

HOW THE BRAIN WORKS: THE NEXT GREAT SCIENTIFIC REVOLUTION

David Hestenes

Arizona State University, Tempe, AZ 85287

In spite of the enormous complexity of the human brain, there are good reasons to believe that only a few basic principles will be needed to understand how it processes sensory input and controls motor output. In fact, the most important principles may be known already! These principles provide the basis for a definite mathematical theory of learning, memory, and behavior.

1. Introduction

I am here to tell you that another major scientific revolution is well under way, though few scientists are aware of it even in fields where the revolution is taking place. In the past decade, a mathematical theory has emerged that bridges the gap between neurophysiology and psychology, providing penetrating insights into brain mechanisms for learning, memory, motivation, and the organization of behavior. It promises a formulation of fundamental principles of psychology in terms of mathematical laws as precise and potent as Newton's laws in physics. If the current theory is on the right track, then we can expect it to develop at an accelerating pace, and the revolution may be over by the turn of the century. We will then have a coherent mathematical theory of brain mechanisms that can explain a great range of phenomena in psychology, psychophysics, and psychophysiology.

To say that this conceptual revolution in psychology will be over is not to imply that all problems in psychology will be solved. It is merely to assert that the fundamental laws and principles of explanation in psychology will be established. To work out all their implications will be an endless task. So has it been in physics, where the laws of classical and quantum mechanics have been well established for some time, but even the classical theory continues to produce surprises. So has it been with the recent revolution in biology brought about by breaking the genetic code; though some principles of genetic coding are undoubtedly still unknown, the available principles are sufficient to provide the field with a unified theoretical perspective and determine the modes of acceptable explanation. Biology is now regarded as a more mature science than psychology, but we shall see that it may be easier to give psychology a mathematical formulation.

If indeed a conceptual revolution is under way in psychology and the brain sciences, you may wonder why you haven't heard about it before. Why hasn't it been bannered by Psychology Today or proclaimed by some expert on the Johnny Carson show? Why is it announced here for the first time to an audience of mathematicians, physicists, and engineers? Before these questions can be answered, we need to consider the status and interrelations of the relevant scientific disciplines.

2. The Science of Mind and Brain

Let us adopt the term neuroscience for the science of mind and brain. Neuroscience is the most interdisciplinary of all the sciences, and it suffers accordingly. The whole field has been carved into a motley assortment of subdisciplines that rarely communicate with one another. Consequently, most experts in one branch of neuroscience are profoundly ignorant about even closely related branches. Very few have a well grounded perspective on the field as a whole.

The neurosciences have accumulated overwhelming evidence that the characteristics of behavior observed, manipulated, and analyzed by psychologists are derived from the functioning of animal nervous systems or brains. This is the basis for the scientific conception of mind as a function of the

brain, and it justifies regarding psychology as the science of mind. Despite this modern insight, the traditional academic division between psychology and biology perpetuates an artificial separation between mind and brain in research as well as knowledge.

Research on the structural properties of brains is carried out in the well established fields of neurophysiology and neuroanatomy as well as in a variety of related specialties such as electroencephalography. These fields subservise the medical profession, which, for the most part, pays scant attention to research in psychology and psychophysics. Psychophysiology attempts to bridge the mind-brain gap in research and knowledge, but the connections are academically tenuous.

The established disciplines in neuroscience from neurophysiology to psychology are predominantly empirical in content. They have accumulated a vast store of isolated facts about the structure and function of brains, but little in the way of coherent theory. They comprise the empirical component of neuroscience. The theoretical component of neuroscience is developing in the fledgling field of neural modeling. This field has yet to become a recognized academic discipline, so the academic respectability of anyone who works in it is at risk.

One might expect the established neurosciences to encourage and support neural modeling. But the empiricist suspicion of theory in general and mathematical theory in particular is pervasive in these fields. I recently conversed at length with two capable young assistant professors at prominent universities. One was a neurophysiologist, the other a psychologist. Both wished to pursue research in neural modeling, but neither would dare to mention this to his colleagues or even consider beginning such research until he achieves tenure. As a consequence of this pervasive antitheoretical bias in the neurosciences, only a few tenured mavericks, like the physiologist Walter Freemann at Berkeley, have developed the mathematical skills needed for serious neural modeling. Most of the neural modeling is done by mathematically trained outsiders from engineering, mathematics, and physics.

As the name suggests, neural modeling is concerned with developing mathematical models of neurons and their interactions. The modeling proceeds at two levels, the single neuron level and the neural network level. These levels are concerned with different experimental and theoretical techniques, facts, and issues. The single neuron level is the physiological level of neural modeling, for it involves the complex details of cell and membrane physiology, chemistry, and physics. Neural modeling at this level has a measure of respectability in neurophysiological circles owing to the impressive success of Hodgkin and Huxley, who won a Nobel Prize for modeling and measuring the propagation of electrical signals along the axon of a neuron. The famous Hodgkin-Huxley equations are recognized as a paradigm for neural modeling at the neuron level.

The aim of modeling at the network level is to explain the information processing capabilities of macroscopic brain components as collective properties of a system of interacting neurons. This is the psychophysical level of neural modeling. It correlates the electrical activity of neural networks

with their "mental" processing capabilities. At this level the fine physiological details of single neuron dynamics become unimportant.

Network modeling is the theoretical bridge between the microscopic and the macroscopic levels of brain activity, between neurophysiology and psychology. Consequently it is open to objections by empiricists at both ends who care little about connecting the antipodes. Little wonder that network modeling is mostly ignored by the neuroscience establishment! Little wonder that the establishment is unaware of the theoretical revolution taking place in its own fields!

I am especially pleased to tell this audience about the exciting developments in network modeling, because the field is wide open for theoretical exploration, and I doubt that I could find an audience more qualified to contribute. Neural modelers have yet to employ statistical concepts such as entropy with the skill and sophistication of the scientists gathered here. Indeed, I believe that maximum-entropy theory will play its greatest role yet in the neural network theory of the future.

3. Introduction to Grossberg

While neural modeling is ignored by the neuroscience establishment, the neural modelers tend to ignore one another. A more fragmented field would be hard to find. Almost everyone in the field seems to be pushing his or her own model, although a few small groups of interacting modelers have formed. One could read extensively in the literature without discovering that a coherent, general network theory already exists. The theory has been developed over the past two decades by Stephen Grossberg.

Grossberg is by far the most versatile and prolific of the neural modelers. He has written extensively on nearly every aspect of neural modeling. He has elevated the subject from a collection of isolated models to a genuine mathematical theory with a small set of general principles to guide the modeling of any brain component, and he has worked out many specific applications. He has thus produced the first truly coherent theory of learning, perception, and behavior. His theory provides coherent explanations for a wide range of empirical results from psychology, psychophysics, psychophysiology, and even neuropharmacology. And it makes a number of striking new predictions that are yet to be verified. Right or wrong, Grossberg has produced the first mathematical approach to psychology that deserves to be called a theory in the sense that the term "theory" is used in the physical sciences.

In spite of all this, Grossberg has been overlooked or ignored by most of the neural modeling community as well as the neuroscience establishment. Grossberg is seldom referenced by other neural modelers except for an occasional criticism, and he returns the favor. No doubt my remarks have strained your credulity to the limit, so let me try to explain why Grossberg is not more widely appreciated.

Let us first consider why Grossberg's impact on the neuroscience establishment has been so slight. In recent years, Grossberg has attempted to

reach psychologists with several long reviews of his work. Unfortunately, few psychologists have the background needed to understand the mathematical core of his theory. Aware of this fact, Grossberg has pushed the mathematics into the background and presented a detailed qualitative account of this theory. But verbal arguments lack the logical force of mathematics at the most crucial points. Moreover, few psychologists are convinced that neural mechanisms are needed in psychological theory. So Grossberg's attempt to give a coherent account of the diverse phenomena in psychology looks to many psychologists like a series of extravagant claims of credit for every significant result in the field.

For quite different reasons, Grossberg is likely to be quickly dismissed by experts in neurophysiology. These experts are aware of many uncertainties and complexities in neuron dynamics, so they know that Grossberg's equations cannot be empirically justified by current physiological evidence, even though the questions are not inconsistent with the evidence. They look for "bottom-up" justification from physiology, whereas the main justification for Grossberg's equations is "top-down," from psychology. Grossberg's theory suggests constraints on physiological theory to make it consistent with psychological evidence. It therefore provides a guide for physiological research. But the physiologists are not looking for external guidance. And experts often point to the vast mass of partially digested data about brains as evidence that brains are much too complex for simple explanations. So they are skeptical of Grossberg's claim to explain brain functions with a small number of mathematical laws and organizing principles—too skeptical, probably, to give Grossberg the attention he needs in order to be understood.

It is harder to explain why Grossberg has been ignored by other neural modelers, since they share certain basic ideas about neural modeling and its significance. I believe the main reason is that very few have expended the substantial effort required to understand and evaluate Grossberg's work. Grossberg is not an easy read. To be understood, he must be studied—studied for weeks or months, not merely days. Let me cite my own experience by way of example.

In 1976 I made my initial foray into the neural modeling literature and came away with the disappointing impression that the field was hopelessly far from explaining anything important. I was unmoved by the only article of Grossberg's that I came across at the time. A couple of years ago I encountered a former student of mine, Bob Hecht-Nielsen, who spoke enthusiastically about Grossberg's work and told me how he was using it to design practical devices that learn to classify patterns. Since I had great respect for Bob's judgment, I decided to give Grossberg a closer look. Fortunately, Bob gave me sufficient insight into Grossberg's ideas to overcome the difficulties I again found in understanding his writings. Being a theoretical physicist, I had no difficulties at all with the mathematics in Grossberg's articles. But at first I had trouble in orienting myself to the thrust of his work and even in interpreting the variables in his equations. Interpretation and evaluation of Grossberg's equations require some familiarity with empir-

ical data spanning the entire range from neurophysiology to psychology. I had a much stronger background in psychology and psychophysics than most physicists. Nevertheless, I initially had difficulty in Grossberg's papers distinguishing established empirical fact from tentative conjecture or even wild speculation.

I have since reviewed the experimental results relevant to Grossberg's work sufficiently to be confident that his citations of such results are apt and reliable. Indeed, I am impressed by his judgment in selecting results to explain with his theory. I believe his grasp of the entire empirical domain from neurophysiology to psychology is unsurpassed. But I am most impressed by the way he has gone about developing his theory. His theoretical style has all the elements of the best theoretical work in physics—studies of specific mathematical models in search of general principles, emphasis on functional relations without premature commitment to specific functional forms, and a high level of idealization to isolate major relations among variables, followed by successive elaborations to capture more details. Grossberg is especially clever at developing Gedanken experiments to motivate his theoretical constructions.

I don't expect you to accept my assessment of Grossberg at face value. My purpose here is to introduce you to Grossberg and supply some of the background you will need to make your own evaluation, to help you over the initial credibility barrier as Hecht-Nielsen helped me. The recent publication of Grossberg's collected papers [1982] makes it much easier to approach his work now. But I hope to give you a good idea of what you can expect to find there. I will introduce you to basic ideas generally accepted by neural modelers and discuss the distinctive contributions of Grossberg.

4. Empirical Background

Although the human brain is the most complex of all known systems, we need only a few facts about neurons and neuroanatomy to establish an empirical base for neural network theory. Only a brief summary of those facts can be given here. For more extensive background, see Kandel and Schwartz [1981].

The signal processing unit of the nervous system is the nerve cell or neuron. There are, perhaps, a thousand types of neurons, but most of them have in common certain general signal processing characteristics, which we represent as properties of an ideal neuron model. To have a specific example in mind, let us consider the characteristics of an important type of long range signaling neuron called a pyramid cell (schematized in Fig. 1).

(1) The internal state of a neuron is characterized by an electrical potential difference across the cell membrane at the axon hillock (Fig. 1). This potential difference is called the generating potential. External inputs produce deviations in this potential from a baseline resting potential (typically between 70 and 100 mV). When the generating potential exceeds a certain threshold potential, a spike (or action) potential is generated at the hillock and propagates away from the hillock along the axon.

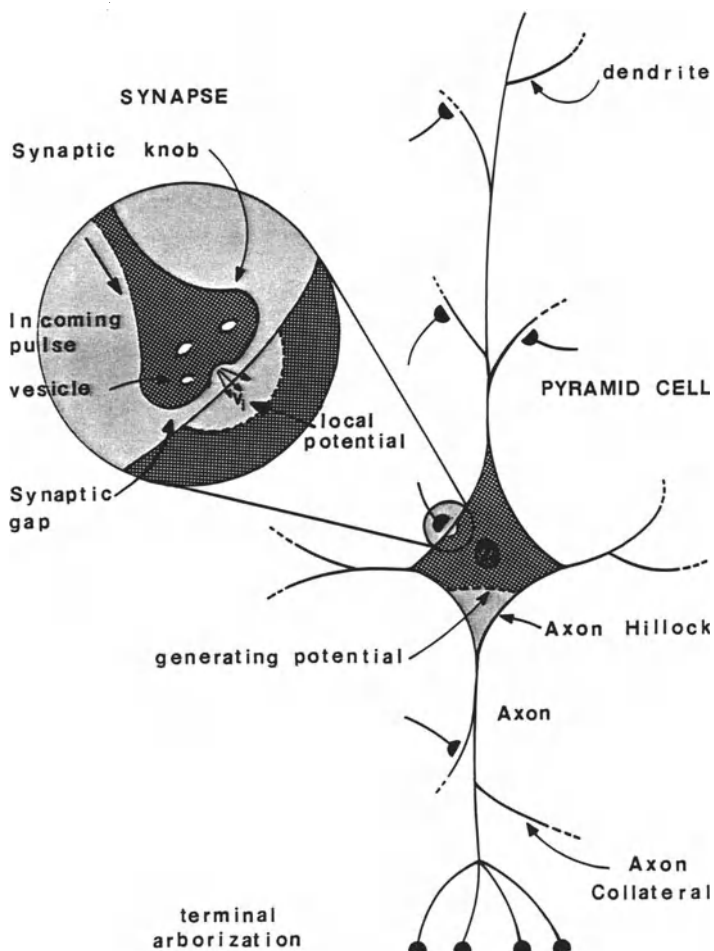


Figure 1. Anatomy of a pyramid cell.

(2) **Axonal signals:** The action potential is a large depolarizing signal (with amplitude up to 110 mV) of brief duration (1 to 10 ms). In a given neuron, every action potential travels with the same constant velocity (typically between 10 and 100 m/s) and undiminished amplitude along all axon collaterals (branches) to their terminal synaptic knobs.

Axonal signals are emitted in bursts of action potentials with pulse frequencies typically in the range between 2 and 400 Hz for cortical pyramid cells or between 2 and 100 Hz for retinal ganglion cells (see below). Single spike potentials are also spontaneously emitted, evidently at random. A single spike is not believed to carry information. It appears that all the information in an axonal signal resides in the pulse frequency of the burst. Thus, the signal can be represented by a positive real number in a limited interval.

(3) Synaptic inputs and outputs: The flow of signals in and out of a neuron is unidirectional. A neuron receives signals from other neurons at points of contact on its dendrites or cell body known as synapses. A typical pyramid cell in the cerebral cortex (see below) receives inputs from about 10^5 different synapses. When an incoming axonal signal reaches the synaptic knob it induces the release of a substance called a neurotransmitter from small storage vesicles. The released transmitter diffuses across the small synaptic gap to the postsynaptic cell, where it alters the local receptor potential across the cell membrane. The synaptic inputs have several important properties:

(a) Quantized transmission: Each spike potential releases (approximately) the same amount of transmitter when it arrives at the synaptic knob.

(b) Temporal summation: Changes in receptor potential induced by successive spike potentials in a burst are additive. Consequently, deviations of the receptor potential from the resting potential depend on the pulse frequency of incoming bursts.

(c) Synaptic inputs are either excitatory or inhibitory, depending on the type of interaction between neurotransmitter and receptor cell membrane. An input is excitatory if it increases the receptor potential or inhibitory if it decreases the receptor potential.

(d) Weighted spatial summation: Input-induced changes in the various receptor potentials of a neuron combine additively to drive a change in the generating potential.

Now let us identify some general anatomical characteristics of the brain to guide our constructions of networks composed of ideal neurons. We can do that by examining a major visual pathway called the geniculostriate system (Fig. 2). Light detected by photoreceptors in the retina drives the production of signals that are transmitted by retinal ganglion cells from the retina to the lateral geniculate nucleus (LGN), which (after some process-

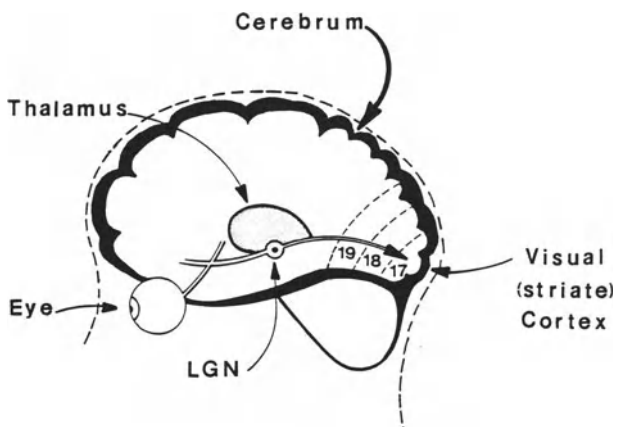


Figure 2. The geniculostriate system.

ing) relays the signal to the visual cortex (also known as the striate cortex or Area 17). From Area 17, signals are transmitted to Areas 18, 19, and other parts of the brain for additional processing.

In Fig. 3 the geniculostriate system is represented as a sequence of three layers of slabs connected by neurons with inputs in one slab and outputs in another slab. There are about 10^6 ganglion cells connecting the retina to the LGN, so we may picture the retina as a screen with 10^6 pixels. Let $I_k(t)$ be the light intensity input detected at the k th pixel at time t . Then the input intensity pattern or image displayed on the retina can be represented as an image vector with $n = 10^6$ components:

$$\mathbf{I}(t) = \{I_1(t), I_2(t), \dots, I_n(t)\} . \quad (1)$$

This vector is filtered (transformed) into a new vector input to the LGN by several mechanisms: (a) the receptive fields (pixel inputs of the ganglion cells) overlap; (b) nearby ganglion cells interact, and (c) the output from each ganglion cell is distributed over the LGN slab by the arborization (branching) of its axon. Later we will see how to model and understand these mechanisms.

Actually, each of the three main slabs in the geniculostriate system is itself composed of several identifiable layers containing a matrix of inter-neurons (neurons with relatively short range interactions). We will take these complexities into account to some extent by generalizing our models

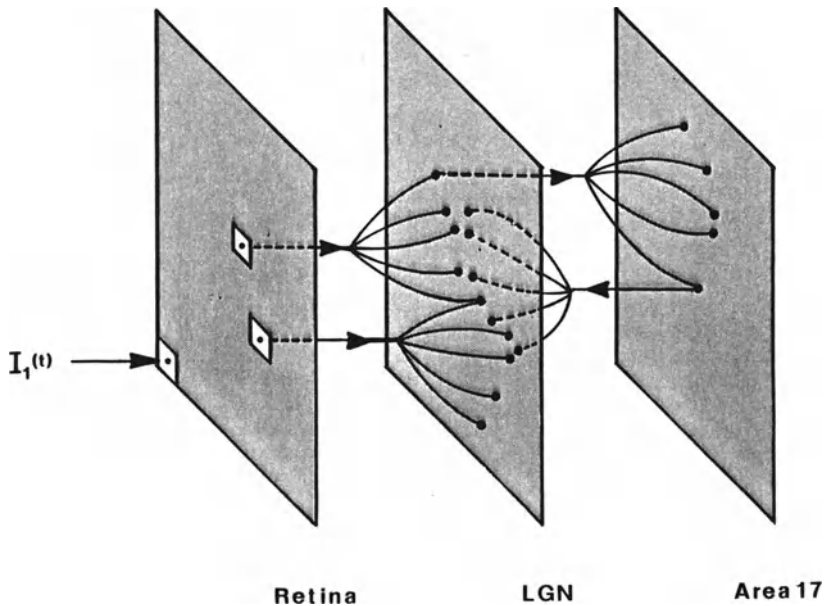


Figure 3. Layered structure of the brain (arrows indicate signal directions).

of slab-connecting neurons. There is little reason to believe that the details we omit from our models will detract from our general conclusions.




The layered structure in Fig. 3 is typical of the organization of major portions of the brain, so we can hope to learn general principles of brain design by understanding its functional significance.

Notice the reciprocal axonal connection from Area 17 back to the LGN. We shall see that the significance of that connection may be especially profound.

5. Neural Variables

We are now prepared to consider Grossberg's general principles for modeling a system of interacting neurons. The i th neuron is represented by a node v_i connected to node v_j by a directed pathway e_{ij} terminating in a synaptic knob N_{ij} as indicated in Table 1. The physical state of each of these three components of a neuron is characterized by a single real-valued state variable, which has a psychological as well as a physiological interpretation. These dual interpretations (mind-brain duality) provide the link between psychology and the brain sciences. The physiological interpretations are fairly evident from what we have said already. We shall see that the psychological interpretations convert the neural network theory into a genuine psychological theory.

Table 1. Neuron Components and Variables

	Node	Directed pathway	Synaptic knob	
Components:	v_i 	e_{ij}	N_{ij} 	v_j 
Variables:	x_i	S_{ij}	z_{ij}	x_j
Name/variable	Physiological interpretation	Psychological interpretation		
Activity x_i	Average generating potential	Stimulus trace or STM trace		
Signal S_{ij}	Average firing frequency	Sampling or performance signal		
Synaptic strength z_{ij}	Transmitter release rate	LTM trace		

The "internal state" of node v_i is described by a variable x_i called the activity of the node. Physiologically, x_i can be interpreted as the deviation of a neuron generating potential from its equilibrium (resting) value. Psychologically, it can be interpreted as a stimulus trace or short term memory (STM) trace. The latter interpretation is especially interesting because it ascribes a definite physical referent for the STM concept. In cognitive psychology, the STM is regarded as some unspecified brain mechanism for the temporary storage of information. For example, a telephone number you have just heard is said to be stored in your STM for a time on the order of 10 seconds, after which it will soon be lost unless you re-store it in STM by rehearsing it or using it in some other way. Note that this way of speaking suggests that the STM is a special component of the brain to which a limited amount of information can be transferred for temporary storage. However, Grossberg claims that STM storage is simply an enhanced activity or activation of neurons somewhere in the brain, different neurons in different places for different concepts stored. He applies the term "STM storage" to any neuron activity x_i that is temporarily maintained at positive values by local feedback loops. Undoubtedly, the cognitive psychologists are unable to probe more than a limited subset of such activated brain states in their studies of STM storage. So Grossberg's STM concept is broader as well as more specific than the conventional STM concept.

A signal propagated along the directed pathway e_{ij} is represented by a nonnegative real variable S_{ij} . In Grossberg's theory S_{ij} is not an independent variable; it is some definite function of the node activity x_i . For the value of the signal when it reaches the synaptic knob we can write

$$S_{ij}(t) = f[x_i(t - \tau_{ij})] b_{ij}, \quad (2)$$

where τ_{ij} is a "time delay constant" and b_{ij} is a "path strength constant" determined by physical properties of the pathway. In general, $f(x_i)$ is a sigmoid function, but for many purposes it can be approximated by the "threshold-linear function"

$$f(x_i) = [x_i - \Gamma_i]^+, \quad (3)$$

where Γ_i is a positive threshold parameter, and $[u]^+ = u$ for $u \geq 0$, $[u]^+ = 0$ for $u < 0$. Thus, the node v_i emits a signal only when its activity x_i exceeds the threshold Γ_i .

The variable S_{ij} [or $f(x_i)$, rather] is to be interpreted physiologically as the average firing frequency of a neuron. Therefore, it does not describe the sudden signal fluctuations in the bursts that are observed experimentally. Because those fluctuations are not believed to carry information, it is reasonable to suppress them in a theory that aims to characterize the information content of neuronal processes.

Psychologically, the variable S_{ij} may be interpreted as an information sampling signal when it is concerned with information input, or as a performance signal when it is concerned with output. We shall see that either

of these interpretations might apply to a signal in a given pathway, depending on the state of the rest of the network.

The coupling of a synaptic knob N_{ij} to a postsynaptic node v_j is characterized by a positive real variable z_{ij} called the synaptic strength. Physiologically, it can be interpreted as the average rate of neurotransmitter release (per unit signal input) at the knob N_{ij} . This interpretation is tentative, however, because the biochemical processes at a synapse are complex and incompletely understood. In any case, a single variable should be sufficient to characterize the signal transmission rate across a synapse, whatever the underlying processes.

We shall see that the variable z_{ij} can be interpreted psychologically as a long term memory (LTM) trace. This is to assert that the long term storage of information in a brain takes place at the synapses, and that learning is a biochemical change of synaptic states. Although one cannot claim yet that this assertion is an established fact, there is considerable evidence to support it, much more than can be mustered to support any alternative hypothesis about the physiological basis for learning and memory. As a working hypothesis, our dual interpretations of the synaptic strength variable has immense implications for psychology. To begin with, we shall see that it has much to tell us about the way brains encode information.

6. Network Field Equations

Having identified the significant components of a neural network and appropriate variables to represent their properties, to complete the formulation of a network theory we need to postulate laws of interaction and equations of motion for the variables. The main facts and hypotheses about neurons that we have already mentioned are accounted for by Grossberg's field equations for a neural network with n nodes:

$$\dot{x}_i = -A_i x_i + \sum_{k=1}^n S_{ki} z_{ki} - \sum_{k=1}^n C_{ki} + I_i(t) \quad (4)$$

$$\dot{z}_{ij} = -B_{ij} z_{ij} + S_{ij} [x_j]^+ , \quad (5)$$

where the overdot denotes a time derivative and $i, j = 1, 2, \dots, n$.

Grossberg's equations are generic laws of neural network theory in the same sense that $F = ma$ is a generic law of Newtonian mechanics. To construct a mathematical model in mechanics from $F = ma$, one must introduce a force law that specifies the function form of F . Similarly, to construct a definite network model from Grossberg's equations, one must introduce laws of interaction that specify the functional dependence of the quantities A_i , S_{ki} , C_{ki} , B_{ij} , and S_{ij} on the network variables x_i and z_{ij} . To investigate these laws of interaction is a major research program within the context of

Grossberg's theory, just as a theoretical and experimental investigation of the force laws "realized" in nature has been a major research program in physics since it was initiated by Newton.

Before examining specific interactions, we should be clear about the general import of Grossberg's equations. Let us refer to Eq. (4) as the activity equation of node v_i . The right-hand side of this equation describes interactions of the node or, if you prefer, inputs to the node. The first thing to note about the equation is the additivity of inputs from different sources, which represents the basic experimental fact of the spatial summation of synaptic inputs to a neuron. The term $I_i(t)$ represents input from sources outside the network, usually some other neurons but sometimes sensory transducers such as photoreceptors in the retina. The other terms represent internal interactions within the network.

The term $-A_i x_i$ characterizes self-interactions of the node v_i . In the simplest case when the node represents a single neuron, A_i is a positive constant, so the term represents the passive decay that is inevitable in a dissipative system. More generally, it will often be convenient to use a single node to represent a lumped subsystem or pool of interneurons coupled to a single output neuron such as a cortical pyramid cell. The pool can be designed with feedback to make the node capable of STM and more complex responses to inputs. The net result can be described simply by making the decay coefficient A_i into some function $A_i = A_i(x_i)$ of the activity x_i . It is reasonable to suppose that, on account of the additivity of interactions, a neuron pool will have an activity equation of the same general form as that of a single neuron, if the time scale for integrating inputs to the pool is sufficiently short. As the theory develops, it should become possible to replace such reasonable assumptions by rigorous "lumping theorems."

For a neuron pool, the activity x_i of the output neuron may be proportional to the number of excited interneurons in a subpopulation of the pool. In that case, it may be more useful to interpret x_i as the number of excited states in the pool rather than as the potential of a single neuron, especially when characterizing the self-interaction properties of the pool.

The term $S_{ki} z_{ki}$ describes an excitatory node-node interaction as indicated by the plus sign preceding it in the activity equation (4). It expresses the effect of node v_k on node v_i mediated by the signal S_{ki} as given by Eq. (2). The synaptic strength z_{ki} plays the role of a variable coupling constant in the activity equation. Typically the time variation of z_{ki} is slow compared to that of x_i . In many cases z_{ki} is essentially constant, and we say that the connection from v_k to v_i is hardwired. This includes the common case when there is no direct connection from v_k to v_i if we regard it as a case with $z_{ki} = 0$. The multiplicative form of the interaction $S_{ki} z_{ki}$ expresses the temporal summation of synaptic inputs. Grossberg describes the role of the synapse by saying that " z_{ki} gates the signal S_{ki} ."

The term C_{ki} in the activity equation (4) describes an inhibitory node-node interaction as indicated by the minus sign preceding it, supplemented by the assumption that $C_{ki} \geq 0$. We can interpret C_{ki} as a signal function similar to S_{ki} . The symmetry as well as the generality of the activity equa-

tion could be increased by allowing for a variable inhibitory coupling constant. However, the available evidence suggests that inhibitory connections are usually (if not invariably) hardwired. So we restrict our considerations to that case and incorporate the fixed coupling constant into the signal function C_{kj} .

From a general theoretical perspective it is crucial to realize that inhibitory interactions are essential for the stability of the network activity equations, just as attractive forces are essential for bound systems in physics. This can be proved as a mathematical theorem of great generality, and we shall see its significance later on in Grossberg's solution to the noise-saturation problem. In spite of the fact that this has been known to some neural modelers for a long time, papers that are flawed by a failure to take stability requirements into account are still being published.

Now let us turn to Eq. (5) and refer to it as the learning equation in anticipation of support for our interpretation of z_{ij} as an LTM trace. We can identify $S_{ij}^+[x_j]^+$ as the learning term in the equation since it drives an increase in z_{ij} . The factor S_{ij}^+ is a nonnegative signal similar to the signal S_{ij} in the activity equation, but possibly differing in its dependence on the presynaptic activity x_i , owing to details of the underlying biochemical processes. The multiplicative form of the learning term implies that learning takes place only when the presynaptic learning signal S_{ij}^+ and the postsynaptic activity x_j are simultaneously positive. Thus, learning at a synapse is driven by correlations between presynaptic and postsynaptic activities.

The term $-B_{ij}z_{ij}$ can be regarded as a forgetting term, describing passive memory decay when B_{ij} is a positive constant. By allowing B_{ij} to be a more general function of the network parameters, Grossberg allows for the possibility of modulating memory loss.

For the case of constant $B_{ij} = \beta$, the learning equation (5) can be given the enlightening integral form

$$z_{ij}(t) = e^{-\beta t} z_{ij}(0) + \int_0^t e^{-\beta(t-\tau)} S_{ij}^+(\tau) [x_j(\tau)]^+ d\tau. \quad (6)$$

This is an integral equation rather than a solution of Eq. (5) because $x_j(\tau)$ depends on $z_{ij}(\tau)$ in the activity equation (4). However, it shows that in the long run the initial STM trace $z_{ij}(0)$ is forgotten and $z_{ij}(t)$ is given by a time correlation function of the presynaptic signal S_{ij}^+ with the postsynaptic activity x_j .

Special cases and variants of Grossberg's equations have been formulated and employed independently by many neural modelers. But Grossberg has gone far beyond anyone else in systematically analyzing the implications of such equations. In doing so, he has transformed the study of isolated ad hoc neural models into a systematic theory of neural networks. But it will be helpful to know more about the empirical status of the learning equation before we survey the main results of Grossberg's theory.

7. Hebb's Law

From the viewpoint of neural network theory, Hebb's law is the fundamental psychophysical law of associative learning. It should be regarded, therefore, as a basic law of psychology. However, most psychologists have never heard of it because they do not concern themselves with the neural substrates of learning.

Hebb [1949] first formulated the law as follows: "If neuron A repeatedly contributes to the firing of neuron B, then A's efficiency in firing B increases." The psychological import of this law comes from interpreting it as the neural basis for Pavlovian (classical) learning. To see how, recall Pavlov's famous conditioning experiment. When a dog is presented with food, it salivates. When the dog hears a bell, it does not salivate initially. But after hearing the bell simultaneously with the presentation of food on several consecutive occasions, the dog is subsequently found to salivate when it hears the bell alone. To describe the experiment in more general terms, when a conditioned stimulus (CS) (such as a bell) is repeatedly paired with an unconditioned stimulus (UCS) (such as food) that evokes an unconditioned response (UCR) (such as salivation), the CS gradually acquires the ability to evoke the UCR.

To interpret this in the simplest possible neural terms, consider Fig. 4. Suppose the firing of neuron B produces the UCR output, and suppose the UCS input fires neuron C, which is coupled to B with sufficient strength to make B fire. Now if a CS stimulates neuron A to fire simultaneously with neuron B, then, in accordance with Hebb's law, the coupling strength z_{AB} between neurons A and B increases to the point where A has the capacity to fire B without the help of C. In actuality, of course, there must be many neurons of types A, B, and C involved in the learning and controlling of a molar behavioral response to a molar stimulus, but our reduction to the interaction of just three neurons assumes that the learning actually takes place at the synaptic level.

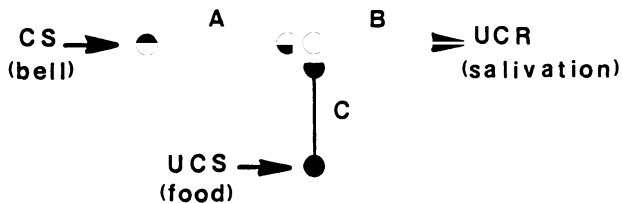


Figure 4. A neural interpretation of Pavlovian learning.

Thus, the molar association strength between stimulus and response that psychologists infer from their experiments is a crude measure of the synaptic coupling strength between neurons in the central nervous system (CNS). The same can be said about all associations among ideas and actions. Thus,

the full import of Hebb's law is this: All associative (long term) memory resides in synaptic connections of the CNS, and all learning consists of changes in synaptic coupling strengths.

This is a strong statement indeed! Although it is far from a proven fact, it is certainly an exciting working hypothesis, and it provides a central theme for research in all the neurosciences. It tells us where to look for explanations of learning and memory. It invites us to do neural modeling.

Ironically, cognitive psychologists frequently dismiss Pavlovian learning as too trivial to be significant in human learning. But Hebb's law tells us that Pavlovian learning is simply an amplified form of the basic neural process underlying all learning. Neural modeling has already advanced far enough to give us good reasons for believing that the most complex cognitive processes will ultimately be explained in terms of the simple neural mechanisms operating in the Pavlovian case.

Direct physiological verification of the synaptic plasticity required by Hebb's law has been slow in coming because the experimental difficulties are extremely subtle and complex. Peripheral connections in the CNS are most easily studied, but they appear to be hardwired as one would expect because the delicate plastic synapses must be protected from destructive external fluctuations. Though some limited experimental evidence for synaptic plasticity exists, there are still considerable uncertainties about the underlying physiological mechanism. There are still doubts as to whether plasticity is due to a pre- or postsynaptic process, though a postsynaptic process is most likely [Stent, 1973].

Considering the experimental uncertainties, many neuroscientists are reluctant to take Hebb's law seriously. They fail to realize that the best evidence for Hebb's law is indirect and theory dependent. Hebb's law should be taken seriously because it is the only available theoretical construct that provides plausible, coherent explanations for psychological facts about learning and memory. Indeed, that is what led Hebb to the idea in the first place. Empiricists may regard such inverse arguments from evidence to theory as unconvincing or even inadmissible, but history shows that inverse arguments have produced the most profound advances in physics, beginning, perhaps, with Newton's law of gravitation. As an example with many parallels to the present case, recall that the modern concept of an atom was developed by a series of inverse arguments to explain observable macroscopic properties of matter. From macroscopic evidence alone, remarkably detailed models of atoms were constructed before they could be tested in experiments at the atomic level. Similarly, we should be able to infer a lot about neural networks from the rich and disorderly store of macroscopic data in psychology. Of course, we should not fail to take into account the available microscopic data about neural structures.

Hebb's original formulation of the associative learning law is too general to have detailed macroscopic implications until it is incorporated in a definite network theory. Grossberg has given Hebb's law a mathematical formulation in his learning equation (5). Hebb himself was not in position to do that, if only because the necessary information about axonal signals was not

available. Of course, Hebb's formulation is vague enough to admit many different mathematical realizations. Grossberg has chosen the simplest realization compatible with the known physiological facts. No doubt Grossberg's law of synaptic plasticity is a crude description of real synaptic activity, but it may be sufficient for the purposes of network theory. In any case, it is good research strategy to study the simplest models first. Recall that Lorentz's classic theory of optical dispersion is based on an electric dipole oscillator model of an atom. The dipole is hardly more than a caricature of a real atom, but Lorentz's dispersion theory was so successful that it is still used today, and failures of the theory were important clues in the development of quantum theory. Grossberg's theory may not characterize neurons any better than a classic dipole characterizes an atom, but it may nevertheless have great success, and any clear failures will be important clues to a better theory. That's how theories progress.

8. The Outstar Learning Theorem

There is one obvious property of neurons that we have not yet incorporated into the network theory, and that is the treelike structure of an axon. What are its implications for information processing? Grossberg's "outstar theorem" shows that the implications are as profound as they are simple.

Consider a slab of noninteracting nodes $V = \{v_1, v_2, \dots, v_n\}$ with a time varying input image $I(t) = [I_1(t), I_2(t), \dots, I_n(t)]$ that drives the nodes above signal threshold. The total intensity of the input is $I = \sum_k I_k$, so we can write $I_k = \theta_k I$ where $\sum_k \theta_k = 1$. Thus, the input has a reflectance pattern $\theta(t) = [\theta_1(t), \theta_2(t), \dots, \theta_n(t)]$.

Now consider a node v_0 with pathways to the slab as shown in Fig. 5. This configuration is called an outstar because it can be redrawn in the symmetrical form of Fig. 6. When an "event" $I_0(t)$ drives v_0 above threshold, a

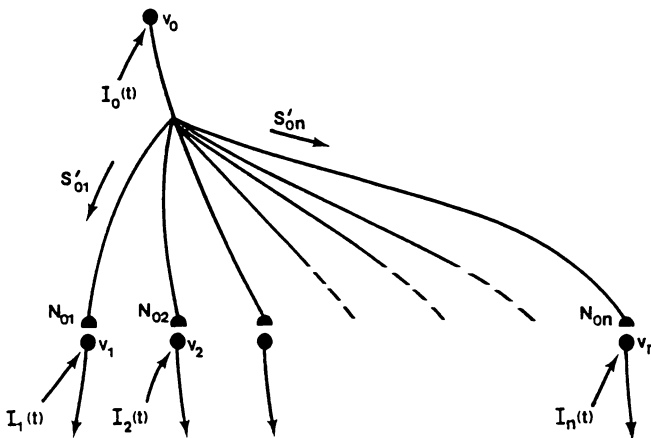


Figure 5. The outstar is the minimal network capable of associative learning.

learning signal S_{0k} is sent to each of the synaptic knobs to trigger a sampling of the activity pattern "displayed" on the slab by driving changes in the synaptic strengths z_{0k} .

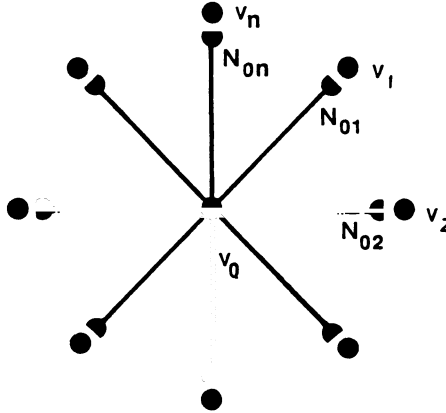


Figure 6. Symmetry of the outstar anatomy.

The outstar learning theorem says that an outstar learns a weighted average of reflectance patterns $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2, \dots, \bar{\theta}_n)$ displayed on the slab in the sense that

$$Z_{0k}(t) \xrightarrow{t \rightarrow \infty} \bar{\theta}_k \tag{7}$$

where

$$Z_{0k} = z_{0k} \left(\sum_k z_{0k} \right)^{-1} . \tag{8}$$

Grossberg calls the Z_{0k} "stimulus sampling probabilities" to emphasize the statistical aspect of the learning process.

The outstar has truly learned the pattern $\bar{\theta}$ in the sense that it can recall the pattern exactly in the following way. Suppose that, at some time after learning, the external slab input $I(t)$ vanishes but $I_0(t)$ is sufficient to stimulate signals S_{0k} from v_0 . The signals S_{0k} read out an activity pattern on the slab that is proportional to the synaptic strengths z_{0k} and hence to the pattern θ_k . When the S_{0k} are sufficiently strong by themselves to drive the slab above threshold, they are called performance signals.

It will be recognized that outstar learning is an instance of Pavlovian learning, where v_0 corresponds to neuron A in Fig. 4 and instead of a single neuron B we have a whole slab of neurons. Accordingly, we can interpret the slab output as the UCR controlled by a UCS input I . When the CS input I_0 is synchronized with the UCS input I , the outstar gradually gains control over the UCR.

To see how the outstar theorem follows from Grossberg's equations, let us consider the simplest case, where the signals $S'_{0k} = S'_0$ are the same for all pathways, and all nodes have identical constant self-interaction coefficients. Then the outstar network equations are

$$\dot{x}_0(t) = -\alpha x_0(t) + I_0(t) \quad (9a)$$

$$\dot{x}_k(t) = -\alpha x_k(t) + S_0(t) z_{0k}(t) + I_k(t) \quad (9b)$$

$$\dot{z}_{0k}(t) = -\beta z_{0k} + S'_0(t) x_k(t) . \quad (9c)$$

The internal processes in a neuron are comparatively fast, so it is often a good approximation to suppose that the relaxation time α^{-1} is short compared to the time variations of the input $I_k(t)$. Then we can use the equilibrium solutions of the activity equations (9b). Therefore, for $S_0 = 0$, Eq. (9b) gives us

$$x_k(t) = \alpha^{-1} I_k(t) = \frac{I(t)}{\alpha} \theta_k(t) . \quad (10)$$

Thus, the activity pattern across the slab is proportional to the reflectance pattern.

Suppose now that the same reflectance pattern θ is repeatedly presented to the slab, so the θ_k are constant but the intensity $I(t)$ may vary wildly in time. For the sake of simplicity, suppose also that the signal S_0 does not significantly perturb the activity pattern across the slab. Then, the integral of the learning equation (9c) has the form of Eq. (6), and gives the asymptotic result

$$z_{0k}(t) = N(t) \theta_k \quad (11a)$$

where

$$N(t) = \int_0^t d\tau e^{-\beta(t-\tau)} S'_0(\tau) \alpha^{-1} I(\tau) . \quad (11b)$$

According to Eq. (11a), the outstar learns the reflectance pattern exactly. The same result is obtained with more mathematical effort even when perturbations of the slab activity pattern are taken into account.

Note that, according to Eq. (11b), the magnitude of $N(t)$, and therefore the rate of learning, is controlled by the magnitudes of the sampling signal $S'_0(t)$ and the total input intensity $I(t)$. Stronger signals, faster learning!

If the θ_k are not constant, the learning equation will still give an asymptotic result of the form (11a) if θ_k is replaced by a suitably defined average θ_k . To see this in the simplest way, suppose that two different but constant patterns $\theta^{(1)}$ and $\theta^{(2)}$ are sampled by the outstar at different times. Then the additivity property of the integral in Eq. (6) allows us to write

$$z_{0k} = \overline{N\theta}_k = N_1\theta_k^{(1)} + N_2\theta_k^{(2)}, \quad (12)$$

where N_1 and N_2 have the same form as N in Eq. (11b), except that the integration is only over the time intervals when $\theta_k^{(1)}$ (or $\theta_k^{(2)}$) is displayed on the slab. Thus, the pattern θ stored in the outstar LTM is a weighted average of sampled patterns. Note that this is a consequence of formulating Hebb's law specifically as a correlation function.

Outstar learning has a number of familiar characteristics of human learning. If the outstar is supposed to learn pattern $\theta^{(1)}$, and $\theta^{(2)}$ is an error, then according to Eq. (12), repeated sampling of $\theta^{(1)}$ will increase the weight (N_1/N) of $\theta^{(1)}$ and $\overline{\theta} \rightarrow \theta^{(1)}$. Thus outstar learning is "error correcting" and "practice makes perfect." Or if $\theta^{(1)}$ is presented and sampled with sufficient intensity, the outstar can learn $\theta^{(1)}$ on a single trial. Thus, "memory without practice" is possible. On the other hand, if an outstar repeatedly samples new patterns, then the weight of the original pattern decreases. So we have an "interference theory of forgetting." Forgetting is not merely the result of passive decay of LTM traces.

Grossberg has proved the outstar theorem for much more general functional forms of the network equations than we have considered. For purposes of the general theory, it is of great importance to determine the variations in physical parameters and network design that are compatible with the pattern learning of the outstar theorem.

The outstar theorem is the most fundamental result of network theory because it tells us precisely what kind of information is encoded in LTM, namely, reflectance patterns. It goes well beyond Hebb's original insight in telling us that a single synaptic strength z_{0k} has no definite information content. All LTM information is in the pattern of relative synaptic strengths z_{0k} defined by Eq. (8).

To sum up, we have learned that the outstar network has the following fundamental properties:

- o Coding: The functional unit of LTM is a spatial pattern.
- o Learning: A pattern is encoded by a stimulus sampling signal.
- o Recall: Probed read-out of LTM into STM by a performance signal.
- o Factorization: A pattern is factored into a reflectance pattern, which is stored, and the total intensity I , which controls the rate of learning.

The outstar is a universal learning device. Grossberg has shown how to use this one device to model every kind of learning in the CNS. The kinds of learning are distinguished by the different interpretations given to the components of the outstar network, which, in turn, depend on how the components fit into the CNS as a whole. To appreciate the versatility of the outstar, let us briefly consider three examples of major importance:

Top-down expectancy learning: Suppose the slab $V = \{v_1, v_2, \dots, v_n\}$ represents a system of sensory feature detectors in the cerebral cortex. A visual (or auditory) event is encoded as an activity pattern $x = (x_1, x_2, \dots, x_n)$ across the slab, where x_k represents the relative importance of the k th feature. The outstar node v_0 can learn this pattern, and when the LTM pattern is played back on V , it represents a prior expectancy of a sensory event.

Motor learning and control. Suppose the slab V represents a system of motor control cells, so each v_k excites a particular muscle group and the activity x_k determines its rate of contraction. Then the outstar command node v_0 can learn to control the synchronous performance of a particular motion with factored rate control modulated by the strength of the performance signal.

Temporal order encoded as spatial order. If the slab V represents a sequence of codes for items on a list (such as a phone number) and the relative magnitudes of the activities x_k reflect the temporal order of the items, then the outstar node v_0 can learn the temporal order.

9. The Network Modeling Game

Once the outstar is recognized as the fundamental device for learning and memory, it becomes evident that information is represented by spatial patterns in the CNS. Information is encoded in STM activity patterns and stored in LTM synaptic strength patterns. Information is processed by filtering, combining, and comparing patterns with patterns and more patterns. Thus, we come to formulate the Network Modeling Game as follows: Explain all learning and behavior with modular network models composed of outstars. The term "behavior" is to be understood here in the broadest sense of any output pattern. The emphasis on outstars is an important refinement of the game introduced by Grossberg.

The aim of the Network Modeling Game is to reduce psychology to a theory of neural mechanisms. Grossberg has been playing the game for a long time and has built up an impressive record of victories. He plays the game systematically by formulating a sequence of design problems for neural networks. For each design problem he finds a minimal solution that provides a design principle for constructing a network with some specific pattern processing capability. He has already developed too many design principles for us to review them all here. But we will look at a few of the most basic design problems and solutions to see how the game gets started.

We will consider the following network design problems:

- o Pattern registration. Considering the limited dynamic range of a neuron, how can a slab be designed to register a well defined pattern if the input fluctuates wildly?
- o STM storage. How can an evanescent input be stored temporarily as an activity pattern for further processing or transfer to permanent LTM storage?
- o Code development. How can a network learn to identify common features and classify different patterns?
- o Code protection and error correction. How can a network continue to learn new information without destroying what it has already learned?
- o Pattern selection. How does a network decide what is worth learning?

9.1. Pattern Registration and STM Storage

A neuron has a limited dynamic range; its sensitivity is limited by noise at low activity and by saturation at high activity. How, then, can neurons maintain sensitivity to wide variations in input intensities? Grossberg calls this the noise-saturation dilemma, and he notes that it is a universal problem that must be solved by every biological system. Therefore, by studying the class of neural networks that solve this problem, we may expect to discover a universal principle of network design. Grossberg has attacked this problem systematically, beginning with a determination of the simplest possible solution.

Grossberg has proved that the minimal network solving the noise-saturation dilemma is a slab of nodes with activity equations

$$\dot{x}_i = -Ax_i + (B - x_i) I_i - (x_i + C) \sum_{k \neq i} I_k, \quad (13)$$

where $i = 1, 2, \dots, n$, and A, B are positive constants while C may be zero or a positive constant. It is easy to see that the form of Eq. (13) limits solutions to the finite range $B \geq x_i(t) \geq -C$, whatever the values of inputs $I_k(t)$.

Two things should be noted about the form of Eq. (13). First, the excitatory input I_i to each node v_i is also fed to the other nodes as an inhibitory input as indicated in Fig. 7. A slab with this kind of external interaction is called an on-center off-surround network. Second, both excitatory and inhibitory interactions include terms of two types, one of the form $B I_i$ and the other of the form $x_i I_k$. Grossberg calls the first type an additive interaction and the second type a shunting interaction. Accordingly, he calls a network characterized by Eq. (13) a shunting on-center off-surround network.

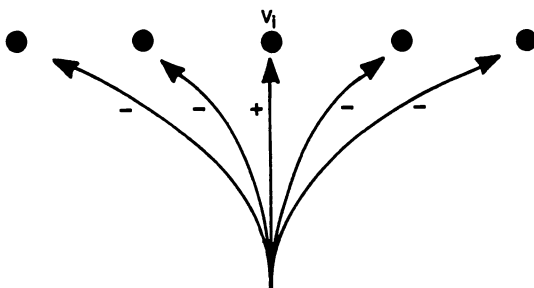


Figure 7. A nonrecurrent on-center off-surround anatomy.

This network can be regarded as a simple model of a retina. The existence of on-center off-surround interactions in the retina is experimentally well established in animals ranging from the mammals to the primitive horse-

shoe crab. The nodes in our model correspond roughly to the retinal ganglion cells with outputs in the LGN as we noted in Sec. 4. The distribution of inhibitory inputs among the nodes is biologically accomplished by a layer of interneurons in the retina called horizontal cells. Our model lumps these interneurons together with the ganglion cells in the nodes. It describes the signal processing function of these interneurons without including unnecessary details about how this function is biologically realized.

It is well known among neuroscientists that intensity boundaries in the image input to an on-center off-surround network are contrast-enhanced in the output. So it is often concluded that the biological function of such a network is contrast enhancement of images. But Grossberg has identified a more fundamental biological function, namely, to solve the noise-saturation dilemma. However, shunting interactions are also an essential ingredient of the solution. Many neural modelers still do not recognize this crucial role of shunting interactions and deal exclusively with additive interactions. It should be mentioned that the combination of shunting and additive interactions appears in cell membrane equations that have considerable empirical support.

To see how Eq. (13) solves the noise saturation dilemma, we look at the steady-state solution, which can be put into the form

$$x_i = \frac{(B + C)I}{A + I} \left(\theta_i - \frac{C}{B + C} \right). \quad (14)$$

Here, as before, the θ_i are the reflectances and I is the total intensity of the input. This solution shows that the slab has the following important image processing capabilities:

(1) Factorization of pattern and intensity. Hence, for any variations in the intensity I , the activities x_i remain proportional to the reflectances θ_i [displaced by the small constant $C(B + C)^{-1}$]. Thus, the slab possesses automatic gain control, and the x_i are never saturated by large inputs.

This factorization matches perfectly with the outstar factorization property, though they have completely different physical origins. Since an outstar learns a reflectance pattern only, it needs to sample from an image slab that displays reflectance patterns with fidelity. Thus, we have here a minimal solution to the pattern registration problem.

(2) Weber law modulation by the factor $I(A + I)^{-1}$. Weber's law is an empirical law of psychophysics that has been found to hold in a wide variety of sensory phenomena. It is usually given in the form $\Delta I/I = \text{constant}$, where ΔI is the "just noticeable difference" in intensity observed against a background intensity I . This form of Weber's law can be derived from Eq. (14) with a reasonable approximation.

(3) Featural noise suppression. The constant $C(B + C)^{-1}$ in Eq. (14) describes an adaptation level, and $B \gg C$ in vivo. The adaptation level $C(B + C)^{-1} = n^{-1}$ is especially significant, for then if $\theta_i = n^{-1}$, one gets $x_i = 0$. In other words, the response to a perfectly uniform input is com-

pletely quenched. It has, in fact, been experimentally verified that when the human retina is exposed to a perfectly uniform illumination, the subject suddenly sees nothing but black. Evidently, the human visual system detects only deviations from a uniform background.

(4) Normalization. From Eq. (14), the total activity is given by $x = \sum_k x_k = [B - (n-1)C](A+1)^{-1}$, which is independent of n if $C = 0$ or $C(B+C)^{-1} = n^{-1}$. In that case, the total activity has an upper bound that is independent of n and l , and we say that the activity is normalized. For some purposes it is convenient to interpret a normalized activity pattern as a probability distribution.

A slab with the pattern processing capabilities that we have just mentioned can perform a variety of different functions as a module in a larger network. It can be used, for example, to compare patterns. Two distinct inputs I_i and J_j produce a composite input $I_i + J_j$. If the patterns mismatch, then the peaks of one pattern will tend to fall on the troughs of the other and the composite will tend to be uniform and so be quenched by the noise saturation property. But if the patterns match, they will be amplified. Indeed, if they match perfectly, then Eq. (14) gives

$$x_i = \frac{(B+C)(I+J)}{A+I+J} \left(\theta_i - \frac{C}{B+C} \right), \quad (15)$$

where I and J are the intensities of the two patterns. The slab output, quenched or amplified, amounts to a decision as to whether the patterns are the same or different. This raises questions about the criteria for pattern equivalence that can be answered by more elaborate network designs. Note that this competitive pattern matching device compares reflectance patterns rather than intensity patterns.

From his analysis of the minimal solution to the noise-saturation dilemma, Grossberg identifies the use of shunting on-center off-surround interactions as a general design principle solving the problem of accurate pattern registration. He then employs this principle to construct more general solutions with additional pattern processing capabilities. He introduces recurrent (feedback) interactions to give the slab an STM storage capability. By introducing distance-dependent interactions he gives the network edge-enhancement capabilities. Beyond this, he develops general theorems about the form of interactions that produce stable solutions of the network equations.

With the general network design principle for accurate pattern registration in hand, we are prepared to appreciate how Grossberg uses it to solve the STM storage problem. The main idea is to introduce interactions among the nodes in a slab in a way that is compatible with accurate pattern registration. Accordingly, we introduce an excitatory self-interaction (on-center) for each node and inhibitory lateral interactions (off-surround), as indicated in Fig. 8. Such a network is said to be recurrent because some of its output is fed back as input.

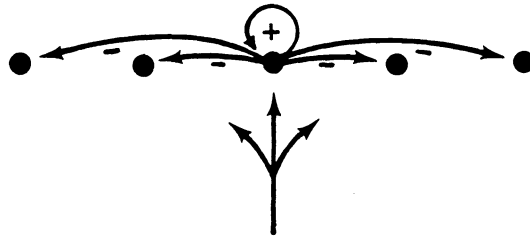


Figure 8. A recurrent on-center off-surround anatomy.

The minimal solution to the STM storage problem is given by the network equations

$$\dot{x}_i = -Ax_i + (B - x_i)[I_i + f(x_i)] - (x_i + C)[J_i + \sum_{k \neq i} f(x_k)] , \quad (16)$$

where $I_i \geq 0$ and $J_i \geq 0$ are excitatory and inhibitory inputs respectively, and $f(x_k)$ is the signal from the k th node. Grossberg has systematically classified signal functions $f(x)$ according to the pattern processing characteristics they give to the network. His main result is that a sigmoid form for the signal function is essential for stability of the network. In spite of this definitive result and its important implications, many modelers continue to work exclusively with linear feedback models.

Besides solving the STM storage problem, the network equations (16) endow the slab with additional pattern processing capabilities. Grossberg has proved that the sigmoid signal function has a definite quenching threshold (QT). Activities below the QT are quenched while those above the QT are sustained. Moreover, the network can easily be given a variable QT controlled by external parameters acting on the interneurons that carry the feedback. This improves the pattern-matching capabilities of the slab, which we have already mentioned. The variable QT provides a tunable criterion for pattern equivalence that can be used as a partial pattern-matching mechanism.

9.2. Pattern Classification and Code Development

Now that we know how to hardwire a slab for accurate pattern registration and STM storage, we are ready to connect slabs into larger networks capable of global pattern processing. Here we shall see how to design a two-slab network that can learn to recognize similar patterns and distinguish between different patterns. This network is a self-organizing system capable of extracting common elements from the time-varying sequence of input patterns and developing its own pattern classification code. Thus, the

network learns from experience. No programmer is necessary. We call such a network an adaptive pattern classifier.

The simplest version of an adaptive pattern classifier is a feed-forward network composed of two slabs with the specifications listed in Fig. 9. To emphasize their functions, let's refer to slabs S_1 and S_2 as the image slab and feature slab, respectively. As indicated in Fig. 9, each node in the image slab is connected to the nodes in the feature slab by pathways terminating in plastic (modifiable) synapses. The configuration is identical to that of the outstar discussed earlier, but we shall not refer to it as an outstar, because its function is not to learn patterns, as we shall see.

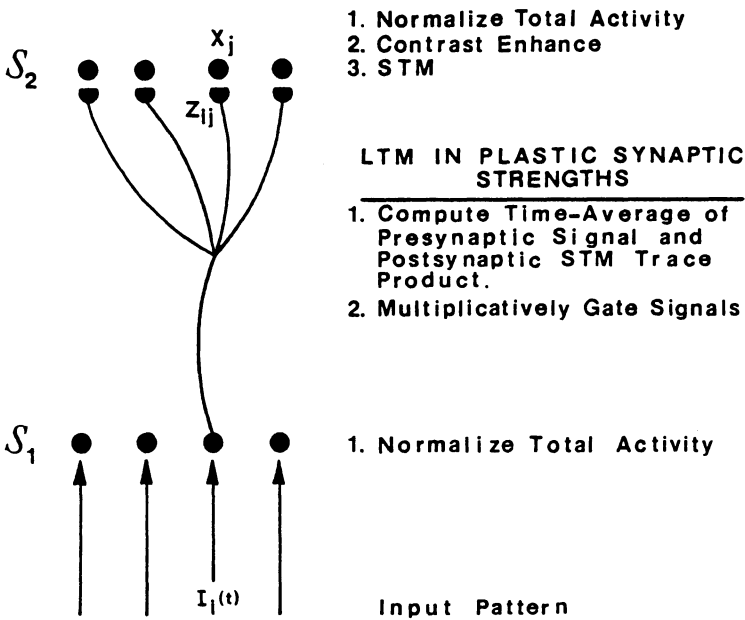


Figure 9. A feed-forward adaptive pattern classifier.

To have a definite example in mind, we can interpret slab S_1 as the LGN and slab S_2 as the visual cortex in Figs. 2 and 3. The input to S_1 can be regarded as the visual input to the eye after some "preprocessing" in the retina. Assuming that the relaxation time for nodes in S_1 is sufficiently short, the activity pattern across S_1 will be proportional to the reflectance pattern $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ of the input. In other words, the input image θ is accurately registered on S_1 . That's why we call S_1 the image slab.

When an image is registered on S_1 , each node v_k in S_2 receives a signal S_{ik} from every node u_i in S_1 , as indicated in Fig. 10. The signals S_{ik} are gated by synaptic strengths z_{ik} when they arrive to produce a total gated input to v_k :

$$T_k = \sum_{i=1}^n S_{ik}z_{ik} = S_k \cdot z_k, \quad (17)$$

where the sum is over all nodes in the image slab, and $S_k = (S_{1k}, S_{2k}, \dots, S_{nk})$, $z_k = (z_{1k}, z_{2k}, \dots, z_{nk})$.

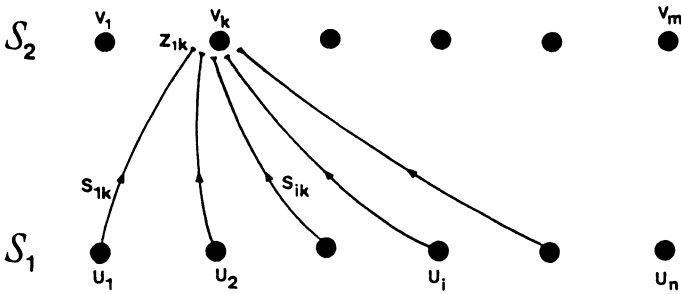


Figure 10. Gated image input to a feature detector.

The excitatory inputs T_k induce a pattern of activities x_k on the feature slab. However, the nodes v_k compete for the input by lateral inhibitory interactions as in Fig. 8. The greater the activity of one node, the more it inhibits the activities of its neighbors. Consequently, nodes with the largest inputs tend to increase their activities and suppress the activities of others until equilibrium is reached. The normalization property limits the total activity of the slab. So an increase in the activity of one node implies a decrease in the activities of others. The normalization property, therefore, implies that the activities of only a limited number of nodes can be driven above their signal thresholds. Thus, the input singles out or chooses these nodes. If the threshold is sufficiently high, then only the node with the greatest activity will be chosen. For the sake of simplicity, let's limit our considerations to this case.

A variety of different input patterns can maximally activate the same node v_k . The node v_k evidently responds to some common feature of these patterns, so we call it a feature detector, and we interpret its output as a signal that a pattern with this feature has been detected. To see how the feature is described mathematically, we look at the activity equation for v_k , which must have the general form of Eq. (4). After v_k has been chosen and the slab is in equilibrium, the lateral interactions C_{ik} in Eq. (4) vanish, and the activity x_k will be proportional to the gated input T_k given by Eq. (17). Also, the signal S_{ik} is a function of the reflectance θ_i , and it will simplify our job of interpretation if we assume that the function is linear, though this

is by no means necessary. Thus, the activity of v_k is given by

$$x_k = \lambda \theta \cdot z_k, \quad (18)$$

where λ is a positive constant. This enables us to describe pattern classification in more specific terms.

The feature detector v_k classifies (recognizes as equivalent) all patterns in the set

$$P_k = \{ \theta : \theta \cdot z_k > \max(\mu, \theta \cdot z_j) \text{ for all } j \neq k \}, \quad (19)$$

where $\mu > 0$ defines a recognition threshold. Clearly, the pattern class is determined by the vectors z_j , so we call them classifying vectors. But note that the pattern class P_k depends on the whole set of classifying vectors and not on z_k alone. The size of Σ determines how similar patterns must be to be classified by the same v_k .

Since the product $\theta \cdot z_k$ can never be negative, the set P_k defined by Eq. (19) is convex in the sense that the pattern $\alpha \theta_1 + (1-\alpha) \theta_2$ is in P_k if θ_1 and θ_2 are in P_k and $0 \leq \alpha \leq 1$. This convex set defines the feature identified by v_k .

The classification rule in Eq. (19) partitions the set of all input patterns into mutually exclusive and exhaustive convex subsets $P_0, P_1, P_2, \dots, P_M$, where, for $1 \leq k \leq M$, v_k detects patterns in the set P_k , and P_0 is the set of undetectable patterns that cannot drive any v_k over threshold. Thus, the feed-forward pattern classifier is capable of categorical perception. The number of feature detectors m determines the maximum number of categories $M \leq m$. The choice of classifying vectors z_k determines how different the categories can be.

The boundaries between the P_j are categorical boundaries. If a pattern input is deformed across a boundary, then there will be a sudden shift in category perception, as occurs when we view those ambiguous figures exhibited in introductory psychology textbooks. The definition of a category P_k by Eq. (19) shows that the boundary of P_k depends on the whole set of classifying vectors and not on z_k alone. Thus, the representation of a single "word" depends on the entire network "vocabulary."

Now that we know how patterns are classified, we are ready to see how new classifications can be learned. When a particular feature detector v_k detects a pattern, it is driven by the gated image as indicated in Fig. 10. This network system is called an instar because it can be redrawn in the symmetrical form of Fig. 11. It differs from the outstar of Fig. 6 only in the signal direction. This is expressed by saying that outstar and instar are dual to one another. Their anatomical duality is matched by a functional duality, the duality of recall and recognition. An outstar can learn to recall a given pattern but cannot recognize it whereas an instar can learn to recognize a pattern but not recall it. The outstar is blind. The instar is dumb.

While the instar to the feature detector v_k is operating, the input pattern drives changes in the classification vector z_k determined by the learn-

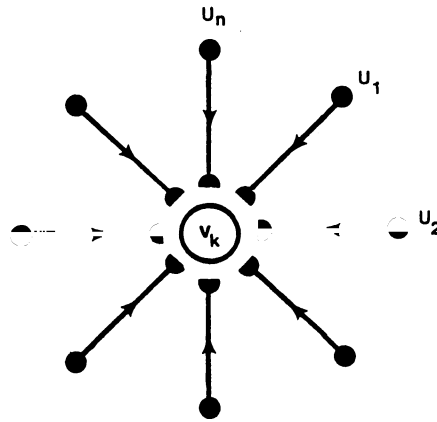


Figure 11. Symmetry of the instar anatomy.

ing equation (4). Noting that normalization will keep the feature detector activity at a fixed value $x_k = \mu^{-1}C_k$ and assuming, for simplicity, that the learning signal has the linear form $S_{ik} = \mu \theta_i$, we can put the instar learning equations into the vectorial form

$$\dot{\mathbf{z}}_k = -\beta \mathbf{z}_k + C_k \boldsymbol{\theta} , \tag{20}$$

where β is a positive constant.

For constant reflectance, Eq. (20) integrates to

$$\mathbf{z}_k(t) = \mathbf{z}_k(0)e^{-\beta t} + (C_k/\beta)(1 - e^{-\beta t})\boldsymbol{\theta} , \tag{21}$$

which holds no matter how wildly the input intensity $I(t)$ fluctuates. For $t \gg \beta^{-1}$, this reduces to $\mathbf{z}_k = \beta^{-1}C_k \boldsymbol{\theta}$. Thus, the classification vector \mathbf{z}_k aligns itself with the input reflectance vector $\boldsymbol{\theta}$. More generally, it can be shown that the classification vector \mathbf{z}_k aligns itself asymptotically with a weighted average $\bar{\boldsymbol{\theta}}$ of reflectance vectors that activate the feature detector v_k . This is the instar code development theorem. It is dual to the outstar learning theorem. Like the outstar theorem, it can be proved rigorously under more general assumptions than we have considered here. Note that the instar factorizes pattern and intensity just like the outstar, but the physical mechanism producing the factorization is quite different in each case. For the instar, factorization is produced by lateral interactions in the image slab.

The code development theorem tells us that an adaptive classifier tunes itself to the patterns it "experiences" most often. When a single classifying vector is tuned by experience, it shifts the boundaries of all the categories P_k defined by Eq. (19). "Dominant features" will be excited most often, so they will eventually overwhelm less salient features. Thus, the adaptive

classifier exhibits a progressive sharpening of memory. The degree of sharpening can be manipulated by varying the QT. The higher the QT, the more sharply tuned the evolving code.

A computer simulant of a feed-forward additive pattern classifier was constructed and tested by Christopher von der Malsburg [1973] with a striking result. He endowed the system with a random set of initial classifying vectors (synaptic strengths). Then he "trained" the system by presenting it with a sequence of nine different input patterns in a "training session." Each pattern displayed on the two-dimensional image slab was a straight line with a particular orientation. He followed the tuning of feature detectors with each training session. His main results after 100 sessions are displayed in Fig. 12. Each point on the figure represents one of the 169 feature detectors in Malsburg's feature slab. A line through a point indicates that the detector is optimally sensitive to a line image with that orientation. Points without lines through them do not respond significantly to any of the patterns. The figure shows many feature detectors for each pattern instead of a single one as in our discussion. That is because Malsburg's hardwired inhibitory interactions were limited in range to near neighbors, so widely separated nodes hardly affect one another. This is more realistic than the simpler case we considered.

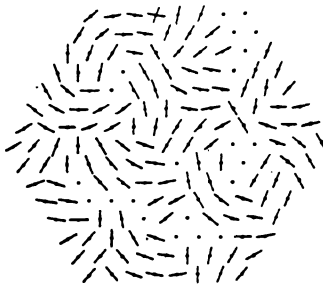


Figure 12. Selective sensitivity of simple cells in the striate cortex.

The striking thing about Fig. 12 is its "swirling" short range order. The points tend to be organized in curves with gradual changes in the direction of the "tangent vector" as one moves from one point to the next. This is striking because it is qualitatively the same as experimental results for which Hubel and Wiesel received a Nobel prize in 1981. Into the striate cortexes (Fig. 2) of cats and monkeys, Hubel and Wiesel inserted microelectrodes that can detect the firing of single neurons called "simple cells." Simple cells are selectively responsive to bars of light with particular orientations that are moved across an animal's retina. Moreover, they found that the responses of neighboring simple cells are related as in Fig. 12. Malsburg's computer experiment suggests an explanation for their results, and so provides one among many pieces of evidence that brains actually process patterns in accordance with the principles we have discussed.

The Adaptive Classifier is a general-purpose pattern processing device. We have discussed the processing of spatial patterns by way of illustration, but Grossberg shows how the same device can be used to process temporal patterns. Moreover, by connecting classifiers in series, one can construct a device that can decompose patterns into features and assemble the features into a hierarchy of perceptual units [Fukushima, 1980]. Computer simulations like Malsburg's show that Adaptive Pattern Classifiers work the way the theory says they should. The development of practical artificial devices for pattern learning and recognition based on the same design principles is under way [Hecht-Nielsen, 1983]. Such devices will be quite different from devices based on the current theory of pattern recognition in the field of artificial intelligence. They will mark the beginning of a new species of computer designed with principles of neural network theory.

9.3. Code Stabilization and Pattern Selection

The feed-forward adaptive pattern classifier of Fig. 9 has a severe limitation. The number of different patterns it can learn cannot exceed the number of nodes in the feature slab. Consequently, when the number of different patterns it experiences approaches the number of feature detectors, the system begins a massive recoding that destroys things it has already learned. This presents us with a new design problem: How can we design a pattern classifier that is plastic enough to learn from experience but stable enough to retain what it has learned? Grossberg calls this the stability-plasticity dilemma. His solution to the dilemma is full of surprises and brings forth a new set of network design problems whose solutions result in network capabilities with increasing similarities to real biological systems. Indeed, Grossberg has developed the theory to the point where it has many implications for animal learning theory in psychology, perhaps only a step or two away from a viable theory of higher order cognitive processes in human beings. Unfortunately, we do not have sufficient space here to do more than indicate the direction of Grossberg's research program.

Grossberg solves the stability-plasticity dilemma by introducing feedback from the feature slab to the image slab. Each feature detector is then the command cell for an outstar that can read out a learned pattern or template on the image slab. An outstar of this sort from the visual cortex to the LGN is indicated in Fig. 3. The template can be interpreted as an expectation; it is the pattern that the feature detector "expects to see." The template is superimposed on an external input image. If the match between template and image is sufficiently close, then the instar signal to the feature detector is amplified and fed back by the outstar to amplify the template. Thus, a feedback loop of sustained resonant activity is set up, and it drives a recoding of the classification vector as well as the template in the direction of the input image. Grossberg calls this resonant state an adaptive resonance. He suggests that every human act of conscious recognition should be interpreted biologically as an adaptive resonance in the brain. This brilliant idea has a host of implications that will surely enable us

to tell someday whether the idea is true or false. For example, adaptive resonances must be accompanied by distinctive electric fields, and the experimental study of such fields is beginning to show promising results. In the meantime, the adaptive resonance idea raises plenty of questions to keep theorists busy.

Adaptive resonances process expected events (that is, recognized patterns). An unexpected event is characterized by a mismatch between template and image. In that case, the mismatch feature detectors must be shut off immediately to avoid inappropriate recoding. Here we have a new design problem for which Grossberg has developed a brilliant solution involving two new neural mechanisms that play significant roles in many other network designs.

The first of these new mechanisms Grossberg calls a gated dipole. The gated dipole is a rapid on-off switch with some surprising properties. There is already substantial indirect evidence for the existence of gated dipoles, but Grossberg's most incisive predictions are yet to be verified. In my opinion, direct observations of gated dipole properties would be as profound and significant scientifically as the recent "direct" detection of electroweak intermediate bosons in physics.

The second new mechanism Grossberg calls nonspecific arousal. It is a special type of signal that Grossberg uses to modulate the quenching threshold of a slab. It is nonspecific in the sense that it acts on the slab as a whole; it carries no specific information about patterns. As Grossberg develops the theory further, he is led to distinguish several types of arousal signals, and their similarities to what we call emotions in people becomes increasingly apparent. Grossberg shows that arousal is essential for pattern processing. This has the implication that emotions play an essential role in the rational thinking of human beings. Reason and emotion are not so independent as commonly believed.

The processing of unexpected events requires more than refinements in the design of adaptive classifiers. It requires additional network components to decide which events are worth remembering and how they should be encoded. To see the rich possibilities opened up by Grossberg's attack on this problem, I refer you to his collected works.

10. Playing the Game

The Classical Mechanics Game is played by Newton's rules. The Relativity Game is played by Einstein's rules. The MAXENT Game is played by Jaynes' rules. If you want to play the neural Network Modeling Game, you had better learn Grossberg's rules or you're not likely to win any of the prizes.

11. References

Scott [1977] discusses single neuron models from the viewpoint of a physicist. The articles by Amari in Metzler [1977] and by Kohonen in Hinton and Anderson [1981] are good samples of work on network modeling by Grossberg's "competitors." The "bottom-up" approach in Freeman [1975] appears to be converging on the same idea of adaptive resonance that came from Grossberg's "top-down" approach, but much theoretical work needs to be done to relate the approaches.

Freeman, W. J. (1975), Mass Action in the Nervous System, Academic Press, New York.

Fukushima, K. (1980), "Neocognitron, a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybernet.* 36, pp. 193-202.

Grossberg, S. (1982), Studies of Mind and Brain, D. Reidel, Dordrecht.

Hebb, D. O. (1949), The Organization of Behavior, Wiley, New York, p. 62.

Hecht-Nielsen, R. (1983), "Neural analog processing," *Proc. SPIE* 360.

Hinton, G. E., and J. A. Anderson, eds. (1981), Parallel Models of Associative Memory, Lawrence Erlbaum, Hillsdale, N.J.

Kandel, B., and J. Schwartz, eds. (1981), Principles of Neural Science, Elsevier/North Holland, New York.

von der Malsburg, Ch. (1973), "Self-organization of orientation sensitive cells in the striate cortex," *Kybernetik* 14, pp. 85-100.

Metzler, J., ed. (1977), Systems Neuroscience, Academic Press, New York.

Scott, A. C. (1977), Neurophysics, Wiley, Interscience, New York.

Stent, G. S. (1973), "A physiological mechanism for Hebb's postulate of learning," *Proc. Nat. Acad. Sci.* 70, p. 997.