988), we obtain that the derivative of E_p with respect to the output signal y_{pj} is

$$\frac{\partial E_p}{\partial y_{pj}} = -y_{pj}$$

But, for calculating the changes in weights to hidden units, it is necessary to get the derivative of the error not as a function of y_{pj}, but of the signal net_{pj} from the hidden layer. Using the chain rule (A1.16), that derivative is the product

$$\frac{\partial E_p}{\partial \text{net}_{pj}} = \frac{\partial E_p}{\partial y_{pj}} \frac{dy_{pj}}{d\text{net}_{pj}}$$

By the last equation, this translates to

$$\frac{\partial E_p}{\partial \text{net}_{pj}} = -y_{pj} \left(\frac{dy_{pj}}{d\text{net}_{pj}} \right)$$

But by (3.24), the output signal is the function f (usually sigmoid) applied to net_{pj}, so that $dy_{pj}/d(\text{net}_{pj}) = f'(\text{net}_{pj})$, where $'$ denotes derivative. If this change in the error with respect to the net signal (which determines a weight change) is called δ_{pj}, then

$$\delta_{pj} = f'(\text{net}_{pj})(t_{pj} - y_{pj}) \tag{3.10a}$$

If the jth unit is instead a hidden unit, then again using the chain rule, we obtain from (3.24), (3.25), and (3.10a) ("δ" denoting partial derivative) that

$$\delta_{pj} = -\frac{\partial E_p}{\partial \text{net}_{pj}} = -f'(\text{net}_{pj}) \left[\frac{\partial E_p}{\partial y_{pj}} \right]$$

If k is the generic index of output units that receive projections from hidden unit j, we obtain, again by the chain rule and previous equations, a value for the previous expression in brackets, namely

$$\frac{\partial E_p}{\partial y_{pj}} = \sum_k \left[\frac{\partial E_p}{\partial \text{net}_{pk}} \right] \left[\frac{d\text{net}_{pk}}{dy_{pj}} \right] = \sum_k \left[\frac{\partial E_p}{\partial \text{net}_{pk}} \right] w_{jk} = -\sum_k \delta_{pk} w_{jk}$$

Combining the above two expressions, we obtain finally that, if unit j is a hidden unit,

$$\delta_{pj} = f_j'(\text{net}_{pj}) \sum_k \delta_{pk} w_{jk} \tag{3.10b}$$

Dynamical Systems: Steady States, Limit Cycles, and Chaos

A *dynamical system* is defined as the movement through time of solution trajectories for a system of differential or difference equations for interacting variables (see, e.g., Hirsch & Smale, 1974, for more details). Each trajectory is described by a vector composed of the values of all the variables in the system at any given time. If n is the number of variables, these vectors can be treated as points in an n-dimensional space. Most of the discussion in this section is about dynamical systems based on differential equations; for difference equations, the mathematics is more difficult, and the system is more likely to exhibit chaotic behavior (see Frauenthal, 1980, Chapter 6, or Smital, 1988, Chapter 3, for one of the classic examples).

Of course, if n is larger than 3, an n-dimensional space is an abstract mathematical object that cannot be drawn. But in many cases (e.g., Anderson et al., 1977; Cohen & Grossberg, 1983; Hopfield, 1982), the network being studied is homogeneous in its structure, so that the number of nodes has little effect on qualitative behavior. For such systems, taking the number n of nodes to be two or three allows one to draw pictures of the dynamics of the network over time (see Figures 4.9 and 4.10). Such qualitative studies of time dynamics can also be useful for networks that are not homogeneous but have homogeneous subnetworks whose activities are described by a time-varying vector. One example is the vector of weights from any given category node to the n feature nodes in an adaptive resonance network (Carpenter & Grossberg, 1987a). Another example is the vector of input-to-hidden-unit weights in a back propagation network (Rumelhart et al., 1986).

A system of differential equations defining a dynamical system usually cannot be solved in closed form, that is, with the solutions expressed as combinations of elementary functions like exponentials, logarithms, polynomials, and trigonometric functions. But, frequently, numerical simulations can be supplemented by theorems about the system's *asymptotic behavior*, that is, what the vector of system variables approaches as time gets large. For most neural networks (as for systems derived from other physical and biological applications), the activities and connection strengths have upper and lower bounds. Hence, the time-varying vector of system variables remains within some "box" or hyper-rectangle in n-dimensional space. The asymptotic behavior of such bounded systems usually falls into one of three categories:

1. *Equilibrium*. The system vector approaches a single point in n-dimensional space (see Figure 4.9). Such a point is called an *equilibrium state* (or *steady state*, or *rest point*, or *critical point*) of the system. Many dynamical systems have only a finite number of possible equilibrium states. In neural networks, each steady state corresponds to a possible stable activity pattern of the network (see Sections 4.2 and 7.1).

2. *Limit cycle*. The system vector approaches a periodic orbit in *n*-dimensional space (see Figure 4.6). In neural networks, periodic orbits are sometimes used to model cyclical processes in the nervous system. Examples include models of reverberating memory in the cortex and thalamus (Wilson & Cowan, 1973); of hallucinations in visual perception (Ermentrout & Cowan, 1980); of circadian rhythm generation in the hypothalamus (Carpenter & Grossberg, 1985); and of rhythmical movements in crustaceans (Selverston, 1976).
3. *Chaos*. In bounded two-dimensional dynamical systems, the Poincaré–Bendixson Theorem (Hirsch & Smale, 1974) shows that convergence to an equilibrium point or to a limit cycle are the only possibilities. The proof of that theorem relies on some facts of two-dimensional geometry (e.g., a curve in the plane has a distinct "inside" and "outside," a result known as the Jordan curve theorem) and is no longer valid in three or more dimensions. In three or more dimensions, the system vector can asymptotically wander through *n*-dimensional space in a fashion that appears to be random but is actually deterministic (e.g., Lorenz, 1963). This phenomenon is widely known as *chaos* and has been suggested as a basis for behavioral variability in nervous systems (e.g., Mpitsos, Burton, Creech, & Soinila, 1988; Skarda & Freeman, 1987).

Some information about qualitative behavior of a dynamical system can be obtained from studying the functions defining the equations. In general, if a system involves n interacting variables $x_1(t), x_2(t), \ldots, x_n(t)$, the rate of change of each of the variables $x_i(t)$ is some function of $x_i(t)$ itself and all the other variables. Most systems defining neural networks are *autonomous*, that is, the functions do not depend on time. Hence, if we call the function f_i, we have a system of differential equations of the form

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \ldots, x_n) \tag{A1.11}$$

For a neural network, the function f_i in (A1.11) denotes the combination of all the excitatory and inhibitory influences on x_i if x_i is a node activity, and of positive and negative influences on x_i if x_i is a connection weight. An equilibrium state is a state, or value of the vector $(x_1, x_2, \ldots, x_n)$, at which all these relative influences are "balanced," that is, $f_i = 0$ for all i, $i = 1, 2, \ldots, n$.

Techniques for studying the qualitative behavior of a system of equations of the form (A1.11) all involve consideration of the functions f_i and their derivatives. Whether f_i is positive or negative at a given point in n-dimensional space determines the direction of change of x_i if the system state reaches that point.

In particular, if $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is an equilibrium point, it is of interest whether solutions of the equations (*trajectories*) starting at points near $\mathbf{x}$

approach **x** or move away from **x** as time gets large. In the former case, **x** is called an *asymptotically stable* equilibrium; in the latter case, **x** is *unstable*. (There is also an intermediate case, an equilibrium that is *stable* but not asymptotically stable. In that case, nearby trajectories stay in the vicinity of **x** without actually approaching **x**.) The stable equilibria are the ones that can actually be reached by the system, and are therefore the ones of interest for applications.

The criteria for stability are discussed in any advanced differential equations textbooks (e.g., Hirsch & Smale, 1974; Miller & Michel, 1982). One of these criteria involves the matrix of partial derivatives of the functions f_i, which is called the *Jacobian matrix* of the system of equations, at **x**. Recall from the last section that the eigenvectors of a matrix **A** are n-dimensional vectors $\mathbf{y}_i$ such that $\mathbf{A}\mathbf{y}_i$ is a constant multiple of $\mathbf{y}_i$. The constant which is multiplied is called an eigenvalue of the matrix **A** (Jordan, 1986a). The equilibrium point **x** is asymptotically stable if all the eigenvalues of the Jacobian matrix at that point (which may be real or complex) have negative real parts, and unstable if any of the eigenvalues have positive real parts. Hence, the eigenvalues indicate the direction of flow of solution trajectories close to an equilibrium.

If no eigenvalues have positive real parts, but some of them are 0 or purely imaginary, the direction of this flow is ambiguous. Hence, under those conditions, one must resort to other methods for determining stability. One of the most important of these is the method of *Lyapunov functions* (see Section 4.2). Recall that a Lyapunov function (sometimes spelled Liapunov or Liapounov) is defined as a function of the system variables that is decreasing along system trajectories. More precisely, let $V(x_1, x_2, \ldots, x_n)$ be any real-valued function of the state vector **x**. Then if $x_1, x_2, \ldots, x_n$ satisfy the system of differential equations (A1.11), the chain rule (see the last subsection) shows that the derivative of V along solutions of the system is

$$\frac{dV}{dt} = \sum_{i=1}^{n} \left(\frac{\partial V}{\partial x_i} \right) \left(\frac{dx_i}{dt} \right) \tag{A1.12}$$

The expression on the right-hand side of (A1.12) is a function of the vector **x**. If this expression is always nonpositive over the range of state vectors reachable by the system, this means that the function V is a Lyapunov function, always nonincreasing along trajectories. Under those conditions, a variety of theorems constrains the motion of trajectories to approach equilibria.

Integrate-and-Fire (AKA Leaky Integrator) Models

Integrate-and-fire models (the simplest version also known as *leaky integrator* or *leaky integrate-and-fire*) are designed to model an individual neuron that follows electrical circuit laws including effects of capacitance and resistance until it reaches a threshold potential, and then it is reset to produce spikes. This

is a model of a neuron's dynamics and can be combined with a wide variety of models for interactions between neurons.

There are many variations on this integrate-and-fire theme (see Gerstner & Kistler, 2002, or Dayan & Abbott, 2005). So this book just outlines one of the simpler linear forms of the model. The current that drives the membrane voltage is divided into two components related to membrane resistance and capacitance. Let the current $I(t)$ be split into

$$I(t) = I_R + I_C \tag{A1.13}$$

The resistive current I_R can be calculated by Ohm's law as voltage across the resistor divided by resistance, that is, $I_R = u_R/R$. In turn u_R equals the voltage u across the membrane minus the resting voltage u_{rest}. The membrane capacitance C equals q/u, where q is the charge. Since current is the derivative of charge over time, $IC = dq/dt = C\,du/dt$. Combining all these terms into Equation (A1.13) yields

$$I(t) = \frac{u(t) - u_{rest}}{R} + C\frac{du}{dt} \tag{A1.14}$$

Let $\tau_m = RC$. Then multiplying (A1.14) by R yields the standard equation

$$\tau_m \frac{du}{dt} = -\left(u(t) - u_{rest}\right) + RI(t) \tag{A1.15}$$

$u(t)$ is called the *membrane potential* and τ_m the *membrane time constant* of the neuron.

Equation (A1.15) is the equation for a *passive membrane*, that is, a membrane that never has action potentials (spikes). Its solution is an exponentially decaying response to an external input pulse. Yet to make it an integrate-and-*fire* model, spiking needs to be added. This is done in the simplest manner possible: spikes are assumed to occur at times when the value of the membrane potential $u(t)$ reaches a threshold θ. If that firing time is called $t^{(f)}$, then the neuron is assumed to spike between $t^{(f)}$ and $t^{(f)}+\delta$ for some small value δ, with the form of the spike not described explicitly. After that interval δ, the membrane potential is reset to its resting value u_{rest}; hence,

$$u(t^{(f)} + \delta) = u_{rest} \tag{A1.16}$$

The last two equations, (A1.15) and (A1.16), define the generic integrate-and-fire model.

Note

1 Recall from Section 3.2 that a linear system means one where the changes in node activities are directly proportional to the influences from other nodes or outside inputs.

APPENDIX 2

BASIC FACTS OF NEUROBIOLOGY

The neural network modeler should at least have a working knowledge of neurobiology on two levels. One is the level of neurons (brain or nerve cells), including their component parts – axons, dendrites, cell bodies or somata, and synapses – and the chemical transmitters at synapses. This includes knowledge of how conduction of nerve impulses is affected by the actions of the various ions in and around the cells. The other is the level of brain regions, their cognitive functions and the pathways between them. This should, at best, include some knowledge of how the nervous system has evolved from invertebrates to fish to other mammals to humans, both structurally and functionally.

This appendix gives an extremely cursory summary of those biological facts that are, in my opinion, essential for the modeler to know. I refer the reader to other books whose coverage of these areas is far more detailed. Good general textbooks on all aspects of neuroscience include Shepherd (1983/1994) and Kandel, Schwartz, and Jessell (2000). Shepherd's book is particularly strong on the cognitive aspects of neural structures. Carlson (2007) and Kandel, Schwartz, and Jessell (2000) are good sources for the physiology relevant to behavioral and cognitive functions. Katz (1966) gives detailed, and still timely, descriptions of fundamental electrical and chemical processes at the neuronal level. A more recent and succinct treatment of these neuronal processes is found in Byrne and Schultz (1994). There are also many good textbooks on neuroanatomy of different brain regions, including Nauta and Feirtag (1986) and Waxman (2000).

The Neuron

Whereas the functional organization of the nervous system differs profoundly between squid, fish, rats, and humans, the organization of individual neurons differs much less. Hence, some classic studies on invertebrates have contributed greatly to our knowledge of general, including mammalian, neurophysiology. Particularly important is the work of Young (1936), Cole and Hodgkin (1939) and others on the giant axon of the squid, which activates an escape reflex.

Figure A2.1 shows a schematic neuron. The main parts of it are the dendrites ("small branches"), which often receive electrical signals from other cells; the soma, or cell body, which sums electrical potentials from many dendrites and also contains the cell's nucleus; and the axon, which conducts electrical signals and transmits them to other cells. The picture of Figure A2.1 is not universally accurate. Sometimes the axons are much shorter, relative to the other cell components, than the one shown there; short axons are particularly common in association areas of the human cerebral cortex. Also, the dendrites can sometimes be "senders" as well as "receivers." Still, this figure illustrates a "generic neuron" fairly well, and most of the exposition in this section assumes a neuron of this type.

The significant variable for information transmission in a neuron is the electrical potential across the membrane of the cell's axon. This potential is determined by the intracellular and extracellular concentrations of three single-element ions, potassium (K^+), sodium (Na^+), and chloride (Cl^-), along with some compound ions. There are two distinct phases of transmembrane electrical activity, the *resting membrane potential* while the cell is not conducting an electrical impulse, and the *action potential* or actual "nerve impulse." The

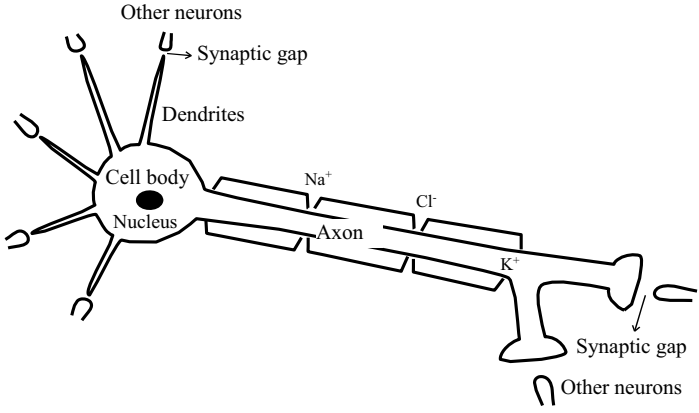


FIGURE A2.1 Schematic neuron. Main parts (axon, cell body or soma, dendrites, synapses) are labeled. Characteristic ions are shown where they are most prevalent, inside or outside the membrane. See text for details.

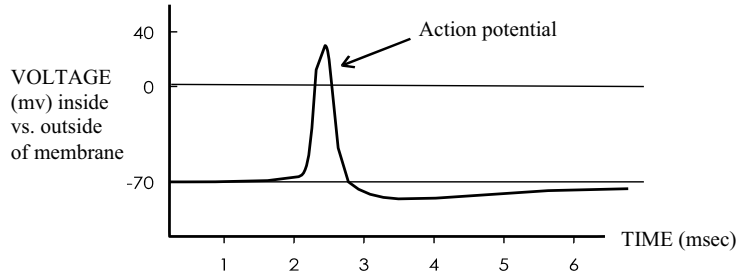


FIGURE A2.2 The action potential recorded across the membrane of a squid giant axon.
Source: Thompson, 1967, modified from Hodgkin & Huxley, 1939; reprinted by permission.

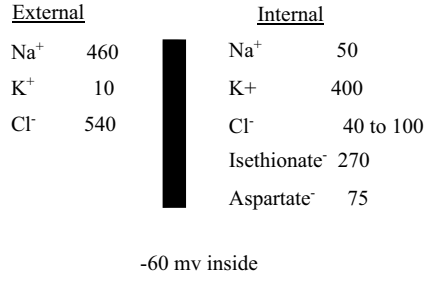


FIGURE A2.3 Ion concentrations (millimoles per liter) and potential difference across the membrane of the squid giant axon during the resting potential phase.
Source: Reprinted from Bernard Katz, *Nerve, Muscle, and Synapse*, copyright 1966, with permission of McGraw-Hill Publishing Company.

action potential, which is propagated down the axon, amounts to a reversal of electrical polarity from the resting phase; the inside is about 60 to 70 millivolts (mV) negative to the outside during the resting potential, but about 40 mV positive during the action potential. Figure A2.2 shows a typical action potential, also called a *spike* because of its shape.

Figure A2.3 lists the intracellular and extracellular concentrations of various ions in the squid axon at the time of the resting potential. Since the inside is electrically negative relative to the outside, it is somewhat surprising that the positive Na^+ (sodium) ion should be more concentrated on the outside, rather than rushing in to neutralize the polarity. This occurs because an active metabolic "pump" keeps most of the sodium outside in the resting state. The enzyme responsible for this pump was identified by Jens Christian Skou, which won him the 1997 Nobel Prize in Chemistry (see, e.g., Skou, 1998). The processes that initiate the action potential temporarily shut off this pump, causing a reversal of the membrane potential.

The cautionary note must be added, however, that impulses (action potentials) are *not* the only means of communication between neurons. There

is also communication by the simple spread of electrical potentials across neurons via synapses, which is called *passive electrotonic spread*. This passive conduction has been found recently to play an important role, particularly in short-distance communication (see Shepherd, 1983, p. 102 for discussion). The potentials recorded by the electroencephalogram (EEG) result from this passive spread.

The conduction of the nerve impulse, as Helmholtz and others in the last century discovered, is too slow to be merely electrical transmission as through a wire. Hence, this conduction must involve an active biochemical process. How the neuron changes from the resting to the excited state was essentially discovered in a series of experiments described, and quantitatively analyzed, by Hodgkin and Huxley (1952).

Early research on the squid giant axon showed that the action potential still occurs even if all the *axoplasm* (protoplasm on the inside of the axon) is squeezed out. It was concluded that the action potential is a *membrane* phenomenon. This research also showed that the action potential depends strongly on the presence of sodium ions in the extracellular medium. It was concluded that the potential change results from inward movement of sodium ions. The resting membrane is much less permeable to sodium ions than to potassium or chloride ions, but the action potential generation (excitation) process increases its permeability to sodium, allowing that inward movement of ions to take place.

The process of action potential generation is partially described by Shepherd (1983):

> a crucial property of the Na^+ conductance . . . is that it is involved in a positive feedback relation with the membrane depolarization. When the membrane begins to be depolarized, it causes the Na^+ conductance to begin to increase, which depolarizes the membrane further, which increases Na^+ conductance, and so on.
>
> (p. 107)

The *depolarization* Shepherd refers to denotes change in the inward membrane potential in the positive direction. If the cell membrane is depolarized from its resting state, either by an impulse from another neuron or by direct stimulating current, the cell will revert to its resting potential unless it reaches a *threshold* transmembrane voltage – typically in the neighborhood of – 40 mv inside in the case of the squid axon. If the cell potential does reach that threshold, the aforementioned positive feedback will take place, ultimately leading to an action potential. This is the biological basis for all-or-none impulses (see Section 2.1).

An important consequence of the sodium-permeability mechanism is that the frequency of impulse generation is limited. For one or a few milliseconds

after an impulse, no additional impulses can be generated; this is called the *absolute refractory period* (see the discussion of continuous models in Chapter 2). For several more milliseconds afterwards, there is a *relative refractory period* in which an action potential can be initiated, but the strength of current required to initiate an impulse is larger than normal.

The refractory periods result from the same ionic mechanism that terminates the impulse, namely, following of sodium conductance increase by an increase in the membrane's conductance of potassium (K^+). This leads to a movement of potassium ions outward, which reduces transmembrane potential back toward the resting level and thereby reduces sodium conductance. In addition, during the absolute refractory period, this outward flow of K^+ makes the membrane potential even more negative than normal, which further decreases the sodium conductance. The cell does not fully recover its ability to generate impulses until both ionic conductances are back to their resting levels.

Many, but not all, axons, particularly the longer ones, are covered by an electrically insulating layer known as the *myelin sheath*. This sheath is made of cells of a different type than neurons, known as *glial cells*. The action potential spreads like a wave down the axon, in a single direction (from dendrites toward outgoing synapses). In the case of myelinated axons, the conduction is along the outer membrane, "jumping" between holes in the myelin sheath known as the *nodes of Ranvier*, a process called *saltatory conduction*.

Thus far we have talked about movement of ions and conduction of electrical activity within a single neuron. As the impulse moves toward a synapse between two neurons, different processes take over.

Synapses, Transmitters, Messengers, and Modulators

The current view of nervous system organization did not become widely accepted until early in the twentieth century, with the work of Cajal and Sherrington (see Cajal, 1990, and Sherrington, 1906/1947, for summaries). Before the work of those two pioneers, there were disputes between adherents of two doctrines, the "neuron doctrine," which held that the nervous system is composed of distinct cells, and the "reticular doctrine," which held that all the fibrous processes are continuous with each other. The "neuron doctrine" won with the discovery that many pairs of cells that are functionally interconnected are actually physically separated. This separation is known as the *synaptic gap*; its width is of the order of one to a few μ ($1\ \mu = 10^{-6}$ meters).

The number of different types of junctions between cells is quite large (see, e.g., Shepherd, 1983, pp. 73–75). There are varying distances between cells, and there are both electrical synapses (where the action potential travels between cells by direct electrical conduction) and chemical synapses (where

the conduction is mediated by a chemical transmitter that affects ionic conductances). The most characteristic junction type, particularly in mammalian brains, is the chemical synapse.

The chemical synapse, unlike some of the other kinds of junctions, is unidirectional. The transmitting neuron is called the *presynaptic* cell, and the receiving neuron is called the *postsynaptic* cell. The difference between presynaptic and postsynaptic neurons is indicated by the greater thickness of the presynaptic membrane, and the presence at the presynaptic side of swellings called *vesicles* which contain packets of chemical transmitter. A similar organization occurs at neuromuscular junctions, with an area of muscle playing the role of "postsynaptic neuron."

Even among chemical synapses, there is a dizzying variety. For example, there are two types, 1 and 2, with different vesicle shapes. It was once thought that most Type 1 synapses are excitatory and Type 2 synapses are inhibitory; this is still a useful heuristic, though it now admits a considerable number of exceptions. Also, synapses can be from the presynaptic axon to the postsynaptic dendrite (*axodendritic*); from axon to axon (*axoaxonic*); from axon to cell body (*axosomatic*); or, when the dendrite actually carries an action potential, from dendrite to dendrite (*dendrodendritic*). Of those cases, the axodendritic is the most common.

Typical chemical synapses, through whatever transmitter substance they release, cause a passive (i.e., without an action potential) increase or decrease in the postsynaptic membrane potential, starting at the junction point (most commonly on a dendrite). In the case of excitatory synapses, the passive depolarization is termed the *excitatory postsynaptic potential* (EPSP). Similarly, inhibitory synapses cause a *hyperpolarization* (negative change in inward membrane potential, the opposite of depolarization) called the *inhibitory postsynaptic potential* (IPSP).

An EPSP occurs because transmitter causes a net inward movement of positive charge, by increasing membrane conductance to Na^+, K^+, and possibly other positive ions such as the calcium ion (Ca^{++}). The EPSP might or might not be large enough to depolarize the postsynaptic neuron to its firing threshold. In fact, the postsynaptic neuron typically has thousands of dendrites receiving synapses from different presynaptic cells. Hence, its firing depends on the sum of EPSPs from these different dendrites minus the sum of IPSPs from other dendrites. IPSPs can occur as a result of increased conductance either for outward movement of positive charge (K^+) or for inward movement of negative charge (Cl^-). Some typical postsynaptic potentials are shown in Figure A2.4.

The first chemical neurotransmitter substance to be discovered was acetylcholine, which was identified by Loewi (1921) as the substance used by the vagus nerve to decrease the heart rate. Subsequently, acetylcholine was found to be the transmitter substance used in other nerves connecting the brain

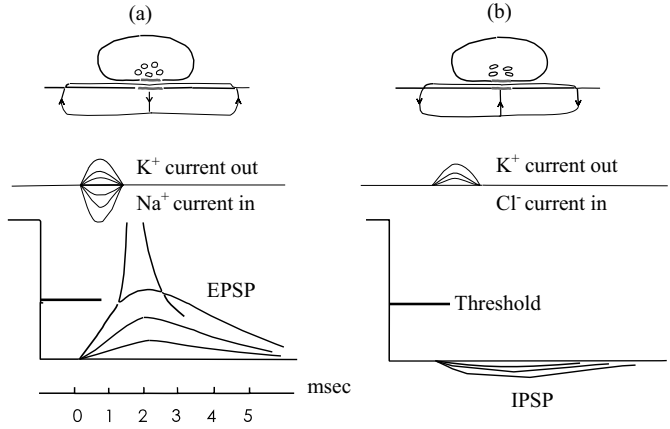


FIGURE A2.4 Typical postsynaptic potentials. Above: pre- and postsynaptic terminals, with net positive current flows shown by arrows for depolarizing (a) and hyperpolarizing (b) actions. Middle: time course of ionic current flows. Below: recordings of a typical EPSP (in a) and IPSP (in b).

Source: Reprinted from Shepherd, 1983, with permission of Oxford University Press.

to internal organs (autonomic nerves) and at the junctions between nerves and skeletal muscles. Later, it was also found to be one of the commonest transmitter substances at both excitatory and inhibitory synapses in the brain.

In addition to acetylcholine (ACh), the most important neural transmitters are dopamine (DA); norepinephrine (NE, also known as noradrenaline or NA); serotonin (5HT); gamma-amino butyric acid (GABA); glutamate (GLU); and glycine (GLY). These transmitters can be excitatory or inhibitory, with the exception of GLY and GABA, which are almost always inhibitory, and GLU, which is always excitatory. In the cortex, GLU is the main excitatory transmitter and GABA the main inhibitory transmitter.

The *monoamine* transmitters – DA, NE, and 5HT – are broadcast by certain midbrain regions out to vast areas of the cortex and the limbic system and have sometimes been found to modulate GLU or GABA connections. There has been considerable literature on the cognitive effects of those three modulators, which with many variations has pointed to overarching roles for each in cognitive functioning. Dopamine is important for learning, and acting on, associations of particular sensory stimuli and motor actions with reward (see the later sections of Chapter 6). Norepinephrine is important for arousal and initiation of behavior. Serotonin is important for stabilization of emotional reactions and for reality discrimination. Acetylcholine is also a widely broadcast modulator originating from a different midbrain region called the *nucleus basalis of Meynert*, and plays a key role in attention.

1.	*Anatomical*: presence of the substance in appropriate amounts in presynaptic processes.
2.	*Biochemical*: presence and operation of enzymes that synthesize the substance in the presynaptic neuron and processes, and remove or inactivate the substance at the synapse.
3.	*Physiological*: demonstration that physiological stimulation causes the presynaptic terminal to release the substance, and that iontophoretic application of the substance to the synapse in appropriate amounts mimics the natural response.
4.	*Pharmacological*: drugs that affect the different enzymatic or biochemical steps have their expected effects on synthesis, storage, release, action, inactivation, and reuptake of the substance.

TABLE A2.1 Criteria for deciding whether a given substance is a neurotransmitter.

Source: Reprinted from Shepherd, 1983/1994, with permission of Oxford University Press.

Space does not permit review of the complex chemical reactions involved in synaptic transmission. In general, the sequence of events is that presynaptic depolarization increases the movement of the calcium ion (Ca^{++}) near the synaptic gap, which in turn stimulates the release of transmitter from vesicles. This description is particularly good for *cholinergic* synapses, the ones using acetylcholine as their transmitter. Calcium plays a variety of other roles in neurochemistry. For example, Ca^{++} and the cyclic nucleotides (cAMP and cGMP) play "second messenger" roles in plastic changes at synapses (see the review by Byrne, 1987, discussed in Section 3.1, and the model of Gingrich & Byrne, 1985, discussed in Section 6.3.1).

Not all chemical substances present in the brain are actual neurotransmitters. Table A2.1 summarizes the criteria that are generally agreed upon for a substance to be considered a transmitter. Many substances that are not, or not known to be, neurotransmitters play other important modulating roles in cellular reactions. Among these are the cyclic nucleotides discussed above, and the neuroactive peptides (see Pert, 1986; Pert & Dienstfrey, 1988). The peptides include endorphins (morphine-like substances) that are associated with positive reinforcement.

Invertebrate and Vertebrate Nervous Systems

Among invertebrates, nerve cells controlling movements in response to particular stimuli appear in most of the multicellular phyla, starting with the coelenterates (jellyfish and medusas). While learning has been studied in flatworms, the best developed invertebrate central nervous systems are in mollusks and arthropods. These nervous systems do not have brains in the sense that

vertebrates do, but possess several *ganglia*, which are defined as concentrated areas involving several sensory and motor processing units.

Invertebrates are probably not capable of quite the same complexity of neural processing as are vertebrates. Yet invertebrate preparations have yielded basic multicellular studies of learning and conditioning (see Byrne, 1987, for a review) and of rhythmical firing patterns (e.g., Selverston, 1976).

The vertebrate nervous system has a basic plan that has persisted in spite of major evolutionary changes. It is divided into the peripheral nervous system, consisting of nerves with connections to the rest of the body, and the central nervous system, consisting of the spinal cord and the brain. The peripheral nervous system has two main parts: skeletal and autonomic, the latter comprising nerve fibers that affect, and receive sensations from, internal organs.

In the evolution of the vertebrate brain, going from lampreys (not proper vertebrates but chordates) up to humans, the characteristic divisions of forebrain, midbrain, and hindbrain are consistently maintained. In primates and in cetaceans (whales, dolphins, and porpoises), however, the forebrain balloons outward and then develops various folds (known as *convolutions* or *gyri*) to become the six-layered cerebral cortex. This cortex performs ever more sophisticated integrative functions, in feedback with the subcortical structures that change much less across species.

Functions of Vertebrate Subcortical Regions

Some of the more important large brain regions below the cerebral cortex are shown schematically in Figure A2.5. The following gross subdivisions are of functional importance:

Pons and *medulla* – just above the spinal cord;
Midbrain – just above the pons and medulla (the pons, medulla, and midbrain are usually considered to constitute the *brainstem*);
Thalamus – deep inside the forebrain;
Hypothalamus – below the thalamus;
Limbic system – forming a border around much of the forebrain and midbrain;
Cerebellum – in back of the pons and underneath the rear (occipital) area of the cortex;
Basal ganglia – at the base of the forebrain.

We now briefly review some current knowledge of the cognitive functions of each of these large regions. This review should provide the reader with an intuitive "landscape," rather than revealed truths. The cautionary note must be added, of course, that complex behaviors involve circuits rather than isolated

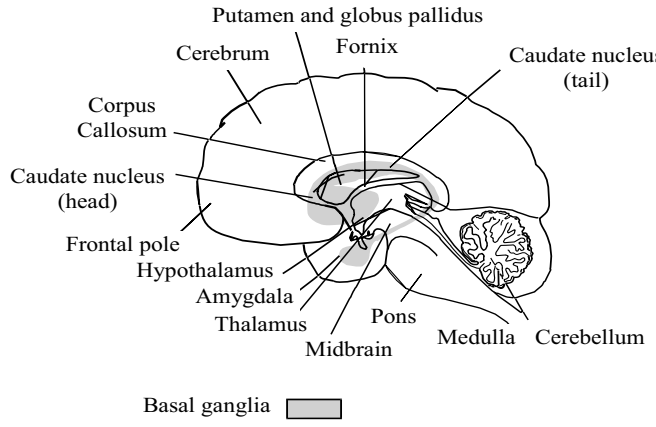


FIGURE A2.5 Medial view of the brain, showing locations of some of its major subdivisions. The amygdala is part of the limbic system emotional circuitry (see Figure A2.6), and the fornix is a pathway linking the hypothalamus with parts of the limbic system. The corpus callosum is a pathway linking the two cerebral hemispheres.

Source: Adapted from Thompson, 1967, by permission.

"centers." Also, just because stimulating a region promotes a behavior, or lesioning that region suppresses the behavior, it need not follow that the region's primary function is to perform that behavior (see Churchland, 1986, for a discussion).

The pons and medulla include some fibers and cell nuclei from the autonomic nervous system. These areas, along with the midbrain, are also the locus of the *reticular activating system* (or *reticular formation*), which is involved in the regulation of sleep, waking, and arousal. (Recall the inclusion of "nonspecific arousal" in some neural networks discussed in earlier chapters, such as the one shown in Figure 3.7.) As Truex and Carpenter (1969) state:

> The term "reticular formation" is a somewhat vague designation given a variety of special connotations; it originated in anatomy to describe portions of the brain stem core characterized structurally by a wealth of cells of various sizes and type . . . enmeshed in a complicated fiber network.
>
> (p. 316)

The reticular formation is usually considered to include some of the sources of modulating monoamine transmitters.

The thalamus is composed of different cell nuclei, most of which have one-to-one feedback connections with some part of the cortex. This includes areas of the cortex devoted to specific senses. For example, the lateral geniculate body of the thalamus is a way station for visual inputs from the optic nerve going to the visual cortex, while the medial geniculate body of the thalamus plays a similar function for the auditory cortex. It also includes multisensory association areas of the cortex; the mediodorsal nucleus of the thalamus, for example, has a one-to-one relationship with the prefrontal cortex (furthest forward section of the frontal lobes). Many areas of the thalamus also have strong connections with the limbic system and hypothalamus, which are involved in emotional expression and processing of visceral information. The hypothalamus has extensive connections with the endocrine system. Hence, particular areas of the hypothalamus are involved in either feeding or mating behavior. The lateral hypothalamic area, for example, is part of a consummatory circuit for eating that also involves areas of the brainstem, some of which use the *catecholamines* (DA and NE) as neurotransmitters. The lateral hypothalamus is also a region whose direct electrical stimulation is rewarding (Olds, 1955). The ventromedial hypothalamus is opposite in effect to the lateral; stimulation of this area produces satiety and its lesion produces overeating. These two hypothalamic areas provided some of the inspiration for motivational dipoles in neural networks (Grossberg, 1972b; see Section 3.3).

The limbic system has been implicated in the emotional expression that accompanies such behavior as feeding, copulation, and aggression. This system includes several subregions going from the hippocampus and amygdala, just under the temporal lobes, through the cingulate gyrus, which is sometimes considered part of the cortex itself, to the septum, closer to the frontal. The full details of emotional expression involve a circuit linking the limbic system with parts of the hypothalamus, midbrain, and thalamus. Precise conclusions about the role of each part of the circuit are lacking, in spite of a tremendous amount of data. Neural network models are likely to make contributions to sorting out all these findings.

The cerebellum and basal ganglia are both involved in different aspects of motor control. Reflex movements in vertebrates need involve only the spinal cord and parts of the brainstem. But for voluntary, adaptive movements, other centers are necessary, including the cerebellum, basal ganglia, and motor cortex. A partial schematic of motor control pathways in the brain and their connections with the spinal cord is shown in Figure A2.6.

The cerebellum has been implicated in contributing to the control of three large functions: muscle tone, balance, and sensorimotor coordination. It has been the subject of many neural network models because its cell types and connections are easily identifiable and repeatable across species, and because its location makes it fairly accessible to neurophysiological study (see Eccles et al., 1967).

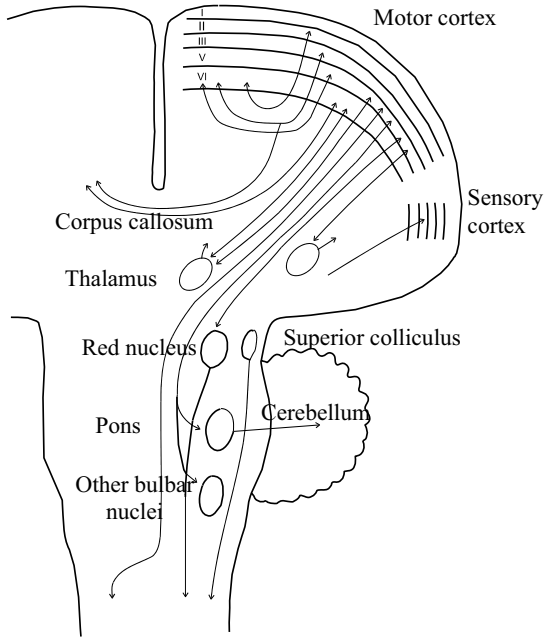


FIGURE A2.6 Summary of brain motor control pathways descending from the cortex toward the spinal cord. Roman numerals denote layers of the cortex.

Source: Reprinted from Shepherd, 1983/1994, with permission of Oxford University Press.

The basal ganglia consist of the *striatum*, which includes the *caudate* and *putamen*, and the *globus pallidus*. The globus pallidus is the output end of this region, and projects to an area of the midbrain called the *substantia nigra*. The role of these areas in movement was discovered when it was noted that lesions in several parts of this region lead to characteristic motor disorders. Parkinson's disease is associated with degeneration of the dopamine (DA) input from the substantia nigra to the striatum; for this reason, the disease is often treated with the drug L-DOPA, which enhances dopamine activity. Cell degeneration in the striatum is found in Huntington's chorea, characterized by involuntary jerking movements. And lesions in the putamen and globus pallidus have been found with athetosis, characterized by slow writhing movements. The basal ganglia are extensively connected with the motor cortex and also, both directly and through the thalamus, with the prefrontal cortex.

Functions of the Mammalian Cerebral Cortex

The cerebral cortex, as noted in Figures A2.6 and A2.7, is the youngest brain region in the evolutionary sense. The well-developed six-layered cortex is

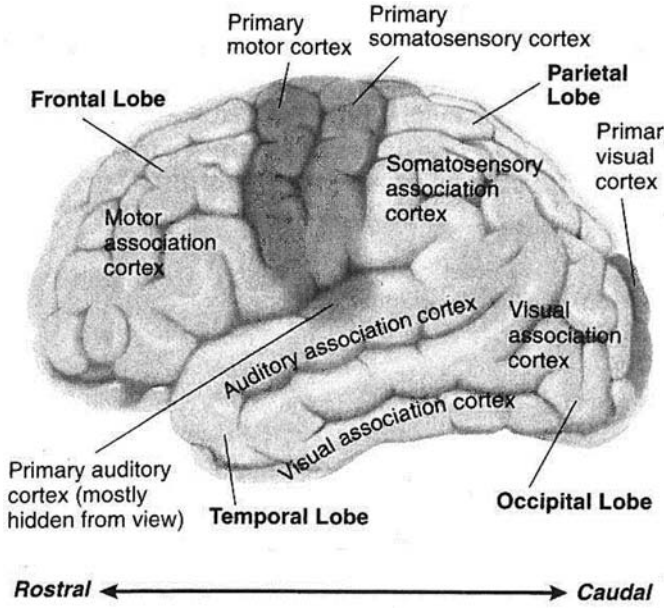


FIGURE A2.7 Schematic drawing of lateral and medial surfaces of the human brain, highlighting subdivisions of the cerebral cortex. "Primary" and "higher order," for sensory cortices, refer to processing stages, the primary being closest (synaptically) to the sensory input. "Primary motor" and "premotor" are motor control stages, the primary motor being closest to the motor output.

Source: Adapted by permission of the publisher from Carlson, 2007, with the permission of Pearson Education, Inc.

present only in mammals, and one area, the prefrontal cortex, is six-layered only in primates. Also, as one moves up the scale of mammals, the cortex becomes ever more folded, with *sulci* (the plural of *sulcus*) or depressions, and *gyri* or areas of the surface between the sulci.

Figure A2.7 shows the major subdivisions of the primate cerebral cortex. The primary motor cortex (see Figure A2.6) is directly in front of the central sulcus. The somatosensory cortex, composed of subareas responding to touch or pressure at various parts of the body, is directly behind that same sulcus. The body is represented unequally in the somatosensory cortex, with face and hands having proportionately larger representation than other areas. The primary visual cortex is in the occipital lobe and the primary auditory cortex in the temporal lobe.

The visual and auditory cortices are several synapses away from their corresponding sense organs (retina and cochlea). The olfactory sense has a more primitive circuit: the receptors project directly into the olfactory bulb, which is in a part of the cerebral cortex having only two layers.

Finally, all of the sensory areas of the cortex send axons to association areas of the cortex – most of the temporal lobe and all of the parietal and frontal lobes. Many of these association areas have specific functions; for example, the regions known as Wernicke's area (in the temporal lobe) and Broca's area (in the frontal lobe near the boundary of the temporal) are important components of circuits for speech and language. Some specialized functions of the orbital part of the prefrontal cortex (furthest forward area of the frontal lobe), which is the only part of association cortex with extensive connections to the limbic system and hypothalamus, are discussed in Section 9.4.

REFERENCES

Chapters in parentheses denote chapters *of this book* in which references appear. An asterisk before the reference indicates it is not discussed in the body of the text but listed as an additional source at the end of the chapter.

Aarts, E., Roelofs, A., & van Turennout, M. (2008). Anticipatory activity in anterior cingulate cortex can be independent of conflict and error likelihood. *Journal of Neuroscience*, *28*, 4671–4678. (Ch. 5)

*Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, *11*, 91–101. (Ch. 3)

Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1996). A Widrow-Hoff learning rule for a generalization of the linear auto-associator. *Journal of Mathematical Psychology*, *40*, 175–182. (Ch. 2, 8)

Abdi, H., Valentin, D., & O'Toole, A. J. (1997). A generalized autoassociator model for face processing and sex categorization: From principal components to multivariate analysis. In D. S. Levine & W. R. Elsberry (Eds.), *Optimality in biological and artificial networks?* (pp. 317–337). Mahwah, NJ: Lawrence Erlbaum Associates. (Ch. 8)

Abeles, M. (1991). C*orticonics: Neural circuits of the cerebral cortex*. Cambridge, UK: Cambridge University Press. (Ch. 2)

*Adorján, P., Barna, G., Érdi, P., & Obermayer, K. (1999). A statistical neural field approach to orientation selectivity. *Neurocomputing*, *26–27*, 313–318. (Ch. 7)

*Adorján, P., Levitt, J. B., Lund, J. S., & Obermayer, K. (1999). A model for the intracortical origin of orientation preference and tuning in macaque striate cortex. *Visual Neuroscience*, *16*, 303–318. (Ch. 7)

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences*, *13*, 266–271. (Ch. 9).

Alexander, G. E., DeLong, M. R., & Strick, P. F. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381. (Ch. 9)

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14, 1338–1344. http://doi.org/10.1038/nn.2921. (Ch. 9)

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: A computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation*, *27*, 2354–2410. doi:10.1162/NECO_a_00779. (Ch. 9)

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'École Américaine. *Econometrica*, 21, 503–546. (Ch. 9)

AlQaudi, B., Levine, D. S., & Lewis, F. L. (2015). Neural network model of decisions on the Asian disease problem. *Proceedings of International Joint Conference on Neural Networks 2015*, 1333–1340. (Ch. 9)

Amari, S.-I. (1971). Characteristics of randomly connected threshold element networks and network systems. *Proceedings of the IEEE*, *59*, 35–47. (Ch. 2)

Amari, S.-I. (1972). Characteristics of random nets of analog neuron-like elements. *IEEE Transactions on Systems, Man, and Cybernetics*, *2*, 643–657. (Ch. 2)

Amari, S.-I. (1974). A method of statistical neurodynamics. *Kybernetik*, *14*, 201–215. (Ch. 2)

Amari, S.-I. (1977a). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77–87. (Ch. 4, 7)

Amari, S.-I. (1977b). Neural theory of association and concept formation. *Biological Cybernetics*, *26*, 175–185. (Ch. 7)

Amari, S.-I. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, *42*, 339–364. (Ch. 7)

Amari, S.-I., & Arbib, M. A. (1977). Competition and cooperation in neural nets. In J. Metzler (Ed.), *Systems neuroscience* (pp. 119–165). New York: Academic. (Ch. 4)

Amari, S.-I., & Arbib, M. A. (Eds.) (1982). *Competition and cooperation in neural nets*. *Lecture Notes in Biomathematics*, Vol. 45. New York: Springer-Verlag. (Ch. 4)

Amari, S.-I., & Takeuchi, M. (1978). Mathematical theory of category detecting nerve cells. *Biological Cybernetics*, *29*, 127–136. (Ch. 4)

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review A*, *32*, 1007–1018. (Ch. 2)

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, *12*, 505–519. (Ch. 9)

Anderson, B. J., Lee, S., Thompson, J., Steinmetz, J., Logan, C., Knowlton, B., Thompson, R. F., & Greenough, W. T. (1989). Increased branching of spiny dendrites of rabbit cerebellar Purkinje neurons following associative eyeblink conditioning. *Society for Neuroscience Abstracts*, *15*, 640. (Ch. 2)

Anderson, J. A. (1968). A memory storage model utilizing spatial correlation functions. *Kybernetik*, *5*, 113–119. (Ch. 3)

Anderson, J. A. (1970). Two models for memory organization using interacting traces. *Mathematical Biosciences*, *8*, 137–160. (Ch. 3, Appendix 1)

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences*, *14*, 197–220. (Ch. 3, 8)

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review*, *80*, 417–438. (Ch. 3)

*Anderson, J. A. (1993). The BSB model: A simple non-linear associative network. In M. Hassoun (Ed.), *Associative neural memories: Theory and implementation* (pp. 77–102). Oxford: Oxford University Press. (Ch. 8)

*Anderson, J. A. (1998). Seven times seven is about fifty. In D. Scarborough & S. Sternberg (Eds.), *An invitation to cognitive science* (Vol. 4). Cambridge, MA: MIT Press. (Ch. 8)

Anderson, J. A., & Mozer, M. (1981). Categorization and selective neurons. In G. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 213–236). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

Anderson, J. A., & Murphy, G. L. (1986). Psychological concepts in a parallel system. *Physica D*, *22*, 318–336. (Ch. 2, 8)

*Anderson, J. A., Pellionisz, A., & Rosenfeld, E. (1990). *Neurocomputing 2: Directions for research*. Cambridge, MA: MIT Press. (Ch. 2)

*Anderson, J. A., & Rosenfeld, E. (1988). *Neurocomputing: Foundations of research*. Cambridge, MA: MIT Press. (Ch. 2)

Anderson, J. A., & Silverstein, J. W. (1978). Reply to Grossberg. *Psychological Review*, *85*, 597–603. (Ch. 8)

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451. (Ch. 4, 8, Appendix 1)

*Anderson, J. A., Spoehr, K. T., & Bennett, D. J. (1994). A study in numerical perversity: Teaching arithmetic to a neural network. In D. S. Levine & M. Aparicio, IV (Eds.), *Neural networks for knowledge representation and inference* (pp. 311–335). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

*Anderson, J. A., & Sutton, J. P. (1995). The network of networks model. *Proceedings of the World Congress on Neural Networks, Washington, DC* (Vol. 1, pp. 145–152). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

*Anderson, J. A., & Sutton, J. P. (1997). High-performance computing and neural and physiological processes. *Behavior Research Methods, Instruments, & Computers*, *29*, 67–77. (Ch. 8)

Angéniol, B., DeLaCroix Vaubois, G., & LeTexier, J.-Y. (1988). Self-organizing feature maps and the traveling salesman problem. *Neural Networks*, *1*, 289–293. (Ch. 7)

Anninos, P. A. (1972a). Mathematical models of memory traces and forgetfulness. *Kybernetik*, *10*, 165–167. (Ch. 2)

Anninos, P. A. (1972b). Cyclic modes in artificial neural nets. *Kybernetik*, *11*, 5–14. (Ch. 2)

Anninos, P. A., Beek, B., Csermely, T. J., Harth, E. M., & Pertile, G. (1970). Dynamics of neural structures. *Journal of Theoretical Biology*, *26*, 121–148. (Ch. 2)

Annis, R. C., & Frost, B. (1973). Human visual ecology and orientation anisotropies in acuity. *Science*, *182*, 729–731. (Ch. 4)

Aosaki, T., Graybiel, A. M., & Kimura, M. (1994). Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science*, *265*, 412–415. (Ch. 5)

*Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin and Review*, 22, 916–943. (Ch. 7)

Armony J. L. (2005). Computational models of emotion. In: *Proceedings of International Joint Conference on Neural Networks, Montreal, Canada* (pp. 1598–1602). Piscataway, NJ: IEEE. (Ch. 6)

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., & LeDoux, J. E. (1995). An anatomically constrained neural network model of fear conditioning. *Behavioral Neuroscience*, *109*, 246–257. (Ch. 6)

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., & LeDoux, J. E. (1997). Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in Cognitive Sciences*, *1*, 28–34. (Ch. 6)

Armony, J. L., Servan-Schreiber, D., Romanski, I. M., Cohen, J. D., & LeDoux, J. E. (1997). Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats. *Cerebral Cortex*, *7*, 157–165. (Ch. 6)

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, *21*, 1399–1407. (Ch. 5)

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481. (Ch. 8)

*Ashby, F. G., & Crossley, M. J. (2011). A computational model of how cholinergic interneurons protect striatal-dependent learning. *Journal of Cognitive Neuroscience*, *23*, 1549–1566. (Ch. 8)

*Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). Frost: A distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*, 1728–1743. (Ch. 9)

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*, 632–656. (Ch. 8)

Ashby, F. G., & Hélie, S. (2011). A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. *Journal of Mathematical Psychology*, *55*, 273–289. (Ch. 1)

Ashby, W. R., Foerster, H. von, & Walker, C. C. (1962). Instability of pulse activity in a net with threshold. *Nature*, *196*, 561–562. (Ch. 2)

*Ashe, J., Lungu, O. V., Basford, A. T., & Lu, X. (2006). Cortical control of motor sequences. *Current Opinion in Neurobiology*, *16*, 213–221. (Ch. 5)

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *99*, 13172–13177. (Ch. 5, 9)

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2003). Neural activity in prefrontal cortex during copying geometrical shapes. I. Single cell studies. *Experimental Brain Research*, *150*, 127–141. (Ch. 5, 9)

Averbeck, B. B., Crowe, D. A., Chafee, M. V., & Georgopoulos, A. P. (2003). Neural activity in macaque prefrontal cortex during copying geometrical shapes. II. Decoding shape segments from neural ensembles. *Experimental Brain Research*, *150*, 142–153. (Ch. 5, 9)

Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press. (Ch. 9)

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*, 2082–2099. (Ch. 5, 9)

*Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. S*cience*, *344*, 424–427. (Ch. 5)

Banquet, J. P., & Grossberg, S. (1987). Probing cognitive processes through the structure of event-related potentials during learning: An experimental and theoretical analysis. *Applied Optics*, *26*, 4931–4946. (Ch. 5, 7)

Bapi, R. S., & Levine, D. S. (1994). Modeling the role of the frontal lobes in performing sequential tasks. I. Basic structure and primacy effects. *Neural Networks*, *7*, 1167 1180. (Ch. 9)

Bapi, R. S., & Levine, D. S. (1997). Modeling the role of the frontal lobes in sequential task performance. II. Classification of sequences. *Neural Network World*, *1/97*, 3–28. (Ch. 9)

Barlow, H. B., Blakemore, C., & Pettigrew, J. D. (1967). The neural mechanism of binocular depth discrimination. *Journal of Physiology (London)*, *193*, 327–342. (Ch. 4, 5)

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*, 215–233. (Ch. 9)

Barto, A. G., & Anandan, P. (1985). Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15, 360–375. (Ch. 8)

Barto, A. G., & Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioural Brain Research*, *4*, 221–235. (Ch. 6)

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 835–846. (Ch. 6, 8)

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 539–602). Cambridge, MA: MIT Press. (Ch. 2)

Bear, M. F. (2003). Bidirectional synaptic plasticity: From theory to reality. *Philosophical Transactions of the Royal Society: Biological Sciences*, *358*, 649–655. (Ch. 3)

Bear, M. F., Cooper, L. N., & Ebner, F. F. (1987). A physiological basis for a theory of synapse modification. S*cience*, *237*, 42–48. (Ch. 2, 7).

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7–15. (Ch. 5, 8)

Bechara, A., Damasio, H., & Damasio, A. R. (2003). Role of the amygdala in decision making. *Annals of the New York Academy of Sciences*, *985*, 356–369. (Ch. 5)

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, *19*, 5473–5481. (Ch. 5)

Bechtel, W., Mandik, P., Mundale, J., & Stufflebeam, R. S. (2001). *Philosophy and the neurosciences: A reader* (Part V). Oxford, UK: Blackwell. (Ch. 1)

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Taylor & Francis. (Ch. 1)

Bednar, J. A., & Miikkulainen, R. (2000). Tilt aftereffects in a self-organizing model of the primary visual cortex. *Neural Computation*, *12*, 1721–1740. (Ch. 4, 8)

Beiser, D. G., & Houk, J. C. (1998). Model of cortical-basal ganglionic processing: Encoding the serial order of sensory events. *Journal of Neurophysiology*, *79*, 3168–3188. (Ch. 9)

Bengio, Y., & Lee, H. (2015). Editorial introduction to the Neural Networks special issue on Deep Learning of Representations. *Neural Networks*, *64*, 1–3. (Ch. 8)

Berger, T. W., Berry, S. D., & Thompson, R. F. (1986). Role of the hippocampus in classical conditioning of aversive and appetitive behaviors. In R. L. Isaacson & K. H. Pribram (Eds.), *The hippocampus* (Vol. 4, pp. 203–239). New York: Plenum. (Ch. 6)

Berger, T. W., & Thompson, R. F. (1978). Neuronal plasticity in the limbic system during classical conditioning of the rabbit nictitating membrane response: I. The hippocampus. *Brain Research*, *145*, 323–346.

*Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, *10*, 108–121. (Ch. 9)

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology*, *191*, 391–431. (Ch. 5, 6)

Berridge, K. C., & Robinson, T. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Review*, *28*, 309–369. (Ch. 5, 6)

Berthier, N. E., & Moore, J. W. (1986). Cerebellar Purkinje cell activity related to the classically conditioned nictitating membrane response. *Experimental Brain Research*, *63*, 341–350. (Ch. 5)

*Bertin, M., Schweighofer, N., & Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Networks*, *20*, 668–675. (Ch. 6)

Beurle, R. L. (1956). Properties of a mass of cells capable of regenerating pulses. *Philosophical Transactions of the Royal Society of London, Series B*, *250*, 55–84. (Ch. 2)

Bi, G.-Q., & Poo, M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*, 139–166. (Ch. 3)

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*, 32–48. (Ch. 3, 4, 7).

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., & Busemeyer, J. R. (2010). Sequential learning models for the Wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of Mathematical Psychology*, *54*, 5–13. (Ch. 9)

Blair, K., Marsh, A. A., Morton, J., Vythilingam, M., Jones, M., Mondillo, K., et al. (2006). Choosing the lesser of two evils, the better of two goods: Specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice. *Journal of Neuroscience*, *26*, 11379–11386. (Ch. 5)

Blakemore, C., Carpenter, R. H. S., & Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*, *228*, 37–39. (Ch. 4)

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, *228*, 477–478. (Ch. 3, 4)

Blazis, D. E. J., Desmond, J. E., Moore, J. W., & Berthier, N. E. (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: A real-time variant of the Sutton-Barto model. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 176–186). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 6)

Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, *361*, 31–39. (Ch. 3)

Bliss, T. V. P., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology (London)*, *232*, 331–356. (Ch. 1, 3)

Blomfield, S. (1974). Arithmetical operations performed by nerve cells. *Brain Research*, *69*, 115–124. (Ch. 4)

*Bogacz, R. (2009). Optimal decision making theories. In L. Dreher (Ed.), *Handbook of reward and decision making* (pp. 373–397). Oxford: Academic Press. doi:10.1016/B978-0-12-374620-7.00018-2. (Ch. 9)

*Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*, 442–477. doi:10.1162/neco.2007.19.2.442. (Ch. 9)

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*, 111–121. (Ch. 9)

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, *22*, 1504–1529. (Ch. 9)

Bolles, R. C. (1975). *A theory of motivation* (2nd ed.). New York: Harper and Row. (Ch. 6)

Borisyuk, R., Kazanovich, Y., Chik, D., Tikhanoff, V., & Cangelosi, A. (2009). A neural model of selective attention and object segmentation in the visual scene: An approach based on partial synchronization and star-like architecture of connections. *Neural Networks*, 22, 707–719. (Ch. 9)

*Botvinick, M. M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113. (Ch. 5)

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652. (Ch. 5)

*Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Chartered territory and new frontiers. *Cognitive Science*, *38*, 1249–1285. Doi:10.1111/cogs.12126. (Ch. 9)

*Botvinick, M. M., Niv. Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262–280. (Ch. 9)

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201–233. (Ch. 8, 9)

Botvinick, M. M., & Plaut, D. C. (2009). Empirical and computational support for context-dependent representations of serial order: Reply to Bowers, Damian, and Davis (2009). *Psychological Review*, *116*, 998–1002. (Ch. 9)

Boureau, Y.-L., & Dayan, P. (2011). Opponency revisited: Competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology Reviews*, *36*, 74–97; doi:10.1038/npp.2010.151. (Ch. 5)

Bourne, J. M., & Harris, K. M. (2008). Balancing structure and function at hippocampal dendritic spines. *Annual Review of Neuroscience*, *31*, 47–67.

Bowers, J. S. (2002). Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand. C*ognitive Psychology*, *45*, 413–445. (Ch. 9)

Bowers, J. S. (2017). Grandmother cells and localist representations: A review of current thinking. *Language, Cognition, and Neuroscience*, *32*, 257–273. (Ch. 2, 8)

Bowers, J. S., Damian, M. F., & Davis, C. J. (2009). A fundamental limitation of the conjunctive codes learned in PDP models of cognition: Comment on Botvinick and Plaut (2006). *Psychological Review*, *116*, 986–997. (Ch. 9)

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389–414. doi:10.103/a0026450. (Ch. 9)

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, *29*, 610–611. (Ch. 8)

Boyce, W. E., & DiPrima, R. C. (2017). *Elementary differential equations and boundary value problems* (11th ed.). Hoboken, NJ: Wiley. (Appendix 1)

Bradski, G., & Grossberg, S. (1995). Fast-learning VIEWNET architectures for recognizing three-dimensional objects from multiple two-dimensional views. *Neural Networks*, *8*, 1053–1080. (Ch. 7)

Braun, M. (2012). *Differential equations and their applications: An introduction to applied mathematics* (3rd ed.). New York: Springer-Verlag. (Appendix 1)

*Braunlich, K., Gomez-Lavin, J., & Seger, C. A. (2015). Frontoparietal networks involved in categorization and item working memory. *NeuroImage*, *107*, 146–162. (Ch. 5)

Braver, T. S., & Ruge, H. (2006). Functional neuroimaging of executive function. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (2nd ed.) (pp. 307–348). Cambridge, MA: MIT Press. (Ch. 5)

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*, 619–639. (Ch. 5)

Brindley, G. S. (1967). The classification of modifiable synapses and their use in models of conditioning. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *168*, 361–376. (Ch. 6)

Brindley, G. S. (1969). Nerve net models of plausible size that perform many simple learning tasks. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *174*, 173–191. (Ch. 6)

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. S*cience*, *307*, 1118–1121. (Ch. 5)

Brown, J. W., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, *19*, 10502–10511. (Ch. 2, 6, 7)

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*, 471–510. (Ch. 9)

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational neural model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, *55*, 37–85. (Ch. 9)

Brozović, M., Abbott, L. F., & Andersen, R. A. (2008). Mechanism of gain modulation at single neuron and network levels. *Journal of Computational Neuroscience*, *25*, 158–168. (Ch. 2)

Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations*. New York: Cambridge University Press. (Ch. 7)

Bullier, J., Hupé, J. M., James, A., & Girard, P. (1996). Functional interactions between areas Vl and V2 in the monkey. *Journal of Physiology, Paris*, *90*, 217–220. (Ch. 5, 9)

Bullock, D. (2004). Adaptive neural models of queuing and timing in fluent action. *Trends in Cognitive Sciences*, *8*, 426–433. (Ch. 9)

Bullock, D., Fiala, J. C., & Grossberg, S. (1994). A neural model of timed response learning in the cerebellum. *Neural Networks*, *7*, 1101–1114. (Ch. 6)

Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, *95*, 49–90. (Ch. 2, 3, 7, 9)

Bullock, D., & Grossberg, S. (1989). VITE and FLETE: Neural modules for trajectory formation and postural control. In W. A. Hershberger (Ed.), *Volitional action* (pp. 253–298). Amsterdam: North-Holland/Elsevier. (Ch. 2, 9)

Bullock, D., Tan, C. O., & John, Y. J. (2009). Computational perspectives on forebrain microcircuits implicated in reinforcement learning, action selection, and cognitive control. *Neural Networks*, *22*, 757–765. (Ch. 5)

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. C*ognitive, Affective & Behavioral Neuroscience*, *4*, 564–579. (Ch. 7)

Buonomano, D. V., Baxter, D. A., & Byrne, J. H. (1990). Small networks of empirically derived adaptive elements simulate some higher-order features of classical conditioning. *Neural Networks*, *3*, 507–523. (Ch. 6)

Burden, R. L., & Faires, J. D. (2005). *Numerical analysis* (8th ed.). Belmont, CA: Thompson Brooks/Cole. (Appendix 1)

Burnod, Y. (1988). *An adaptive neural network: The cerebral cortex*. Paris: Masson. (Ch. 2)

Busemeyer, J. G., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks*, *19*, 1047–1058. (Ch. 9)

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*, 253–262. (Ch. 9)

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. (Ch. 9)

Byrne, J. H. (1987). Cellular analysis of associative learning. *Physiological Reviews*, *67*, 329–439. (Ch. 3, 5, Appendix 2)

Byrne, J. H., LaBar, K. S., LeDoux, J. E., Schafe, G. E., & Thompson, R. F. (2014). Learning and memory: Basic mechanisms. In J. H. Byrne, R. Heidelberger, & M. N. Waxham, *From molecules to networks: An introduction to cellular and molecular neuroscience* (3rd ed.) (pp. 591–637). Amsterdam: Elsevier. (Ch. 3, 5)

Byrne, J. H., & Schultz, S. G. (1994). *An introduction to membrane transport and bioelectricity* (3rd ed.). New York: Lippincott-Raven. (Appendix 2)

Cabeza, R., & Kingstone, A. (Eds.) (2006). *Handbook of functional neuroimaging of cognition*. Cambridge, MA: MIT Press. (Ch. 5)

Cajal, S. Ramon y (1990). *New ideas on the structure of the nervous system in man and vertebrates*. English translation by N. Swanson and L. W. Swanson of S. Ramon y Cajal, *Les nouvelles idées sur la structure des centres nerveux chez l'homme et chez les vertèbres*. Cambridge, MA: MIT Press. (Appendix 1)

Calabresi, P., Picconi, B., Tozzi, A., & DiFilippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. T*rends in Neurosciences*, *30*, 211–219. (Ch. 5)

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, *43*, 9–64. (Ch. 9)

*Canavier, C. C., Baxter, D. A., Clark, J. W., Jr., & Byrne, J. H. (1993). Nonlinear dynamics in a model neuron provide a novel mechanism for transient synaptic inputs to produce long-term alterations of postsynaptic activity. *Journal of Neurophysiology*, *69*, 2252–2257. (Ch. 6)

*Canavier, C. C., Baxter, D. A., Clark, J. W., Jr., & Byrne, J. H. (1994). Multiple modes of activity in a model neuron suggest a novel mechanism for the effects of neuromodulators. *Journal of Neurophysiology*, *72*, 872–882. (Ch. 6)

*Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*, 1333–1336. (Ch. 4)

Carlson, N. R. (2007). *Physiology of behavior* (9th ed.). Boston: Pearson. (Appendix 2)

Carpenter, G. A. (1994). A distributed outstar network for spatial pattern learning. *Neural Networks*, *7*, 159–168. (Ch. 3)

*Carpenter, G. A. (1997). Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks*, *10*, 1473–1494. (Ch. 8)

*Carpenter, G. A., & Gaddam, S. G. (2010). Biased ART: A neural architecture that shifts attention toward previously disregarded features following an incorrect prediction. *Neural Networks*, *23*, 435–451. (Ch. 8)

Carpenter, G. A., & Grossberg, S. (1985). A neural theory of circadian rhythms: Split rhythms, after-effects, and motivational interactions. *Journal of Theoretical Biology*, *113*, 163–223. (Appendix 1)

Carpenter, G. A., & Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54–115. (Ch. 4, 6, 7, 8, 9, Appendix 1)

Carpenter, G. A., & Grossberg, S. (1987b). ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, *26*, 4919–4930. (Ch. 4, 6, 7, Appendix 1)

Carpenter, G. A., & Grossberg, S. (1989). Search mechanisms for adaptive resonance theory (ART) architectures. *International Joint Conference on Neural Networks* (Vol. I, pp. 201–205). Piscataway, NJ: IEEE. (Ch. 7)

Carpenter, G. A., & Grossberg, S. (1990). ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architecture. *Neural Networks*, *3*, 129–152. (Ch. 7)

Carpenter, G. A., Grossberg, S., & Lesher, G. W. (1998). The What-and-where Filter: A spatial mapping neural network for object recognition and image understanding. *Computer Vision and Image Understanding*, *69*, 1–22. (Ch. 2, 7)

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, *3*, 698–713. (Ch. 8)

Carpenter, G. A., Grossberg, S., & Mehanian, C. (1989). Invariant recognition of cluttered scenes by a self-organizing ART architecture: CORT-X boundary segmentation. *Neural Networks*, *2*, 169–181. (Ch. 7)

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588. (Ch. 7, 8)

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991a). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771. (Ch. 7, 8)

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991b). Fuzzy ART: An adaptive resonance algorithm for rapid, stable classification of analog patterns (Technical Report CAS/CNS-91-006). *International Joint Conference on Neural Networks* (Vol. II, pp. 411–416). Piscataway, NJ: IEEE. (Ch. 7)

*Carpenter, G. A., & Markuzon, N. (1998). ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks*, *11*, 323–336. (Ch. 8)

*Carpenter, G. A., Milenova, B. L., & Noeske, B. W. (1998). Distributed ARTMAP: A neural network for fast distributed supervised learning. *Neural Networks*, *11*, 793–813. (Ch. 8)

*Carpenter, G. A., & Olivera, S. (2012). Self-organizing ARTMAP rule discovery. *Neural Networks*, *25*, 161–177. (Ch. 8)

Carpenter, G. A., & Ross, W. D. (1995). ART-EMAP: A neural network architecture for learning and prediction by evidence accumulation. *IEEE Transactions on Neural Networks*, *6*, 805–818. (Ch. 7)

Carpenter, G. A., & Tan, A.-H. (1995). Rule extraction: From neural architecture to symbolic representation. *Connection Science*, *7*, 3–27. (Ch. 7)

Casasent, D., & Psaltis, D. (1976). Position, rotation, and invariant optical correlations. *Applied Optics*, *15*, 1793–1799. (Ch. 7)

*Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*, 277–284. (Ch. 5)

Chelazzi, L, Miller, E., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature (London)*, *363*, 345–347. (Ch. 9)

Chevalier, G., & Deniau, J. M. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, *13*, 277–280. (Ch. 4, 8)

*Chevalier, N., Martis, S. B., Curran, T., & Munakata, Y. (2015). Metacognitive processes in executive control development: The case of reactive and proactive control. *Journal of Cognitive Neuroscience*, *27*, 1125–1136. (Ch. 5)

Chey, J., Grossberg, S., and Mingolla, M. (1997). Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction. *Journal of the Optical Society of America*, *14*, 2570–2594. (Ch. 4)

*Chey, J., Grossberg, S., and Mingolla, M. (1998). Neural dynamics of motion processing and speed discrimination. *Vision Research*, *38*, 2769–2786. (Ch. 4)

Choe, Y. (2004). The role of temporal parameters in a thalamocortical model of analogy. *IEEE Transactions on Neural Networks*, *15*, 1071–1082. (Ch. 9)

Choe, Y., & Miikkulainen, R. (2004). Contour integration and segmentation with self-organized lateral connections. *Biological Cybernetics*, *90*, 75–88. doi 10.1007/s00422–003–0435–5. (Ch. 2, 8)

Chowdhury, D. (1986). *Spin glasses and other frustrated systems*. Princeton, NJ: Princeton University Press. (Ch. 2)

Christoff, K., & Gabrieli, J. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, *28*, 168–186. (Ch. 5, 7, 8, 9)

Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., & Mädler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. *Brain Research*, *1286*, 94–105. (Ch. 5, 7, 8)

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press. (Appendix 2)

Cofer, C. N. (1972). *Motivation and emotion*. Glenview, IL: Scott, Foresman. (Ch. 6)

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361. (Ch. 9)

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45–77. (Ch. 2, 9)

*Cohen, J. D., Usher, M., & McClelland, J. L. (1998). A PDP approach to set size effects within the Stroop task: Reply to Kanne, Balota, Spielerand, Faust (1998). *Psychological Review*, *105*, 188–194. (Ch. 9)

Cohen, M. A. (1988). Sustained oscillations in a symmetric cooperative-competitive neural network: Disproof of a conjecture about content-addressable memory. *Neural Networks*, *1*, 217–221. (Ch. 4)

Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*, 815–826. (Ch. 3, 4, Appendix 1)

Cohen, M. A., & Grossberg, S. (1984). Some global properties of binocular resonances: Disparity matching, filling-in, and figure-ground synthesis. In P. Dodwell & T. Caelli (Eds.), *Figural synthesis* (pp. 117–152). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 4)

Cohen, M. A., & Grossberg, S. (1986). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping and competition for short term memory. *Human Neurobiology*, *5*, 1–22. (Ch. 9)

Cohen, M. A., & Grossberg, S. (1987). Masking fields: A massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, *26*, 1866–1891. (Ch. 4, 8, 9)

Cohen, M. A., Grossberg, S., & Stork, D. (1987). Recent developments in a neural model of real-time speech analysis and synthesis. *IEEE First International Conference on Neural Networks* (Vol. IV, pp. 443–453). San Diego: IEEE/ICNN. (Ch. 9)

Cole, K. S., & Hodgkin, A. L. (1939). Membrane and protoplasm resistance in the squid giant axon. *Journal of General Physiology*, *22*, 671–687. (Appendix 2)

Coleman, S., Brown, V. R., Levine, D. S., & Mellgren, R. L. (2005). A neural network model of foraging decisions made under predation risk. *Cognitive, Affective, and Behavioral Neuroscience*, *5*, 434–451. (Ch. 9)

Collingridge, G. L. (2003). The induction of N-methyl-D-aspartate receptor-dependent long-term potentiation. *Philosophical Transactions of the Royal Society of London Series B*, *358*, 635–641. doi:10.1098/rstb.2002. 1241. (Ch. 3)

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*, 190–229. http://dx. doi. org/10.1037/a0030852. (Ch. 9)

Collins, A. G. E., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, *10*, e1001293. doi:10.1371/journal.pbio.1001293.g001. (Ch. 9)

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247. (Ch. 8)

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256. doi:10.1037//0033–295X.108.1.204. (Ch. 9)

*Commons, M. E., Grossberg, S., & Staddon, J. E. R. (Eds.). (1991) *Neural network models of conditioning and action*. Hillsdale, NJ: Erlbaum. (Ch. 6)

*Contreras-Vidal, J. L., & Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Comparative Neuroscience*, *6*, 191–214. (Ch. 9)

Contreras-Vidal, J. L., & Stelmach, G. E. (1995). A neural model of basal ganglia-thalamocortical relations in normal and parkinsonian movement. *Biological Cybernetics*, *73*, 467–476. (Ch. 6)

*Corchs, S., & Deco, G. (2002). Large-scale neural model for visual attention: Integration of experimental single-cell and FMRI data. *Cerebral Cortex*, *12*, 339–348. (Ch. 9)

Cowan, J. D. (1970). A statistical mechanics of nervous activity. In M. Gerstenhaber (Ed.), *Lectures on mathematics in the life sciences* (Vol. 2, pp. 1–57). Providence, RI: American Mathematical Society. (Ch. 2)

*Coultrop, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-takes-all competition via lateral inhibition. *Neural Networks*, *4*, 47–54. (Ch. 4)

Crick, F., & Koch, C. (1990). Towards a neurophysiological theory of consciousness. S*eminars in the Neurosciences*, *2*, 263–275. (Ch. 2)

Cruthirds, D. R., Gove, A., Grossberg, S., Mingolla, E., Novak, N., & Williamson, J. (1992). Processing of synthetic aperture radar images by the boundary contour system. *International Conference on Neural Networks* (Vol. IV, pp. 414–419). Piscataway, NJ: IEEE. (Ch. 4)

Dalenoort, G. J. (1983). Grossberg's "cells" considered as cell assemblies. *The Behavioral and Brain Sciences*, *6*, 662–663. (Ch. 4)

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam. (Ch. 5)

*Davelaar, E. J. (2007). Sequential retrieval and inhibition of parallel (re)activated representations: A neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, *15*, 51–71. (Ch. 9)

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*, 713–758. doi:10.1037/a0019738. (Ch. 9)

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*, 603–616. (Ch. 5, 8)

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711. (Ch. 6)

Dawes, R. (1992). Perfect memory. In D. S. Levine & S. J. Leven (Eds.), *Motivation, emotion, and goal direction in neural networks* (pp. 411–425). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 7)

*Dawson, M. R. W. (2008). Connectionism and classical conditioning. *Comparative Cognition & Behavior Reviews*, *3*, 1–115. (Ch. 6)

Dayan, P. (2001). Motivated reinforcement learning. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 13* (pp. 11–18). Cambridge, MA: MIT Press. (Ch. 6)

Dayan, P. (2007). Bilinearity, rules, and prefrontal cortex. *Frontiers in Computational Neuroscience*, *1*, 1–14. (Ch. 6)

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*, 213–219. (Ch. 6)

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press. (Preface, Appendix 1).

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, *14*, 473–492. (Ch. 5, 6)

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, and Behavioral Neuroscience*, *8*, 429–453. (Ch. 9)

*Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*, 1153–1160. (Ch. 6)

Dayhoff, J., Hameroff, S., Swenberg, C. E., & Lahoz-Beltra, R. (1993). The neuronal cytoskeleton: A complex system that subserves neural learning. In K. H. Pribram (Ed.), *Rethinking neural networks: Quantum fields and biological data* (pp. 389–442). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

Deco, G., & Rolls, E. T. (2005a). Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex*, *15*, 15–30. (Ch. 6)

*Deco, G., & Rolls, E. T. (2005b). Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology*, *76*, 236–256. (Ch. 9)

Dehaene, S., & Changeux, J.-P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, *1*, 244–261. (Ch. 9)

Dehaene, S., & Changeux, J.-P. (1991). The Wisconsin card sorting test: Theoretical analysis and modeling in a neural network. *Cerebral Cortex*, *1*, 62–79. (Ch. 9)

Dehaene, S., Changeux, J.-P., & Nadal, J. (1987). Neural networks that learn temporal sequences by selection. *Proceedings of the National Academy of Sciences*, *84*, 2727–2731. (Ch. 9)

DeMartino, B., Kumaran, D., Seymour, B., & Dolan, R. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*, 684–687. (Ch. 5)

DeNeys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brain detects we're biased. *Psychological Science*, *19*, 483–489. (Ch. 5)

Denton, S. E., & Kruschke, J. K. (2006). Attention and salience in associative blocking. *Learning & Behavior*, *34*, 285–304. (Ch. 8)

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin & Review*, *15*, 780–786. (Ch. 8)

*de Pinho, M., Mazza, M., & Roque, A. C. (2006). A computational model of the primary auditory cortex exhibiting plasticity in the frequency representation. *Neurocomputing*, *70*, 3–8. (Ch. 6)

Deregowski, J. B. (1973). Illusion and culture. In R. L. Gregory & G. H. Gombrich (Eds.), *Illusions in nature and art* (pp. 161–192). New York: Scribner. (Ch. 4)

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. (Ch. 5, 8)

Dev, P. (1975). Perception of depth surfaces in random-dot stereograms: A neural model. *International Journal of Man-Machine Studies*, *7*, 511–528. (Ch. 4)

DeWeerd, P., Peralta, M. R., Desimone, R., & Ungerleider, L. G. (1999). Loss of attentional stimulus selection after extrastriate cortical lesions in macaques. *Nature Neuroscience*, *2*, 753–758. (Ch. 9)

DeYoe, E. A., & Van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. T*rends in Neurosciences*, *11*, 219–226. (Ch. 5)

Dickinson, A., & Balleine, B. S. W. (2002). The role of learning in motivation. In C. R. Gallistel (Ed.), *Stevens' handbook of experimental psychology* (Vol. 3, pp. 497–533). New York: Wiley. (Ch. 6)

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, *33*, 5797–5805. (Ch. 6)

Dizon, M. J., & Khodakhah, K. (2011). The role of interneurons in shaping Purkinje cell responses in the cerebellar cortex. *Journal of Neuroscience*, *31*, 10463–10473. (Ch. 4)

Doboli, S., Brown, V. R., & Minai, A. A. (2009). A conceptual neural model of idea generation. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 723–729). Atlanta, GA. (Ch. 9)

Doboli, S., Minai, A. A., & Brown, V. R. (2007). Adaptive dynamic modularity in a connectionist model of context-dependent idea generation. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007* (pp. 2183–2188). (Ch. 9)

Dominey, P., Arbib, M. A., & Joseph, J. P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *Journal of Cognitive Neuroscience*, *7*, 311–336. (Ch. 9)

*Donohue, S. E., Green, J. J., & Woldorff, M. G. (2015). The effects of attention on the temporal integration of multisensory stimuli. *Frontiers in Integrative Neuroscience*, *9*, April 23ArtID: 32. (Ch. 5)

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*, 1481–1486. (Ch. 9)

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural Networks*, *12*, 961–974. (Ch. 6)

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*, 1347–1369. (Ch. 6)

*Dragoi, V. (1997). A dynamic theory of acquisition and extinction in operant learning. N*eural Networks*, *10*, 201–229. (Ch. 6)

Dranias, M., Grossberg, S., & Bullock, D. (2008). Dopaminergic and non-dopaminergic value systems in conditioning and outcome-specific revaluation. *Brain Research*, *1238*, 239–287. (Ch. 3, 6, 8)

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley and Sons. (Ch. 3, 7)

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*, 475–483. (Ch. 5)

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, *5*, 94–98. (Ch. 9)

*Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000a). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology*, *83*, 1733–1750. (Ch. 9)

*Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000b). Neurocomputational models of working memory. *Nature Neuroscience*, *3* (Suppl.), 1184–1191. (Ch. 9)

Eccles, J. C., Ito, M., & Szentagothai, J. (1967). *The cerebellum as a neuronal machine*. New York: Springer. (Ch. 4, Appendix 2)

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics*, *60*, 121–130. (Ch. 2, 4)

Edelman, G. M. (1987). *Neural Darwinism*. New York: Basic Books. (Ch. 2, 7)

Edelman, G. M., & Reeke, G. N., Jr. (1982). Selective networks capable of representative transformation, limited generalizations, and associative memory. *Proceedings of the National Academy of Sciences*, *79*, 2091–2095. (Ch. 7)

Edwards, C. H., Jr., & Penney, D. E. (2007/2014). *Calculus and analytic geometry: Early transcendentals* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall. International Edition published by Pearson in 2014. (Appendix 1)

Egelman, D. M., Person, C., & Montague, P. R. (1998). A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, *10*, 623–630. (Ch. 9)

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. A*nnual Review of Neuroscience*, *30*, 123–152. (Ch. 5, 8)

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the XXVII annual conference of the Cognitive Science Societ*y. Mahwah, NJ: Lawrence Erlbaum Associates. (Ch. 9)

*Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, *25*, 245–286. (Ch. 9)

Ellias, S. A., & Grossberg, S. (1975). Pattern formation, contrast control, and oscillations in the short-term memory of shunting on-center off-surround networks. *Biological Cybernetics*, *20*, 69–98. (Ch. 4, 7)

Elliott, R., Agnew, Z., & Deakin, J. F. W. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, *27*, 2213–2218. (Ch. 5)

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. (Ch. 8, 9)

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, *49*, 709–724. (Ch. 9)

Eriksson, P. S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A. M., Nordborg, C., Peterson, D., & Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine*, *4*, 1313–1317. (Ch. 3)

Ermentrout, G. B., & Cowan, J. D. (1980). Large scale spatially organized activity in neural nets. *SIAM Journal on Applied Mathematics*, *38*, 1–21. (Ch. 4, Appendix 1)

Farajidavar, A., Levine, D. S., Kohn, N. W., & Paulus, P. B. (2010). Modeling the beneficial effects of incubation in creative brainstorming. W*CCI 2010 IEEE World Congress on Computational Intelligence*, 2722–2727. (Ch. 9)

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55. doi:10.1146/annurev.neuro.051508.135516 . (Ch. 3).

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254. (Ch. 1)

Feldmeyer, D., Egger, V., Lubke, J., & Sakmann, B. (1999). Reliable synaptic connections between pairs of excitatory layer 4 neurones within a single "barrel" of developing rat somatosensory cortex. *Journal of Physiology (London)*, *521*, 169–190. (Ch. 4)

Felleman. D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. C*erebral Cortex*, *1*, 1–47. (Ch. 5, 7)

Fender, D. H., & Julesz, B. (1967). Extension of Panum's fusional area in binocularly stabilized vision. *Journal of the Optical Society of America*, *57*, 819–830. (Ch. 4)

*Feng, C., Luo, Y.-J., & Krueger, F. (2015). Neural signatures of fairness_related normative decision making in the ultimatum game: A coordinate_based meta_analysis. *Human Brain Mapping*, *36*, 591–602. (Ch. 5)

Ferster, D., & Koch, C. (1987). Neuronal connections underlying orientation selectivity in cat visual cortex. *Trends in Neurosciences*, *10*, 487–492. (Ch. 7)

Fiala, J. C., Grossberg, S., & Bullock, D. (1996). Metabotropic glutamate receptor activation in cerebellar Purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response. *Journal of Neuroscience*, *16*, 3760–3774. (Ch. 6)

Finkel, L. H., & Edelman, G. M. (1985). Interaction of synaptic modification rules within populations of neurons. *Proceedings of the National Academy of Sciences*, *82*, 1291–1295. (Ch. 7)

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902. (Ch. 5, 6)

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioral and Brain Functions*, *1*, 7–11. (Ch. 5, 6)

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71. (Ch. 4, 8)

*Francis, G., & Grossberg, S. (1996). Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours. *Vision Research*, *36*, 149–173. (Ch. 4)

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*, 300–326. (Ch. 3, 9)

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience*, *1*, 137–160. (Ch. 6, 9)

Franklin, D. J., & Grossberg, S. (2017). A neural model of normal and abnormal learning and memory consolidation: adaptively timed conditioning, hippocampus, amnesia, neurotrophins, and consciousness. *Cognitive, Affective, and Behavioral Neuroscience*, *17*, 24–76. doi:10.3758/s13415–016–0463-y. (Ch.6, 7, 9)

Frauenthal, J. C. (1980). *Introduction to population modeling*. Boston, MA: Birkhauser. (Appendix 1)

Freeman, W. J. (1972a). Waves, pulses, and the theory of neural masses. *Progress in Theoretical Biology*, *2*, 87–165. (Ch. 2)

Freeman, W. J. (1972b). Linear analysis of the dynamics of neural masses. *Annual Review of Biophysics and Bioengineering*, *1*, 225–256. (Ch. 2)

Freeman, W. J. (1975a). *Mass action in the nervous system*. New York: Academic Press. (Ch. 2)

Freeman, W. J. (1975b). Parallel processing of signals in neural sets as manifested in the EEG. *International Journal of Man-Machine Studies*, *7*, 347–369. (Ch. 2)

Freeman, W. J. (1983). Experimental demonstration of (shunting networks,(the (sigmoid function,(and (adaptive resonance(in the olfactory system. *The Behavioral and Brain Sciences*, *6*, 665–666. (Ch. 4)

Freeman, W. J. (1992). Tutorial on neurobiology: From single neurons to brain chaos. *International Journal of Bifurcation and Chaos*, *2*, 451–482. (Ch. 2)

Freeman, W. J., & Skarda, C. A. (1990). Representations: Who needs them? In J. L. McGaugh, N. Weinberger, & G. Lynch (Eds.), Brain *organization and memory: Cells, systems, and circuits* (pp. 375–380). New York: Oxford. (Ch. 1)

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*, 128–135. (Ch. 8)

Freud, S. (1895/1953). Project for a scientific psychology. In *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 1, pp. 283–397). London: Hogarth Press. (Ch. 2)

*Fries, P., Womelsdorf, T., Oostenveld, R., & Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. *Journal of Neuroscience*, *28*, 4823–4835. (Ch. 5)

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, *360*, 343–346. (Ch. 4)

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, *20*, 121–136. (Ch. 7)

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–204. (Ch. 2, 7, 8)

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, *15*, 455–469. (Ch. 8)

Fukushima, K. (2013). Artificial vision by multi-layered neural networks: Neocognitron and its advances. *Neural Networks*, *37*, 103–119. (Ch. 7)

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformation and shifts in position. *Pattern Recognition*, *15*, 455–469. (Ch. 7)

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*, 61–78. (Ch. 9)

Fuster, J. M. (1985a). The prefrontal cortex, mediator of cross-temporal contingencies. *Human Neurobiology*, *4*, 169–179. (Ch. 5)

Fuster, J. M. (1985b). The prefrontal cortex and temporal integration. In A. Peters and E. G. Jones (Eds.), *Cerebral cortex* (Vol. 4, pp. 151–177). New York: Plenum. (Ch. 5)

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe* (3rded.). Philadelphia: Lippincott-Raven. (Ch. 5, 8)

Fuster, J. M., Bauer, R. H., & Jervey, J. P. (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey during cognitive tasks. *Experimental Neurology*, *77*, 679–694. (Ch. 4)

Fuster, J. M., & Bressler, S. L. (2012). Cognit activation: a mechanism enabling temporal integration in working memory. *Trends in Cognitive Sciences*, *16*, 207–217. (Ch. 2)

*Garagnani, M., & Pulvermüller, F. (2013). Neuronal correlates of decisions to speak and act: Spontaneous emergence and dynamic topographies in a computational model of frontal and temporal areas. *Brain and Language*, *127*, 75–85. (Ch. 9)

Gazzaniga, M. S. (Editor-in-chief) (2009). *The cognitive neurosciences* (4th edition). Cambridge, MA: MIT Press. (Ch. 5)

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*, 771–781. doi:10.1016/j.visres.2010.09.027. (Ch. 9)

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, *27*, 379–402. doi:10.106/@0364–0213 (03)00009–0. (Ch. 9)

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711–724. (Ch. 9)

Gelperin, A., Hopfield, J. J., & Tank, D. W. (1985). The logic of *Limax* learning. In A. Selverston (Ed.), *Model neural networks and behavior*. New York: Plenum. (Ch. 6)

Geman, S. (1979). Some averaging and stability results for random differential equations. *SIAM Journal on Applied Mathematics*, *36*, 86–105. (Ch. 2)

Geman, S. (1980). The law of large numbers in neural modeling. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology* (pp. 91–105). Providence, RI: American Mathematical Society. (Ch. 2)

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. (Ch. 4)

Gentry, R. D. (1978). *Introduction to calculus for the biological and health sciences*. Reading, MA: Addison-Wesley. (Appendix 1)

*Georgiopoulos, M., Heileman, G. E., & Huang, J. (1991). Properties of learning related to pattern diversity in ART 1. *Neural Networks*, *4*, 751–757. (Ch. 7)

*Georgiopoulos, M., Heileman, G. E., & Huang, J. (1992). The N-N-N conjecture in ART 1. *Neural Networks*, *5*, 745–753. (Ch. 7)

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning and Behavior*, *40*, 255–268. (Ch. 6)

Gerstner, W., & Abbott, L. F. (1997). Learning navigational maps through potentiation and modulation of hippocampal place cells. *Journal of Computational Neuroscience*, *4*, 79–94. (Ch. 3)

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*, 76–78. (Ch. 3)

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1999). Hebbian learning of pulse timing in the barn owl auditory system. In Maass, W., & Bishop, C. M. (Eds.) (1999), *Pulsed neural networks* (pp. 353–377). Cambridge, MA: MIT Press. (Ch. 3)

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. New York: Cambridge University Press. (Appendix 1)

Ghashghaei, H. T., & Barbas, H. (2002). Pathways for emotion: Interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience*, *115*, 1261–1279. (Ch. 5)

*Ghose, G. M., & Freeman, R. D. (1997). Intracortical connections are not required for oscillatory activity in the visual cortex. *Visual Neuroscience*, *14*, 963R–979R. (Ch. 4)

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect, and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, *20*, 453–467. (Ch. 4)

Giesbrecht, B., Kingstone, A., Handy, T. C., Hopfinger, J. B., & Mangun, G. R. (2006). Functional neuroimaging of attention. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (2nd ed., pp. 85–111). Cambridge, MA: MIT Press. (Ch. 5)

*Giles, C. L., Kuhn, G. M., & Williams, R. J. (1994). Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Networks*, *5*, 153–156. (Ch. 9)

Gingrich, K. J., & Byrne, J. H. (1985). Simulation of synaptic depression, posttetanic potentiation, and presynaptic facilitation of synaptic potentials from sensory neurons mediating gill-withdrawal reflex in *Aplysia*. *Journal of Neurophysiology*, *53*, 652–669. (Ch. 6, Appendix 2)

Gingrich, K. J., & Byrne, J. H. (1987). Single-cell model for associative learning. *Journal of Neurophysiology*, *57*, 1705–1715. (Ch. 6)

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585–595. (Ch. 6)

Glimcher, P. W., Dorris, M. C., & Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and Economic Behavior*, *52*, 213–256. (Ch. 9)

Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science*, *306*, 447–452. (Ch. 9)

*Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195. (Ch. 8, 9)

Gluck, M. A., & Bower, G. H. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247. (Ch. 8, 9)

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*, 491–516. (Ch. 6)

*Gluck, M. A., Myers, C., & Meeter, M. (2005). Cortico-hippocampal interaction and adaptive stimulus representation: A neurocomputational theory of associative learning and memory. *Neural Networks*, *18*, 1265–1279. (Ch. 6)

Goel, P., & Gelperin, A. (2006). A neuronal network for the logic of *Limax* learning. *Journal of Computational Neuroscience*, *21*, 259–270. doi:10.1007/s10827–006–8097–7. (Ch. 6)

Golden, R. (1986). The brain-state-in-a-box model is a gradient descent algorithm. *Journal of Mathematical Psychology*, *30*, 73–80. (Ch. 4)

Goldman-Rakic, P. S. (1984). Modular organization of prefrontal cortex. *Trends in Neurosciences*, *7*, 419–429. (Ch. 4)

Gorchetchnikov, A., & Hasselmo, M. E. (2005). A simple rule for spike-timing-dependent plasticity: Local influence of AHP current. *Neurocomputing*, *65–66*, 885–890. (Ch. 3)

Gorchetchnikov, A., Versace, M., & Hasselmo, M. E. (2005). A model of STDP based on spatially and temporally local information: Derivation and combination with gated decay. *Neural Networks*, *18*, 458–466.

*Gori, S., Giora, E., Yazdanbakhsh, A., & Mingolla, E. (2011). A new motion illusion based on competition between two kinds of motion processing units: The Accordion Grating. *Neural Networks*, *24*, 1082–1092. (Ch. 9)

Gould, E. (2007). How widespread is adult neurogenesis in mammals? *Nature Reviews Neuroscience*, *8*, 481–488. (Ch. 3)

Graham, N. (1981). The visual system does a crude Fourier analysis of patterns. In S. Grossberg (Ed.), *Mathematical Psychology and Psychophysiology* (pp. 1–16). Providence, RI: American Mathematical Society. (Ch. 4)

Gray, C. M., König, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, *338*, 334–337. (Ch. 2, 4)

Gray, C. M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences*, *86*, 1698–1702. (Ch. 2, 4)

Gray, J. A., & Smith, P. T. (1969). An arousal-decision model for partial reinforcement and discrimination learning. In R. M. Gilbert & N. S. Sutherland (Eds.), *Animal discrimination learning* (pp. 243–272). New York: Academic Press. (Ch. 3)

Graybiel, A. M. (1991). Basal ganglia – Input, neural activity, and relation to the cortex. *Current Opinion in Neurobiology*, *1*, 644–651. (Ch. 6)

Green, J. T., & Woodruff-Pak, D. S. (2000). Eyeblink classical conditioning: Hippocampal formation is for neutral stimulus associations as cerebellum is for association-response. *Psychological Bulletin*, *126*, 138–158. doi:10.J037//0033–2909.126.1.138. (Ch. 5)

Greenberg, A. S., Esterman, M., Wilson, D., Serences, J. T., & Yantis, S. (2010). Control of spatial and feature-based attention in frontoparietal cortex. *Journal of Neuroscience*, *30*, 14330–14339. (Ch. 5)

Greenspan, D., & Casulli, V. (1993). *Numerical analysis for applied mathematics, science, and engineering* (2nd ed.). Reading, MA: Addison-Wesley. (Appendix 1)

*Greve, A., Donaldson, D. I., & van Rossum, M. C. W. (2010). A single-trace dual-process model of episodic memory: A novel computational account of familiarity and recollection. *Hippocampus*, *20*, 235–251. (Ch. 9)

Grieve, K. L., & Sillito, A. M. (1995). Non-length-tuned cells in layers II/III and IV of the visual cortex: The effect of blockade of layer VI on responses to stimuli of different lengths. *Experimental Brain Research*, *104*, 12–20. (Ch. 9)

Griffith, J. S. (1963a). On the stability of brain-like structures. *Biophysical Journal*, *3*, 299–308. (Ch. 2)

Griffith, J. S. (1963b). A field theory for neural nets, I: Derivation of the field equations. *Bulletin of Mathematical Biophysics*, *25*, 111–120. (Ch. 2)

Griffith, J. S. (1965). A field theory for neural nets, II: Properties of the field equations. *Bulletin of Mathematical Biophysics*, *27*, 187–195. (Ch. 2)

Grimson, W. E. L. (1983). To have your edge and fill-in too. *Behavioral and Brain Sciences*, *4*, 666–667. (Ch. 4)

Grossberg, S. (1968a). A prediction theory for some non-linear functional-differential equations, I. Learning of lists. *Journal of Mathematical Analysis and Applications*, *21*, 643–694. (Ch. 3, 6, Appendix 1)

Grossberg, S. (1968b). A prediction theory for some non-linear functional-differential equations, II. Learning of patterns. *Journal of Mathematical Analysis and Applications*, *22*, 490–522. (Ch. 3, 6)

Grossberg, S. (1969a). Embedding fields: A theory of learning with physiological implications. *Journal of Mathematical Psychology*, *6*, 209–239. (Ch. 3, 6)

Grossberg, S. (1969b). On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics*, *1*, 319–350. (Ch. 3, 6)

Grossberg, S. (1969c). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I. *Journal of Mathematics and Mechanics*, *19*, 53–91. (Ch. 3)

Grossberg, S. (1970a). Neural pattern discrimination. *Journal of Theoretical Biology*, *27*, 291–337. (Ch. 3, 4)

Grossberg, S. (1970b). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Mathematics*, *49*, 135–166. (Ch. 3)

Grossberg, S. (1971). On the dynamics of operant conditioning. *Journal of Theoretical Biology*, *33*, 225–255. (Ch. 6, 8)

Grossberg, S. (1972a). Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt (Ed.), *Delay and Functional Differential Equations and Their Applications* (pp. 121–160). New York: Academic Press. (Ch. 3)

Grossberg, S. (1972b). A neural theory of punishment and avoidance. I. Qualitative theory. *Mathematical Biosciences*, *15*, 39–67. (Ch. 2, 3, 6, Appendix 2)

Grossberg, S. (1972c). A neural theory of punishment and avoidance. II. Quantitative theory. *Mathematical Biosciences*, *15*, 253–285. (Ch. 2, 3, 6)

Grossberg, S. (1972d). Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, *10*, 49–57. (Ch. 4, 7)

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, *52*, 213–257. (Ch. 4, 7)

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, *18*, 263–327. (Ch. 6)

Grossberg, S. (1976a). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics*, *21*, 145–159. (Ch. 4, 7)

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134. (Ch. 7)

Grossberg, S. (1976c). Adaptive pattern classification and universal recoding: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, *23*, 187–202. (Ch. 7)

Grossberg, S. (1978a). Competition, decision and consensus. *Journal of Mathematical Analysis and Applications*, *66*, 470–493. (Ch. 4)

Grossberg, S. (1978b). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (Vol. 5, pp. 233–374). (Ch. 9)

Grossberg, S. (1978c). Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz, and Jones. *Psychological Review*, *85*, 592–596. (Ch. 8)

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1–51. (Ch. 9)

Grossberg, S. (Ed.). (1982). S*tudies in mind and brain*. Boston: Reidel. (Ch. 7)

Grossberg, S. (1982a). Processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. *Psychological Review*, *89*, 529–572. (Ch. 6)

Grossberg, S. (1982b). A psychophysiological theory of reinforcement, drive, motivation, and attention. *Journal of Theoretical Neurobiology*, *1*, 286–369. (Ch. 6)

Grossberg, S. (1983). The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *Behavioral and Brain Sciences*, *4*, 625–692. (Ch. 4)

Grossberg, S. (1984). Some psychophysiological and pharmacological correlates of a developmental, cognitive, and motivational theory. In R. Karrer, J. Cohen, & P. Tueting (Eds.), *Brain and information: Event related potentials* (pp. 58–142). New York: New York Academy of Sciences. (Ch. 7)

Grossberg, S. (1987a). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23–63. (Ch. 4, 7)

Grossberg, S. (Ed.). (1987b). *The adaptive brain* (Vols. 1 and 2). New York: Elsevier. (Ch. 7)

Grossberg, S. (1987c). Cortical dynamics of three-dimensional form, color, and brightness perception, I. Monocular theory. *Perception and Psychophysics*, *41*, 87–116. (Ch. 4)

Grossberg, S. (1987d). Cortical dynamics of three-dimensional form, color, and brightness perception, II. Binocular theory. *Perception and Psychophysics*, *41*, 117–158. (Ch. 3, 4)

Grossberg, S. (1988). *Neural networks and natural intelligence*. Cambridge, MA: MIT Press.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. Perception and Psychophysics, 55, 48–120. (Ch. 4)

Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex. *Spatial Vision*, *12*, 163–186. (Ch. 7, 8)

Grossberg, S. (2000). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences*, *4*, 233–246. (Ch. 9)

*Grossberg, S. (2001). Linking the laminar circuits of visual cortex to visual perception: Development, grouping, and attention. *Neuroscience & Biobehavioral Reviews*, *25*, 513–526. (Ch. 9)

*Grossberg, S. (2003). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*, 47–76. (Ch. 9)

Grossberg, S. (2006). My interests and theoretical method. www.cns.bu.edu/Profiles/Grossberg/ GrossbergInterests.pdf. (Ch. 1)

*Grossberg, S. (2007). Towards a unified theory of neocortex: Laminar cortical circuits for vision and cognition. *Progress in Brain Research*, *165*, 79–104. (Ch. 9)

*Grossberg, S. (2017). Towards solving the Hard Problem of Consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Networks*, *87*, 38–95. (Ch. 9)

Grossberg, S., Bullock, D., & Dranias, M. R. (2008). Neural dynamics underlying impaired autonomic and conditioned responses following amygdala and orbitofrontal lesions. *Behavioral Neuroscience*, *122*, 1100–1125. (Ch. 6, 8)

Grossberg, S., Carpenter, G. A., & Ersoy, B. (2005). Brain categorization: Learning, attention, and consciousness. *Proceedings of International Joint Conference on Neural Networks, Montreal, Canada* (pp. 1609–1614). (Ch. 7, 8)

Grossberg, S., & Grunewald, A. (1997). Cortical synchronization and perceptual framing. *Journal of Cognitive Neuroscience*, *9*, 117–132. (Ch. 9)

Grossberg, S., & Gutowski, W. (1987). Neural dynamics of decision making under risk: Affective balance and cognitive-emotional interactions. *Psychological Review*, *94*, 300–318. (Ch. 9)

Grossberg, S., & Kuperstein, M. (1986/1989). *Neural dynamics of adaptive sensory-motor control: Ballistic eye movements*. Amsterdam: Elsevier/ North-Holland. (Expanded edition by Pergamon Press). (Ch. 3, 9)

Grossberg, S., & Levine, D. S. (1975). Some developmental and attentional biases in the contrast enhancement and short-term memory of recurrent neural networks. *Journal of Theoretical Biology*, *53*, 341–380. (Ch. 4, 7)

Grossberg, S., & Levine, D. S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, interstimulus interval, and secondary reinforcement. *Applied Optics*, *26*, 5015–5030. (Ch. 3, 6).

Grossberg, S., Levine, D. S., & Schmajuk, N. A. (1992). Associative learning and selective forgetting in a neural network regulated by reinforcement and attentive feedback. In D. S. Levine & S. J. Leven (Eds.), *Motivation, Emotion, and Goal Direction in Neural Networks* (pp. 37–62). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 6)

Grossberg, S., & Merrill, J. W. L. (1992). A neural network model of adaptively time reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, *1*, 3–38. (Ch. 6)

Grossberg, S., & Merrill, J. W. L. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience*, *8*, 257–277. (Ch. 6)

Grossberg, S., & Mingolla, E. (1985a). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological Review*, *92*, 173–211. (Ch. 4)

Grossberg, S., & Mingolla, E. (1985b). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception and Psychophysics*, *38*, 141–171. (Ch. 4, 7)

*Grossberg, S., & Mingolla, E. (1987). Neural dynamics of surface perception: Boundary webs, illuminants, and shape-from-shading. *Computer Vision, Graphics, and Image Processing*, *37*, 116–165. (Ch. 4)

Grossberg, S., & Mingolla, E. (1993). Neural dynamics of motion perception: Direction fields, apertures, and resonant grouping. *Perception and Psychophysics*, *53*, 243–278. (Ch. 4)

Grossberg, S., Mingolla, E., & Williamson, J. (1995). Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation. *Neural Networks*, *8*, 1005–1028. (Ch. 4)

Grossberg, S., & Pearson, L. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: Toward a unified theory of how the cerebral cortex works. *Psychological Review*, *115*, 677–732. doi:10.1037/a0012618. (Ch. 9)

Grossberg, S., & Raizada, R. D. S. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*, 1413–1432. (Ch. 9)

Grossberg, S., & Rudd, M. E. (1989). A neural architecture for visual motion perception: Group and element apparent motion. *Neural Networks*, *2*, 421–450. (Ch. 4, 6)

Grossberg, S., & Rudd, M. E. (1992). Cortical dynamics of visual motion perception: Short-range and long-range apparent motion. *Psychological Review*, *99*, 78–121. (Ch. 4)

Grossberg, S., & Schmajuk, N. A. (1987). Neural dynamics of attentionally-modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, *15*, 195–240. (Ch. 2, 3, 6, 9)

Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*, 79–102. (Ch. 6)

Grossberg, S., & Seidman, D. (2006). Neural dynamics of autistic behaviors: Cognitive, emotional, and timing substrates. *Psychological Review*, *113*, 483–525. (Ch. 6)

Grossberg, S., & Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks*, *4*, 453–466. (Ch. 4, 8)

Grossberg, S., & Stone, G. O. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, *93*, 46–74. (Ch. 7, 9)

*Grossberg, S., & Todorović, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena. *Perception and Psychophysics*, *43*, 241–277. (Ch. 4)

Grossberg, S., & Versace, M. (2008). Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Research*, *1218*, 278–312. (Ch. 7, 8)

Grossberg, S., & Williamson, J. R. (2001). A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual groupings and learning. *Cerebral Cortex* (New York, N.Y.), *11*, 37–58. (Ch. 9)

*Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*, 43–53. (Ch. 9)

*Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621. (Ch. 9)

Guigon, E., Dorizzi, B., Burnod, Y., & Schultz, W. (1995). Neural correlates of learning in the prefrontal cortex of the monkey: A predictive model. *Cerebral Cortex*, *5*, 135–147. (Ch. 8. 9)

*Gurney, K., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*, 401–410. (Ch. 9)

*Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, 84, 411–423. (Ch. 9)

Halsband, U., Matsuzaka, Y., & Tanji, J. (1994). Neuronal activity in the primate supplementary, pre-supplementary and premotor cortex during externally and internally instructed sequential movements. *Neuroscience Research*, *20*, 149–155. (Ch. 9)

*Harley, T. (1996). Connectionist modeling of the recovery of language functions following brain damage. *Brain and Language*, 52, 7–24. (Ch. 9)

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528. doi:10.1037/0033–295X.106.3.491. (Ch. 9)

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. doi:10.1037/0033–295X.111.3.662. (Ch. 9)

Harmon, L. D., & Lewis, E. R. (1968). Neural modeling. *Physiological Reviews*, *46*, 513–591. (Ch. 2)

Harth, E. M., Csermely, T. J., Beek, B., & Lindsay, R. D. (1970). Brain functions and neural dynamics. *Journal of Theoretical Biology*, *26*, 93–120. (Ch. 2)

Hartline, H. K., & Ratliff, F. (1957). Inhibitory interactions of receptor units in the eye of *Limulus*. *Journal of General Physiology*, *40*, 351–376. (Ch. 2, 4)

Hawkins, R. D., Abrams, T. W., Carew, T. J., & Kandel, E. R. (1983). A cellular mechanism of classical conditioning in *Aplysia*: activity-dependent amplification of presynaptic facilitation. *Science*, *219*, 400–405. (Ch. 6).

Hawkins, R. D., & Kandel, E. R. (1984). Is there a cell biological alphabet for simple forms of learning? *Psychological Review*, *91*, 375–391. (Ch. 6)

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience and Biobehavioral Reviews*, *34*, 701–720. (Ch. 6)

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley, 1949. (Ch. 2, 3, 5, 6)

Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). *Psychological Review*, *62*, 243–254. (Ch. 6)

Hecht-Nielsen, R. (1986). Performance limits of optical, electro-optical, and electronic neurocomputers. In H. Szu (Ed.), *Hybrid and Optical Computing* (pp. 277–306). Bellingham, WA: SPIE, Vol. *634*. (Ch. 1)

*Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 623–627. (Ch. 4)

Heidelberger, R., Shouval, H., Zucker, R. S., & Byrne, R. S. (2014). Synaptic plasticity. In J. H. Byrne, R. Heidelberger, & M. N. Waxham, *From molecules to networks: An introduction to cellular and molecular neuroscience* (3rd ed.) (pp. 533–561). Amsterdam: Elsevier. (Ch. 3)

*Heileman, G. E., Georgiopoulos, M., & Abdallah, C. (1995). A dynamical adaptive resonance architecture. *IEEE Transactions on Neural Networks*, *6*, 873–889. (Ch. 7)

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review*, *117*, 994–1024. (Ch. 9)

*Hélie, S., Paul, E. J., & Ashby, F. E. (2012). A neurocomputational account of cognitive deficits in Parkinson's disease. *Neuropsychologia*, *50*, 2290–2302. (Ch. 7)

Henderson, S. E., & Norris, C. J. (2013). Counterfactual thinking and reward processing: An fMRI study of responses to gamble outcomes. *NeuroImage*, *64*, 582–589.

Herd, S. A., Banich, M. T., & O'Reilly, R. C. (2006). Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience*, *18*, 22–32. (Ch. 9)

*Herd, S. A., Krueger, K. A., Kriete, T. E., Huang, T., Hazy, T. E., & O'Reilly, R. C. (2013). Strategic cognitive sequencing: A computational cognitive neuroscience approach. *Computational Intelligence and Neuroscience*, 2013 ArtID: 149329. (Ch. 9)

Herd, S. A., O'Reilly, R. C., Hazy, T. E., Chatham, C. H., Brant, A. M., & Friedman, N. P. (2014). A neural network model of individual differences in task switching abilities. *Neuropsychologia*, *62*, 375–389. (Ch. 9)

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539. (Ch. 9)

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S., & Doya, K. (1999). Parallel neural networks for learning sequential procedures. T*rends in Neurosciences*, *22*, 464–471. (Ch. 5)

Hinton, G. E. (1987). *Connectionist learning procedures* (Tech. Rep. No. CMU-CS-87–115). Pittsburgh: Carnegie Mellon University, Computer Science Department. (Ch. 6)

Hinton, G. E. (2007). Learning multiple layers of representation. Trends in Cognitive Sciences, 11, 428–434. (Ch. 7)

Hinton, G. E., & Becker, S. (1990). An unsupervised learning procedure that discovers surfaces in random-dot stereograms. *International Joint Conference on Neural Networks* (Vol. 1, pp. 218–222). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 4)

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural information processing systems, 1987* (pp. 358–366). New York: American Institute of Physics. (Ch. 8)

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554. (Ch. 2, 8)

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. I, pp. 282–317). Cambridge, MA: MIT Press, 1986. (Ch. 4)

Hirose, Y., Yamashita, K., & Hijiya, S. (1991). Back-propagation algorithm which varies the number of hidden units. *Neural Networks*, *4*, 61–66. (Ch. 8)

Hirsch, M. W. (1982). Systems of differential equations which are competitive or cooperative. I: Limit sets. *SIAM Journal on Mathematical Analysis*, *13*, 167–179. (Ch. 4)

Hirsch, M. W. (1984). The dynamical systems approach to differential equations. *Bulletin of the American Mathematical Society*, *11*, 1–64. (Ch. 4)

Hirsch, M. W., & Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra*. New York: Academic Press. (Appendix 1)

Hirsch, H. V. B., & Spinelli, D. N. (1970). Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science*, *168*, 869–871. (Ch. 3, 4, 7)

Hirsch, J. A., & Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, *11*, 1800–1809. (Ch. 4, 8)

Hirsch, M. W. (1990). On the Amari-Takeuchi theory of category formation. *International Joint Conference on Neural Networks* (Vol. I, pp. 297–300). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 4)

*Hitch, G. J., Flude, B., & Burgess, N. (2009). Slave to the rhythm: Experimental tests of a model for verbal short-term memory and long-term sequence learning. *Journal of Memory and Language*, *61*, 97–111. (Ch. 5)

Hodgkin, A. L. (1964). *The conduction of the nervous impulse*. Springfield, IL: C. C. Thomas. (Ch. 4)

Hodgkin, A. L., & Huxley, A. F. (1939). Action potentials recorded from inside nerve fibre. *Nature*, 144, 710–711.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, *117*, 500–544. (Appendix 2)

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79, 2554–2558. (Ch. 2, 3, 4, 8)

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*, 3088–3092. (Ch. 3, 4)

Hopfield, J. J., & Tank, D. W. (1985). (Neural(computation of decisions in optimization problems. *Biological Cybernetics*, *52*, 141–152. (Ch. 4)

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: a model. *Science*, *233*, 625–633. (Ch. 4)

*Horn, D., & Opher, I. (1996). Temporal segmentation in a neural dynamical system. *Neural Computation*, *8*, 375–391. (Ch. 4)

Horn, D., Sagi, D., & Usher, M. (1992). Segmentation, binding, and illusory conjunctions. *Neural Computation*, *3*, 510–525. (Ch. 4)

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366. (Ch. 8)

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation* (pp. 287–319). New York: Academic Press. (Ch. 9)

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge: MIT Press. (Ch. 6)

*Hoyer, P. O., & Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, *42*, 1593–1605. (Ch. 9)

*Huang, T.-R., & Grossberg, S. (2010). Cortical dynamics of contextually cued attentive visual learning and search: Spatial and object evidence accumulation. *Psychological Review*, *117*, 1080–1112. (Ch. 9)

Hubel, D. H., & Livingstone, M. S. (1987). Segregation of form, color, and stereopsis in primate Area 18. *Journal of Neuroscience*, *7*, 3378–3415. (Ch. 5)

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154. (Ch. 2, 3, 4, 7)

Hubel, D. H., & Wiesel, T. N. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, *26*, 994–1002. (Ch. 5)

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*, 229–298. (Ch. 2, 3, 4, 7)

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, *195*, 215–243. (Ch. 5)

*Hui, S., & Zak, S. H. (1992). Dynamical analysis of the Brain-State-in-a-Box (BSB) neural model. *IEEE Transactions on Neural Networks*, *3*, 86–94. (Ch. 8)

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton. (Ch. 2, 3, 6)

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466. (Ch. 9)

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. P*sychological Review*, *110*, 220–264. (Ch. 9)

*Humphreys, G., & Muller, H. (1993). Search via recursive rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, *25*, 43–110. (Ch. 9)

*Humphries, M. D., Stewart, R. D., & Gurney, K. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *Journal of Neuroscience*, *26*, 12921–12942. (Ch. 9)

Intraub, H. (1985). Visual dissociation: An illusory conjunction of pictures and forms. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 431–442. (Ch. 4)

Ito, M., Sukurai, M., & Tongroach, P. (1982). Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *Journal of Physiology*, *324*, 113–134. (Ch. 5)

Iyer, L. R., Minai, A. A., Doboli, S., Brown, V. R., Levine, D. S., & Paulus, P. B. (2009). Neural dynamics of idea generation and the effects of priming. *Neural Networks*, *22*, 674–686. (Ch. 9)

*Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, *17*, 2443–2452. (Ch. 6)

Jani, N. G., & Levine, D. S. (2000). A neural network theory of proportional analogy-making. *Neural Networks*, *13*, 149–183. (Ch. 9)

Jansma, J. M., Ramsey, N. F., Slagter, H. A., & Kahn, R. S. (2001). Functional anatomical correlates of controlled and automatic processing. *Journal of Cognitive Neuroscience*, *13*, 730–743. (Ch. 5)

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*, 535–547. (Ch. 6)

John, Y. J., Bullock, D., Zikopoulos, B., & Barbas, H. (2013). Anatomy and computational modeling of networks underlying cognitive-emotional interaction. *Frontiers in Human Neuroscience*, *7*, 1–26. doi:19.3389/fnhum.2013.00101. (Ch. 6)

*Johnson, M. R., & Johnson, M. K. (2009). Top-down enhancement and suppression of activity in category-selection extrastriate cortex from an act of reflective attention. *Journal of Cognitive Neuroscience*, *21*, 2320–2327. (Ch. 5)

Jordan, M. I. (1986a). An introduction to linear algebra in parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 365–422). Cambridge, MA: MIT Press. (Ch. 8, Appendix 1)

Jordan, M. I. (1986b). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Lawrence Erlbaum Associates (Ch. 8, 9)

Julesz, B. (1960). Binocular depth perception of computer-generated patterns. *The Bell System Technical Journal*, *37*, 1125–1162. (Ch. 4)

Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press. (Ch. 4)

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux. (Ch. 6, 9)

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. (Ch. 9)

*Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., & Eliasmith, C. (2017). A spiking neuron model of word associations for the Remote Associates Test. *Frontiers in Psychology*, *8*, ArtID: 99. (Ch. 9)

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts. (Ch. 6)

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.) (2000). *Principles of neural science* (4th ed.). New York: McGraw-Hill. (Appendix 1)

Kandel, E. R., & Tauc, L. (1965). Heterosynaptic facilitation in neurones of the abdominal ganglion of *Aplysia depilans*. *Journal of Physiology (London)*, *181*, 1–27. (Ch. 1, 3, 6)

Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*, 48–64. (Ch. 4)

Kant, J.-D. (1995). Categ_ART: A neural network for automatic extraction of human categorization rules. *ICANN 95 Proceedings* Vol. 2, pp. 479–484). (Ch. 7)

Kant, J.-D. (1996). *Modélisation et mise en œuvre de processus cognitifs de catégorisation à l'aide d'un réseau connexionniste*. Unpublished doctoral dissertation, Université de Rennes I. (Ch. 7)

Kaplan, G. G., Şengör, N. S., Gürvit, H., Genç, İ., & Güzeliş, C. (2006). A composite neural network model for perseveration and distractibility in the Wisconsin card sorting test. *Neural Networks*, *19*, 375–387. (Ch. 9)

Kaplan, G. G., Şengör, N. S., Gürvit, H., & Güzeliş, C. (2007). Modelling the Stroop effect: A connectionist approach. *Neurocomputing*, *70*, 1414–1423. (Ch. 9)

Kaski, S., & Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks*, *7*, 973–984. (Ch. 7)

Katchalsky, A., Rowland, V., & Blumenthal, R. (Eds.). (1974). Dynamic patterns of brain cell assemblies. *Neurosciences Research Program Bulletin*, *12*, 3–187. (Ch. 2)

Kastner, S., & Ungerleider, L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*, 315–341. (Ch. 5)

Katz, B. (1966). *Nerve, muscle, and synapse*. New York: McGraw-Hill. (Ch. 2, 4, Appendix 2)

*Kawato, M. (1995). Bi-directional theory approach to integration. In T. Inui & J. L. McClelland (Eds.), *Attention and Performance*. Cambridge, MA: MIT Press. (Ch. 9)

*Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, *57*, 169–185. (Ch. 9)

*Kawato, M., Isobe, M., Maeda, Y., & Suzuki, R. (1988). Coordinates transformation and learning control for visually-guided voluntary movement with iteration: a Newton-like method in function space. *Biological Cybernetics*, *59*, 161–177. (Ch. 9)

*Kawato, M., & Samejima, K. (2007). Efficient reinforcement learning: Computational theories, neuroscience and robotics. *Current Opinion in Neurobiology*, *17*, 205–212. (Ch. 9)

*Kepecs, A., van Rossum, M. C. V., Song, S., & Tegner, J. (2002). Spike-timing-dependent plasticity: Common themes and divergent vistas. *Biological Cybernetics*, *87*, 446–458. (Ch. 3)

Kernell, D. (1965). The adaptation and the relation between discharge frequency and current strength of cat lumbosacral motoneurones stimulated by long-lasting injected currents. *Acta Physiologica Scandinavica*, *65*, 65–73. (Ch. 2)

Kilmer, W., McCulloch, W. S., & Blum, J. (1969). A model of the vertebrate central command system. *International Journal of Man-Machine Studies*, *1*, 279–309. (Ch. 2, 4, 6)

Kilmer, W., & Olinski, M. (1974). Model of a plausible learning scheme for CA3 hippocampus. *Kybernetik*, *16*, 133–143. (Ch. 2, 4)

*Kim, S., Choi, Y., & Lee, M. (2015). Deep learning with support vector data description. *Neurocomputing*, *165*, 111–117. (Ch. 8)

*Kim, S., Hwang, J., Seo, H., & Lee, D. (2009). Valuation of uncertain and delayed reward in primate prefrontal cortex. *Neural Networks*, *22*, 294–304. (Ch. 6, 8)

Kimberg, D. Y., & Farah, M. J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General*, *122*, 411–428. (Ch. 9)

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680. (Ch. 4)

Kirkwood, A. & Bear, M. F. (1994). Hebbian synapses in visual cortex. *Journal of Neuroscience*, *14*, 1634–1645. (Ch. 3)

Klapp, S. T. (1995). Motor response programming during simple and choice-reaction time – The role of practice. *Journal of Experimental Psychology – Human Perception and Performance*, *21*, 1015–1027. (Ch. 9)

Klopf, A. H. (1972). *Brain function and adaptive systems: A heterostatic theory* (Res. Rep. AFCRL-72-0164). Bedford, MA: Air Force Cambridge Research Laboratories. (Ch. 3)

Klopf, A. H. (1979). Goal-seeking systems from goal-seeking components. *Cognition and Brain Theory Newsletter*, *3*, 2. (Ch. 3)

Klopf, A. H. (1982). *The hedonistic neuron*. Washington, DC: Hemisphere. (Ch. 3, 6)

Klopf, A. H. (1986). A drive-reinforcement model of single neuron function: An alternative to the Hebbian neuronal model. In J. S. Denker (Ed.), *Neural networks for computing* (pp. 265–270). AIP Conference Proceedings. New York: American Institute of Physics, Vol. *151*. (Ch. 2, 3, 6)

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, *16*, 85–125. (Ch. 6)

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 616–637. (Ch. 8)

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *16*, 373–381. (Ch. 9)

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749. (Ch. 5, 8)

Knutson, B., & Peterson, R. (2005). Neurally reconstructing expected utility. *Games and Economic Behavior*, *52*, 305–315. (Ch. 5)

Koch, C., & Crick, F. (1994). Some further ideas regarding the neuronal basis of awareness. In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 93–109). Cambridge, MA: MIT Press. (Ch. 2)

*Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Toward the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227. (Ch. 4)

Koechlin, E. (2014). An evolutionary computational theory of prefrontal executive function in decision making. *Philosophical Transactions of the Royal Society B*, *369*, 20130474. http://dx.doi.org/10.1098/rstb.2013.0474. (Ch. 9)

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*, 1181–1185. http://doiorg/10.112/science.1088545. (Ch. 9)

Koechlin, E., & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision making. *Science*, *318*, 594–598. (Ch. 5, 9)

*Kogo, N., & Wagemans, J. (2013). The "side" matters: How configurality is reflected in completion. *Cognitive Neuroscience*, *4*, 31–45. (Ch. 5)

Kohonen, T. (1977). *Associative memory – A system-theoretical approach*. New York: Springer. (Ch. 3, 8, Appendix 1)

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69. (Ch. 7)

Kohonen, T. (1984/1995). *Self-organization and associative memory*. Berlin: Springer-Verlag. Reprinted in 1988, 1989, and 1995. (Ch. 3, 7)

Kohonen, T. (1993). Physiological interpretation of the self-organizing mapping algorithm. *Neural Networks*, *6*, 895–905. (Ch. 8)

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer-Verlag. (Ch. 3, 7)

Kohonen, T., & Oja, E. (1976). Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics*, *21*, 85–95. (Ch. 3)

Kohonen, T., Lehtio, P., Rovamo, J., Hyvarinen, J., Bry, K., & Vainio, L. (1977). A principle of neural associative memory. *Neuroscience*, *2*, 1065–1076. (Ch. 3)

Kohonen, T., Reuhkala, E., Makisara, K., & Vainio, L. (1976). Associative recall of images. *Biological Cybernetics*, *22*, 159–168. (Ch. 3)

Kosko, B. (1986a). Differential Hebbian learning. In J. S. Denker (Ed.), *Neural Networks for Computing* (pp. 265–270). AIP Conference Proceedings. New York: American Institute of Physics, Vol. *151*. (Ch. 3, 6)

Kosko, B. (1986b). Differential Hebbian learning. In J. S. Denker (Ed.), AIP Conference Proceedings: *Vol. 151. Neural Networks for Computing* (pp. 265–270). New York: American Institute of Physics. (Ch. 3, 6)

Kosko, B. (1987a). Adaptive bidirectional associative memories. *Applied Optics*, *26*, 4947–4960. (Ch. 3, 6)

Kosko, B. (1987b). Competitive adaptive bidirectional associative memories. *IEEE First International Conference on Neural Networks* (Vol. II, pp. 759–766). San Diego: IEEE/ICNN. (Ch. 3, 6)

Kosko, B. (1987c). Competitive adaptive bidirectional associative memories. *IEEE First International Conference on Neural Networks* (Vol. II, pp. 759–766). San Diego, CA: IEEE/ICNN. (Ch. 3, 6)

Kosko, B. (1987d). Constructing associative memory. *Byte*, 137–144. (Ch. 3)

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*, 49–60. (Ch. 3)

Kozma, R., & Freeman, W. J. (2009). The KIV model of intentional dynamics and decision making. N*eural Networks*, *22*, 277–285. (Ch. 2)

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. L., & Camerer, C. F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *Journal of Neuroscience*, *29*, 2188–2192. (Ch. 5)

Krimer, L. S., & Goldman-Rakic, P. S. (1991). Prefrontal microcircuits: Membrane properties and excitatory input of local, medium, and wide arbor interneurons. *Journal of Neuroscience*, *21*, 3788–3796. (Ch. 4)

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. (Ch. 7, 8)

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083–1119. (Ch. 8)

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 120–152). Cambridge, UK: Cambridge University Press. (Ch. 8)

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*, 337–347. (Ch. 5)

Kuffler, S. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*, 37–68. (Ch. 4, 7)

*LaBerge, D. (1990). Thalamic and cortical mechanisms of attention suggested by recent positron emission tomographic experiments. *Journal of Cognitive Neuroscience*, *2*, 358–372. (Ch. 4)

*LaBerge, D., Carter, M., & Brown, V. R. (1992). A network simulation of thalamic circuit operations in selective attention. *Neural Computation*, *4*, 318–331. (Ch. 4)

*Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience*, 26, 2894–2906. (Ch. 4)

Lashley, K. (1929). *Brain mechanisms and intelligence*. Chicago: University of Chicago Press. (Ch. 2)

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley (Reprinted in F. A. Beach, D. O. Hebb, C. T. Morgan, & H. W. Nissen (Eds.), *The neuropsychology of Lashley*, pp. 506–528, New York: McGraw-Hill, 1960). (Ch. 5, 8)

LeCun, Y. (1985). Une procédure d'apprentissage pour réseau a seuil asymétrique. *Proceedings of Cognitiva 85Paris* (pp. 599–604). (Ch. 2, 3, 8).

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436–444. doi:10.1038/nature14539. (Ch. 2, 8)

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–441. (Ch. 8)

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. *Proceedings of Advances in Neural Information Processing Systems*, 396–404. (Ch. 8)

LeDoux, J. (1996). *The emotional brain*. New York: Simon and Schuster.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*, 155–184. (Ch. 3, 5, 6)

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, *54*, 95–103. (Ch. 8)

Lengyel, M., Kwag, J., Paulsen, O., & Dayan, P. (2005). Matching storage and recall: Hippocampal spike timing-dependent plasticity and phase response curves. (Ch. 3)

Leven, S. J., & Levine, D. S. (1987). Effects of reinforcement on knowledge retrieval and evaluation. *IEEE First International Conference on Neural Networks* (Vol. II, pp. 269–279). San Diego: IEEE/ICNN. (Ch. 9)

Leven, S. J., & Levine, D. S. (1996). Multiattribute decision making in context: a dynamical neural network methodology. *Cognitive Science*, *20*, 271–299. (Ch. 7, 9)

Levine, D. S. (1975). *Studies in the transformation and storage of patterns in reverberating neural networks*. Unpublished doctoral dissertation, Massachusetts Institute of Technology. (Ch. 4)

Levine, D. S. (1983). Neural population modeling and psychology: A review. *Mathematical Biosciences*, *66*, 1–86. (Ch. 1, 2, 3, 7)

Levine, D. S. (1988). Survival of the synapses. *The Sciences* November-December, 46–52. (See also the ensuing letters in the July-August, 1989 issue). (Ch. 7)

Levine, D. S. (1989). The third wave in neural networks. A*I Expert*, December, 26–33. (Ch. 1)

Levine, D. S. (1996). Modeling dysfunction of the prefrontal executive system. In J. A. Reggia, E. Ruppin, & R. Berndt (Eds.), *Neural modeling of brain and cognitive disorders* (pp. 413–439). Singapore: World Scientific. (Ch. 8)

Levine, D. S. (1999). What are neural networks, and what can they contribute to psychology? *Psychline*, *3*(2), 23–30. (Ch. 1)

Levine, D. S. (2012). Neural dynamics of affect, gist, probability, and choice. C*ognitive Systems Research*, *15–16*, 57–72. doi:10.1016/j. Cogsys.2011.07.002. (Ch. 6, 7, 8)

Levine, D. S. (2016). Toward a neuro-developmental theory of decision attribute weighting. *Proceedings of IJCNN 2016*. (Ch. 9)

*Levine, D. S., & Brown, V. R. (2007). Uses (and abuses?) of inhibition in network models. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in cognition* (pp. 281–303). Washington, DC: American Psychological Association. (Ch. 4)

Levine, D. S., Chen, K.-Y., & AlQaudi, B. (2017). Neural network modeling of business decision making. *Proceedings of IJCNN 2017*, 206–213. (Ch. 9)

Levine, D. S., & Grossberg, S. (1976). Visual illusions in neural networks: Line neutralization, tilt after-effect, and angle expansion. *Journal of Theoretical Biology*, *61*, 477–504. (Ch. 4)

*Levine, D. S., Brown, V. R., & Shirey, T. (Eds.). (2000). *Oscillations in neural systems*. Mahwah, NJ: Lawrence Erlbaum Associates. (Ch. 4)

Levine, D. S., Mills, B. A., & Estrada, S. (2005). Modeling emotional influences on human decision making under risk. IEEE: *Proceedings of International Joint Conference on Neural Networks, August 2005*, 1657–1662. (Ch. 9)

Levine, D. S., Parks, R. W., & Prueitt, P. S. (1993). Methodological and theoretical issues in neural network models of frontal cognitive functions. *International Journal of Neuroscience*, *72*, 209–233. (Ch. 9)

*Levine, D. S., & Perlovsky, L. I. (2008). A network model of rational versus irrational choices on a probability maximization task.*2008 IEEE World Congress on Computational Intelligence* (pp. 2821–2825). (Ch. 9)

Levine, D. S., & Prueitt, P. S. (1989). Modeling some effects of frontal lobe damage: Novelty and perseveration. *Neural Networks*, *2*, 103–116. (Ch. 6, 7, 9)

Levy, W. B., & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*, 791–797.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error propagation for deep learning. *Nature Communications*, *7:12376*. doi:10.1038/ncomms13276. (Ch. 2, 7)

*Lillo, W. E., Miller, D. C., Hui, S., & Zak, S. H. (1995). Synthesis of brain-state-in-a-box (BSB) based associative memories. *IEEE Transactions on Neural Networks*, *4*, 730–737. (Ch. 8)

Linsker, R. (1986a). From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences*, *83*, 7508–7512. (Ch. 7)

Linsker, R. (1986b). From basic network principles to neural architecture: emergence of orientation-sensitive cells. *Proceedings of the National Academy of Sciences*, *83*, 8779–8783. (Ch. 7)

Linsker, R. (1986c). From basic network principles to neural architecture: emergence of orientation columns. *Proceedings of the National Academy of Sciences*, *83*, 8390–8394. (Ch. 7)

Litt, A., Eliasmith, C., & Thagard, P. (2006). Why losses loom larger than gains: Modeling neural mechanisms of cognitive-a_ective interaction. In R. Sun & N. Miyake (Eds.), *Proceedingsofthe28thannualconference of the Cognitive Science Society* (pp. 495–500). Mahwah, NJ: Lawrence Erlbaum Associates. (Ch. 9)

Litt, A., Eliasmith, C., & Thagard, P. (2008). Neural affective decision theory: Choices, brains, and emotions. *Cognitive Systems Research*, *9*, 252–273. (Ch. 9)

Livingstone, M. S., & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, *4*, 309–356. (Ch. 4, 5)

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*, 145–163. (Ch. 5, 6)

*Lo, J. T. (2010). Functional model of biological neural networks. *Cognitive Neurodynamics*, *4*, 295–313. (Ch. 7)

Loewi, O. (1921). Über humorale Übertragbarkeit der Herznervenwirkung. *PflügersArchivfür die Gesamte Physiologie*, *189*, 239–242. (Appendix 2)

Lorenz, E. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, *20*, 130–141. (Appendix 1)

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. C*ognitive, Affective, & Behavioral Neuroscience*, *7*, 90–108. (Ch. 8)

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332. (Ch. 8)

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning and Behavior*, *40*, 305–319. (Ch. 6)

Maass, W., & Bishop, C. M., eds. (1999). *Pulsed neural networks*. Cambridge, MA: MIT Press. (Ch. 3)

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. S*cience*, *288*, 1835–1838. (Ch. 5)

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*, 737–749. (Ch. 8)

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298. (Ch. 6)

Mackintosh, N. J. (1983). Conditioning *and associative learning*. New York: Oxford University Press. (Ch. 6)

Maddox, W. T, & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70. (Ch. 8)

*Maertens, M., & Pollmann, S. (2007). Illusory contours do not pass through the "blind spot." *Journal of Cognitive Neuroscience*, *19*, 91–101. (Ch. 5)

*Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective & Behavioral Neuroscience*, *9*, 343–364. (Ch. 5)

*Maia, T. V., & Cleeremans, A. (2005). Consciousness: Converging insights from connectionist modeling and neuroscience. *Trends in Cognitive Sciences*, *9*, 397–404. (Ch. 9)

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the USA*, *101*, 16075–16080. (Ch. 9)

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*, 5–21. (Ch. 3)

Mannella, F., Gurney, K., & Baldassarre, G. (2013). The nucleus accumbens as a nexus between values and goals in goal-directed behavior: A review and a new hypothesis. *Frontiers in Behavioral Neuroscience*, *7*, Article 135, 1–29. doi:10.3389/fnbeh.2013.00135. (Ch. 6)

*Mannella, F., Koene, A., & Baldassarre, G. (2009). Navigation via Pavlovian conditioning: A robotic bio-constrained model of autoshaping in rats. In *Proceedings of the Ninth International Conference on Epigenetic Robotics (EpiRob 2009)*, Vol. *146*, November 12–14 (Lund University), 97–104. (Ch. 9)

*Mannes, C. (1992). A neural network model of spatio-temporal pattern recognition, recall, and timing. *International Joint Conference on Neural Networks* (Vol. 4, pp. 109–114). Piscataway, NJ: IEEE. (Ch. 9)

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*, 213–215. (Ch. 3)

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman. Reissued by MIT Press with authors D. Marr, T. A. Poggio, and S. Ullman in 2010. (Ch. 4)

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, Series B*, *207*, 187–217. (Ch. 4)

Marr, D., & Poggio, T. (1977a). From understanding computation to understanding serial circuitry. *Neurosciences Research Program Bulletin*, *15*, 470–488. (Ch. 4)

Marr, D., & Poggio, T. (1977b). *A theory of human stereo vision*. Massachusetts Institute of Technology, A. I. Memo #451. (Ch. 4)

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London, Series B*, *204*, 301–328. (Ch. 4)

Marr, D., & Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London, Series B*, *211*, 151–180. (Ch. 4)

*Marriott, S., & Harrison, R. F. (1995). A modified fuzzy ARTMAP architecture for the approximation of noisy mappings. *Neural Networks*, *8*, 619–641. (Ch. 8)

*Marshall, J. A. (1990). Self-organizing neural networks for perception of visual motion. *Neural Networks*, *3*, 45–74. (Ch. 4)

*Marshall, J. A. (1995). Adaptive pattern recognition by self-organizing neural networks: Context, uncertainty, multiplicity, and scale. *Neural Networks*, *8*, 335–362. (Ch. 9)

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457. (Ch. 6, 8)

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407. (Ch. 9)

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific memory. *Journal of Experimental Psychology: General*, *114*, 159–188. (Ch. 8)

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing*. Cambridge, MA: MIT Press (Ch. 8)

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–165). New York: Academic Press. (Ch. 7)

McCormick, D. A., & Thompson, R. F. (1984). Cerebellum: Essential involvement in the classically conditioned eyelid response. *Science*, 223, 296–299. (Ch. 5)

McCulloch, W. S. (1965). *Embodiments of mind*. Cambridge, MA: MIT Press. (Ch. 1)

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. (Ch. 1, 2, 4)

McDonnell, J. V., & Gureckis, T. M. (2011). Adaptive clustering. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 220–252). Cambridge, UK: Cambridge University Press. (Ch. 8)

McGuire, B. A., Gilbert, C. D., Rivlin, P. K., & Wiesel, T. N. (1991). Targets of horizontal connections in macaque primary visual cortex. *Journal of Comparative Neurology*, *305*, 370–392. (Ch. 4, 9)

*McKinstry, J. L., Seth, A. K., Edelman, G. M., & Krichmar, J. L. (2008). Embodied models of delayed neural responses: Spatiotemporal categorization and predictive motor control in brain based devices. *Neural Networks*, *21*, 553–561. (Ch. 9)

Meeter, M., Jehee, J., & Murre, J. (2007). Neural models that convince: Model hierarchies and other strategies to bridge the gap between behavior and the brain. *Philosophical Psychology*, *20*, 749–772. (Ch. 1, 9)

*Meeter, M., Myers, C., & Gluck, M. A. (2005). Integrating incremental learning and episodic memory models of the hippocampal region. P*sychological Review*, *112*, 560–585. (Ch. 6, 9)

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*, 423–429. (Ch. 9)

Mengov, G., Egbert, H., Pulov, S., & Georgiev, K. (2008). Emotional balances in experimental consumer choices. *Neural Networks*, *21*, 1213–1219. (Ch. 9)

Mengov, G. (2014). Person-by-person prediction of intuitive economic choice. *Neural Networks*, *60*, 232–245. (Ch. 9)

*Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention, and control: A network model of insula function. *Brain Structure and Function*, *215*, 655–667. (Ch. 5)

Mercado, E., III, Myers, C. E., & Gluck, M. A. (2000). Modeling auditory cortical processing as an adaptive chirplet transform. *Neurocomputing*, *32–33*, 913–919. (Ch. 7)

Mercado, E., III, Myers, C. E., & Gluck, M. A. (2001). A computational model of mechanisms controlling experience-dependent reorganization of representational maps in auditory cortex. C*ognitive, Affective, & Behavioral Neuroscience*, *1*, 37–55. (Ch. 7)

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115. (Ch. 8)

*Metcalfe, J. (1994). A computational modeling approach to novelty monitoring, metacognition, and frontal lobe dysfunction. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 137–156). Cambridge, MA: MIT Press. (Ch. 9)

Miikkulainen, R., Bednar, J. A., Choe, Y., & Sirosh, J. (2005). *Computational maps in the visual cortex*. New York: Springer.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. (Ch. 5, 9)

Miller, R. K., & Michel, A. N. (1982). *Ordinary differential equations*. New York: Academic Press. (Appendix 1)

*Miller, N. Y., & Shettleworth, S. J. (2008). An associative model of geometry learning: A modified choice rule. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 419–422. (Ch. 6)

Milner, B. (1964). Some effects of frontal lobectomy in man. In J. M. Warren & K. Akert (Eds.), *The frontal granular cortex and behavior* (pp. 313–334). New York: McGraw-Hill. (Ch. 5, 9)

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review*, *81*, 521–535. (Ch. 2, 4)

Minai, A. A., & Levy, W. B. (1993). Sequence learning in a single trial. *INNS World Congress on Neural Networks* (Vol. 2, pp. 505–508). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 3, 9)

Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, *49*, 8–30. Reprinted in E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*, McGraw-Hill, 1963. (Ch. 2)

Minsky, M. L., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press. (Ch. 2, 7)

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, *72*, 1024–1027. (Ch. 5)

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, J., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518, 529–541 (Ch. 8)

Monchi, O., & Taylor, J. G. (1999). A hard wired model of coupled frontal working memories for various tasks. *Information Sciences Journal*, *113*, 221–243. (Ch. 9)

Monchi, O., Taylor, J. G., & Dagher, A. (2000). A neural model of working memory processes in normal subjects, Parkinson's disease and schizophrenia for fMRI design and predictions. *Neural Networks*, *13*, 953–973. (Ch. 9)

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*, 265–284. (Ch. 9)

Montague, P. R., Dayan P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947. (Ch. 6, 8)

Montalvo, F. S. (1975). Consensus versus competition in neural networks. *International Journal of Man-Machine Studies*, *7*, 333–346. (Ch. 4)

Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, D. E. J., Sutton, R. S., & Barto, A. G. (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing, and interstimulus intervals. *Behavioural Brain Research*, *21*, 143–154. (Ch. 6)

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*, 782–784. (Ch. 5)

Morishita, I., & Yajima, A. (1972). Analysis and simulation of networks of mutually inhibiting neurons. *Kybernetik*, *11*, 154–165. (Ch. 4)

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli. *Journal of Neurophysiology*, *70*, 909–919. (Ch. 5)

Motter, B. C. (1994). Neural correlates of attentive selection for colors or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*, 2178–2189. (Ch. 9)

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology*, *20*, 408–434. (Ch. 2, 3, 4, 5)

Moustafa, A. A., Gilbertson, M. W., Orr, S. P., Herzallah, M. M., Servatius, R. J., & Myers, C. E. (2013). A model of amygdala-hippocampal-prefrontal interaction in fear conditioning and extinction in animals. *Brain and Cognition*, *81*, 29–43. (Ch. 6)

Mpitsos, G. J., Burton, R. M., Creech, H. C., & Soinila, S. O. (1988). Evidence for chaos in spike trains of neurons that generate rhythmic motor patterns. *Brain Research Bulletin*, *21*, 529–538. (Appendix 1)

Mukherjee, K. (2010). A dual system model of preferences under risk. *Psychological Review*, *177*, 243–255. (Ch. 9)

Mushiake, H., Inase, M., & Tanji, J. (1990). Selective coding of motor sequence in the supplementary motor area of the monkey cerebral cortex. *Experimental Brain Research*, *82*, 208–210. (Ch. 5)

Mushiake, H., Masahiko, I., & Tanji, J. (1991). Neuronal activity in the primate premotor, supplementary, and precentral motor cortex during visually guided and internally determined sequential movements. *Journal of Neurophysiology*, *66*, 705–718. (Ch. 5)

*Myers, C. E., Ermita, B. R. Harris, K., Hasselmo, M., Solomon, P., & Gluck, M. A. (1996). Computational model of cholinergic disruption of septohippocampal activity in classical eyeblink conditioning. N*eurobiology of Learning & Memory*, *66*, 51–66. (Ch. 6)

*Myers, C. E., Ermita, B. R., Hasselmo, M., & Gluck, M. A. (1998). Further implications of a computational model of septohippocampal cholinergic modulation in eyeblink conditioning. *Psychobiology*, *26*, 1–20. (Ch. 7)

Myers, C. E., & Gluck, M. A. (1994). Context, conditioning, and hippocampal representation. *Behavioral Neuroscience*, *108*, 835–847. (Ch. 6)

*Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences—A computational approach. *Journal of Cognitive Neuroscience*, *13*, 626–647. (Ch. 9)

Nakamura, K., Sakai, K., & Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. *Journal of Neurophysiology*, *80*, 2671–2687. (Ch. 5)

*Naqvi, N., Shiv, B., & Bechara, A. (2006). The role of emotion in decision making: a cognitive neuroscience perspective. *Current Directions in Psychological Science*, *15*, 260–264. (Ch. 5)

Nass, M. M., & Cooper, L. N. (1975). A theory for the development of feature detecting cells in the visual cortex. *Biological Cybernetics*, *19*, 1–18. (Ch. 3, 7)

Nauta, W. J. H. (1971). The problem of the frontal lobe: A reinterpretation. *Journal of Psychiatric Research*, *8*, 167–187. (Ch. 5)

Nauta, W. J. H., & Feirtag, M. (1986). *Fundamental neuroanatomy*. New York: Freeman. (Appendix 2)

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe deficits. *Cortex*, *12*, 313–324. (Ch. 9)

*Newell, B. R, & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*, 1–18. (Ch. 5)

*Newman, E. L., Gupta, K., Climer, J. R., Monaghan, C. K., & Hasselmo, M. E. (2012). Cholinergic modulation of cognitive processing: Insights drawn from computational models. *Frontiers in Behavioral Neuroscience*, *6*, Jun 13, ArtID: 24. (Ch. 9)

Nigrin, A. (1993). *Neural networks for pattern recognition*. Cambridge, MA: MIT Press. (Ch. 7, 8)

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154. (Ch. 6)

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*, 6. (Ch. 6)

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4, pp. 1–18). New York: Plenum Press. (Ch. 5)

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. (Ch. 8)

Nosofsky, R. M., Little, D. R., & James, T. W. (2012). Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proceedings of the National Academy of Sciences*, *109*, 333–338. (Ch. 5, 8)

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*, 247–255. (Ch. 7)

Nosofsky, R. M., & Zaki, S. R (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal* of Experimental Psychology: Learning, Memory, and Cognition, 28, 924–940. (Ch. 7)

Nowlan, S. J., & Sejnowski, T. J. (1994). Filter selection model for motion segmentation and velocity integration. *Journal of the Optimal Society of America*, *11*, 3711–3200. (Ch. 4)

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal* of Neuroscience, 15, 1195–1214. (Ch. 4)

*Nyhus, E., & Barceló, F. (2009). The Wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: A critical update. *Brain and Cognition*, *71*, 437–451.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York: Oxford University Press. (Ch. 9)

Oaksford, M., & Chater, N. (2009). Précis of *Bayesian rationality: The probabilistic approach to human reasoning*. *Behavioral and Brain Sciences*, 32, 69–120. (Ch. 9)

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. N*ature Neuroscience*, *4*, 95–102. (Ch. 5)

*O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. S*cience*, *304*, 452–454. (Ch. 5)

*Ö_men, H., & Gagné, S. (1990). Neural network architectures for motion perception and elementary motion detection in the fly visual system. *Neural Networks*, *3*, 487–505. (Ch. 4)

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University Press. (Ch. 7)

Olds, J. (1955). Physiological mechanisms of reward. In M. Jones (Ed.), *Nebraska symposium on motivation* (pp. 73–142). Lincoln: University of Nebraska Press. (Ch. 6, Appendix 2)

Olds, J., & Milner, P. M. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, *47*, 419–427. (Ch. 6)

*Orban, P., Peigneux, P., Lungu, O., Albouy, G., Breton, E., Laberenne, F., Benali, H., Maquet, P., & Doyon, J. (2010). The multifaceted nature of the relationship between performance and brain activity in motor sequence learning. *NeuroImage*, *49*, 694–702. (Ch. 5)

O'Reilly, R. C. (1996a). Biologically plausible error-driven learning using local activation differences: The Generalized Recirculation Algorithm. *Neural Computation*, *8*, 895–938. (Ch. 8)

O'Reilly, R. C. (1996b). *The Leabra model of neural interactions and learning in the neocortex*. Unpublished PhD dissertation, Carnegie Mellon University. (Ch. 7)

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*, 283–328. (Ch. 6, 8)

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007). PVLV: The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, *121*, 31–49. (Ch. 6)

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press. (Ch. 6)

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389–397. (Ch. 6, 8)

Otto, T., & Eichenbaum, H. (1992). Neuronal activity in the hippocampus during delayed nonmatch to sample performance in rats: Evidence for hippocampal processing in recognition memory. *Hippocampus*, *2*, 323–334. (Ch. 8)

Owen, A. M., James, M., Leigh, P. N., Summers, B. A., Marsden, C. D., Quinn, N. P., Lange, K. W., & Robbins, T. W. (1992). Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain*, *115*, 1727–1751. (Ch. 9)

Page, M., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761–781. (Ch. 9)

*Palm, G. (1982). *Neural assemblies: An alternative approach*. New York: Springer-Verlag. (Ch. 2)

Palmeri, T. J., Nosofsky, R. M. & McKinley, S. K. (1994). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 548–568. (Ch. 8)

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*, 6235–6242. (Ch. 6)

Parker, D. B. (1985). *Learning-logic (TR-47)*. Cambridge, MA: Massachusetts Institute of Technology, Center for Computational Research in Economics and Management Science. (Ch. 2, 3, 8)

Paulus, P. B., Levine, D. S., Brown, V. R., Minai, A. A., & Doboli, S. (2010). Modeling ideational creativity in groups: Connecting cognitive, neural, and computational approaches. *Small Group Research*, 41, 688–724. (Ch. 9)

Pavlov, I. P. (1927). *Conditioned reflexes* (V. Anrep, Translator). London: Oxford University Press. (Ch. 2, 3, 5, 6)

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552. (Ch. 6)

*Penke, M., & Westermann, G. (2006). Broca's area and inflectional morphology: Evidence from Broca's aphasia and computer modeling. *Cortex*, *42*, 563–576. (Ch. 9)

*Penner, M., & Mizumori, S. (2011). Neural systems analysis of decision making during goal-directed navigation. *Progress in Neurobiology*, *96*, 96–135. doi:10.1016/j.pneurobio.2011.08.010. (Ch. 6)

Perez, R., Glass, L., & Shlaer, R. (1974). Development of specificity in the cat visual cortex. *Journal of Mathematical Biology*, *1*, 275–288. (Ch. 7)

*Perfetti, R. (1995). A synthesis procedure for BSB neural networks. *IEEE Transactions on Neural Networks*, *6*, 1071–1080. (Ch. 8)

Perrett, S. P., Ruiz, B. P., & Mauk, M. D. (1993). Cerebellar cortex lesions disrupt learning-dependent timing of conditioned eyelid responses. *Journal of Neuroscience*, *13*, 1708–1718. (Ch. 5)

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP_ model of reading aloud. *Psychological Review*, *114*, 273–315. doi:10.1037/0033–295X.114.2.273. (Ch. 9)

Pert, C. B. (1986). The wisdom of the receptors: Neuropeptides, the emotions, and the body mind. *Advances*, *3*(*3*), Summer 1986, 8–16. (Appendix 2)

Pert, C. B., & Dienstfrey, H. (1988). The neuropeptide network. *Annals of the New York Academy of Sciences*, *521*, 189–194. (Appendix 2)

Peterhans, E., & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *Journal of Neuroscience*, *9*, 1749–1763. (Ch. 5, 8)

Petersen, S. E., van Mier, H., Fiez, J. A., & Raichle, M. E. (1998). The effects of practice on the functional anatomy of task performance. *Proceedings of the National Academy of Sciences*, *USA*, *95*, 853–860. (Ch. 5)

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a "low road" to "many roads" of evaluating biological significance. *Nature Reviews Neuroscience*, *11*, 773–782. (Ch. 5, 6)

*Phaf, R., & Heijden, H. V. D., & Hudson, T. (1990). Slam: A connectionist model for attention in visual selection tasks. C*ognitive Psychology*, *11*, 273–341. (Ch. 9)

Pilly, P. K., & Grossberg, S. (2012). How do spatial learning and memory occur in the brain? Coordinated learning of entorhinal grid cells and hippocampal place cells. *Journal of Cognitive Neuroscience*, *24*, 1031–1054. (Ch. 6)

Pineda, F. J. (1987). Generalization of backpropagation to recurrent neural networks. *Physical Review Letters*, *59*, 2229–2232. (Ch. 8)

Pineda, F. J. (1989). Recurrent backpropagation and the dynamical approach to adaptive neural computation. *Neural Computation*, *1*, 161–172. (Ch. 8)

Pineda, F. J. (1995). Recurrent backpropagation networks. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 99–135). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 2)

*Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115. (Ch. 9)

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500. (Ch. 9)

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., & Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, *391*, 580–584. (Ch. 9)

*Porr, B., Saudargiene, A., & Wörgötter, F. (2004). Analytical solution of spike-timing dependent plasticity based on synaptic biophysics. In S. Thrun, L. Saul, & B. Schölkopf, *Advances in neural information processing systems (16)*. Cambridge, MA: MIT Press. (Ch. 3)

*Portilla, J., Strela, V., Wainwright, M. J., & Simoncelli, E. P. (2003). Image denoising using Gaussian scale mixtures in the wavelet domain. *IEEE Transactions on Image Processing*, *12*, 1338–1351. (Ch. 9)

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304–308. (Ch. 8)

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. A*nnual Review of Neuroscience*, *13*, 25–42. (Ch. 5)

Pribram, K. H. (1961). A further experimental analysis of the behavioral deficit that follows injury to the primate frontal cortex. *Journal of Experimental Neurology*, *3*, 432–466. (Ch. 9)

Pribram, K. H. (1973). The primate frontal cortex (Executive of the brain. In K. H. Pribram & A. R. Luria (Eds.), *Psychophysiology of the frontal lobes* (pp. 293–314). New York: Academic Press. (Ch. 5)

Pribram, K. H. (1991). *Brain and perception: Holonomy and structure in figural processing*. Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 4)

Prigogine, I. (1969). Structure, dissipation, and life. *First International Conference, Theoretical Physics and Biology, Versailles* (pp. 23–52). Amsterdam: North-Holland. (Ch. 2)

*Provost, J.-S., & Monchi, O. (2015). Exploration of the dynamics between brain regions associated with the default_mode network and frontostriatal pathway with regards to task familiarity. *European Journal of Neuroscience*, *41*, 835–844. (Ch. 5)

Quintana, J., & Fuster, J. M. (1992). Mnemonic and predictive functions of cortical neurons in a memory task. *NeuroReport*, *3*, 721–724. (Ch. 5)

*Raijmakers, M. E. J., van der Maas, H. L. J., & Molenaar, P. C. M. (1996). Numerical bifurcation analysis of distance-dependent on-center off-surround shunting neural networks. *Biological Cybernetics*, *75*, 495–507. (Ch. 7)

*Raijmakers, M. E. J., & Molenaar, P. C. M. (1997). Exact ART: A complete implementation of an ART network. *Neural Networks*, *10*, 649–669. (Ch. 7)

Rainville, E. D., Bedient, P. E., & Bedient, R. E. (2014). *Elementary differential equations*. Pearson International Edition. (Appendix 1)

Raizada, R. D. S., & Grossberg, S. (2001). Context-sensitive binding by the laminar circuits of V1 and V2: A unified model of perceptual grouping, attention, and orientation contrast. *Visual Cognition*, *8*, 431–466. (Ch. 7, 8)

Raizada, R. D. S., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex*, *13*, 100–113. (Ch. 9)

Rall, W. (1955). Experimental monosynaptic input-output relations in the mammalian spinal cord. *Journal of Cellular and Comparative Physiology*, *46*, 413–437. (Ch. 2)

Ramachandran, V. (1990). Interaction between motion, depth, color, and form: The utilitarian theory of perception. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 346–360). Cambridge, UK: Cambridge University Press. (Ch. 9)

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*, 193–202. (Ch. 5, 8)

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, *42*, 2–13. (Ch. 8)

Ranti, C., Chatham, C. H., & Badre, D. (2015). Parallel temporal dynamics in hierarchical cognitive control. *Cognition*, *142*, 205–229. (Ch. 9)

*Rao, R. P. N., & Sejnowski, T. J. (2000). Predictive sequence learning in recurrent neocortical circuits. In S. A. Solla, T. K. Leen, & K. R. Muller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 164–170). Cambridge, MA: MIT Press. (Ch. 3)

*Rao, R. P. N., & Sejnowski, T. J. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. (Ch. 3)

Rashevsky, N. (1960). *Mathematical biophysics* (Vol. II). New York: Dover. (Ch. 2)

Ratliff, F. (1965). *Mach bands: Quantitative studies of neural networks in the retina*. San Francisco: Holden-Day. (Ch. 4)

*Raymond, J. L., Baxter, D. A., Buonomano, D. V., & Byrne, J. H. (1992). A learning rule based on empirically-derived activity-dependent neuromodulation supports operant conditioning in a small network. *Neural Networks*, *5*, 789–803. (Ch. 6)

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998a). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences*, *95*, 747–750. (Ch. 5)

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998b). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*, 420–428. (Ch. 5)

Reber, P. J., Wong, E. C., & Buxton, R. B. (2002). Comparing the brain areas supporting non-declarative categorization and working memory. *Cognitive Brain Research*, *14*, 245–257. (Ch. 5)

Reeke, G. N., & Edelman, G. M. (1984). Selective networks and recognition automata. *Annals of the New York Academy of Sciences*, *426*, 181–201. (Ch. 7)

Reeke, G. N., & Edelman, G. M. (1987). Selective neural networks and their implications for recognition automata. *International Journal of Supercomputer Applications*, *1*, 44–69. (Ch. 7)

Reeve, J. (1997). *Understanding motivation and emotion* (2nd ed.) Fort Worth: Harcourt Brace. (Ch. 6)

*Reeves, A., & Sperling, G. (1986). Attention gating in short-term visual memory. *Psychological Review*, *93*, 180–206. (Ch. 4)

*Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131. (Ch. 9)

*Reggia, J. A., D'Autrechy, C., Sutton, G. G., III, & Weinrich, M. (1992). A competitive distribution theory of neocortical dynamics. *Neural Computation*, *4*, 287–317. (Ch. 4)

Reilly, D. L., Cooper, L. N., & Elbaum, C. (1982). A neural model for category learning. *Biological Cybernetics*, *45*, 35–41. (Ch. 7)

*Reithler, J., van Mier, H. I., & Goebel, R. (2010). Continuous motor sequence learning: Cortical efficiency gains accompanied by striatal functional reorganization. *NeuroImage*, *52*, 263–276. (Ch. 5)

Reeve, J. (1997). *Understanding motivation and emotion* (2nd ed.). Fort Worth: Harcourt Brace. (Ch. 6)

Reisberg, D. (2016). *Cognition: Exploring the science of the mind* (6th ed.). New York: W. W. Norton. (Ch. 1)

Rescorla, R. A., & Wagner, A. B. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts. (Ch. 3, 6, 7)

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, *4*, 249–262. doi:10.1002/bdm.3960040403. (Ch. 9)

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107. (Ch. 9)

Reynolds, J. H., Chelazzi, L, & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, *19*, 1736–1753. (Ch. 5, 8)

Reynolds, J., Nicholas, J., Chelazzi, L., & Desimone, R. (1995). Spatial attention protects macaque V2 and V4 cells from the influence of non-attended stimuli. *Society for Neuroscience Abstracts*, *21*, 1759. (Ch. 9)

Rhodes, B. J. (2000). *Learning-driven changes in the temporal characteristics of serial movement performance: A model based on cortico-cerebellar cooperation*. Unpublished Ph. D. dissertation, Boston University. (Ch. 9)

Rhodes, B. J., & Bullock, D. (2002). *Neural dynamics of learning and performance of fixed sequences. I: Data interpretation and model architecture*. Boston University Technical Report CAS/CNS-2002–005. (Ch. 5, 8)

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. A. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*, *23*, 699–746. (Ch. 9)

Ricart, R. (1992). Neuromodulatory mechanisms in neural networks and their influence on interstimulus interval effects in Pavlovian conditioning. In D. S. Levine & S. J. Leven (Eds.), *Motivation, Emotion, and Goal Direction in Neural Networks*. Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 6)

Ricker, S., & Bouton, M. (1996). Reacquisition following extinction in appetitive conditioning. A*nimal Learning and Behavior*, *24*, 423–436. (Ch. 6)

*Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1919–1025. (Ch. 9)

Rilling, J. K., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, *35*, 395–405. (Ch. 5)

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*, 23–48. (Ch. 5)

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport*, 15, 2539–2543. (Ch. 5)

Rivest, F., Kalaska, J. F., & Bengio, Y. (2014). Conditioning and time representation in long short-term memory networks. *Biological Cybernetics*, *108*, 23–48. (Ch. 6)

*Roberts, P. D. (1999). Computational consequences of temporarily asymmetric learning rules: I. Differential Hebbian learning. *Journal of Computational Neuroscience*, *7*, 235–246. (Ch. 3)

Robson, J. G. (1975). Receptive fields: Neural representation of the spatial and intensive attributes of the visual image. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. V, pp. 82–116). New York: Academic Press. (Ch. 4, 5)

Robinson, T. E., & Kolb, B. (1999). Alterations in the morphology of dendrites and dendritic spines in the nucleus accumbens and prefrontal cortex following repeated treatment with amphetamine or cocaine. *European Journal of Neurology*, *11*, 1598–1604. (Ch. 5)

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multi-alternative decision field theory: A dynamic connectionist model of decision-making. *Psychological Review*, *108*, 370–392. (Ch. 9)

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*, 376–381. (Ch. 5)

Rogers, T. T., & McClelland, J. R. (2004). S*emantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press. (Ch. 8)

Rogers, T. T., & McClelland, J. R. (2011). Semantics without categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 88–119). Cambridge, UK: Cambridge University Press. (Ch. 8)

Rolls, E. T. (2000). Memory systems in the brain. *Annual Review of Psychology*, *51*, 599–630. (Ch. 5)

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition*, *55*, 11–29. (Ch. 5, 6)

*Rolls, E. T., & Deco, G. (2015). Stochastic cortical neurodynamics underlying the memory and cognitive changes in aging. *Neurobiology of Learning and Memory*, *118*, 150–161. (Ch. 9)

Rose, D., & Blakemore, C. (1974). Analysis of orientation selectivity in the cat's visual cortex. *Experimental Brain Research*, *20*, 1–17. (Ch. 4)

Rosenberg, C. R., & Sejnowski, T. J. (1986). The spacing effect on NETtalk, a massively-parallel network. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 72–89). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

Rosenblatt, F. (1962). *Principles of neurodynamics*. Washington, DC: Spartan Books. (Ch. 1, 2, 3, 4, 8)

Rosenfield, I. (1988). *The invention of memory*. New York: Basic Books. (Ch. 7)

Rosenkilde, C. E., Bauer, R. H., & Fuster, J. M. (1981). Single cell activity in ventral prefrontal cortex of behaving monkeys. *Brain Research*, *209*, 375–394. (Ch. 4)

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*, 185–190. (Ch. 9)

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences USA*, *102*, 7338–7343. (Ch. 9)

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). San Diego: Academic Press. (Ch. 7)

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press. (Ch. 2, 3, 8)

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94. (Ch. 9)

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986a). *Parallel distributed processing* (Vols. 1 and 2). Cambridge, MA: MIT Press. (Ch. 1, 2, 3, 4, 8)

Rumelhart, D. E., & McClelland, J. L. (1986b). On learning the past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press. (Ch. 8)

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer and S. Kornblum (Eds.), Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp. 3–30). Cambridge, MA: MIT Press. (Ch. 8)

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*, 75–112. (Ch. 4, 7)

Rustichini, A., & Padoa-Schioppa, C. (2015). A neuro-computational model of economic decisions. *Journal of Neurophysiology*, *114*, 1382–1398. (Ch. 9)

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–1758. (Ch. 5, 8)

*Santos, A., Porto, A., Romero, J., Albó, A., & Pazos, A. (2007). Study of classical conditioning in *Aplysia* through the implementation of computational models of its learning circuit. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*. 119–158. (Ch. 6)

Sawaguchi, T. (1996). Functional modular organization of the primate prefrontal cortex for representing working memory process. *Cognitive Brain Research*, *5*, 157–163. (Ch. 4)

*Schmajuk, N. A. (1997). *Animal learning and cognition: A neural network approach*. New York: Cambridge University Press. (Ch. 6)

*Schmajuk, N. A., & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, *99*, 268–305. (Ch. 6)

*Schmajuk, N. A., Larrauri, J. A., & LaBar, K. S. (2007). Reinstatement of conditioned fear and the hippocampus: An attentional-associative model. *Behavioural Brain Research*, *177*, 242–253. (Ch. 6)

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. (Ch. 2, 8)

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*, 1–66. (Ch. 5)

Schneiderman, N., & Gormezano, I. (1964). Conditioning of the nictitating membrane response of the rabbit as a function of the CS-US interval. *Journal of Comparative and Physiological Psychology*, *57*, 188–195. (Ch. 6)

Schoenbaum, S., Setlow, B., Saddoris, M., & Gallagher, M. (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron*, *39*, 855–867. (Ch. 5)

*Schubotz, R. (2011). Long-term planning and prediction: Visiting a construction site in the human brain. In: W. Welsch et al. (Eds.), *Interdisciplinary anthropology* (pp. 79–104). Heidelberg; Springer-Verlag. doi:10.1007/978-2-642-11668-1_4. (Ch. 5)

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology*, *56*, 1439–1461. (Ch. 5)

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913. (Ch. 5, 6)

*Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, *18*, 2680–2718. (Ch. 9)

*Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L, & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*, 2349–2356. (Ch. 5)

*Segawa, J. A., Tourville, J. A., Beal, D. S., & Guenther, F. H. (2015). The neural correlates of speech motor sequence learning. *Journal of Cognitive Neuroscience*, *27*, 819–831. (Ch. 5)

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review*, 96, 523–568. (Ch. 8, 9)

*Sejnowski, T. J. (1976). On global properties of neuronal interaction. *Biological Cybernetics*, *22*, 85–95. (Ch. 2)

Sejnowski, T. J., & Nowlan, S. J. (1995). A model of visual motion processing in area MT of primates. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 437–449). Cambridge, MA: MIT Press.

Sejnowski, T. J., & Rosenberg, C. R. (1986). *NETtalk: A parallel network that learns to read aloud*. The Johns Hopkins University Electrical Engineering and Computer Science Tech. Rep. JHU/EECS-86/01. (Ch. 8)

*Sejnowski, T. J., & Tesauro, G. (1989). The Hebb rule for synaptic plasticity: Algorithms and implementations. In J. H. Byrne & W. O. Berry (Eds.), *Neural models of plasticity* (pp. 94–103). San Diego: Academic Press. (Ch. 3)

Selfridge, O. G. (1959). PANDEMONIUM: A paradigm for learning. *Proceedings of Symposium on Mechanisation of Thought Processes, National Physics Laboratory, Teddington, England* (pp. 511–529). London: Her Majesty's Stationery Office. (Ch. 2)

Selverston, A. (1976). A model system for the study of rhythmic behavior. In J. C. Fentress (Ed.), *Simpler networks and behavior* (pp. 83–98). Sunderland, MA: Sinauer. (Appendix 1, Appendix 2)

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*, 664–667. (Ch. 2)

Shastri, L. (2001). A computational model of episodic memory formation in the hippocampal system. *Neurocomputing*, *38–40*, 889–897. (Ch. 4, 9)

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences*, 6, 162–168. (Ch. 4, 9)

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494. (Ch. 4)

Shepherd, G. M. (1983). *Neurobiology*. New York: Oxford University Press. A later edition appeared in 1994. (Appendix 2)

Sherrington, C. S. (1906/1947). *The integrative action of the nervous system*. Oxford, UK: Oxford University Press. (Ch. 2, Appendix 2)

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II: Perceptual leaning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190. (Ch. 5)

Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology*, *84*, 2148–2160. (Ch. 5)

*Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: A comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1003–1021. doi:10.1037//0278–7393.27.4.1003. (Ch. 9)

Sillito, A. M., Jones, H. E., Gerstein, G. L., & West, D. C. (1994). Feature-linked synchronization of thalamic relay-cell firing induced by feedback from the visual cortex. *Nature*, *369*, 479–482. (Ch. 5, 9)

*Simen, P, & Cohen, J. D. (2009). Explicit melioration by a neural diffusion model. *Brain Research*, *1299*, 95–117. (Ch. 6)

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, *31*, 5526–5539. (Ch. 6)

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129–138. doi:10.1037/h0042769. (Ch. 9)

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press. (Ch. 1)

*Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, *22*, 123–158. (Ch. 2)

Sirosh, J., & Miikkulainen, R. (1994). Cooperative self-organization of afferent and lateral connections in cortical maps. *Biological Cybernetics*, *71*, 66–78. (Ch. 7)

Sirosh, J., & Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, *9*, 577–594. (Ch. 7)

Skarda, C., & Freeman, W. J. (1987). How brains make chaos to make sense of the world. *Behavioral and Brain Sciences*, *10*, 161–195. (Ch. 2)

Skou, J. C. (1998). Nobel lecture: The identification of the sodium pump. Bioscience Reports, 18, 155–169. (Appendix 2)See comment in PubMed Commons below

Sloviter, R. S., & Brisman, J. L. (1995). Lateral inhibition and granule cell synchrony in the rat hippocampal dentate gyrus. *Journal of Neuroscience*, *15*, 811–820. (Ch. 4)

Smital, J. (1988). *On functions and functional equations*. Bristol, UK: Adam Hilger. (Appendix 1)

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science*, *283*, 1657–1661. (Ch. 5)

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436. (Ch. 5, 7)

Smith, M. C., Coleman S. R., & Gormezano, I. (1969). Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous, and forward CS-US intervals. *Journal of Comparative and Physiological Psycho1ogy*, *69*, 226–231. (Ch. 6)

Smolensky, P. (1986). Harmony theory. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 194–281). Cambridge, MA: MIT Press. (Ch. 4)

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review*, *81*, 119–145. (Ch. 3)

*Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*, 120–154. (Ch. 5)

Song, S., & Abbott, L. F. (2001). Cortical development and remapping through spike timing-dependent plasticity. *Neuron*, *32*, 339–350. (Ch. 3)

*Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*, 919–926.

Sontag, E., & Sussmann, H. (1989). Backpropagation separates when perceptrons do. *International Joint Conference on Neural Networks* (Vol. I, pp. 639–642). Piscataway, NJ: IEEE. (Ch. 8)

Sperling, G., & Sondhi, M. M. (1968). Model for visual luminance detection and flicker detection. *Journal of the Optical Society of America*, *58*, 1133–1145. (Ch. 4)

Sporns, O. (2010). *Networks of the brain*. Cambridge, MA: MIT Press. (Ch. 9)

Sporns, O. (2012). *Discovering the human connectome*. Cambridge, MA: MIT Press. (Ch. 9)

*Staddon, J. E. R., & Zhang, Y. (1991). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. S. Staddon, (Eds.), *Neural network models of conditioning and action* (pp. 279–293). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 6)

*Steinbuch, K. (1961). Die Lernmatrix. *Kybernetik*, *1*, 36–45. (Ch. 2)

*Steinbuch, K. (1990). Die Lernmatrix – The beginning of associative memories. In R. Eckmiller (Ed.), *Advanced Neural Computers* (pp. 21–29). Amsterdam: North-Holland. (Ch. 2).

Stent, G. S. (1973). A physiological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences*, *70*, 997–1001. (Ch. 2)

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1980). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 469–505). Hillsdale, NJ: Erlbaum. (Reprinted from *Information processing in motor control and learning*, pp. 117–152, by G. Stelmach, Ed., 1978, New York: Academic Press). (Ch. 9)

Stevens, K. A. (1983). False dilemmas: Confusion between mechanism and computation. *Behavioral and Brain Sciences*, *4*, 675. (Ch. 4)

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 444–459). Cambridge, MA: MIT Press. (Ch. 2)

Stork, D. G. (1989a). Self-organization, pattern recognition, and adaptive resonance networks. *Journal of Neural Network Computing*, *1*, 26–42. (Ch. 8)

Stork, D. G. (1989b). Is backpropagation biologically plausible? *International Joint Conference on Neural Networks* (Vol. II, pp. 241–246). Piscataway, NJ: IEEE. (Ch. 8)

Stratford, K. J., Tarczy-Hornoch, K., Martin, K. A. C., Bannister, N. J., & Jack, J. J. B. (1996). Excitatory synaptic inputs to spiny stellate cells in cat visual cortex. *Nature*, *382*, 258–261. (Ch. 9)

Stuart, G., Spruston, N., Sakmann, B., & Hauser, M. (1997). Action potential initiation and backpropagation in neurons of the mammalian central nervous system. *Trends in Neurosciences*, *20*, 125–131. (Ch. 8)

*Subagdja, B., & Tan, A.-H. (2015). Neural modeling of sequential inferences and learning over episodic memory. *Neurocomputing*, *161*, 229–242. (Ch. 9)

Sun, R. (2002). *Duality of the mind*. Mahwah, NJ: Lawrence Erlbaum Associates. (Ch. 9)

Sun, R., & Zhang, X. (2006). Accounting for a variety of reasoning data within a cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *18*, 169–191. (Ch. 8, 9)

*Suppes, P., de Barros, J. A., & Oas, G. (2012). Phase-oscillator computations as neural models of stimulus-response conditioning and response selection. *Journal of Mathematical Psychology*, *56*, 95–117. (Ch. 6)

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, *121*, 350–354. (Ch. 6)

Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*, 871–890. (Ch. 6)

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *N*eural Computation, 13, 841–862. (Ch. 2, 6)

*Suter, R. S., Pachur, T., Hertwig, R., Endestad, T., & Biele, G. (2015). The neural basis of risky choice with affective outcomes. *PLoS ONE*, April 1, 2015, 1–22. doi:10.1371/journal. Pone.0122475. (Ch. 5)

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–170. (Ch. 2, 3, 5, 6, 8)

Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In J. W. Moore & M. Gabriel (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press. (Ch. 3, 6)

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press. (Ch. 6, 7, 8, 9)

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*, 1187–1188. (Ch. 5)

*Sutton, J. P., & Breiter, H. C. (1994). Neural scale invariance: An integrative model with implications for neuropathology. *World Congress on Neural Networks* (Vol. 4, pp. 667–672). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 8)

*Swindale, N. W. (1980). A model for the formation of ocular dominance stripes. *Proceedings of the Royal Society of London B*, *208*, 243–264. (Ch. 7)

Swokowski, E. W. (1988). *Calculus with analytic geometry* (4th edition). Boston: PWS-Kent. (Appendix 1)

Szu, H. (1986). Three layers of vector output product neural networks for optical pattern recognition. In H. Szu (Ed.), *Hybrid and optical computing* (pp. 312–330). Bellingham, WA: SPIE, Vol. *634*. (Ch. 7)

*Takebe, K., Nakauchi, S., & Usui, S. (1996). A computational model for color constancy by separating reflectance and illuminant edges within a scene. *Neural Networks*, *9*, 1405–1415. (Ch. 4)

Tan, C. O., & Bullock, D. B. (2008a). Neuropeptide co-release with GABA may explain functional non-monotonic uncertainty responses in dopamine neurons. *Neuroscience Letters*, 430, 218–223. (Ch. 6)

Tan, C. O., & Bullock, D. B. (2008b). A local circuit model of learned striatal and dopamine cell responses under probabilistic schedules of reward. *Journal of Neuroscience*, 28, 10062–10074. (Ch. 6)

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature*, *371*, 413–416. (Ch. 5, 8)

*Taylor, J. G. (1997). Neural networks for consciousness. *Neural Networks*, 10, 1207–1225. (Ch. 9)

*Taylor, J. G. (1999). Towards the networks of the brain: From brain imaging to consciousness. *Neural Networks*, *12*, 943–959. (Ch. 9)

Taylor, J. G., & Alavi, F. N. (1993). A global competitive network for attention. *Neural Network World*, *5*, 477–502. (Ch. 4)

Taylor, J. G., & Alavi, F. N. (1995). A global competitive network. *Biological Cybernetics*, *72*, 233–248. (Ch. 4)

Taylor, J. G., & Alavi, F. N. (1996). A basis for long-range inhibition across cortex. In J. Sirosh & R. Miikkulainen & Y. Choe (Eds.), *Lateral interactions in cortex: structure and function*, Austin, TX: The UCTS Neural Networks Research Group. Electronic book ISBN 0-9647060-0-8. www.cs.utexas.edu/users/nn/web-pubs/htmlbook96. (Ch. 9)

*Taylor, J. G., & Mueller-Gaertner, H.-W. (1997). Non-invasive analysis of awareness. *Neural Networks*, 10, 1185–1194. (Ch. 9)

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin*, *139*, 766–791. doi:10.1037/a0030266. (Ch. 9)

Taylor, N. R., & Taylor, J. G. (2000). Hard-wired models of working memory and temporal sequence storage and generation. N*eural Networks*, *13*, 201–224. (Ch. 9)

Taylor, P. C. J., Rushworth, M. F. S., & Nobre, A. C. (2008). Choosing where to attend and the medial frontal cortex: An fMRI study. *Journal of Neurophysiology*, *100*, 1397–1406. (Ch. 5)

*Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review*, *120*, 1–38. (Ch. 9)

Terman, D., & Wang, D. (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D*, *81*, 148–176. (Ch. 9)

Tesauro, G. J. (1986). Simple neural models of classical conditioning. *Biological Cybernetics*, *55*, 187–200. (Ch. 6)

Thomas, G. B., Jr., & Finney, R. L. (1988). *Calculus and analytic geometry*. Reading, MA: Addison-Wesley. (Appendix 1)

*Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2013). Modeling socioeconomic status effects on language development. *Developmental Psychology*, *49*, 2325–2343. (Ch. 9)

Thomas, E., & French, R. (2017). Grandmother cells: Much ado about nothing. *Language, Cognition, and Neuroscience*, *32*, 342–349. (Ch. 2, 8)

*Thomas, M. S. C., Purser, H. R. M., Tomlinson, S., & Mareschal, D. (2012). Are imaging and lesioning convergent methods for assessing functional specialisation? Investigations using an artificial neural network. *Brain and Cognition*, *78*, 38–49. (Ch. 9)

Thompson, R. F. (1967). *Foundations of Physiological Psychology*. New York: Harper and Row. (Ch. 3, Appendix 2)

Thompson, R. F., Barchas, J. D., Clark, G. A., Donegan, N., Kettner, R. E., Lavond, D. G., Madden, J., Mauk, M. D., & McCormick, D. A. (1984). Neuronal substrates of associative learning in the mammalian brain. In D. L. Aldon & J. Farley (Eds.), *Primary neural substrates of learning and behavioral change* (pp. 71–99). New York: Cambridge University Press. (Ch. 6)

Thompson, R. F., Clark, G. A., Donegan, N. H., Lavond, G. A., Lincoln, D. G., Maddon, J., Mamounas, L. A., Mauk, M. D., and McCormick, D. A. (1987). Neuronal substrates of discrete, defensive conditioned reflexes, conditioned fear states, and their interactions in the rabbit. In I. Gormezano, W. F. Prokasy, and R. F. Thompson (Eds.), *Classical conditioning* (3rd ed.) (pp. 371–399). Hillsdale, NJ: Erlbaum. (Ch. 6)

Thorndike, E. L. (1933). A proof of the Law of Effect. *Science*, *77*, 173–175. (Ch. 6)

*Tijsseling, A. G., & Gluck, M. A. (2002). A connectionist approach to processing dimensional interaction. *Connection Science*, *14*, 1–48. (Ch. 7)

*Tissera, M. D., & McDonnell, M. D. (2016). Deep extreme learning machines: Supervised autoencoding architecture for classification. *Neurocomputing*, *174*, 42–49. (Ch. 8)

*Tlapale, E., Dosher, B. A., & Lu, Z.-L. (2015). Construction and evaluation of an integrated dynamical model of visual motion perception. *Neural Networks*, *67*, 110–120. (Ch. 4)

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189–208. doi:10.1037/h0061626. (Ch. 6)

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*, 515–518. (Ch. 5)

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. N*ature*, *398*, 704–708. (Ch. 5)

Tremblay, L., & Schultz, W. (2000a). Reward-related neuronal activity during go-no go task performance in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*, 1864–1876. (Ch. 5)

Tremblay, L., & Schultz, W. (2000b). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*, 1877–1885. (Ch. 5)

Trommald, M., Hulleberg, G., & Anderson, P. (1996). Long-term potentiation is associated with new excitatory synapses on rat dentate granule cells. *Learning and Memory*, *3*, 218–228. (Ch. 2)

Truex, R. C., & Carpenter, M. B. (1969). *Human neuroanatomy* (6th ed.). Baltimore: Williams and Wilkins. (Appendix 2)

*Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*, 507–545. (Ch. 4)

Tsukahara, N., & Oda, Y. (1981). Appearance of new synaptic potentials at corticorubral synapses after the establishment of classical conditioning. *Proceedings of the Japanese Academy, Series B (Physical and Biological Sciences)*, *57*, 398–401. (Ch. 2, 3)

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 382–403). New York: Academic Press. (Ch. 7)

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. (Ch. 9)

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the rationality of choice. *Science*, *211*, 453–458. (Ch. 5, 9)

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representations of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. (Ch. 9)

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior*. Cambridge, MA: MIT Press. (Ch. 5, 7)

*Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592. doi:10.1037/0033–295X.108.3.550. (Ch. 9)

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*, 757–769. (Ch. 9)

Usher, M., & Niebur, E. (1996). Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention. *Journal of Cognitive Neuroscience*, *8*, 311–327. (Ch. 9)

*Usher, M., & Zakay, D. (1993). A neural network model for attribute-based decision processes. *Cognitive Science*, *17*, 349–396. (Ch. 9)

*Usui, S., Nakauchi, S., & Miyake, S. (1994). Acquisition of the color opponent representation by a three-layered neural network. *Biological Cybernetics*, *72*, 35–41. (Ch. 4)

Uttley, A. M. (1966). The transmission of information and the effect of local feedback in theoretical and neural networks. *Brain Research*, *2*, 21–50. (Ch. 6)

Uttley, A. M. (1970). The informon: A network for adaptive pattern recognition. *Journal of Theoretical Biology*, *27*, 31–67. (Ch. 6)

Uttley, A. M. (1975). The informon in classical conditioning. *Journal of Theoretical Biology*, *49*, 355–376. (Ch. 6)

Uttley, A. M. (1976a). A two-pathway informon theory of conditioning and adaptive pattern recognition. *Brain Research*, *102*, 23–35. (Ch. 6)

Uttley, A. M. (1976b). Simulation studies of learning in an informon network. *Brain Research*, *102*, 37–53. (Ch. 6)

Uttley, A. M. (1976c). Neurophysiological predictions of a two-pathway informon theory of neural conditioning. *Brain Research*, *102*, 55–70. (Ch. 6)

Van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A., & Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Social Cognitive Affective Neuroscience*, *4*, 294–304. (Ch. 5)

*Van der Stigchel, S., Belopolsky, A. V., Peters, J. C., Wijnen, J. G., Meeter, M., & Theeuwes, J. (2009). The limits of top-down control of visual attention. *Acta Psychologica*, 132, 201–212. (Ch. 5)

Verwey, W. B. (1996). Buffer loading and chunking in sequential keypressing. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 544–562. (Ch. 9)

*Vigdor, B., & Lerner, B. (2007). The Bayesian ARTMAP. *IEEE Transactions on Neural Networks*, *18*, 1628–1644. (Ch. 8)

Vlachos, I., Herry, C., Lüthi, A., Aertsen, A., & Kumar, A. (2011). Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala. *PLoS Computational Biology*, 7(3): e1001104. doi:10.1371/journal.Pcbi.1001104. (Ch. 6)

von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience*, *9*, 1731–1748. (Ch. 5, 8)

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, *14*, 85–100. (Ch. 2, 7)

von der Malsburg, C., & Cowan, J. D. (1982). Outline of a theory for the ontogenesis of iso-orientation domains in visual cortex. *Biological Cybernetics*, *45*, 49–56. (Ch. 7)

*von der Malsburg, C., & Schneider, W. (1986). A neural cocktail party processor. *Biological Cybernetics*, *54*, 29–40. (Ch. 4)

von Neumann, J. (1951). Probabilistic logics and the synthesis of reliable organisms from unreliable components. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior, the Hixon Symposium*. New York: Wiley. (Ch. 3)

Wagar, B., & Thagard, P. (2004). Spiking Phineas Gage: A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review*, *111*, 67–79. (Ch. 9)

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50. (Ch. 1)

Wallach, H. (1976). *On perception*. New York: Quadrangle/New York Times Book Company. (Ch. 4)

Wallas, G. (1926). *The art of thought*. New York: Harcourt, Brace. (Ch. 9)

*Walley, R. E., & Weiden, T. D. (1973). Lateral inhibition and cognitive masking: A neuropsychological theory of attention. *Psychological Review*, *80*, 284–302. (Ch. 4)

Walters, E. T., & Byrne, J. H. (1983). Associative conditioning of single sensory neurons suggests a cellular mechanism for learning. *Science*, *219*, 405–408. (Ch. 6)

Wang, D. (2000). On connectedness: A solution based on oscillatory correlation. *Neural Computation*, *12*, 131–139. (Ch. 9)

Wang, D., Buhmann, J., & Malsburg, C. von der (1990). Pattern segmentation in associative memory. *Neural Computation*, *2*, 94–106. (Ch. 4)

Wang. X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*, 955–968. (Ch. 9)

Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research*, *89*, 233–247. (Ch. 5)

Waters, J., Schaefer, A., & Sakmann, B. (2005). Backpropagating action potentials in neurones: measurement, mechanisms and potential functions. *Progress in Biophysics and Molecular Biology*, *87*, 145–170. (Ch. 8)

Waxman, S. G. (2000). *Correlative neuroanatomy* (24th ed.). New York: Lange Medical Books/McGraw-Hill. (Appendix 2)

Weideman, W. E., Manry, M. T., Yau, H. C., & Gong, W. (1995). Comparisons of a neural network and a nearest-neighbor classifier via the numeric handprint recognition problem. *IEEE Transactions on Neural Networks*, *6*, 1524–1535. (Ch. 8)

*Weidner, R., Boers, F., Mathiak, K., Dammers, J., & Fink, G. R. (2010). The temporal dynamics of the Müller-Lyer illusion. *Cerebral Cortex*, *20*, 1586–1595. (Ch. 5)

Weingard, F. S. (1990). Self-organizing analog fields (SOAF). *International Joint Conference on Neural Networks* (Vol. II, p. 34). Hillsdale, NJ: Lawrence Erlbaum Associates. (Ch. 7)

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Unpublished doctoral dissertation, Harvard University. (Ch. 2, 3, 8).

Werbos, P. J. (1988). Backpropagation: past, present, and future. *IEEE International Conference on Neural Networks* (Vol. I, pp. 343–353). San Diego: IEEE. (Ch. 6, 8)

Werbos, P. J. (1992a). The cytoskeleton: Why it may be crucial to human learning and to neurocontrol. *Nanobiology*, *1*, 75–95. (Ch. 8)

Werbos, P. J. (1992b). Approximate dynamic programming for real-time control and neural modeling. In D. A. White & D. A. Sofge (Eds.), *Handbook of intelligent control: Neural, fuzzy and adaptive approaches* (pp. 493–525). New York: Van Nostrand Reinhold. (Ch. 9)

Werbos, P. J. (1993). *The roots of backpropagation: From ordered derivatives to neural networks and political forecasting*. New York: Wiley. (Ch. 1, 2, 3, 8)

Werbos, P. J. (2003). Backpropagation: General principles and issues for biology. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (2nd ed.). Cambridge, MA: MIT Press. (Ch. 8)

Werbos, P. J. (2009). Intelligence in the brain: A theory of how it works and how to build it. *Neural Networks*, *22*, 200–212. (Ch. 9)

*Werbos, P. J. (2012). Neural networks and the experience and cultivation of mind. *Neural Networks*, *32*, 86–95. (Ch. 9)

*Westermann, G., & Ruh, H. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, *119*, 649–667. (Ch. 9)

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*, 59–85. (Ch. 9)

White, H. (1987). Some asymptotic results for back-propagation. *IEEE First International Conference on Neural Networks* (Vol. III, pp. 261–266). San Diego: IEEE/ICNN. (Ch. 4, 8)

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behavioural Brain Research*, *199*, 119–128. (Ch. 7)

Wickens, J. R., & Kotter, R. (1995). Cellular models of reinforcement. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 187–214). Cambridge, MA: MIT Press. (Ch. 5)

Widrow, B. (1962). Generalization and information storage in networks of adaline neurons. In M. C. Yovits, G. T. Jacobi, & G. D. Goldstein (Eds.), *Self-organizing systems – 1962* (pp. 437–461). Washington, DC: Spartan Books. (Ch. 2, 8)

Widrow, B. (1987). ADALINE and MADALINE – 1963. *IEEE First International Conference on Neural Networks* (Vol. I, pp. 145–157). San Diego: IEEE/ICNN. (Ch. 2)

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (Stanford Electronics Laboratories Tech. Rep. 1553–1), Stanford University, Stanford, CA. (Ch. 2, 3, 8)

*Widrow, B., Pierce, W. H., & Angell, J. B. (1961). Birth, life, and death in microelectronic systems. I*RE Transactions on Military Electronics*, *4*, 191–201. (Ch. 2)

Wiener, N. (1948). *Cybernetics*. New York: Wiley. (Ch. 2)

Wiener, N. (1954). *The human use of human beings*. New York: Avon Books. (Ch. 1)

*Williamson, J. R. (1996). Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps. *Neural Networks*, *9*, 881–897. (Ch. 8)

Willshaw, D. J., & Malsburg, C. von der (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London B*, *194*, 431–445. (Ch. 7)

Willshaw, D. J., & Malsburg, C. von der (1979). A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem. *Philosophical Transactions of the Royal Society of London B*, *287*, 203–243. (Ch. 7)

Wilson, H. R. (1975). A synaptic model for spatial frequency adaptation. *Journal of Theoretical Biology*, *50*, 327–352. (Ch. 2, 7)

Wilson, H. R., & Bergen, J. R. (1979). A four-mechanism model for spatial vision. *Vision Research*, *19*, 19–32. (Ch. 4)

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*, 1–24. (Ch. 4)

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, *13*, 55–80. (Ch. 4, Appendix 1)

*Wokke, M. E., Vandenbroucke, A. R. E., Scholte, H. S., & Lamme, V. A. F. (2013). Confuse your illusion: Feedback to early visual cortex contributes to perceptual completion. *Psychological Science*, *24*, 63–71. (Ch. 5)

Wong, R., & Harth, E. (1973). Stationary states and transients in neural populations. *Journal of Theoretical Biology*, *40*, 77–106. (Ch. 2)

*Wu, R., Scerif, G., Aslin, R. N., Smith, T. J., Nako, R., & Eimer, M. (2013). Searching for something familiar or novel: Top-down attentional selection of specific items or object categories. *Journal of Cognitive Neuroscience*, *25*, 719–729. (Ch. 5)

*Wurbs, J., Mingolla, E., & Yazdanbakhsh, A. (2013). Modeling a space-variant cortical representation for apparent motion. *Journal of Vision*, *13*, ArtID 2. doi:10.1167/13.10.2.

*Xue, C., & Liu, F. (2014). Structured synaptic inhibition has a critical role in multiple-choice motion-discrimination tasks. *Journal of Neuroscience*, *34*, 13444–13457. (Ch. 4)

*Yazdanbakhsh, A., & Grossberg, S. (2004). Fast synchronization of perceptual grouping in laminar visual cortical circuits. Neural Networks, 17, 707–718. (Ch. 9)

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin and Review*, *12*, 387–402. (Ch. 9)

Yechiam, E., Stout, J. C., Busemeyer, J. R., Rock, S. L., & Finn, P. P. (2005). Individual differences in the response to forgone payoffs: An examination of high functioning drug abusers. *Journal of Behavioral Decision Making*, *18*, 97–100. doi:10.1002/bdm.487. (Ch. 9)

Yeo, C. H., Hardiman, M. J., & Glickstein, M. (1985a). Classical conditioning of the nictitating membrane response of the rabbit. I. Lesions of the cerebellar nuclei. *Experimental Brain Research*, *60*, 87–98. (Ch. 5)

Yeo, C. H., Hardiman, M. J., & Glickstein, M. (1985b). Classical conditioning of the nictitating membrane response of the rabbit. II. Lesions of the cerebellar cortex. *Experimental Brain Research*, *60*, 99–113. (Ch. 5)

Yeung, N., & Nieuwenhuis, S. (2009). Dissociating response conflict and error likelihood in anterior cingulate cortex. *Journal of Neuroscience*, *29*, 14506–14510. (Ch. 5)

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. (Ch. 5)

Young, J. Z. (1936). Structures of nerve fibers and synapses in some invertebrates. *Cold Spring Harbor Symposia in Quantitative Biology*, *4*, 1–6. (Appendix 2)

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS ONE*, *2*, e1128. (Ch. 5)

Zak, P. J., Kurzban, R., Ahmadi, S., Swerdloff, R. S., Park, J. et al. (2009). Testosterone administration decreases generosity in the ultimatum game. *PLoS ONE*, *4*, e8330. (Ch. 5)

*Zhang, J. (2009). Adaptive learning via selectionism and Bayesianism. Part I: Connection between the two. *Neural Networks*, *22*, 220–228. (Ch. 6)

Zhao, C., Deng, W., & Gage, F. H. (2008). Mechanisms and functional implications of adult neurogenesis. *Cell*, *132*, 645–660. doi:10.1016/j. Cell.2008.01.033. (Ch. 3)

*Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Frontiers in Psychology*, *4*, August 20, ARTID: 515. (Ch. 8)

AUTHOR INDEX

Page numbers in *italics* indicate items in figures or tables.

SUBJECT INDEX

Pages listed in **bold** are those where the concept is either defined or given an extensive treatment. Pages listed in *italics* indicate figures or tables.