Psychological Review

VOLUME 84 NUMBER 5 SEPTEMBER 1977

Distinctive Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model

James A. Anderson, Jack W. Silverstein, Stephen A. Ritz, and Randall S. Jones
Brown University

A previously proposed model for memory based on neurophysiological considerations is reviewed. We assume that (a) nervous system activity is usefully represented as the set of simultaneous individual neuron activities in a group of neurons; (b) different memory traces make use of the same synapses; and (c) synapses associate two patterns of neural activity by incrementing synaptic connectivity proportionally to the product of pre- and postsynaptic activity, forming a matrix of synaptic connectivities. We extend this model by (a) introducing positive feedback of a set of neurons onto itself and (b) allowing the individual neurons to saturate. A hybrid model, partly analog and partly binary, arises. The system has certain characteristics reminiscent of analysis by distinctive features. Next, we apply the model to "categorical perception." Finally, we discuss probability learning. The model can predict overshooting, recency data, and probabilities occurring in systems with more than two events with reasonably good accuracy.

In the beginner's mind there are many possibilities, but in the expert's there are few.
—Shunryu Suzuki
1970

I. Introduction

If we knew some of the organizational principles of brain tissue, it might be possible to make a few general statements about how the brain works in a psychological sense. There is a close relation between available building blocks and the performance that can be realized easily with those component parts in most systems, and we should expect the same to be true for the nervous system.

Let us consider an example of some importance to psychologists. In current theories of cognition and perception one often finds explanations of phenomena in terms of what are essentially little computer programs, complete with flow charts and block diagrams. Certainly, the desire to decompose complex mental events into simpler basic units follows the strategy that has been triumphantly successful in the physical sciences.

Even a poor understanding of brain organization might be of value in placing some kinds of limits on these elementary operations.

All of us are somewhat familiar with digital computers and, particularly, with computer programs. In some psychological models, the elementary instructions that we use to tell a computer what to do seem to serve as the model for brain function. Assumed are "comparisons," "scans," "lists," "decisions," and other computer-like operations.

However, computers are made with fast, reliable, binary electronic components that are designed to operate by executing very quickly a long series of simple operations. When an elementary operation takes a fraction of a microsecond and when internal noise is not a

major problem, this is a very successful technique.

But most neuroscientists agree at present that the brain is a slow, intrinsically parallel, analog device that contains a great deal of internal noise from a variety of sources. It is very poorly suited to the accurate execution of a long series of simple operations. What is it good at?

The brain is best adapted to interacting, highly complex, spatially distributed parallel operations. Adjectives such as "distributed," "parallel," and "holographic" are sometimes used to describe operations of this type.

Since neurons are very slow compared to electronic devices—on the order of a few milliseconds at the very fastest—we should expect there to be time for only a few of the elementary operations that compose the instructions for a "brain computation." There would be neither the speed, equipment, nor accuracy for enormous strings of simple operations.

We may conclude that the elementary operations used by the brain are of a much more powerful and different kind than the simpler instructions familiar to us from our experience with computers.

In this article we will present the outlines of a theory that is suggested by the anatomy and physiology of the brain and that is realized by arrays of simple, parallel, analog elements that are meant to be an oversimplification of the neurons in a mammalian neocortex. The model is associative and distributed, and we shall see that its elementary operations are of a very different type from simple logical operations. Systems that are intrinsically parallel have a number of pronounced and unfamiliar properties, as well as some impressive capabilities.

When we propose a model that tries to join together information from several fields, we have difficulties when we try to verify it. At present, we simply do not know enough about the detailed connectivity and synaptic properties of the brain to do more than make our models in qualitative agreement with what is known of the neurobiology. Similarly, the psychological data are useful when it comes to testing general approaches, but they often do not allow unequivocal tests of the details of a theory. Thus, it seems to us best to try to make our neurally based models refined enough to fit, in detail, a few experiments—just to show it can be done. But we would also like to point out, in a more impressionistic way, areas of agreement between theory and observation elsewhere. We have deliberately made our theories extremely simple, perhaps unrealistically so, because if we can make adequate models with very simple theories, surely it will be possible to do better when more complex and/or more realistic assumptions are made. We feel that many different versions of distributed memories can be made to give results similar to those we describe here.

We will first discuss some necessary theoretical background. We will present a simple version of a distributed, associative memory. We will then modify the simple, linear model to incorporate positive feedback of a set of neurons on itself. We will show that feedback gives rise to behavior that is reminiscent of the analysis of an input in terms of what are called "distinctive features," a type of analysis that is commonly held to be of great importance in perception. We will then apply the model to two widely differing psychological phenomena. We shall discuss in detail the classic set of experiments generally described as "probability learning," and we shall discuss more generally the perceptual phenomenon called "categorical perception."

These diverse effects, which at first sight might seem to be very complicated and involve much information processing, may be

This work was supported in part by grants from the Ittleson Family Foundation to the Center for Neural Studies, Brown University, and by a Public Health Service General Research Support Grant from the Division of Biological and Medical Sciences, Brown University. Jack W. Silverstein was supported by the National Science Foundation Grant MCS75-15153-A01.

Some material derived from this paper has been presented at the "Workshop on Formal Theories in Information Processing," Stanford University, Stanford, California, June 1974. Some material was also presented at the 7th Mathematical Psychology meeting, Ann Arbor, Michigan, August 1974; at the annual meeting of the Society for Neuroscience, St. Louis, Missouri, October 1974, and at the 9th Mathematical Psychology Meeting, New York, August 1976.

Requests for reprints should be sent to James A. Anderson, Center for Neural Studies, Department of Psychology, Box 1853, Brown University, Providence, Rhode Island 02912.

explainable by a single, rather simple, set of assumptions.

II. Theoretical Development

In the past few years a number of related realizations of distributed memories applied to brain models have been put forward by several groups of investigators (Anderson, 1970, 1972; Cooper, 1974; Grossberg, 1971; Kohonen, 1972, 1977; Little & Shaw, 1975; Willshaw, Buneman, & Longuet-Higgins, 1969). One form of the central learning assumptions of these models was first proposed by Hebb (1949) but, as is often the case in an active field of science, many of the fundamental assumptions of the models have been arrived at independently by different groups.

In our development here, we will follow the notation and basic assumptions of Anderson (1968, 1970, 1972, 1977; Cooper, 1974). This version of a distributed, associative memory is formally exceptionally simple. It is easy to work with and may give a first approximation to some of the common properties of many distributed models.

We start by making two central assumptions. First, nervous system activity can be most usefully represented as the set of simultaneous individual neuron activities in a group of neurons. Neuron "activity" is considered to be related to a continuous variable, the average firing frequency. Patterns of *individual* activities are stressed, because properties of particular neuron activities need not be related to each other for the system to function. Indeed, some evidence (Noda & Adey, 1970) indicates that interneuronal spike activity correlations of nearby cells recorded with the same electrode in parietal ("association") cortex may be quite low when the brain is doing "interesting" things. They found that in an animal in REM (rapid eye movement) sleep or in the awake, alert state, correlations were very low, whereas the same pairs of cells had highly correlated discharges in deep sleep. The same appears to be true in hippocampus (Noda, Manohar, & Adey, 1969).

Recent work by Creutzfeldt, Innocenti, and Brooks (1974) seems to suggest that most cells in primary visual cortex, even those close to one another in the same cortical column, are not strongly coupled together, again implying a good deal of individuality of cell response. The individuality of cell responses in auditory cortex has been remarked upon (Goldstein, Hall, & Butterfield, 1968). Morrell, Hoeppner, and de Toledo (1976) studied single units in cat and rabbit parastriate cortex. They found that nearby units responding to the same stimulus (recorded with the same electrode) often differed in the type and direction of alteration of their discharges when a stimulus configuration and a cutaneous shock were paired in a Pavlovian conditioning paradigm. They comment that their data "provide little support for the notion of coherent changes in large neuronal populations" (p. 448).

Thus we are dealing, to a first approximation, with a system of individualistic cells, each with its own properties. Although cells near to one another may show similar response properties (for example, orientation or binocularity in visual cortex), each cell behaves differently from its neighbors when studied in detail. We will be concerned with the behavior of large groups of cells, but this does not mean that each cell is doing the same thing. Indeed, most distributed systems do not work well if cells all respond to the same inputs in the same ways, since diversity of cell properties allows for better operation. Simple redundancy is the most uninteresting way to provide reliability.

This assumption allows us to represent these large-scale activity patterns as vectors of high dimensionality with independent components. All the models we work with will use these vectors as the elementary units. We will show that it is possible to develop theories where these complex activity patterns, representing discharges of very many neurons, can act as basic units that combine and interact in relatively simple ways.

As our second major assumption, we hold that different memory traces (sometimes called "engrams"), corresponding to these large patterns of individual neuron activity, interact strongly at the synaptic level so that different traces are not separate in storage. Considerable physiological evidence supports this idea. As Sir John Eccles comments, "each neurone and even each synaptic junction are built into many engrams. The systematic study

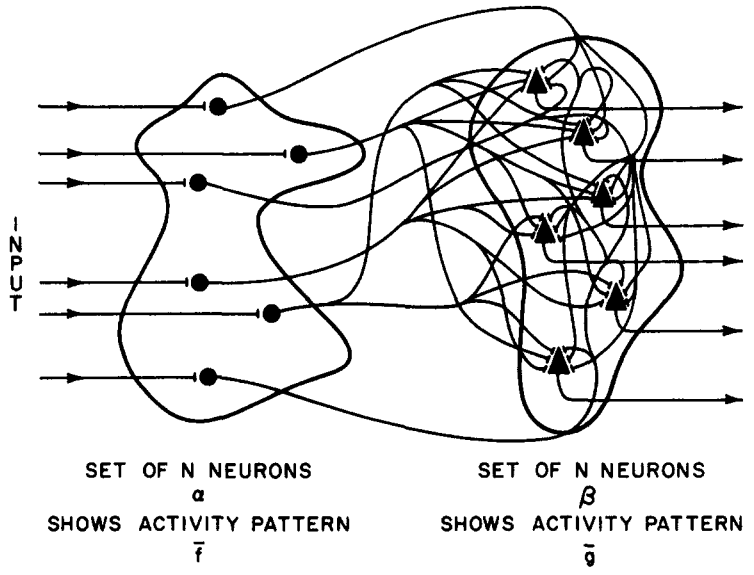

Figure 1. We consider the properties of sets of N neurons, α and β. Close inspection of this spaghetti-like drawing will reveal that every neuron in α projects to (i.e., has a synapse with) every neuron in β. Since this drawing, where $N = 6$, understates the size and connectivity of the nervous system by several orders of magnitude, it can be seen that single neurons and single synapses may have little effect on the discharge patterns of the group as a whole. Properties of such large, interconnected systems can sometimes be modeled simply with the techniques of linear algebra, the approach taken in the text. ($\bar{f}$ and $\bar{g}$ indicate vectors.)

of the responses of individual neurones in the cerebrum, cerebellum, and in the deeper nuclei of the brain is providing many examples of this multiple operation" (Eccles, 1972, p. 59).

Association Model

Let us assume we have two groups of N neurons, α and β (see Figure 1). We will assume that every neuron in α projects to every neuron in β. This clearly unrealistic assumption is not made in Anderson (1972). A detailed discussion of some of the mathematical aspects of this model are found in Cooper (1974) and Nass and Cooper (1975).

To proceed further, we must describe how the activity of a neuron reflects its synaptic inputs. For many cells, this can be extremely complex. However for some systems rather simple relations are found. We shall assume at first that neurons are simple linear analog integrators of their inputs, and we shall see what kinds of models evolve from this assumption. There is evidence supporting this assumption in a few well-studied systems. In the lateral inhibitory system of *Limulus*, rather good linear integration is found (Knight, Toyoda, & Dodge, 1970; Ratliff, Knight, Dodge, & Hartline, 1974). Linear transmission, according to Mountcastle (1967), holds for many mammalian sensory systems, once beyond an initial nonlinear transduction.

We assume there is a synaptic strength, a_{ij}, which couples the jth neuron in α with the ith neuron in β. Thus, subject to our assumption of linear integration, we can write the following: If $f(j)$ is the activity shown by neuron j in α and if $g(i)$ is the activity shown by neuron i in β at a given time, then

$$g(i) = \sum_j a_{ij} f(j). \tag{1}$$

Let us now consider the following situation: The set of neurons α shows an activity pattern, a vector, $\mathbf{f}_1$, the set of all the $f(j)$. The set of neurons β shows an activity pattern, a vector, $\mathbf{g}_1$. We wish to associate the pattern $\mathbf{f}_1$ with the pattern $\mathbf{g}_1$ so that later presentation of $\mathbf{f}_1$ alone will give rise to $\mathbf{g}_1$ in the set β. Let us assume that initially our set of synaptic connectivities a_{ij} is zero.

We can ask what detailed local information could influence a synaptic junction to allow storage of memory. Locally available information includes the presynaptic activity and, we shall assume, the postsynaptic activity. Let us make about the simplest assumption for synaptic learning that allows for pre- and postsynaptic interaction at the synaptic level. Let us postulate, as our essential learning assumption, that *to associate pattern* $\mathbf{f}_1$ *in* α *with pattern* $\mathbf{g}_1$ *in* β *we need to change the set of synaptic weights according to the product of presynaptic activity at a junction with the activity of the postsynaptic cell.*

For convenience, let

$$\|\mathbf{f}_1\| = 1, \quad \text{where} \quad \|\mathbf{f}\| = [\sum_{i=1}^{N} f(i)^2]^{1/2}.$$

This quantity is usually called the length of the vector. The change in the ijth synapse is given by $f_1(j)g_1(i)$. The set of connections form a matrix $\mathbf{A}_1$ given by

$$\begin{aligned}\mathbf{A}_1 &= \mathbf{g}_1\mathbf{f}_1{}^T \\ &= [\mathbf{g}_1 f_1(1), \mathbf{g}_1 f_1(2) \cdots \mathbf{g}_1 f_1(N)], \quad (2)\end{aligned}$$

where T is the transpose operation (throughout this paper all vectors will be assumed to be N-dimensional column vectors).

Assume that after we have "printed" the set of connectivities $\mathbf{A}_1$, pattern of activity $\mathbf{f}_1$ arises in α. Then we see that activity in β is given by

$$\mathbf{A}_1\mathbf{f}_1 = \mathbf{g}_1(\mathbf{f}_1{}^T\mathbf{f}_1) = \|\mathbf{f}_1\|^2\mathbf{g}_1 = \mathbf{g}_1; \quad (3)$$

so we have $\mathbf{g}_1$ appearing as the pattern of activity in β, which corresponds to our definition of association.

It is very unlikely that these sets of neurons exist only to associate a single set of activity patterns. Let us assume that we have K sets of associations, $(\mathbf{f}_1,\mathbf{g}_1)$, $(\mathbf{f}_2,\mathbf{g}_2)$, $\ldots$, $(\mathbf{f}_K,\mathbf{g}_K)$, each generating a matrix of synaptic increments, $\mathbf{A}_k$. Then, since we have assumed that a single synapse participates in storing many traces, we form an overall connectivity matrix $\mathbf{A}$ given by

$$\mathbf{A} = \sum_k \mathbf{A}_k.$$

Let us assume that the $\mathbf{f}_i$ are mutually orthogonal, that is, that the inner product of $\mathbf{f}_i$ with $\mathbf{f}_j$, defined as

$$\mathbf{f}_i{}^T\mathbf{f}_j = \sum_{s=1}^{N} f_i(s)f_j(s),$$

is zero for $i \neq j$. Then, if pattern $\mathbf{f}_j$ is impressed on the set of neurons α, we have

$$\begin{aligned}\mathbf{A}\mathbf{f}_j &= \sum_k \mathbf{A}_k\mathbf{f}_j \\ &= \mathbf{A}_j\mathbf{f}_j + \sum_{k\neq j} \mathbf{A}_k\mathbf{f}_j \\ &= \mathbf{g}_j + \sum_{k\neq j} \mathbf{g}_k(\mathbf{f}_k{}^T\mathbf{f}_j) \\ &= \mathbf{g}_j.\end{aligned}$$

Thus, the system associates perfectly. If the $\mathbf{f}$s are not orthogonal the system will produce noise as well as the correct association, but the system is often quite usable. Actually, the "mistakes" made by this system are often as interesting as the "correct" responses (see Anderson, 1977).

Orthogonality is an effective way of dealing with the notion of independence of inputs. If two inputs are orthogonal, then there is no interaction between them; that is, the response of the system to one input is in no way influenced by the other input. It is as if the inputs are going through completely different mechanisms.

A Numerical Example

To show how this system works, let us construct a simple, eight-dimensional system. Assume we have the four orthogonal input vectors, $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$, and $\mathbf{f}_4$ shown in Table 1. These vectors are normalized Walsh functions, which are digital versions of sine and cosine functions. We choose arbitrary output vectors, $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, and $\mathbf{g}_4$ and wish to make the associations between pairs of vectors $(\mathbf{f}_i,\mathbf{g}_i)$. Note that we need place no restrictions on the $\mathbf{g}$s. Thus $\mathbf{g}_1$ and $\mathbf{g}_2$ are orthogonal, and $\mathbf{g}_3$ and $\mathbf{g}_4$ are very close together. They also vary considerably in length: $\mathbf{g}_1$ is 2.24 units long, $\mathbf{g}_2$ is 3.61 units long, and $\mathbf{g}_3$ and $\mathbf{g}_4$ are 4.47 units long. The $\mathbf{A}_1$ matrix, the association matrix between $\mathbf{f}_1$ and $\mathbf{g}_1$, is shown in Table 2. We have not shown the other three matrices, $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$, but their construction should be clear from $\mathbf{A}_1$. The resulting sum of all four matrices is shown in the second part of Table 2.

Table 1
Input and Output Vectors Used in the Numerical Example Given in the Text

Type of vectors

Input

$$\mathbf{f}_1 = \frac{1}{\sqrt{8}}\begin{pmatrix}1\\1\\1\\1\\-1\\-1\\-1\\-1\end{pmatrix} \quad \mathbf{f}_2 = \frac{1}{\sqrt{8}}\begin{pmatrix}-1\\-1\\1\\1\\-1\\-1\\1\\1\end{pmatrix} \quad \mathbf{f}_3 = \frac{1}{\sqrt{8}}\begin{pmatrix}1\\-1\\1\\-1\\1\\-1\\1\\-1\end{pmatrix} \quad \mathbf{f}_4 = \frac{1}{\sqrt{8}}\begin{pmatrix}-1\\-1\\1\\1\\1\\1\\-1\\-1\end{pmatrix}$$

Output

$$\mathbf{g}_1 = \begin{pmatrix}1\\0\\-1\\0\\1\\-1\\-1\\0\end{pmatrix} \quad \mathbf{g}_2 = \begin{pmatrix}-1\\2\\0\\-1\\-1\\-1\\-1\\2\end{pmatrix} \quad \mathbf{g}_3 = \begin{pmatrix}3\\0\\-1\\-1\\-2\\0\\-1\\2\end{pmatrix} \quad \mathbf{g}_4 = \begin{pmatrix}4\\0\\-1\\-1\\-1\\0\\0\\1\end{pmatrix}$$

Inspection of this matrix shows little sign of the component parts that went to make it up, and asking about the value of any particular element is pointless in relation to the information stored in the matrix. However, calculation will show that

$$\mathbf{A}\mathbf{f}_i = \mathbf{g}_i$$

for all four input vectors, which shows that indeed such a simple matrix can "learn" four essentially arbitrary associations.

One of the most important aspects of these models is their similarity to a filter in the strict sense of a system which responds weakly to an input that has not been learned (i.e., to

Table 2
Matrices Associating Pairs of Vectors in the Numerical Example

Matrix

$$\mathbf{A}_1 = \mathbf{g}_1\mathbf{f}_1^T$$

$$\mathbf{A}_1 = \frac{1}{\sqrt{8}}\begin{pmatrix}1&1&1&1&-1&-1&-1&-1\\0&0&0&0&0&0&0&0\\-1&-1&-1&-1&1&1&1&1\\0&0&0&0&0&0&0&0\\1&1&1&1&-1&-1&-1&-1\\-1&-1&-1&-1&1&1&1&1\\-1&-1&-1&-1&1&1&1&1\\0&0&0&0&0&0&0&0\end{pmatrix}$$

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \mathbf{A}_4$$

$$\mathbf{A} = \frac{1}{\sqrt{8}}\begin{pmatrix}1&-5&7&1&7&1&-3&-9\\-2&-2&2&2&-2&-2&2&2\\-1&1&-3&-1&-1&1&1&3\\1&3&-3&-1&-1&1&-1&1\\1&5&-3&1&-3&1&-3&1\\0&0&-2&-2&2&2&0&0\\-1&1&-3&-1&1&3&-1&1\\-1&-5&5&1&1&-3&3&-1\end{pmatrix}$$

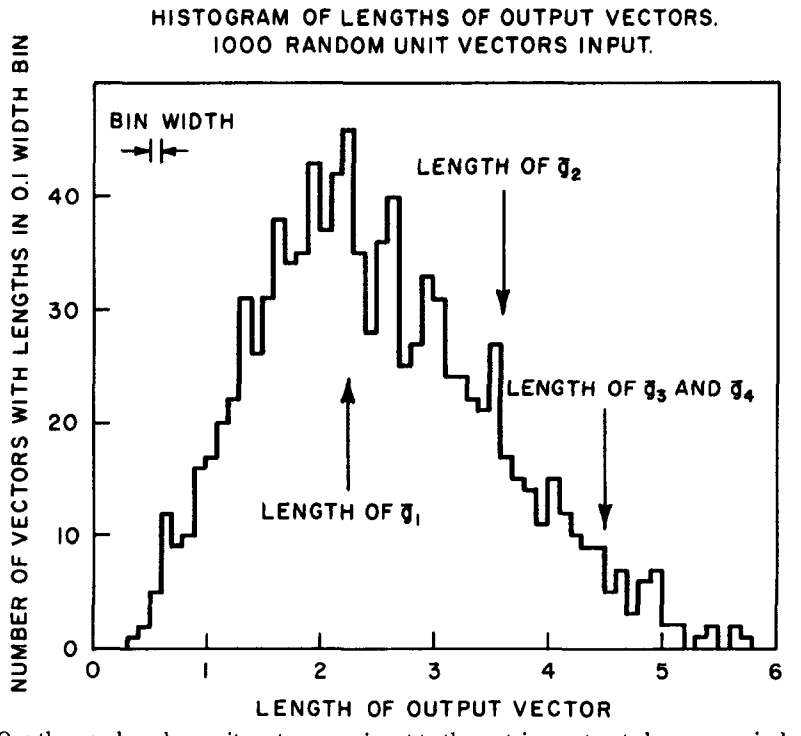

Figure 2. One thousand random unit vectors were input to the matrix constructed as a numerical example in the bottom half of Table 2. The figure shows a histogram of the lengths of the 1,000 resulting output vectors. The arrows point to the lengths of the actual output vectors associated with the four inputs used in the matrix. ($\vec{g}_1$, $\vec{g}_2$, $\vec{g}_3$, and $\vec{g}_4$ indicate vectors.)

which the filter is not "tuned") but which responds strongly to an input that has been seen before. We can demonstrate this property with our matrix. Suppose we have a random unit vector as input, that is, a vector generated by the uniform distribution on the unit sphere in eight-dimensional space. We should expect, if the system is filter-like, that there will be a short vector appearing at the output, on the average. We used a computer to generate 1,000 random unit vectors and looked at the lengths of the resulting output vectors. The distribution of output lengths is shown in Figure 2. The lengths of the actual stored associations—the $\mathbf{g}$s—are shown by arrows in the figure. Strictly on the basis of length, almost no outputs due to random vectors were as long as $\mathbf{g}_3$ and $\mathbf{g}_4$ (96% were shorter), and 85% of the outputs were shorter than $\mathbf{g}_2$. Even $\mathbf{g}_1$, which was half the length of $\mathbf{g}_3$ and $\mathbf{g}_4$, was longer than 45% of the random outputs. Thus, even with as crude a measure of filter characteristic as output length, in a very low dimensionality system, a fairly good job of discrimination can be made between old and new inputs. Introduction of even a very few elementary mechanisms to "sharpen up" the response from the system, such as cascades of filters or positive feedback (discussed in Section IV), can make these matrices good filters with an interesting "cognitive" structure to them.

We should observe, however, that the system can make mistakes. Noise is inherent in the system. The histogram shows that there are a few inputs that can give a larger output than given by any of the inputs the system has learned. Thus, we can see that a useful type of analysis for such systems is statistical, and many aspects of their behavior can be studied well with the techniques of communication theory and decision theory. If such a distributed system is indeed present in our cortex, possibly we can see why such statistical methods are so successful in accounting for many of the phenomena found for the psychology of even our highest mental functions.

III. Biological and Psychological "Features"

Introduction

The association model just presented is a good associator, and it has been applied elsewhere to several sets of experimental data (see Anderson, 1977; Nass & Cooper, 1975). A simple variant of the model, relying on the filter characteristics of a similar system, has been used to propose an explanation for some of the data arising in the Sternberg list-scanning experiment (Anderson, 1973).

We hope now to show that an extension of the model has interesting qualitative similarities to some important characteristics of human perception. We hope to show that we can represent noisy inputs to the system in terms of their "distinctive features" and that the features that the model generates are both cognitively significant, in that they are the most useful for discriminating among members of a stimulus set, and, at the same time, are most strongly represented in the output from the system. Thus, what appears to be a highly structured and analytical approach to perception—distinctive feature analysis—can be explained as the result of the operation of a highly parallel, analog system with feedback.

We shall interweave the theoretical discussion with a very brief discussion of the psychological theory and a little data from the neurosciences. This occasionally awkward means of presentation is intended to convey the close interdependence between theory and data from several fields.

Let us make clear all we hope to do. At this time, we do not have the data from either psychology or neuroscience to decisively test the theory we are to present. However, we hope to show that there are striking qualitative similarities between the structure of the theory and the structure that many feel typifies some kinds of perception.

What is a Distinctive Feature?

That there are entities called distinctive features and that these entities are somehow of importance in perception is a commonly accepted belief is psychology at present. A recent elementary textbook (Lindsay & Norman, 1972) builds a large part of the first few chapters around the idea of features, and this approach is central to Neisser's very influential book, *Cognitive Psychology* (1967). Since the early 1950s, there has been strong evidence from linguistics that phonemes could be characterized as being represented by the presence or absence of a small set (12 in the original analysis) of distinctive features (Jakobsen, Fant, & Halle, 1961). Each phoneme was uniquely represented by its own particular set of features. The use of a good set of features is an excellent type of preprocessing and is a commonly used practical pattern recognition technique. The great reduction in dimensionality of the stimulus allows the system to discard irrelevant information and eliminate noise, while making later stages simpler and more reliable, since they have to cope with less complex inputs.

How Might the Brain Do Feature Analysis?

Feature analysis seems to be a strategy used by the brain. How is this analysis performed?

The simplest way might be to have, somewhere, neurons that respond when, and only when, a particular distinctive feature appears. These would then be true "feature detecting" neurons. We might point out that in this case, discharges corresponding to different features would be orthogonal in the sense discussed in the previous section, since different features would give rise to activity patterns with nonzero values in different sets of elements.

Barlow (1972) has argued strongly that this is truly the way the brain works. He stated a number of what he calls "dogmas" about the relation between brain and perception, proposing that sensory systems are organized so as to achieve as complete a representation of the stimulus as possible with the smallest number of discharging neurons. He estimated that as few as 1,000 cells in visual cortex may fire in response to a complex visual stimulus. He proposed that what we call "perception" may correspond to the activity of a small number of high level, very selective neurons, each of which "says something of the order of complexity of a word" (p. 385). Informal discussions indicate to us that many neurophysiologists are sympathetic to this point of view.

Biology and Psychology: Audition

Suppose we look at published feature lists for spoken language (see Lindgren, 1965) and for letter perception (Gibson, 1969; Laughery, 1969). The features proposed by these and other writers are psychological features. They are quite complex. Lindgren describes the acoustic characteristics of the distinctive features proposed for spoken language. Almost every feature is characterized by complex changes in wide, often ill-defined bands of frequencies. For example, the "vocalic versus non-vocalic" feature is characterized by "presence versus absence of a sharply defined formant structure" (Lindgren, 1965, p. 55). The "nasal versus oral" feature is described as "spreading the available energy over wider (versus narrower) frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional (nasal) formants" (Lindgren, 1965, p. 55). The only feature that seems to correspond to a well-defined set of frequencies is the "voiced versus voiceless" feature, which describes the presence or absence of vocal cord vibration. Even here, though, men, women, and children have characteristic vocal cord frequencies differing over a two-octave range, all presumably capable of exciting the voicing feature. The complicated structure of the vocal tract would usually be expected to give rise to equally complicated variations in frequencies of resonances with changes in geometry.

The neurophysiology of the auditory system is very complex and not well understood at present. Although the lower levels of the auditory pathway seem to be primarily frequency analyzers, neurons in auditory cortex are highly individualistic and variable in their responses. Many cells in primary auditory cortex have very sharply tuned responses, but others have quite wide bandwidths.

If species-specific vocalizations are used to stimulate cells, the picture is no simpler. Wollberg and Newman (1972) recorded from the auditory cortex of squirrel monkeys, using recordings of species-specific calls as stimuli. They reported, "some cells responded with temporally complex patterns to many vocalizations. Other cells responded with simpler patterns to only one call. Most cells lay between these two extremes" (p. 212).

Funkenstein and Winter (1973; Winter & Funkenstein, 1973) did a similar experiment with a wider range of squirrel monkey vocalizations. They also found a variety of cell responses, from a small percentage of cells that responded only to particular calls, to cells that responded to particular frequencies in any context—noise, pure tones or calls.

Evans (1974) reviews a number of experiments and describes the bewildering variety of response types observed. Even in research on an animal as unintelligent as a bullfrog, which has a behaviorally important mating call with two spectral peaks, Frishkopf, Capranica, and Goldstein (1968) did not uncover cells that responded only to both peaks presented simultaneously, a property which would be required of a "mating call detector." Although the frequency responses of cells in the frog auditory system were commonly tuned to one or the other spectral peak, the investigators found no cells in the midbrain and medulla that put the two peaks together.

Biology and Psychology: Vision

The feature lists that are proposed for recognition of capital letters are deceptively simple. In the lists, one typically finds proposed features such as vertical line segments at left, center, or right; horizontal line segments at top, middle, or bottom; or curve slants or parallel lines (Laughery, 1969).

With a list of this type, it is only a small conceptual leap to identify these psychological features with groups of particular cells of the type known to exist in primary visual cortex which show orientation and edge sensitivity (Hubel & Wiesel, 1962, 1968). It should be emphasized that single cells in primary visual cortex do *not* show the requisite selectivity in the sense that they respond to features and only features. Cells in primary visual cortex are quite selective, but they respond to many aspects of the stimulus. Schiller, Finlay, and Volman (1976), in perhaps the most careful quantitative study of single-cell response properties in monkey striate cortex, make the following comments in the abstract of the

summary paper:

1. Several statistical analyses were performed on 205 S-type and CX-type cells which had been completely analyzed on 12 response variables: orientation tuning, end stopping, spontaneous activity, response variability, direction selectivity, contrast selectivity for flashed or moving stimuli, selectivity for interaction of contrast and direction of stimulus movement, spatial-frequency selectivity, spatial separation of subfields responding to light increment or light decrement, sustained/transient response to flash, receptive-field size, and ocular dominance.
2. Correlation of these variables showed that within any cell group, these response variables vary independently. (p. 1362)

(S and CX cells correspond roughly to the familiar simple versus complex distinction, but with more precise definition.)

Besides finding support of our earlier assumption of the great individuality of cortical neurons, we can see that single cells can have their discharges modified by a wide range of aspects of the stimulus. Here also we do not seem to find cells that respond only to psychological features, although there are cells with very pronounced selectivities.

A Regrettable Misapprehension

It is apparent from reading the literature in this area that the word "feature" as used by psychologists and by neuroscientists has come to mean different things. When a psychologist discusses features, what seems to be meant is a complex kind of perceptual atom which is independent of other atoms and constitutes an elementary unit out of which perception is built. The feature lists that have been proposed for both letter perception and speech perception involve many different aspects of the input stimulus. Their simplicity is deceiving when considered in light of the properties of the single cells of the nervous system.

It is also apparent, regrettably, that when a neurobiologist refers to a "feature detector," he is typically referring to a single neuron which displays a certain amount of selectivity in its discharge, often for the biologically important and relevant aspects of the stimulus. This does not mean that this cell has the specificity of response to be a detector of the psychological feature. Something more is involved.

IV. A Theoretical Approach to Psychological Features

There is an implicit neural model—that formulated clearly by Barlow—in the identification of single-cell properties with features. There is an alternative point of view that uses the observed single-cell selectivities to provide selectivity to the psychological features, but now as part of well-defined activity patterns that *use* cell properties.

McIlwain (1976) makes the important point that the large receptive fields often seen in the higher levels of the visual system and the response of the cells to many aspects of a stimulus do not necessarily mean that the overall system lacks precision. Our distributed system using the output of the entire group of cells can be very precise, in that the discharge pattern due to one input can be reliably differentiated from that arising from a different stimulus, even though there are many cells in the group that may respond to both inputs. As a relevant example, the associative network presented in Section II of this article is completely interconnected. A cell in the second set of neurons can respond to any cell in α, the first set. A single cell has a large receptive field and is very unselective. However, we showed that the output patterns of all the cells in β can be made to respond strongly only to particular inputs, and thus the system displays considerable selectivity.

Can we make a neural model analyze its inputs in terms of features? It is clear that distinctive feature analysis has an important learned component. In the perception of written letters, this is obvious. In spoken language, which is much more biologically determined than are written letters, Eimas, Siqueland, Jusczyk, and Vigorito (1971) have shown that some aspects of linguistic distinctive features are both built in and modifiable. For example, a category boundary for voice onset time, which appears to correspond to the voicing–voiceless feature, is present in the human infant. Yet this feature boundary can be modified in the adult, depending on the phonetic structure of the language spoken (Eimas & Corbit, 1973).

Our theoretical aim is to reduce the dimensionality of the stimulus so that a very com-

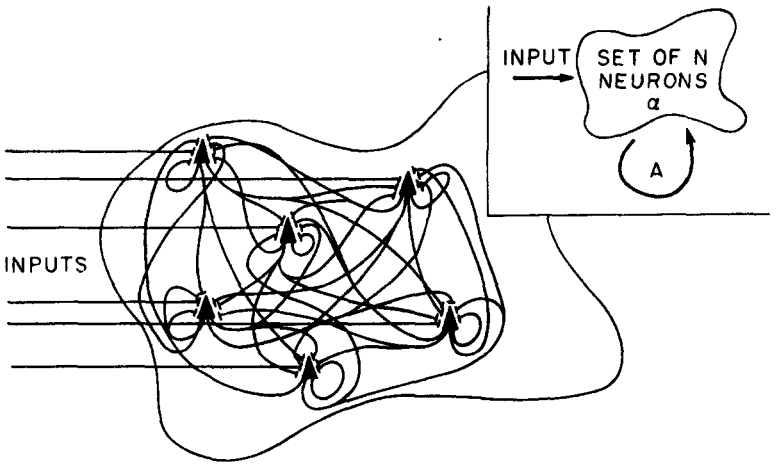

Figure 3. A group of neurons feeds back on itself. Again, note that $N = 6$, as in Figure 1. Note that each cell feeds back to itself as well as to its neighbors.

plicated stimulus, exciting perhaps millions of selective cells, could act as if only a small number of independent elements were involved.

A Model

Assume we have an associative system (Section II) which couples a set of neurons, α, to itself, instead of to a different set of neurons. We again make the approximation that every neuron projects to every other neuron. Let us assume that this feedback connection is through a matrix, **A**, of synaptic connectivities. Figure 3 shows this situation.

Let us consider the case where a pattern of activity on the set α is coupled to itself. The increment in synaptic strength, Δa_{ij} is proportional to the product of the activity shown by the ith neuron, $f(i)$, and the jth neuron, $f(j)$. Note that

$$\Delta a_{ij} = \Delta a_{ji}. \tag{4}$$

This means that **A** is what is called a symmetric matrix.

This implies, in turn, the existence of N mutually orthogonal vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$ such that

$$\mathbf{A}\mathbf{e}_i = \lambda_i \mathbf{e}_i, \quad i = 1, \ldots, N,$$

where each λ_i is a real number. The $\mathbf{e}_i$s are called eigenvectors of **A**, while each λ_i is called the eigenvalue associated with $\mathbf{e}_i$. Since there are N mutually orthogonal eigenvectors, every vector is a linear combination of the eigenvectors. An important consequence of this is that **A** is completely determined by its sets of eigenvectors and corresponding eigenvalues.

In many systems, the eigenvalues and eigenvectors are of great importance. Our system is no exception. Let us consider how a matrix, starting from zero, would develop.

Assume we present K orthonormal inputs, that is, mutually orthogonal unit vectors, each input $\mathbf{f}_i$ appearing k_i times. Then by our associative model, each $\mathbf{f}_i$ is an eigenvector of **A** with corresponding eigenvalue k_i, since

$$\mathbf{A} = \sum_{i=1}^{K} k_i \mathbf{f}_i \mathbf{f}_i^T$$

and

$$\mathbf{A}\mathbf{f}_j = \Big[\sum_{i=1}^{K} k_i \mathbf{f}_i \mathbf{f}_i^T\Big]\mathbf{f}_j = k_j \mathbf{f}_j, \quad j = 1, \ldots, K.$$

If $K < N$, the remaining $N - K$ eigenvectors of **A** have zero eigenvalues. We see then, in this case, that the eigenvectors of **A** with large eigenvalues will tend to correspond to com-

monly presented patterns, and the eigenvalue (in a more complex system) will be at least a rough estimate of the frequency of presentation.

Suppose now the inputs are arbitrary except that the average input

$$\sum_{i=1}^{K} k_i \mathbf{f}_i / \sum_{i=1}^{K} k_i$$

is zero. This important assumption is acknowledgment of the fact that inhibition and excitation are equally important and equally prominent in the operation of the nervous system. We make the above assumption in the form it takes because it is very convenient for our mathematical interpretation, but the exact conditions that occur in the nervous system are beyond our present knowledge.

With this assumption, $\mathbf{A}$ is then a scalar multiple of the sample covariance matrix of the inputs, since the latter is given by

$$\sum_{i=1}^{K} p_i \mathbf{f}_i \mathbf{f}_i^T - \Big[\sum_{i=1}^{K} p_i \mathbf{f}_i\Big]^T \Big[\sum_{i=1}^{K} p_i \mathbf{f}_i\Big], \qquad (5)$$

where $p_i = k_i / \sum k_i$, $i = 1, \ldots, K$. We should point out that this matrix is positive semi-definite, that is, that all the eigenvalues are greater than or equal to zero.

From principal components analysis, we find that the eigenvectors of $\mathbf{A}$ are related very strongly to the inputs in the following way. We will use terms from probability theory. Let $\mathbf{f}$ be the random vector which takes the values $\mathbf{f}_i$ with probability p_i, $i = 1, \ldots, K$. Let $\mathrm{cov}(\mathbf{f})$ denote the covariance matrix of $\mathbf{f}$, given by Equation 5. The main result from principal components analysis states that any unit vector $\mathbf{u}$ that maximizes the variance of the random variable $c = \mathbf{u}^T\mathbf{f}$ over all unit vectors must be an eigenvector $\mathbf{u}_1$ of $\mathrm{cov}(\mathbf{f})$ with the largest eigenvalue λ_1. The variance of $c_1 = \mathbf{u}_1{}^T\mathbf{f}$ turns out to be λ_1. The maximum variance of $\mathbf{u}^T\mathbf{f}$ over all unit vectors orthogonal to $\mathbf{u}_1$ is the second largest eigenvalue λ_2 of $\mathrm{cov}(\mathbf{f})$, and it must occur at the corresponding eigenvector. This maximal principle follows through for all the eigenvalues where at the jth step we maximize all of the unit vectors orthogonal to the $j - 1$ eigenvectors already established. The random vector $\mathbf{f}$ can then be expressed as

$$\mathbf{f} = \sum_{i=1}^{N} c_i \mathbf{u}_i, \quad c_i = \mathbf{u}_i{}^T\mathbf{f}.$$

The c_i values turn out to be mutually uncorrelated, and we see that

$$\mathrm{var}\|\mathbf{f}\|^2 = \mathrm{var} \sum_{i=1}^{N} c_i \mathbf{u}_i{}^T \Big[\sum_{i=1}^{N} c_i \mathbf{u}_i\Big]$$

$$= \sum_{i=1}^{N} \mathrm{var}\, c_i = \sum_{i=1}^{N} \lambda_i.$$

Each input then is a linear combination of the eigenvectors of $\mathbf{A}$, that is,

$$\mathbf{f}_i = \sum_{j=1}^{N} c_j{}^i \mathbf{u}_j,$$

where, qualitatively, the eigenvectors with large eigenvalues account for most of the differences between inputs. Furthermore, no correlation exists between $c_{j_1}{}^i$ and $c_{j_2}{}^i$, where $j_1 \neq j_2$. The eigenvectors of $\mathbf{A}$ can therefore be considered as the basic components of the inputs.

As an additional comment, we know from linear algebra that the eigenvectors give the maximum responses from the system. Over all vectors of unit length (that is, all vectors $\mathbf{x}$, such that $\|\mathbf{x}\| = 1$), the largest value of $\|\mathbf{Ax}\|$ occurs when $\mathbf{x}$ is the eigenvector $\mathbf{e}_i$ corresponding to the largest eigenvalue. The next largest value of $\|\mathbf{Ax}\|$ over unit vectors orthogonal to $\mathbf{e}_i$ occurs when $\mathbf{x}$ is the eigenvector $\mathbf{e}_j$ corresponding to the second largest eigenvalue, and so on.

Significance of Feedback

Let us consider what this system might do to an input to the set of neurons from the sensory receptors or from an earlier stage of processing.

Suppose the input is composed of one of the eigenvectors of the feedback matrix that has a large positive eigenvalue. This means that the activity pattern will pass through the feedback matrix unchanged in direction. It will add algebraically to what is already going on in the set of neurons. Since the eigenvalue is positive, it will add to ongoing activity, that is, to the eigenvector. The larger amplitude pattern will be fed back again and the output from the

feedback matrix will again add to the activity in the set of neurons. This is positive feedback, and the amplitude of the eigenvector will grow. Depending on the details of the system, it may grow without bound or merely show a longer and stronger response, but it will be increased in strength relative to other patterns. Consider an input which contains a contribution from an eigenvector with small or zero eigenvalue. Positive feedback will not significantly enhance this pattern, and the amplitude may increase very slowly or not at all.

Thus the input pattern, after a while, will tend to be composed of only the components of the original input that have large positive eigenvalues, and only these components will participate in further processing. We have just seen that the eigenvalue is in some sense a measure of the importance of the particular eigenvector in discriminating among different members of the items the system has learned, so the patterns with large eigenvalues are the most important patterns for the system. Since this is exactly the behavior we want from distinctive features, *let us specifically identify the eigenvectors of the feedback matrix with large positive eigenvalues as the distinctive features of the system.* We see that in all important aspects of their behavior they act as we would like distinctive features to act.

Let us note as well that a similar technique is used in pattern recognition and statistics. We have previously mentioned the similarity of this analysis to principal components analysis, and many pattern recognition tasks use very similar techniques because of their theoretical optimality (Young & Calvert, 1974, Chap. 6).

The operation of this system gives us good insight into the profound practical differences between the brain and a digital computer. The reason such pattern recognition techniques are not used more widely is economic: Excessive computation time is required because of the large matrix operations involved. However, we see that an adaptive parallel feedback system with highly interconnected analog elements can process an input so it most strongly weights its features, according to importance, in only one step. We shall discuss this process further in the next few sections.

Some Comments on Neurobiology

The idea of a distributed feedback network is consistent with much we know about the physiology and anatomy of cerebral cortex. We should emphasize as well that the actual pathways may sometimes be more complex than our model. For example, inhibition seems often to be accomplished in mammals by inhibitory interneurons. This extra neuron need not affect the mathematics of the model.

There are a number of feedback systems in cortex and thalamus. Perhaps the most attractive candidate to perform operations like those we discuss in this article is the very rich network of recurrent collaterals of cortical pyramidal cells. Recurrent collaterals are axonal fibers which branch off the axon of a pyramidal cell, loop back into the nearby gray matter, and synapse extensively with the dendrites of nearby pyramids over a range of several millimeters. Globus and Scheibel (1967) comment that the recurrent collaterals of pyramids are the most common class of fibers in neocortex.

Freeman and his collaborators (see Freeman, 1975, for a detailed review) have worked extensively, both experimentally and theoretically, on the electrical activity and connections of prepyriform cortex (olfactory cortex) and olfactory bulb in cats. This primitive cortex may show in simple form the connections that are more highly developed in neocortex. Freeman has had considerable success in applying linear systems analysis to these networks. He has incorporated in some of his models the type of excitatory feedback that we have suggested as a basis for feature analysis. An excitatory collateral system from prepyriform pyramids onto nearby pyramids has been described anatomically, as has a similar collateral system in hippocampus, another primitive cortex. The anatomy of these connections, and other recurrent systems in neocortex, as well as some of the physiology, is reviewed in Shepherd (1974).

Higher order loops, from one region of cortex to another and back, or from thalamus to cortex and back, are also common in the brain and may also be candidates for the psychologically significant feedback loops we would like to find. However, the physiological

data on these systems, other than those suggesting their existence and importance, are presently sparse. Further discussion of evidence for physiological mechanisms that may participate in feedback interactions is given elsewhere (Anderson, 1977, Note 1).

More Detailed Study of Feedback

In the study described in detail in Anderson (1977), we used linear systems analysis to obtain exact solutions of an interesting case. We showed that the response of the feedback system had the properties we claimed for it. Suppose we represent an input as a weighted sum of eigenvectors of the feedback matrix, **A**. Since the eigenvectors are orthogonal, they can serve as a basis set. Suppose the input is then presented to the system. We showed that after a period of time, the relative weights of the eigenvectors changed, and that the eigenvectors with large positive eigenvalues were much more heavily represented in the activity of the set of neurons than were the eigenvectors with smaller eigenvalues. We also showed that the response of the system to eigenvectors with large positive eigenvalues lasts longer. Some variants of the model have regions of stability as well; that is, the response of the system dies back to zero or remains bounded as long as the largest eigenvalue does not exceed a certain critical value. Above this value, the system "blows up"; that is, the amplitude of the activity pattern increases without bound. In this calculation, the time constants of the feedback system of the brain were quite long relative to the duration of the neural activity representing the sensory input.

In this article, we consider a slightly different model. We assume, essentially, that the time constant of the feedback is fast compared to the sensory input. Thus, feedback and current activity directly add in the same time period.

As a speculative comment, we observe that the first model bears a certain impressionistic relation to the way visual processing in reading has been conjectured to be performed. A visual input is received and coded with a great burst of activity from the sensory neurons, which then become relatively quiet; however, processing continues for 200 or 300 msec before the eyes jump to a new location in a saccade. Then the process is repeated. We might conjecture that the auditory system—particularly in speech perception—is following a somewhat different strategy. Phonemes follow one another in quick succession, tens of milliseconds apart, and the nervous system must respond more quickly to a constantly changing input.

There are many ways we could set up the feedback model we shall discuss for the remainder of this article, but we initially chose one that was exceptionally convenient for computer simulations. Our general philosophy of modeling is always to try simple things first. We suspect that similar but more complex variants will not show very different qualitative behavior.

The dynamics of the system are assumed to occur in discrete time. Let $\mathbf{x}(t)$ denote the activity vector (the "state vector") at time t. Integer values are taken by t. The activity at time $t+1$ is assumed to be the sum of the activity at time t and the action of the feedback matrix on the activity at time t. The summing of the output of the feedback system and the activity at this time is assumed to occur in the same time quantum, which implies that their time courses are comparable. Thus, we have

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) = (\mathbf{I} + \mathbf{A})\mathbf{x}(t), \quad (6)$$

where $\mathbf{I}$ is the identity matrix. Throughout this discussion, we let $\mathbf{A}$ be fixed. The system is, indeed, a positive feedback system, due to the fact that all eigenvalues of $\mathbf{A}$ are nonnegative. To see this, let $\mathbf{e}_1, \ldots, \mathbf{e}_N$ denote the orthonormal eigenvectors of $\mathbf{A}$ with corresponding nonnegative eigenvalues λ_i, $i = 1, \ldots, N$. Then $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{i=1}^{N} \lambda_i \mathbf{e}_i \mathbf{e}_i^T.$$

Let

$$\mathbf{x}(t) = \sum_{i=1}^{N} x_i \mathbf{e}_i.$$

Then

$$\begin{aligned} \mathbf{x}(t+1) &= (\mathbf{I} + \mathbf{A})\mathbf{x}(t) \\ &= \Big[\sum_{i=1}^{N} (1 + \lambda_i)\mathbf{e}_i \mathbf{e}_i^T\Big] \sum_{j=1}^{N} x_j \mathbf{e}_j \\ &= \sum_{i=1}^{N} x_i (1 + \lambda_i)\mathbf{e}_i, \end{aligned} \quad (7)$$

so that

$$\|\mathbf{x}(t+1)\|^2 = \Big[\sum_{i=1}^{N} x_i(1+\lambda_i)\mathbf{e}_i\Big]^T \sum_{j=1}^{N} x_j(1+\lambda_j)\mathbf{e}_j$$

$$= \sum_{i=1}^{N} x_i^2(1+\lambda_i)^2 \geq \sum x_i^2 = \|\mathbf{x}(t)\|^2. \quad (8)$$

Thus, the length of the activity vector is nondecreasing at every step. If $\mathbf{x}(t)$ is made up only of eigenvectors of $\mathbf{A}$ with zero eigenvalues, then from Equation 6 we see that $\mathbf{x}(t+1) = \mathbf{x}(t)$, so if the system starts at one of these points it stays there for all time. All other vectors will respond to the system. In fact, for these vectors strict inequality holds in Equation 8.

At this point we must introduce an important feature into the model, one that will break with the linearity that we have assumed up to now. We pointed out that in one version of the feedback model certain positive eigenvalues of $\mathbf{A}$ were stable, in that the system activity did not grow indefinitely large as time progressed. In the present version of the model, the same is not true; activity will grow without bound. Unfortunately, the desirable features of positive feedback are exactly the ones that cause catastrophes. This is inappropriate behavior for a system that requires unquestioned stability at all times. The cases of unstable neuronal discharge that we know of give rise to highly pathological seizure states. The normally functioning brain seems to be extremely stable and resistant to "runaway," which is a very important observation in view of the powerful excitatory mechanisms that the brain contains.

The simplest way of containing the activity of the system is to use the fact that neurons have limits on their activities: They cannot fire faster than some frequency (usually around several hundred spikes per sec, and in some auditory units, as fast as 1,000 spikes per sec), and they cannot fire slower than 0 spikes per sec. Thus, there are positive and negative limits on firing rate.

Suppose we incorporate this property into our model. A particular activity pattern is a point in a very high dimensionality space. The coordinate axes correspond to activities of individual neurons. Thus, putting limits on firing frequency corresponds to putting the allowable activity patterns into a box. Since large, high-dimensionality vectors describing the system are often called "state vectors," we name this model the "brain-state-in-a-box" model, with the associated image of a state vector, a point in space, buzzing around like a bee under the influence of input and feedback.

We formalize this situation by assuming (possibly unrealistically) that the limits are symmetrical around the origin. That is, saturation in the system is achieved by keeping the system on or inside the cube in N-dimensional space defined by $x_i = \pm C$; $i = 1, \ldots, N$, where x_i is the activity of the ith neuron.

Applying this assumption to our model, at each time step, the activity vector first undergoes the change given by Equation 6, and then each coordinate which is either greater than C or less than $-C$ is replaced by C or $-C$, respectively. Using the maximum (max) and minimum (min) functions, we can write the dynamics of the saturating system for the ith element of $\mathbf{x}$, at time $t+1$, as

$$x(t+1)_i = \max(-C, \min(C, [(\mathbf{I}+\mathbf{A})x(t)]_i)), \quad i = 1, \ldots, N.$$

Our primary interest in the cube is the corners, that is, points of the form

$$C(\pm 1, \pm 1, \ldots, \pm 1)^T.$$

There are 2^N different corners. Suppose $\mathbf{x}_0$ is a corner with the property that each coordinate of $\mathbf{A}\mathbf{x}_0$ is nonzero and carries the same sign as the corresponding coordinate of $\mathbf{x}_0$. Then it follows that there exists a neighborhood N of $\mathbf{x}_0$ (for our purposes, a neighborhood of $\mathbf{x}_0$ can be thought of as a ball centered about $\mathbf{x}_0$) such that if the activity vector ever lands in the intersection of N and the cube, it eventually reaches $\mathbf{x}_0$ and stays there for all future time. Points of this type are called stable. If some of the coordinates of $\mathbf{A}\mathbf{x}_0$ are zero, then $\mathbf{x}_0$ can still be stable if there is a neighborhood N of $\mathbf{x}_0$ such that each element $\mathbf{x}$ of N, where $\mathbf{x} \neq \mathbf{x}_0$, satisfies the above condition. It is easy to see that if a corner is stable its antipodal corner is stable.

If an eigenvector of $\mathbf{A}$ lies along a diagonal

of the cube, then the two corresponding corners are stable. Moreover, all points (except 0) sufficiently close to the eigenvector will wind up in either corner. The collection of points in the cube that are attracted to a stable corner $\mathbf{x}_0$ is called the region of stability of $\mathbf{x}_0$. We will use these regions (where all points are considered equivalent, since they all end in the same corner) in our applications in the next two sections.

Our experience with computer simulations and our intuitions suggest that the qualitative behavior of the system is straightforward. Suppose we start off with an activity vector which is receiving powerful positive feedback; that is, without limits the vector would grow indefinitely. The vector lengthens until it reaches one of the walls of the box; that is, one of its component neurons reaches the firing limit. The vector will try to get longer, but it cannot escape from the box. Thus, it will head for a corner, where it will stay if the corner is stable. It can be shown that in many cases only some corners are stable.

An Important Special Case

In some of the calculations we shall perform in the next two sections, we must specify how the eigenvalues change with time. In one important special case, this is extremely simple.

We observe that in a saturating model, the final state of the system is always a corner. Thus, if we increment matrix $\mathbf{A}$ by the final state given by that corner, the incremental set of synaptic changes will always be the same.

If the corners are eigenvectors, which could be the case after a long time learning only corners, then the eigenvalues will be increased by a constant amount every time that the corner corresponding to that eigenvector appears. This can be seen easily. If $\mathbf{f}$ is both an eigenvector and a corner, then $\mathbf{Af} = \lambda\mathbf{f}$. The incremental change in $\mathbf{A}$ due to learning is given by $\Delta\mathbf{A} = \eta\mathbf{f}\,\mathbf{f}^T$, where η is a learning parameter. If $\mathbf{f}$ is normalized so $\|\mathbf{f}\| = 1$, then

$$\Delta\mathbf{Af} = \eta\mathbf{f}(\mathbf{f}^T\mathbf{f}) = \eta\mathbf{f}$$

and

$$(\mathbf{A} + \Delta\mathbf{A})\mathbf{f} = (\lambda + \eta)\mathbf{f}. \qquad (9)$$

A Further Comment

Mathematically, the covariance matrix defined in Equation 5 is positive semidefinite; that is, all the eigenvalues are either zero or positive. One reason for this is the presence of large positive values on the main diagonal, the a_{jj}. This is so, since if we have K stored inputs with each input f_i appearing k_i times, then

$$a_{jj} = \sum_{i=1}^{K} k_i f_i^2(j).$$

This term corresponds to the feedback of a neuron on itself. Although so-called "autapses" are found occasionally in cortical pyramidal cells (van der Loos & Glaser, 1972), most of the physiologically studied cases where a cell feeds back on itself involve special neural circuits, often inhibitory. A well-known example is the Renshaw cell system, a special class of cells providing inhibitory feedback to spinal motor neurons. We need not be restricted to the values given above for the main diagonal of the matrix. We can let these values be zero, if we wish, as we show in Appendix C.

Zero Activity Level

The model presented previously predicts that almost all neurons will be firing at either maximum or minimum rate. Clearly, in a real nervous system, many, if not most, neurons probably will not respond to a given stimulus, although a sizable fraction may participate in the activity pattern. There are several ways we can have a number of zero elements in our activity vectors. Even in the model given here, where feedback is very powerful and connectivity is complete, input patterns will occasionally give rise to a zero activity level of a neuron in the set if all the eigenvectors with nonzero eigenvalues have zero in the coordinate corresponding to that neuron. If the space is even moderately filled with eigenvectors (i.e., if K is somewhat comparable to N and many eigenvalues are nonzero) or if noise is introduced into the system, then this is very uncommon. However, if we assume any one of several mechanisms—thresholds of feedback, adaptation, restricted connectivity—we can easily produce a system where many elements

in the N-dimensional space remain at zero, even in the final state.

In the previously mentioned paper by Barlow (1972), it was pointed out that cortical neurons seem to respond in a characteristic way. A linear model would make no restrictions on the activity pattern. That is, we could have activity patterns containing a great many small changes in activity or a few large ones. But the nervous system seems to have chosen the latter kind of response pattern. We see, if we make appropriate assumptions, that we may restrict saturation to a set of the most "important" neurons for the discrimination of the input stimulus.

Cells in cortex are rather selective in that they do not respond to most inputs but have a strong response to some. Our brain-state-in-a-box model has something of this aspect to it, in that in the final state cells may be fully on, fully off, or not participating in the activity pattern at all. It would be of interest to look at cortical neurons in the awake, behaving animal in light of our proposed model.

V. Categorical Perception

The best evidence for distinctive feature analysis comes from linguistics. Therefore, it seemed natural to us to try to apply this model to find if it agreed with the kind of perceptual analysis that occurs in speech perception.

Preprocessing

Distinctive features are usually viewed as a system for efficient preprocessing, whereby a noisy stimulus is reduced to its essential characteristics and decisions are made on these. We showed that the feature vectors arising in the model are akin to those found in principal components analysis and are indeed the most useful patterns for these kinds of discriminations. More, however, is suggested by the model. The brain-state-in-a-box model suggests that the final output of such a system is a stable state corresponding to a corner of the high-dimensional hypercube formed by saturating neural activity patterns. A good preprocessor should put a noisy input into a more or less noise-free standard form for use in later stages of perception. Analysis and noise-free resynthesis of the signal is not necessary if the whole system gives directly as output a noise-free standard form. Features may never occur by themselves, for example, but only serve the function of "steering" the noisy input into the appropriate corner.

Categorical Perception

One of the characteristics that seems to be found in speech perception is "categorical perception." When stimuli of an artificially constructed set vary continuously across a feature dimension, the listener is experimentally found to have difficulties making discriminations within categories. Thus, voice onset time—the time when the vocal cords start to vibrate— is the physical feature used in making the /p/ versus /b/ discrimination. It is found that the listener cannot discriminate well among stimuli classified as /p/ even though voice onset times may vary considerably. Conversely, discriminations across a category boundary are very good. Two stimuli differing only slightly acoustically, with small differences in voice onset time, are well discriminated if they happen to fall on different sides of the category boundary. It has been suggested (see review by Studdert-Kennedy, 1975, for references to the literature) that incoming speech stimuli are analyzed by both a "categorizer" and a "precategorical acoustic store" (PAS) analogous to iconic memory in vision. It is found experimentally that consonants, particularly stop consonants, display much stronger categorical perception than do vowels. The reason for this, it is suggested, is that consonants are so short in duration that the sensory information contained in the PAS decays too quickly to be used to make discriminations, while the categorical information is much more stable. This hypothesis predicts that vowels, for example, will show more aspects of categorical perception if they are degraded in noise or shortened in duration, both of which would primarily distort the PAS. The predicted result is found. Thus categorical perception may be a very important—indeed, characteristic—property of language perception.

The Brain-State-in-a-Box Model

The brain-state-in-a-box model presented here is a model of categorical perception in a rather pure sense. All points in a region are classified together (differences vanish) and all points in different regions, even if initially very near each other, are classified apart. Thus, no discriminations are possible within the regions, and perfect discrimination occurs between regions.

In previous work (Anderson, 1977; Anderson, Silverstein, & Ritz, in press) we showed that it was possible to use the saturating model to categorically perceive "vowels,"—that is, vectors corresponding to the outputs from a bank of frequency filters when spoken vowels were the input. For the simulation, we used an eight-dimensional space (i.e., an eight-dimensional vector representing each vowel) and showed that if we let our feedback matrix **A** learn according to the rules presented earlier, the system would eventually come to act as a good preprocessor. By this, we meant that nine initial vowel codings, derived from experimental data and often starting close to one another, would, after the operation of feedback, be associated with separate stable corners. The system "learned" to do this with about 20,000 total presentations of the nine vowels in the set of stimuli. Initially the system classified different vowels in the same corner and took many computer iterations (i.e., passages through the feedback system according to Equation 6), but after 20,000 learning trials, all vowels had their own final corner and the system performed correct classification after only seven iterations.

This simultaneous increase in both accuracy and speed of classification struck us as a good demonstration of what we feel are the practical virtues of such a system for perception.

A Computer Simulation

Testing categorical perception experimentally involves two parts: identification and discrimination. A set of artificial stimuli are first constructed which vary smoothly from one speech sound to the other. Sometimes it is easy to construct a continuously graded set of stimuli—for example, voice onset time, where the physical feature is quite clear-cut—but often it is more difficult, as for some of the consonant-vowel formant transitions. The essence of the effect is that subjects do not *hear* the continuous variation, but instead perceive an abrupt shift from one sound to the other.

For the *identification* experiment, subjects are presented with different stimuli and asked to say which phoneme they hear. Most of the published *discrimination* data have relied on the psychophysical technique called "ABX discrimination." Two stimuli—A and B—which differ by a few milliseconds in voice onset time are presented. The subject is then presented with a third stimulus—either A or B—and is required to say whether the third stimulus matches the first or the second. Thus, if the subject cannot make a discrimination he must guess, which means he will be right 50% of the time, on the average.

We decided that the most straightforward computer simulation we could do would be to simply duplicate these experiments, using the outputs from the saturating neural model, and to show that it behaved in a way which looked like the human data. Since we know very little, to say the least, about the neurophysiology of speech perception, it seems to us premature to do more than point out general similarities of the model with the phenomenon of categorical perception.

We used an eight-dimensional system again, which seemed to us large enough to be indicative of the behavior of a real system, yet small enough to be manageable and of reasonable cost. We assumed we had two eigenvectors of the feedback matrix with nonzero eigenvalues pointing toward two corners. Corner A was $(1,1,1,1,-1,-1,-1,-1)$, and Corner B was $(1,1,-1,-1,1,1,-1,-1)$. This situation is shown in Figure 4. To perform our simulations, we had to have a set of stimuli which varied smoothly from one feature to the other. We picked 16 equally spaced points along the unit sphere through the plane containing the two eigenvectors and the origin.

In a real system, of course, there would be noise as well as other eigenvectors. Our simulations would show perfect categorization if there were no noise. We added zero-mean Gaussian noise to each of the eight components of the initial position. In the figures, *SD* repre-

sents the value of the standard deviation of this noise. Figure 4 shows the average length of the noise vector added to the starting point. Note that an *SD* of .4 corresponds to an average added noise about 1 unit long, or a distance as long as that from the origin to the starting point. This is a great deal of added noise. An *SD* of .1 corresponds roughly to the distance between initial positions two steps apart. This seemed to us to be a reasonable range of lengths to cover, and one that might be found in a perception system under normal operating conditions. The noise vector is not constrained to lie on the plane containing the eigenvectors and the origin. The geometry of this model is very complicated.

Results of the Simulations

We did a Monte Carlo simulation of 100 presentations of each starting point with added random noise. When the final state of the system was Corner A, we called it Response A,

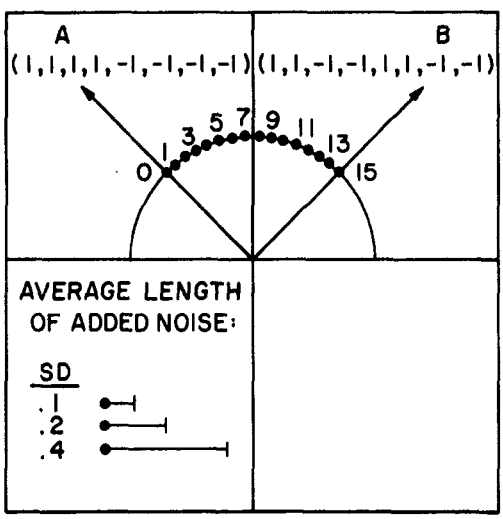

Figure 4. In the simulation of categorical perception presented in Section V, an eight-dimensional system is constructed with two eigenvectors which point toward two corners. Limits of saturation are two units from the origin. Starting positions for test inputs are equally spaced points along the unit circle, numbered 0 to 15. Zero-mean Gaussian noise of standard deviation (*SD*) is added to each of the eight components. The average resulting lengths of the noise vectors corresponding to different *SD*s are shown in the lower-left quadrant. This figure portrays a two-dimensional slice of an eight-dimensional space.

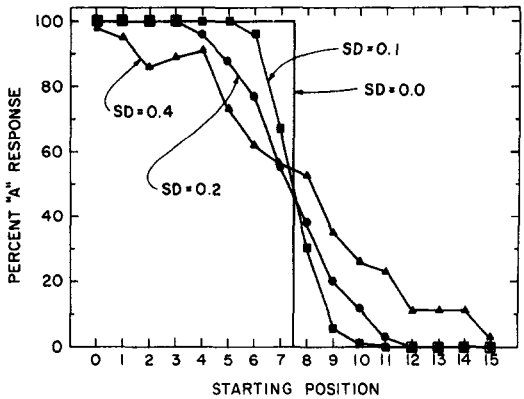

Figure 5. In the "identification" experiment, a "response" corresponds to the corner in which an input ends. The curves show the results of adding various amounts of random noise to a given starting position. One hundred trials were used for each point.

and similarly for B. Figure 5 shows the results of this simulation for several values of *SD*. If there is no added noise (i.e., $SD = 0$), the system categorizes perfectly. If there is noise of *SD* equal to .4, there is a nearly linear decrease in the probability of Response A as the starting point moves from 0 to 15. Intermediate values of noise produced curves which look very much like the published data presented in Studdert-Kennedy (1975), Pisoni (1971), or Eimas and Corbit (1973).

The real test of a categorical perceiver is the difficulty it has performing discriminations within categories. Since experiments often use an ABX paradigm, we simply did so in our simulation, although it was expensive in terms of computer time. We took as an initial input a starting point at, say, point n. We then added random noise and noted which corner appeared as the final state. We repeated the process for an input at point $n + 4$ or $n + 6$. We again noted the final state. Then we randomly chose the first or second starting point, added different noise, repeated the process, and attained a third final state. Finally we determined whether the final state agreed with what it was supposed to ("correct") or whether it was in error. For several combinations of corners, it was necessary to require the program to guess, corresponding to a forced choice. In this case, it was correct or incorrect with a probability of .50. We did Monte Carlo simulations using only 40 trials per point to save

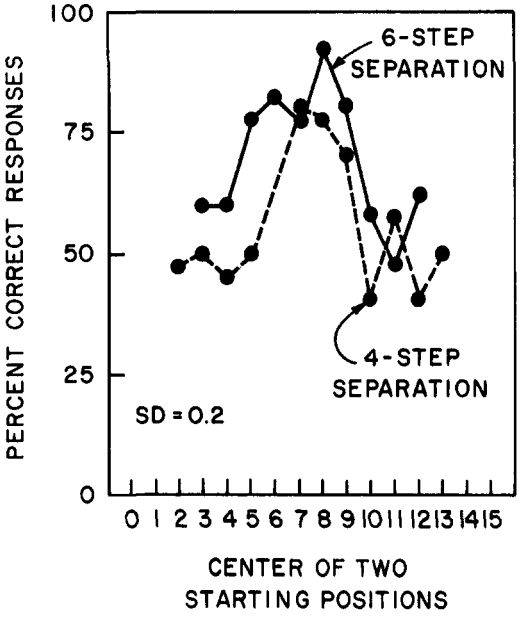

Figure 6. The "discrimination" experiment had the computer perform an ABX experiment. Two starting points four or six steps apart corresponded to A and B. The program noted which final state appeared for each input. A third input—X—was chosen from either A or B, and the computer decided whether it was classified correctly, incorrectly, or whether guessing was necessary. Added Gaussian noise had standard deviation of .2 units. For each point, 40 responses were averaged.

computer money. The results are given in Figure 6.

We have shown only data for an added noise of *SD* equal to .2 units, since this value had identification functions that looked appropriate to us in light of the experimental data. The discrimination functions look similar to what is seen—discrimination shows a pronounced peak when the category boundary separates the starting points and shows a drop if both points start on the same side of the boundary.

An interesting by-product of the identification simulation was an estimate of the "reaction time" required for the system to classify an input as one or the other corner. We simply counted the number of iterations required to saturate the system and attain the final state. This number was averaged over the 100 trials for each point and is plotted in Figure 7 for *SD* equal to .2. Decreasing noise seemed to slightly increase average categorization time. Maximum change in the number of steps required between conditions was around 20%, and the shape of the curve was quite similar in all cases.

The required number of steps was about twice as great when the stimulus started from a point near the category boundary. An effect like this has been observed. Data from Pisoni and Tash (1974) for an experiment using voice onset time showed a change in reaction time from about 475 msec in the centers of the categories to about 575 msec at the category boundary, as determined by the identification function. The distribution of reaction times was symmetrical around the category boundary, as is ours.

Adaptation

Since we potentially have a learning system, and since we have already shown in our previous discussion how the eigenvalues change when the system is learning eigenvectors pointed toward corners (see Equation 9), it seemed an obvious extension of our simulations to look briefly at the effects of "adaptation" on categorical perception.

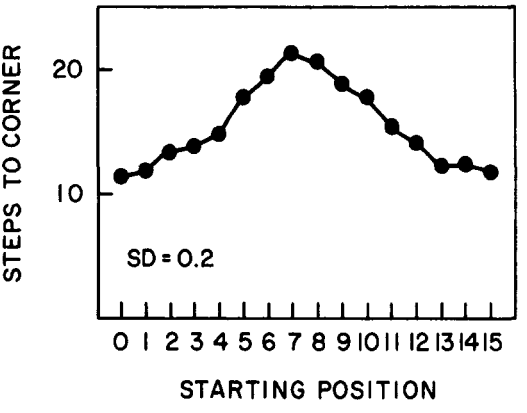

Figure 7. The number of computer iterations required for all components of the system to saturate is plotted here. Added noise had a standard deviation of .2. As starting position varied, the number of steps required to saturate increased near the boundary. Different noise conditions showed very similar patterns. If each iteration or its equivalent takes about the same time to perform in a real nervous system, then this graph could be interpreted as a rough indicator of the pattern of reaction times that would be observed in categorical perception as an input stimulus is moved from one category to another across a boundary. Each point is the average of 100 responses.

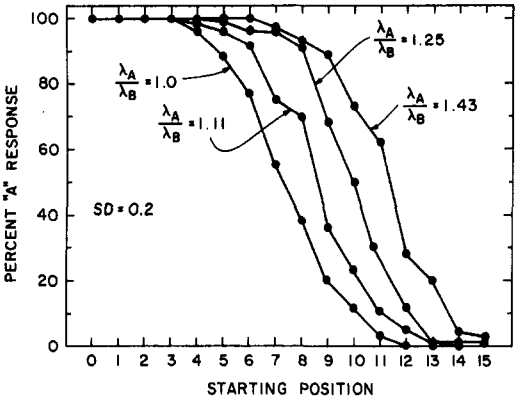

Figure 8. When the eigenvalues associated with the eigenvectors are not equal, the identification function shifts. If we assume repeated presentation of an adapting stimulus causes synaptic "antilearning," then this simulation corresponds to how we would expect our system to behave if one feature vector is adapted. Each point is the average of 100 responses.

Cooper (1975) has reviewed some of the adaptation literature and points out some of the general features experimentally observed. Adaptation is typically produced experimentally by having the subject listen to a minute of the adapting stimulus at a presentation rate of 2 per sec. After adaptation, the identification function for the unadapted stimulus (A) indicates a shift toward the adapted stimulus (B). The curve for the identification function is displaced parallel to its initial position; as Cooper comments, "the slopes of the identification functions obtained after adaptation were as steep as the slopes of functions obtained in the unadapted state" (p. 26). There was evidence of crossed series adaptation; that is, subjects who adapted to voicing in one pair of phonemes also adapted (to a slightly lesser extent) to another pair of phonemes differing in the same feature.

Our simple simulation could not test crossed series adaptation because our system only had two eigenvectors. However, we can easily check the shift in identification function.

We assumed that "adaptation" was equivalent to synaptic "antilearning," or learning according to Equation 2 with negative sign. This is *not* simple fatigue but a more complex and subtle process involving precise synaptic change at many synapses. By Equation 9, we see that adaptation causes the eigenvalue of the adapted stimulus to decrease slightly. The eigenvectors did not shift direction. In our initial simulation (see Figure 5) we assumed the two eigenvalues were equal. For the simulation of adaptation we used exactly the same program but decreased the eigenvalue λ_B. The results for several different ratios of eigenvalue are shown in Figure 8. The noise had *SD* equal to .2 units. The ratio $\lambda_A/\lambda_B = 1.0$ was included to correspond to the initial simulation. It can be seen that the data look very much like parallel displacements of the unadapted curve toward Response B, exactly as seen in the experimental data.

It might be mentioned that we did virtually no searching for parameters in most of these simulations, and the simulations used many of the same programs as those used for probability learning in the next section.

Conclusions

We suggest that the simulations presented here are quite good replicas of the major experimental findings of categorical perception. Our previous work with the simulation of vowel preprocessing, coupled with the work discussed here, suggests that models for speech perception might consider using some such ideas as positive feedback, saturation, and synaptic learning, which seem to be responsible for the interesting effects in our simulations.

This simulation shows clearly some of the features we feel may be typical of natural systems constructed with distributed, parallel arrays of interconnected analog elements. The system acts like an adaptive filter, where the filter characteristics are determined by the past history of the filter. The system does not "analyze" its inputs in the sense that a computer or logician might analyze them, by dissecting them into component parts, but it simply responds to them. However, the response of the system is determined by its past, so its analysis becomes meaningful in terms of this past.

VI. Probability Learning

We shall consider here the set of experiments usually called "probability learning." We shall make a direct application of the model

previously described and show that it can provide a model for this seemingly remote application.

The ability to estimate the probability of occurrence of an event with a random component—whether or not it will rain, who will win an election, what the stock market will do—is obviously important in daily life. It has also served as the basis of a large body of work in experimental psychology. Probability learning has been studied in a number of ways. A classic experimental technique involves a prediction by the subject as to which of two lights will be turned on. Typically, a subject will sit facing the lights. He will be asked, usually immediately after a signal, which light will turn on after a brief interval. The two lights are usually turned on randomly.

The general results are quite consistent from experiment to experiment. The subjects will tend to "match" the probabilities of the events; that is, if the left light occurs with a probability of .8, the subject, after a number of trials, will predict that light about 80% of the time. This result is not what one would expect if one were to assume that the subject was trying to maximize his chances of successful prediction. If he were, then he would choose the most probable light all of the time. Of course, by providing appropriately large payoffs, it is possible to encourage the subjects to change strategies, but, if left to themselves, there is a strong tendency to match probabilities in most simple experimental situations.

Statistical Learning Theory

The prediction of probability matching is a simple consequence of statistical learning theory, which is one of the reasons for the great interest in the effect. Derivation of this result is given in a number of places (see Estes & Straughan, 1954, Estes, 1957, and the other papers collected in Neimark & Estes, 1967).

Let us briefly sketch some of the important aspects of the derivation. We assume there are a number of alternative responses, A_1, A_2, $\cdots A_r$, which can be made by the subject predicting, for example, which of several lights will turn on. The response is followed by an event—$E_1, E_2, \cdots E_r$—which is the actual turning on of one of the lights. Learning theory assumes that the actual event will change the future probability of predicting that event. Suppose the probability of a response, A_j, on the nth-trial is given by $p_{n,j}$. If an event, E_j, occurs, then the probability of Response A_j is increased. In early formulations of the learning rule, the correctness or incorrectness of a prediction determined the subsequent learning. However, it seems that only the occurrence of a particular event is required to increase the associated response probability. As Estes (1972) says in a review article, "the mathematical operator applied on each trial is determined solely by the information received by the subject" (p. 82) as long as special rewards for success and failure are not present. This was shown directly by Reber and Millward (1968). They showed that mere observation of the event lights, in the absence of prediction, produced probability matching in the subjects.

The quantitative rule that the increase in probability of A_j will follow if Event E_j occurs can be derived from statistical sampling theory, or can arise from other assumptions, and follows the form,

$$p_{j,n+1} = (1 - \theta)p_{j,n} + \theta. \tag{10}$$

The parameter, θ, is an important learning parameter in statistical learning theory, and $0 \leq \theta \leq 1$. If Event E_j does not occur, then the probability of making Response A_j in the future will be given by

$$p_{j,n+1} = (1 - \theta)p_{j,n}. \tag{11}$$

Qualitatively, these expressions make the probability of a response tend to *increase* after the associated event, and tend to *decrease* after another event. If we consider the behavior of the probabilities, the expected value for $p_{i,n+1}$ is given by

$$\mathrm{E}(p_{i,n+1}) = (1 - \theta)\mathrm{E}(p_{i,n}) + \theta\pi_{i,n}, \tag{12}$$

where $\pi_{i,n}$ is the current value of the actual probability of occurrence of the Event E_i. Thus, after the first trial,

$$\mathrm{E}(p_{i,2}) = (1 - \theta)\mathrm{E}(p_{i,1}) + \theta\pi_{i,1},$$

after the second trial,

$$\begin{aligned}\mathrm{E}(p_{i,3}) &= (1 - \theta)\mathrm{E}(p_{i,2}) + \theta\pi_{i,2} \\ &= (1 - \theta)^2\mathrm{E}(p_{i,1}) \\ &\quad + \theta(1 - \theta)\pi_{i,1} + \theta\pi_{i,2},\end{aligned}$$

and so on.

For the important special case, often used in experiments, where the probability of an Event E_i is fixed at some value, π_i, the expected value of the probability of making Response A_i after n trials is given by

$$E(p_{i,n}) = \pi_i - (\pi_i - p_{i,1})(1 - \theta)^{n-1}. \quad (13)$$

The second term of this expression will go to zero as n increases if θ is greater than zero, that is, if the system learns at all. Thus, the asymptotic value of $E(p_{i,n})$ is given by the probability of the event, π_i, which is essentially what is observed. This result is completely independent of the initial probability of the response and the learning parameter, θ.

Problems

This elegant result agrees well with many aspects of the data. However, the actual data present some difficulties for the simple theory, which has led to a number of attempts to modify or extend the simple model.

In some of these extensions, subjects are assumed to memorize past sequences and to predict the event that occurred in the past. There are other approaches, involving "hypothesis testing" or other, generally "intelligent" behavior on the part of the subject.

We will assume here that the tendency of the subjects to look for regularities and sequencies in the data interferes with, and is separate from, an underlying straightforward learning phenomenon with no cognitive component.

There are two problems with the simple model, which can be seen in most of the experimental data.

First, when there are only two alternative events, the subjects match probabilities closely but they do not do so exactly. Myers (1976) says in a recent review that the response probability "consistently overshoots" exact matching, and the overshooting has been seen "by almost anyone who has run subjects for more than 300 trials" (p. 173). Overshooting is not large, usually at most a few percent in the two-choice experiment, but it is significant and thus a problem for theorists. The overshooting becomes much larger when there are more than two alternatives (Neimark & Estes, 1967, p. 261). Estes (1964) suggests that overshooting may be a function, to a certain extent, of the instructions given the subject. Myers comments that "investigators who are determined to do so can produce overshooting in a variety of ways; undershooting is somewhat more difficult to achieve, but possible" (p. 175).

Exact probability matching is a good first approximation. However, it is clear that a model should have sufficient flexibility to explain the small but consistent deviations above and below exact matching that are a feature of the experimental data.

Second, a matter that could either be considered aesthetic or substantive, depending on perspective, is the value of the learning parameter, θ. Although θ does not appear in the expression for the asymptote, it determines the rate and trajectory with which the response probability will approach its asymptotic value, because the factor $(1 - \theta)^{n-1}$ appears in the second term of Equation 13. Given the time course of response probabilities, it is possible to estimate θ. When different probability conditions are run with the same subjects, θ does not appear to be constant (Neimark & Estes, 1967, pp. 257–258). The change in θ may be considerable, even in very similar experiments.

A Neural Model for Probability Learning

We have at hand, in our brain-state-in-a-box model, the means for automatically generating discrete responses. We also have, if we introduce noise into the system, a means of introducing probabilities naturally into the model. Although we can handle any number of events, we will restrict ourselves to two at first.

Let us observe here that there are two parts to this model. The first is the *learning feedback matrix*. We have discussed the properties of this matrix in some detail previously. The second part involves the *dynamics* of the system. The initial state moves, under the influence of feedback, into a corner. How fast it moves and into which corner it moves are given by the properties of the feedback matrix at that time. We assume here that the feedback matrix changes only *after* a stable corner has been reached. This means that the matrix does not learn while the state vector is changing; that

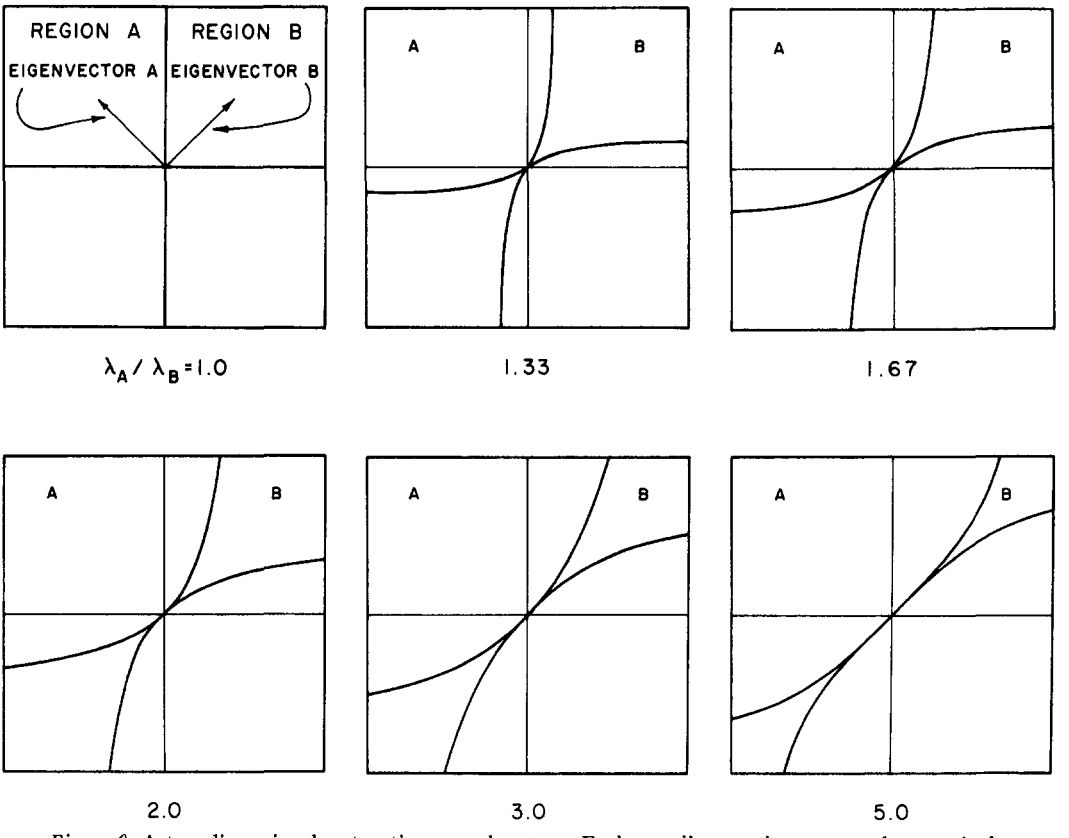

Figure 9. A two-dimensional, saturating neural system. Each coordinate axis corresponds to a single neuron, or the amplitude of a pattern of neuron activity. This activity saturates; thus activity is confined to the interior and edges of the square. There are two eigenvectors in this system, one pointing to each pair of diagonally opposite corners. When feedback is positive, all inputs to points in Region A cause the activity vector to end up in the corner to which eigenvector A points, or in the opposite corner. As the ratio of eigenvalue A to eigenvalue B increases, the size of Region A increases.

is, the time constant of learning is long compared to the dynamics of the system.

Suppose we have a two-dimensional system with the dimensions coupled by a feedback matrix. Thus the two-dimensional limits of saturation form a square. Suppose that the two eigenvectors of the feedback matrix point toward the corners of the square. If we start off at some initial point of the square, positive feedback will drive the system toward one of the corners (see Figure 9).

Let us identify one pair of diagonal corners, associated with one eigenvector, with one response. Let us identify the response associated with corners $(-1, +1)$ and $(+1, -1)$ as Response A and the other pair of corners $(1, 1)$ and $(-1, -1)$ as Response B. Identifying two diagonally opposite corners with one response is made primarily for convenience, and could be avoided if necessary.

This two-dimensional system is not so restrictive as it seems. Assume that one response is associated with one large pattern of neural activity, as it might be in a real system, and the other response is associated with another large pattern, orthogonal to the first. These patterns give rise to two orthogonal vectors in a high-dimensional space. Then by considering the plane through the vectors and the origin, we have a two-dimensional system, with corners associated with eigenvectors. The axes of our two-dimensional system might correspond to amplitudes of complex activity patterns interacting with each other. The tractable case we will discuss might be a

reasonable approximation to a much more complicated, high-dimensionality system.

As we have mentioned, the plane will be divided into two regions. The origin is unstable and need not be considered further. One region will be associated with one response, and the other region with the other. This is shown in Figure 9.

Suppose the subject wishes to make a prediction (i.e., to make one response or the other). There is no lack of random noise in the nervous system. We will assume that the subject initiates the prediction process by starting at some random point on the plane. If the chances of starting at all points on the plane are equal, then by simply calculating the areas associated with each region and dividing by the total area, we know the probabilities of each response.

The calculation of these probabilities is straightforward, though involved, and is presented in Appendix A. The boundaries between regions are simple curves, and the areas of the regions can be calculated exactly, with a simple resulting expression. If λ_A and λ_B are the eigenvalues of the feedback matrix associated with Response A and Response B, respectively, and if

$$\lambda_0 = \frac{\lambda_A}{\lambda_B},$$

then the probability of Response A is given by

$$p_A = \frac{3\lambda_0^2 + \lambda_0^3}{(\lambda_0 + 1)^3}. \tag{14}$$

As the ratio increases, Response A becomes more and more probable. The shapes of the regions for different values of the ratio λ_0 are shown in Figure 9. The probability of Response A versus the ratio of the eigenvalues is shown in Figure 10. Note that p_A is monotonically increasing for $\lambda_A/\lambda_B \geq 1$.

Learning Rules

To make predictions of response probabilities, we must specify only how the eigenvalues change with time. We then have a rule for turning the ratio of eigenvalues into response probabilities.

We observe that in the saturating model, the final state of the system is always a corner.

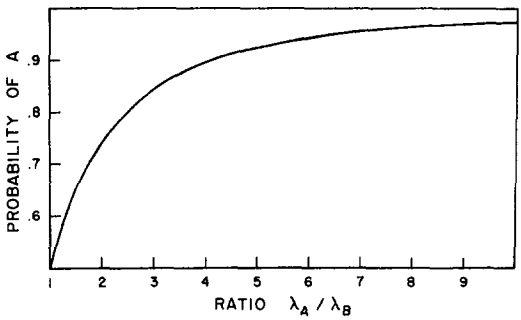

Figure 10. In the model for probability learning presented in the text, it is possible to calculate the probability of each response if the eigenvalue associated with each response is known. Their ratio then gives the probability according to the graph.

Thus, as we showed at the end of Section IV, the eigenvalues follow a very simple learning rule: They increment by a constant amount every time a corner is reached, as shown in Equation 9.

Let us allow that memory decays with time. We also assume that there is a stable part of the eigenvalue which is not affected by either decay or learning. This might correspond to the knowledge, for example, that there are *two* possible responses. Thus, we get a formula for the eigenvalues that is similar to, but not identical with, that used for the probabilities in statistical learning theory.

Let g be a decay factor, $0 \leqq g \leqq 1$, and let η be the amount of increment produced when the system learns. Then, if Event A has just occurred, and if the set of synaptic changes associated with that corner has been added to the feedback matrix, that is, A has been learned, the two eigenvalues for the $n + 1$st trial are given by

$$\lambda_A(n + 1) = 1 + g[\lambda_A(n) - 1] + \eta \tag{15a}$$

and

$$\lambda_B(n + 1) = 1 + g[\lambda_B(n) - 1]. \tag{15b}$$

The first term, "1," is the constant part; the second term is the decay term; and the third part is the increment. If Event B occurs, λ_B receives the increment. Note that the success or failure of the prediction made by the subject does not appear in this scheme.

Since we are concerned only with response probabilities, we are concerned only with the ratio of λ_A and λ_B. In other applications—

prediction of reaction time, for example, in related tasks—we cannot make this assumption and must know the true values of the eigenvalues.

The general similarity of this model to many aspects of stimulus sampling theory should be emphasized. The learning scheme for eigenvalues is a variant of the "linear" learning model. Thus, many of the predictions will be somewhat similar to those of statistical learning theory. The *essential* difference is that here the eigenvalues are what are changing—the size of an eigenvalue becomes a sophisticated measure of trace strength—and that changes in probability of response occur as a result of the operation of a complex process instead of directly.

Asymptotic Behavior

It can be seen that the model is completely defined once we specify the constants and have the sequence of events. We can derive asymptotic behavior of the eigenvalues quite easily in some cases. If there is a constant value of probability of Events A and B, with probabilities π_A and π_B, and if the eigenvalues start at λ_{A_0} and λ_{B_0}, then average values of λ_A and λ_B after n trials are given by

$$E[\lambda_A(n)] = 1 + \frac{\pi_A\eta(1 - g^n)}{1 - g} + g^n(\lambda_{A_0} - 1) \quad (16a)$$

and

$$E[\lambda_B(n)] = 1 + \frac{\pi_B\eta(1 - g^n)}{1 - g} + g^n(\lambda_{B_0} - 1); \quad (16b)$$

and after a very large number of trials, if $g < 1$, the asymptotic values of the averages are given by

$$E(\lambda_A) = 1 + \frac{\pi_A\eta}{1 - g}$$

$$E(\lambda_B) = 1 + \frac{\pi_B\eta}{1 - g}.$$

Since our formula for response probability (Equation 14) requires knowledge of the ratio of the two eigenvalues, it is easy to get an estimate of this value. We can approximate the expected value of the ratio well enough for our purposes by letting

$$E\left(\frac{\lambda_A}{\lambda_B}\right) \cong \frac{E(\lambda_A)}{E(\lambda_B)}. \quad (17)$$

At asymptote, we see that

$$E\left(\frac{\lambda_A}{\lambda_B}\right) \cong \frac{1 - g + \pi_A\eta}{1 - g + \pi_B\eta}.$$

For the values of parameters used here, the approximation is good to better than 1%.

This relation does not predict simple probability matching but a complex relation involving η and g. Using simple calculus, we can see the largest expected value of overshoot will occur when there is no forgetting, that is, when $g = 1$. Then

$$E(\lambda_A/\lambda_B) \approx \frac{\pi_A}{\pi_B}.$$

We can then calculate the estimate of the maximum expected value of probability from Equation 14. The maximum expected overshoot in this case is about 10% when π_A is around .75.

The smallest average asymptotic value of λ_A/λ_B will occur when the second term is zero, that is, when there is no learning. Then

$$\lambda_A/\lambda_B = 1,$$

and the probability of both Response A and B will be .5. Thus, we can predict values of asymptotic probability both above and below probability matching.

It does not seem to be possible to derive exact expressions for the important characteristics of the random variable λ_A/λ_B, such as the expected value and the variance. However, characteristics of the response probability (Equation 14) can, somewhat surprisingly, be calculated, a result shown in Appendix B.

Application to Data

Let us see if we can fit some experimental data with the model as it stands. Our purpose in this section, it must be emphasized, is not primarily to provide a different or quantitatively more satisfactory model for probability learning, but to establish that our neurally based model is capable of handling a wide range of interesting phenomena. We are

not concerned with explaining every aspect of the available data.

A good set of data to use is the series of experiments reported by Friedman et al. (1964). They used an exceptionally large group of subjects—80 Indiana University undergraduates—in a 3-day series of probability learning experiments. In the first 2 days, subjects received sequences of 48 trial blocks. The probabilities of an event changed from block to block ("variable — π series"). Odd numbered blocks had both π_A and π_B equal to .5. Probabilities in the even numbered blocks varied from .1 to .9 in steps of .1, excluding .5. Each subject received all probability conditions during the 2 days. Two different sequences of block probabilities were used, but the first, last, and alternate intermediate blocks were assigned .5 probabilities. On the 3rd day, eight 48-trial blocks were given. The first and last blocks had π_A equal to .5; the middle six blocks had π_A equal to .8.

Friedman et al. (1964) were able to fit the data reasonably well by assuming the basic equations of statistical learning theory, Equations 10 and 11, with a changing θ. The parameter, θ, varied over a considerable range. To fit the transition between a block with π_A equal to .1, to the succeeding block with π_A equal to .5, the best θ was found to be .62. To fit the transition between π_A equal to .4 and π_A equal to .5, θ was found to be .07. Other transitions fell in between, with θ increasing as the difference in probability between blocks increased.

We felt it would be of interest to see if our model would fit the data with a single set of parameters. Our model requires, as does statistical learning theory, detailed knowledge of the stimulus sequences and ordering of blocks to make the best predictions. This data was not available in most cases in the Friedman et al. (1964) paper, although a wealth of carefully gathered and computed averaged data was presented.

Strong learning effects were demonstrated to be present between the beginning and the end of the series, and it was stated that the ordering of blocks had a "highly significant effect" which "severely limits the analyses that can usefully be accomplished with the variable — π sequence" (Friedman et al., 1964, p. 260).

Friedman et al. (1964) were primarily concerned, in the analysis of the variable — π series, with the transitions between a variable probability block and the following .5 probability block, since this allowed them to estimate θ and to get some idea of the general behavior of the statistical learning theory model with respect to the data.

Our model was sufficiently complex and probabilistic to make it difficult to generate simple expressions for some of the averages. Since we had access to a large computer, the most direct way of initially checking the fit of our model to Friedman et al.'s (1964) data was to do a simulation, using 80 computer-generated pseudosubjects and then to compare the results with the real data.

We at first assumed that our pseudosubjects would receive the sequence of stimuli in Friedman et al.'s (1964) Summary Table 7, which has the following sequence of block probabilities: .5, .1, .5, .2, .5, .3, .5, .4, .5, .6, .5, .7, .5, .8, .5, .9, .5. Every pseudosubject received an individual set of events generated with probabilities given by the probability sequences. Eigenvalues associated with responses of the pseudosubjects were calculated according to the learning scheme, and the probabilities of response of each subject were calculated and averaged across subjects. This meant the pseudosubjects did not actually make pseudoresponses, which would be then processed as the responses of the real subjects were, so the data have lost a significant source of variance. The statistics can also be calculated directly by use of the formulas in Appendix B. The resulting average probabilities were then compared with the real data in Friedman et al.'s (1964) Table 7. Figure 11 shows the results.

A crude search was made in order to see which set of parameters produced a simulation most resembling the actual data. The best fitting parameters were found to be $\eta = .3$ and $g = .90$. Parameters were not especially critical. The value of the decay factor, .9, was such as to make the contribution to the current eigenvalue from trials over 48 trials (i.e., a block) in the past negligible, so this simulation could be viewed as a sequence of fits of transi-

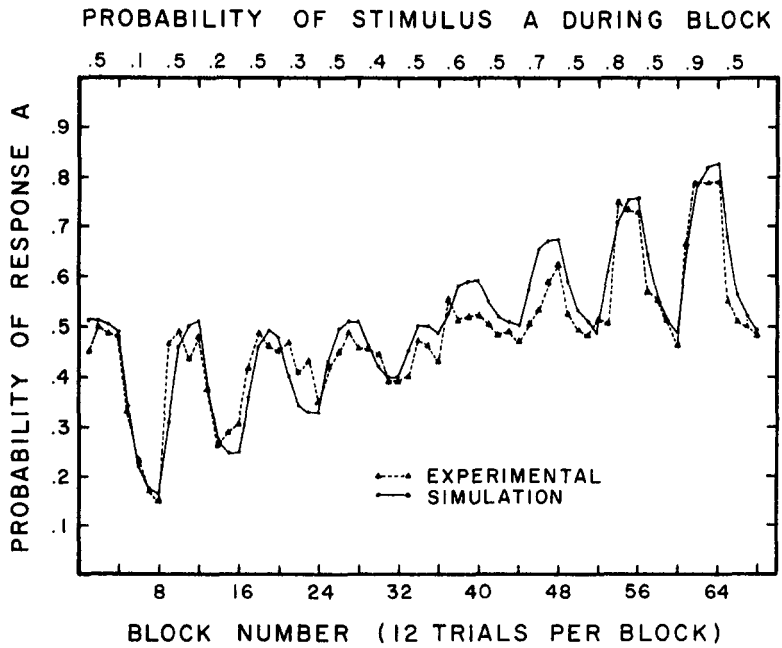

Figure 11. Computer simulation and real data of the first two days of the experiment performed by Friedman et al. (1964). The solid line gives the results of a computer simulation with 80 pseudosubjects receiving random sequences according to the probability schedule. The learning parameter, η, was .3; and the decay parameter, g, was .90. The pseudosubjects' calculated average response probabilities are plotted. The dashed line is the actual experimental data, with 80 subjects making real responses. Subjects did not receive blocks in the sequence given.

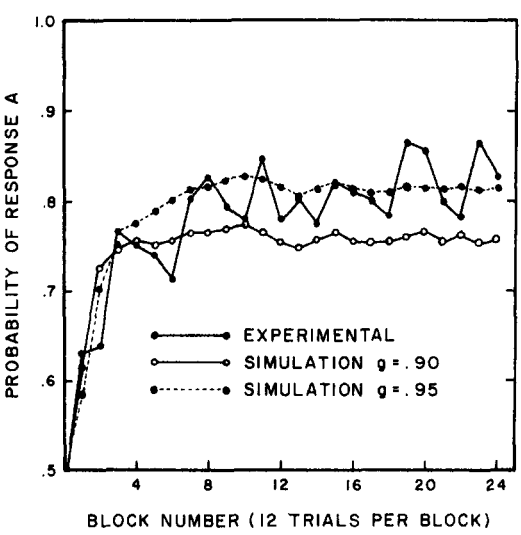

Figure 12. Two computer simulations and real data of the 3rd day of the experiment in Friedman et al. (1964). The learning parameter, η, of the simulation was .3. The decay parameter, g, takes two values, .90 and .95. The experimental data are given by the solid line.

tions to the .5 blocks. Friedman et al. (1964) plotted these transitions (and fitted curves to them) individually in their Figure 4.

Data from the 3rd Day

The fit of the simulated to the actual data was encouraging, although the actual data from the first 2 days of the experiment concealed a demonstrated multitude of extraneous effects, such as event sequence, and subject learning.

The probability sequence on the 3rd day was the same for all the subjects, now well practiced, and provided a good test of the model. Figure 12 shows the fits for data from the 3rd day. The same set of parameters provided a reasonable fit, although the fit was made better by increasing the decay parameter, g, from .90 to .95. Note that both the data and the simulation show a small, but definite, overshoot above .8. The response probability over the last two 48-trial blocks

(12 trial blocks, 25–32) was .82, as opposed to probability matching. The same average for the simulation with g equal to .95 was .81.

In a previous paper (Anderson, 1973), a model similar in many respects to this was proposed for the learning of short lists and for a choice reaction time experiment. A quantity, γ, was defined, and is equal to $\eta/(1-g)$ in the present section. This quantity was found, in fitting the data, to vary from about 2 to about 7. In the probability learning experiments, with η equal to .3 and g equal to .95, $\gamma = 6$; and with η equal to .3 and g equal to .90, $\gamma = 3$, which falls in the same range. Since list learning and probability learning seem at first glance to be very different tasks, this coincidence is interesting.

Recency Curves

Friedman et al.'s (1964) article, and other articles as well, pay special attention to the behavior of response probability when a continuous sequence of a particular event ("a run") occurs. Both statistical learning theory and our theory predict a so-called "positive recency" effect, where the response probability for an event *increases* after a run of that event. When subjects are well practiced, they show this effect very strongly, although they do not always do so during the first few blocks of trials.

Friedman et al. (1964) provide extensive recency data for their subjects during the 3rd day, when $\pi_A = .8$ and $\pi_B = .2$, and at the end of the 2nd day, when π_A and $\pi_B = .5$.

When they fit the data with the simple statistical learning theory model, they find fair fits, but θ again is not constant. In the data from the 3rd day, θ was estimated to be .058 during the transitions to and from the $\pi_A = .8$ condition, while the fit during a run suggested $\theta = .17$.

If we assume the eigenvalues start off, in our model, at asymptote, calculated from the formula and parameters given previously, we can calculate (exactly) the eigenvalues after a particular sequence by use of Equations 15a and 15b and can then calculate the associated probabilities using Equation 14.

When this was done with the data given in Table 11 of Friedman et al. (1964) for runs in the $\pi_A = .8$, 3rd-day condition, and with the data read from Figure 11 for the last block (.5 probability) during Day 2, the fit was not very good for runs of 2, 3, or 4 events, although the match was quite good after a run of 15 or 20 A events.

Our model predicts that response probability can never equal 1.00, even after an infinite run of a particular event. As Reber and Millward (1971) point out, published recency data usually indicate asymptotes well below 1.00. In Reber and Millward's (1971) experiment, which involved a continuously variable event probability which was "tracked" by the subjects, the probability of Response A after a run of up to 10 A events was extraordinarily low and seemed stable at about .78.

In the data from runs of Event A in the $\pi_A = .8$ condition of the Friedman et al. (1964) study, values of response probability also seemed well away from 1.0. Since there were, as might be expected, relatively few very long runs, even in a .8 condition, the individual data points are of dubious significance beyond runs of about 10. However, the average probability of Response A for runs of length 17 through 20 in the 48 trial blocks 3, 4, 5, and 6 is .91; that is, there were 211 A responses out of 231 total responses. We can predict what the model with parameters determined previously would give us after 20 trials of the same event simply by following the formula. We find that, with η equal to .3 and g equal to .95, the probability of Response A after a run of 20 is .91, indicating good fit for long runs. However, the failure of the fit for short runs is intriguing and seems also to be implied by the Friedman et al. finding of wide differences in θ.

Short-term Memory and Probability Learning

Perhaps the simplest explanation of this short-run divergence from a theory satisfactory for the long run is simply that there is a highly weighted contribution from the past few events. Possibly, we are seeing an effect that is due to the limited capacity, short-term store, which is nearly universally accepted to exist in human memory.

Millward and Reber (1968) asked subjects about past trials in a probability learning

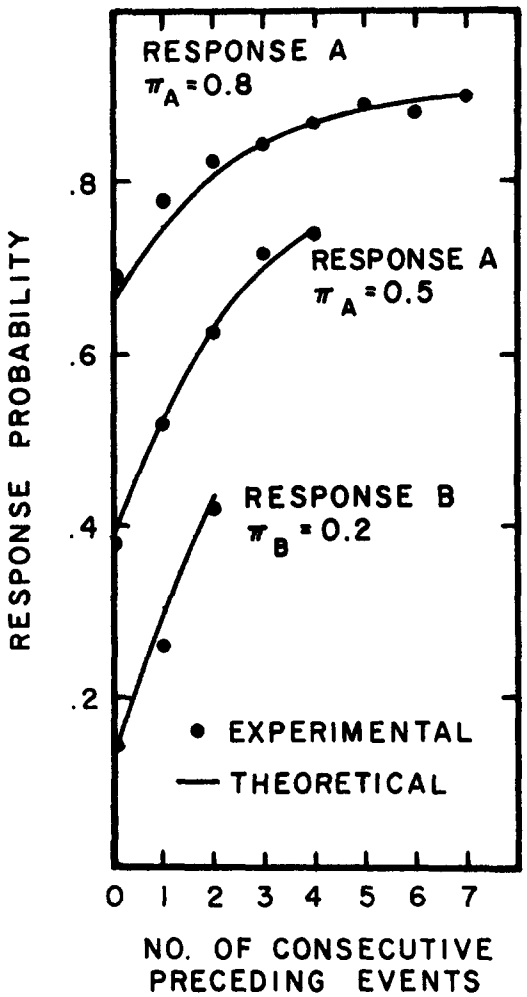

Figure 13. Experimental recency data and theoretical calculations for the experiment in Friedman et al. (1964). Response probabilities are calculated for "runs" of events. A run of zero length means that the preceding event was the other event. These data were simulated by postulating a rapidly decaying short-term trace, in addition to the more slowly decaying trace assumed previously (see text for details).

experiment and found, in most conditions, that "memory for past events goes back about five events and/or four event runs" (p. 988).

Since limited capacity, in our model, can be modeled by assuming it is synonomous with "rapid decay," suppose we postulate that there is a rapidly decaying short-term representation of the preceding three or four events.

There seemed no point in making extensive parameter fits, since we would now have two more parameters to add to our basic model As is well known, with a large computer, enough parameters, and enough persistence, any given set of data can be fitted, so we simply picked some reasonable parameters. We let the short-term decay factor, g_s, be .65, which corresponds to a memory span of three or four. Since we know that a g of .95 fitted data quite well when averaged over 12 trial blocks (which would wash out contributions from a process with a decay factor of .65), we assumed the long-term decay factor, $g_l = .95$.

We assume the two processes are separate, decay separately, and add together to give the magnitude of the eigenvalue, each following an equation like Equation 15a or 15b. Thus, at asymptote, for the eigenvalue associated with Response A, we have

$$\lambda_A = 1 + \frac{\pi_A \eta_s}{1 - g_s} + \frac{\pi_A \eta_l}{1 - g_l}.$$

As a starting point, let us assume that at asymptote there are equal contributions from the short- and long-term processes; that is, the second term equals the third term. We then find that the learning parameter of the short-term process has η_s equal to 1.067 and that of the long-term process, η_l, equal to .1525. This is all the information we need to calculate the average probabilities of the subjects during runs of a particular event. We show the theoretical and experimental curves in Figure 13. The fit is quite good. All these points are fitted with the above parameters, and only the starting points vary. There is a value plotted for a run of zero length, which gives the probability of one response when a run of the other event is about to start. (A run of one event has to start with the other event.) We assume that the eigenvalues at the "-1" trial started at asymptote.

Comments About Variance

Analysis of the variance in these models is rather complicated, since several mechanisms can contribute to the observed experimental variance. First, the raw data is generated by a probabilistic process. Even if we knew the subject's response probabilities exactly, the resulting data from a relatively small number of responses would contribute a significant amount of variance. Second, in our model

Table 3
Comparison Between Published Three-Choice Experimental Data and a Computer Simulation

Source of data	Event p			Simulation			Experiment		
	π_A	π_B	π_C	A	B	C	A	B	C
Gardner[a]	.60	.30	.10	.75	.21	.04	.68	.24	.08
Cotton & Rechtshaffen[b]							.66		
Gardner	.60	.20	.20	.79	.10	.11	.68	.16	.16
Cotton & Rechtshaffen							.66		
Gardner	.70	.20	.10	.84	.11	.05	.80	.13	.07
Cotton & Rechtshaffen							.80		
Gardner	.70	.15	.15	.86	.06	.08	.80	.10	.10
Cotton & Rechtshaffen							.81		
Cole[c]	.44	.33	.22	.52	.32	.16	.51	.31	.17
Cole	.67	.22	.11	.82	.13	.05	.84	.11	.05

[a] Data taken from Gardner (1957). Average of response probabilities for Trials 286–450.
[b] Data taken from Cotton and Rechtshaffen (1958). Average of response probabilities for Trials 286–450.
[c] Date taken from Cole (1965). Average of response probabilities on Trials 501–1,000. A correction procedure was used in this experiment.

there is a contribution due to the intrinsic variance in the probability of a subject's response. This rather involved calculation is given in Appendix B. Third, in our model, intersubject differences in parameters have a profound effect on asymptotic probability, causing a large increase in the variance of the pooled data. This effect is largest when deviations from equal probability are large.

Our guess is that intersubject variation in parameters will turn out to be a major cause of experimental variance, and since we have no idea what kind of distribution of parameters might exist for subjects, we felt it best to do no detailed calculations of variance at this time.

Extension to More Events

The derivation, sketched earlier, of the simplest form of statistical learning theory says that probability matching is independent of the number of stimuli. However, experimentally, overshooting of the most probable stimulus increases greatly as the number of alternatives increases (Estes, 1972). Some data are available for three-choice experiments (Cole, 1965; Cotton & Rechtshaffen, 1958; Gardner, 1957). All these experiments show pronounced overshooting above matching, generally many percent.

Since our model becomes difficult to work with analytically in higher dimensional boxes, we resorted to a Monte Carlo simulation to estimate the response probabilities in a three-choice system. We considered three orthogonal vectors pointing toward corners in a four-dimensional space. (There are orthogonal vectors pointing toward corners only when the dimensionality of the space is divisible by a power of 2.) The orthogonal vectors were $\mathbf{e}_1 = (1, -1, -1, 1)$, $\mathbf{e}_2 = (1, 1, -1, -1)$, and $\mathbf{e}_3 = (1, -1, 1, -1)$. The eigenvalues could be calculated from our asymptotic formula. With eigenvectors and eigenvalues known, the feedback matrix was constructed. Then a random point in the box was chosen (with a uniform distribution) as a starting point. The feedback then forced the system into a corner (see Table 3).

Occasionally stable corners not associated with an eigenvector appeared when the eigenvalues were nearly equal. These corners were ignored in the calculation of probabilities. It is not clear what these extraneous corners might correspond to psychologically—possibly paralyzed uncertainty.

In each Monte Carlo simulation, 1,000 initial random vectors were used, and the number of times a particular corner appeared was counted. This gave an estimate of the probabilities of a particular response. We used the same parameter values we used previously, $\eta = .3$ and $g = .95$. These two values produced only very slight overshooting in the two-choice

system but very pronounced overshooting in the three-choice system.

Table 3 shows the results of the simulations and compares them with the limited experimental data. We find agreement with the actual data is surprisingly good in several cases. In a couple of the cases where the fit is not so good (the 60–20–20 and 60–30–10 conditions of Gardner, 1957, and of Cotton & Rechtshaffen, 1958), the data in their published figures clearly indicate that asymptote had not been reached and the probabilities of the most probable response were still increasing. Cole used the last 500 trials of a 1,000-trial experiment, while the other two workers used average probabilities of Trials 286–450 in a shorter experiment. Cole's experiment incorporated a correction procedure, where subjects continued to predict until they were correct. This should not greatly affect our prediction for the first response probabilities in this experiment.

One might wonder why it would not be possible in the three-choice case to consider Responses B and C, say, to be a single "response" and then to apply the two-choice analysis to them. We showed in our simulations that the two-choice model does not give rise to as much overshooting (for the same parameters) as the multiple-choice model, so this must be an incorrect approach. One reason for this is that we are not using a linear system but one with a high degree of nonlinearity, both in geometry and dynamics. Inputs and outputs cannot be freely combined in a nonlinear system because the superposition principle does not hold. Many "obvious" approaches are incorrect when applied to nonlinear systems, a point to be careful of when analyzing the operation of a system with such spectacular nonlinearities as the brain.

VII. Conclusions

Our aim in this article has been to present a relatively detailed and precise model, which was suggested by the anatomy and physiology of the brain, and to consider some interesting psychological applications. Although it was necessary to oversimplify reality to get a model that we could work with easily, the resulting model was sufficiently rich to have a pronounced structure, and, with a very modest amount of manipulation, gave rise to some testable predictions. We feel that the best way to work with such a very wide-ranging model is to try to fit with reasonable success many different phenomena, rather than to try to explain every detail of a limited body of experimental data.

We have argued that our model provided a theoretical framework for entities that acted and behaved very much like the distinctive features of psychology. A set of neurons with positive feedback tends to analyze its inputs by most heavily weighing the eigenvectors of the feedback matrix with large positive eigenvalues and by suppressing the rest. We also pointed out that these particular eigenvectors are often the most meaningful in terms of the discrimination to be performed, since they contain most of the information allowing discriminations to be made among the simulus set.

When we introduced a saturating nonlinearity—the "brain-state-in-a-box" model—we were able to suggest directly a model for probability learning, which fitted some actual experimental data in reasonable detail. As an application of the model in a different area, we showed that the same model, with slight modifications, acted as a categorical perceiver with properties similar to those seen in recent data.

The model has some obvious shortcomings. Among these are some grevious oversimplications of the physiology, the assumption of a high degree of linearity in some parts of the system, and, later, the assumption of "hard" saturation, with no transition from linearity to saturation. Other necessary details that are needed for an actual operating system are simply ignored. For example, a mechanism must be provided to get the brain state *out* of a corner, once it has gotten *in*. There are a multitude of ways this could be accomplished, all of which could be used together. There could be selective adaptation or habituation of rapidly firing cells and of their synapses; there could be large amounts of noise which could force the system from corner to corner; and, quite probably, there could be a special neural circuit, where rapidly firing cells generate recurrent inhibition to turn them-

selves off. We preferred to ignore this problem here, since we did not need to consider it for the problems we discussed. However, eventually we must make concrete assumptions about how brain states decay as well as how they grow.

Some of the quantitative predictions for probability learning may be fairly sensitive to the assumptions about the geometry of the system. It is quite unlikely that our hypothetical enclosing "box" is a hypercube, with all sides equal. We are not sure how a more general box would affect the behavior of the system. Many of the qualitative properties would remain the same, but the system is sufficiently complex that we cannot say for certain.

There are many immediate extensions of the model. As one example, the probability learning model serves also as a model for choice reaction time. If we assume that instead of starting randomly, we start at a point in the box that is determined to some extent by the stimulus—by initially passing the stimulus through a memory filter of the kind we have been discussing here, for example—we can make a model for choice reaction time. In fact, the model as stated has similarities, which can be made precise and explicit, with random walk models for reaction time, which are very successful in explaining many of the quantitative aspects of reaction time experiments (Link, 1975). If we consider the corner as an absorbing state and the noisy memory filter output evolving in time (with additive noise) as the directed random walk, then the resemblance becomes striking.

As another point, the model strongly suggests that the brain would rather be wrong than undecided; that is, a misperception (the wrong corner) is better than no perception at all. Some misperceptions are clearly more likely than others, reflecting the past learning of the system. Also, the model suggests that brain states in perception move from stable state to stable state, with abrupt transitions between stable states.

All these suggestions can be made precise relatively easily and can be compared with the rich body of data and concepts in experimental psychology. We have tried to show, with theoretical discussion and with several detailed examples, that such an effort may be worthwhile.

Reference Note

1. Anderson, J. A. *What is a distinctive feature?* (Tech. Rep. 74-1). Providence, R.I.: Brown University, Center for Neural Studies, 1974.

References

Anderson, J. A. A memory storage model utilizing spatial correlation functions. *Kybernetik*, 1968, *5*, 113–119.

Anderson, J. A. Two models for memory organization using interacting traces. *Mathematical Biosciences*, 1970, *8*, 137–160.

Anderson, J. A. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 1972, *14*, 197–220.

Anderson, J. A. A theory for the recognition of items from short memorized lists. *Psychological Review*, 1973, *80*, 417–438.

Anderson, J. A. Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, N.J.: Erlbaum, 1977.

Anderson, J. A., Silverstein, J. W., & Ritz, S. A. Vowel pre-processing with a neurally based model. In *Conference Record: 1977 I.E.E.E. International Conference on Acoustics, Speech, and Signal Processing*. Hartford, Conn.: in press.

Barlow, H. B. Single units and sensation: A neuron doctrine for perceptual psychology. *Perception*, 1972, *1*, 371–394.

Cole, M. Search behavior: A correction procedure for three choice probability learning. *Journal of Mathematical Psychology*, 1965, *2*, 145–170.

Cooper, L. N. A possible organization of animal memory and learning. In B. Lundquist & S. Lundquist (Eds.), *Proceedings of the Nobel Symposium on Collective Properties of Physical Systems*. New York: Academic Press, 1974.

Cooper, W. E. Selective adaptation to speech. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1975.

Cotton, J. W., & Rechtshaffen, A. Replication report: Two- and three-choice verbal conditioning phenomena. *Journal of Experimental Psychology*, 1958, *56*, 96–97.

Creutzfeldt, O., Innocenti, G. M., & Brooks, D. Vertical organization in the visual cortex (area 17) in the cat. *Experimental Brain Research*, 1974, *21*, 315–336.

Eccles, J. C. Possible synaptic mechanisms subserving learning. In A. G. Karczmar & J. C. Eccles (Eds.), *Brain and human behavior*. New York: Springer, 1972.

Eimas, P. D., & Corbit, J. D. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 1973, *4*, 99–109.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito,

J. Speech perception in infants. *Science*, 1971, *171*, 303–306.

Estes, W. K. Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika*, 1957, *22*, 113–132.

Estes, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964.

Estes, W. K. Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, 1972, *67*, 81–102.

Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology*, 1954, *47*, 225–234.

Evans, E. F. Neural processes for the detection of acoustic patterns and for sound localization. In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program*. Cambridge, Mass.: MIT Press, 1974.

Freeman, W. J. *Mass action in the nervous system*. New York: Academic Press, 1975.

Friedman, M. P., *et al.* Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford, Calif.: Stanford University Press, 1964.

Frishkopf, L. S., Capranica, R. R., & Goldstein, M. H., Jr. Neural coding in the bullfrog's auditory system: A teleological approach. *Proceedings of the I.E.E.E.*, 1968, *56*, 969–980.

Funkenstein, H. H., & Winter, P. Responses to acoustic stimuli of units in the auditory cortex of awake squirrel monkeys. *Experimental Brain Research*, 1973, *18*, 464–488.

Gardner, R. A. Probability learning with two and three choices. *American Journal of Psychology*, 1957, *70*, 174–185.

Gibson, E. J. *Principles of perceptual learning and development*. New York: Meredith, 1969.

Globus, A., & Scheibel, A. B. Pattern and field in cortical structure: The rabbit. *Journal of Comparative Neurology*, 1967, *131*, 155–172.

Goldstein, M. H., Jr., Hall, J. L. II, & Butterfield, B. O. Single-unit activity in the primary auditory cortex of unanesthetized cats. *Journal of the Acoustical Society of America*, 1968, *43*, 444–455.

Grossberg, S. Pavlovian pattern learning by nonlinear neural networks. *Proceedings of the National Academy of Sciences*, 1971, *68*, 828–831.

Hebb, D. O. *The organization of behavior*. New York: Wiley, 1949.

Hubel, D. H., & Wiesel, T. N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 1962, *160*, 106–154.

Hubel, D. H., & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 1968, *195*, 215–243.

Jakobson, R., Fant, G. G. M., & Halle, M. *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, Mass.: MIT Press, 1961.

Knight, B. W., Toyoda, J. I., & Dodge, F. A., Jr. A quantitative description of the dynamics of excitation and inhibition in the eye of *Limulus*. *Journal of General Physiology*, 1970, *56*, 421–437.

Kohonen, T. Correlation matrix memories. *I.E.E.E. Transactions on Computers*, 1972, *C-21*, 353–359.

Kohonen, T. *Associative memory: A system theoretic approach*. Berlin: Springer-Verlag, 1977.

Laughery, K. R. Computer simulation of short-term memory: A component decay model. In G. H. Bower & J. R. Spence (Eds.), *The psychology of learning and motivation* (Vol. 3). New York: Academic Press, 1969.

Lindgren, N. Machine recognition of human language: Part 2. Theoretical models of speech perception and language. *I.E.E.E. Spectrum*, 1965, *2*, 45–59.

Lindsay, P. H., & Norman, D. A. *Human information processing: An introduction to psychology*. New York: Academic Press, 1972.

Link, S. W. The relative judgment theory of two choice reaction time. *Journal of Mathematical Psychology*, 1975, *12*, 114–135.

Little, W. A., & Shaw, G. L. A statistical theory of short and long term memory. *Behavioral Biology*, 1975, *14*, 115–135.

McIlwain, J. T. Large receptive fields and spatial transformations in the visual system. In R. Porter (Ed.), *International review of physiology: Neurophysiology II* (Vol. 10). Baltimore, Md.: University Park Press, 1976.

Millward, R. B., & Reber, A. S. Event-recall in probability learning. *Journal of Verbal Learning and Verbal Behavior*, 1968, *7*, 980–989.

Morrell, F., Hoeppner, T. J., & de Toledo, L. Mass action re-examined: Selective modification of single elements within a small population. *Neuroscience Abstracts*, 1976, *2*, 448.

Mountcastle, V. B. The problem of sensing and the neural coding of sensory events. In G. C. Quarton, T. Melnechuk, & F. L. Schmitt (Eds.), *The neurosciences*. New York: Rockefeller University Press, 1967.

Myers, J. L. Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 3). Hillsdale, N.J.: Erlbaum, 1976.

Nass, M. M., & Cooper, L. N. A theory for the development of feature detecting cells in visual cortex. *Biological Cybernetics*, 1975, *19*, 1–18.

Neimark, E. D., & Estes, W. K. *Stimulus sampling theory*. San Francisco: Holden-Day, 1967.

Neisser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1967.

Noda, H., & Adey, W. R. Firing of neuron pairs in cat association cortex during sleep and wakefulness. *Journal of Neurophysiology*, 1970, *33*, 672–684.

Noda, H., Manohar, S., & Adey, W. R. Correlated firing of hippocampal neuron pairs in sleep and wakefulness. *Experimental Neurology*, 1969, *24*, 232–247.

Pisoni, D. B. *On the nature of categorical perception of speech sounds*. Unpublished doctoral dissertation, University of Michigan, 1971.

Pisoni, D. B., & Tash, J. Reaction time to comparisons within and across phonetic categories. *Perception & Psychophysics*, 1974, *15*, 285–290.

Ratliff, F., Knight, B. W., Dodge, F. A., Jr., & Hartline, H. K. Fourier analysis of dynamics of excitation and inhibition in the eye of *Limulus:* Amplitude, phase, and distance. *Vision Research,* 1974, *14,* 1155–1168.

Reber, A. S., & Millward, R. B. Event observation in probability learning. *Journal of Experimental Psychology,* 1968, *77,* 317–327.

Reber, A. S., & Millward, R. B. Event tracking in probability learning. *American Journal of Psychology,* 1971, *84,* 85–99.

Schiller, P. H., Finlay, B. L., & Volman, S. F. Quantitative studies of single-cell properties in monkey striate cortex: V. Multivariate statistical analyses and models. *Journal of Neurophysiology,* 1976, *39,* 1362–1374.

Shepherd, G. M. *The synaptic organization of the brain.* New York: Oxford University Press, 1974.

Studdert-Kennedy, M. The nature and function of phonetic categories. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1975.

Suzuki, S. *Zen mind, beginner's mind.* New York: Weatherhill, 1970.

van der Loos, H., & Glaser, E. M. Autapses in neocortex cerebri: Synapses between a pyramidal cell's axon and its own dendrites. *Brain Research,* 1972, *48,* 355–360.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. Non-holographic associative memory. *Nature,* 1969, *222,* 960–962.

Winter, P., & Funkenstein, H. H. The effect of species specific vocalization on the discharge of auditory cortical cells in the awake squirrel monkey. *Experimental Brain Research,* 1973, *18,* 489–504.

Wollberg, Z., & Newman, J. D. Auditory cortex of squirrel monkey: Response patterns of single cells to species-specific vocalizations. *Science,* 1972, *175,* 212–214.

Young, T. Z., & Calvert, T. W. *Classification, estimation, and pattern recognition.* New York: American Elsevier, 1974.

Appendix A

The calculation of size of the regions corresponding to each corner is reasonably straightforward. In our system we have two orthogonal eigenvectors which point in the (1, 1) and (−1, 1) directions. Rotate the system by $-\pi/4$ radians to obtain the normalized eigenvectors

$$\mathbf{e}_x = \begin{pmatrix}1\\0\end{pmatrix}, \quad \mathbf{e}_y = \begin{pmatrix}0\\1\end{pmatrix}.$$

The eigenvalues are unchanged by a rotation; that is,

$$\lambda_A = \lambda_x, \quad \lambda_B = \lambda_y.$$

A point with coordinates, $\mathbf{v} = \begin{pmatrix}x_0\\y_0\end{pmatrix}$ in this rotated frame, then may be expressed as

$$\mathbf{v} = x_0\mathbf{e}_x + y_0\mathbf{e}_y. \tag{A1}$$

If we consider that a feedback cycle occurs once every unit time, t_i, $i = 0, 1, 2, \ldots$, then

$$\mathbf{v}(t_0) = x_0\mathbf{e}_x + y_0\mathbf{e}_y$$

$$\mathbf{v}(t_1) = x_0(\mathbf{e}_x\lambda_x) + y_0(\mathbf{e}_y\lambda_y) + \mathbf{v}(t_0)$$
$$= x_0\mathbf{e}_x(1 + \lambda_x) + y_0\mathbf{e}_y(1 + \lambda_y)$$

$$\mathbf{v}(t_2) = x_0\mathbf{e}_x(1 + \lambda_x)\lambda_x + y_0\mathbf{e}_y(1 + \lambda_y)\lambda_y + \mathbf{v}(t_1)$$
$$= x_0\mathbf{e}_x(1 + \lambda_x)^2 + y_0\mathbf{e}_y(1 + \lambda_y)^2.$$

By induction we get

$$\mathbf{v}(t_n) = x_0\mathbf{e}_x(1 + \lambda_x)^n + y_0\mathbf{e}_y(1 + \lambda_y)^n.$$

Since $\mathbf{e}_x = \begin{pmatrix}1\\0\end{pmatrix}$, $\mathbf{e}_y = \begin{pmatrix}0\\1\end{pmatrix}$, we have

$$\mathbf{v}(t_n) = \begin{pmatrix}x_0(1 + \lambda_x)^n\\y_0(1 + \lambda_y)^n\end{pmatrix}. \tag{A2}$$

In order to get a reasonably simple, general solution for the sizes of the regions, we must make the approximation, which should be quite good if step size is not too large, that time is continuous.

For continuous time we have

$$dx/dt = \lambda_x x$$
$$dy/dt = \lambda_y y.$$

These two equations may be solved to determine the "motion" of a point (x, y) as it is fed back through the system. The solutions are

$$\lambda_x^{-1} \ln x = t + k_1$$
$$\lambda_y^{-1} \ln y = t + k_2.$$

Eliminating t, we find

$$\lambda_x \ln y - \lambda_y \ln x = \text{constant}$$
$$\ln \frac{y^{\lambda_x}}{x^{\lambda_y}} = \text{constant} \quad \text{(A3)}$$
$$y^{\lambda_x} = kx^{\lambda_y} \text{ or } y = k' x^{\lambda_y/\lambda_x}, \quad \text{(A4)}$$

where k' is a constant. This family of curves provides the trajectories that will be followed by an initial point under the influence of feedback. A point with initial coordinates (x_0, y_0) will follow the curve expressed by Equation A4 with

$$k' = \frac{y_0}{x_0^{\lambda_y/\lambda_x}}. \quad \text{(A5)}$$

We must now consider the behavior of these curves at the boundary of the square. A point will travel along the curve given by Equation A4 until it reaches the boundary. At this point all motion in the direction normal to the boundary will be prevented. Only the component of the predicted motion that is tangent to the boundary will contribute to the actual motion. If $\mathbf{t}$ is a normalized tangent vector along the boundary, then

$$\left.\frac{d\mathbf{v}}{dt}\right|_{\substack{\text{along} \\ \text{boundary}}} = \left\{\left(\frac{dx}{dt}\right)\mathbf{e}_x + \left(\frac{dy}{dt}\right)\mathbf{e}_y\right\} \cdot \mathbf{t}. \quad \text{(A6)}$$

We are interested in finding the curve represented by Equation A4 for which a point will reach the boundary and stop. This will occur when the scalar product (dot product) of Equation A6 of the point reached on the boundary is zero. Thus we wish to find the point at which

$$\left(\frac{dx}{dt}\mathbf{e}_x + \frac{dy}{dt}\mathbf{e}_y\right) \cdot \mathbf{t} = 0.$$

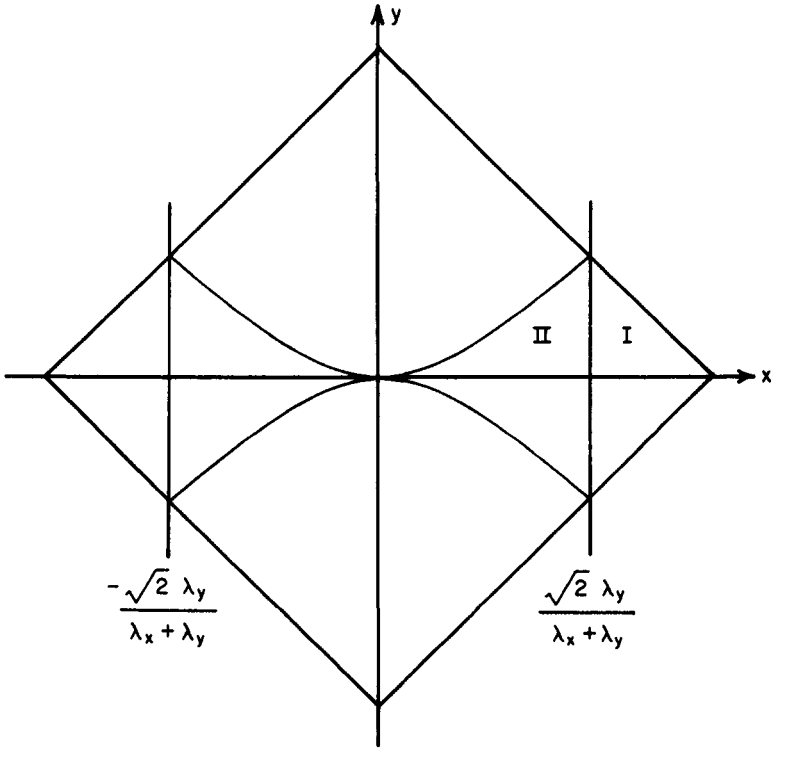

Figure A1. Regions of integration for the calculation of response probabilities for the two-dimensional neural system. The eigenvectors point along the coordinate axes.

Consider only the positive quadrant, with a normalized tangent vector,

$$\mathbf{t} = \frac{-\mathbf{e}_x + \mathbf{e}_y}{\sqrt{2}}.$$

We have then

$$0 = \{x\lambda_x\mathbf{e}_x + y\lambda_y\mathbf{e}_y\} \cdot \left\{\frac{-\mathbf{e}_x + \mathbf{e}_y}{\sqrt{2}}\right\},$$

so that

$$0 = -x\lambda_x + y\lambda_y.$$

Since $x + y = \sqrt{2}$ on the boundary, we have

$$0 = x\lambda_x + x\lambda_y - \sqrt{2}\lambda_y,$$

and therefore

$$x = \frac{\sqrt{2}\,\lambda_y}{\lambda_x + \lambda_y}$$

$$y = \frac{\sqrt{2}\,\lambda_x}{\lambda_x + \lambda_y}.$$

The lines, $x = \sqrt{2}\,\lambda_y/(\lambda_x + \lambda_y)$ and $x = -\sqrt{2}\,\lambda_y/(\lambda_x + \lambda_y)$ determine by their intersection with the boundary four (unstable) equilibrium points. The curves representing all interior points of the square which will move so as to intersect the boundary at one of these equilibrium points are given by Equation A4 with

$$k' = \pm \frac{y_0}{x_0^{\lambda_y/\lambda_x}} = \pm\left(\frac{\sqrt{2}\,\lambda_x}{\lambda_x + \lambda_y}\right)\left(\frac{\lambda_x + \lambda_y}{\sqrt{2}\,\lambda_y}\right)^{\lambda_y/\lambda_x}.$$

The regions of the two-dimensional system are thus determined (see Figure A1).

The area of the sector labelled I and II in Figure A1 can be found by integrating $y = k'x^{\lambda_y/\lambda_x}$ from 0 to $x = \sqrt{2}\,\lambda_y/(\lambda_x + \lambda_y)$ and adding the area of the triangular segment, I. This area times 4 will give the amount of the total area for which a vector will end up in corners (1, 1) or (−1, −1) in the nonrotated system. Since the total area of the square is 4, the area I-II corresponds to the probability of Response A.

If P is the probability of ending up in corners (1, 1) or (−1, −1) then

$$\begin{aligned} P &= \int_0^{\sqrt{2}\lambda_y/(\lambda_x+\lambda_y)} k'x^{\lambda_y/\lambda_x}dx \\ &\quad + \frac{1}{2}\left(\sqrt{2} - \frac{\sqrt{2}\,\lambda_y}{\lambda_x + \lambda_y}\right)\left(\frac{\sqrt{2}\,\lambda_x}{\lambda_x + \lambda_y}\right) \\ &= \frac{k'x^{\lambda_y/\lambda_x+1}}{1 + \lambda_y/\lambda_x}\Bigg|_0^{\sqrt{2}\lambda_y/(\lambda_x+\lambda_y)} + \left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right)^2 \\ &= \left(\frac{\sqrt{2}\,\lambda_x}{\lambda_x + \lambda_y}\right)\left(\frac{\lambda_x + \lambda_y}{\sqrt{2}\,\lambda_y}\right)^{\lambda_y/\lambda_x}\left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right) \\ &\quad \times \left(\frac{\sqrt{2}\,\lambda_y}{\lambda_x + \lambda_y}\right)^{\lambda_y/\lambda_x+1} + \left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right)^2 \\ &= \sqrt{2}\left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right)^2\left(\frac{\sqrt{2}\,\lambda_y}{\lambda_x + \lambda_y}\right) + \left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right)^2 \\ &= \frac{3\lambda_x^2\lambda_y + \lambda_x^3}{(\lambda_x + \lambda_y)^3} \\ &= \frac{3\lambda^2 + \lambda^3}{(\lambda + 1)^3}, \end{aligned} \tag{A7}$$

where $\lambda = \lambda_x/\lambda_y$. This is the formula used to calculate response probabilities in the section on probability learning.

Appendix B

We can write the equation governing λ_A as

$$\lambda_A(n + 1) = 1 + P_{n+1}, \tag{B1}$$

where

$$P_{n+1} = gP_n + \eta I_{n+1}, \quad P_0 = 0, \tag{B2}$$

where the I_ns are independent random variables taking the value 1 with probability π and 0 with probability $1 - \pi$. The solution for Equation B2 is given by

$$P_n = \eta \sum_{j=1}^{n} g^{n-j}I_j \tag{B3}$$

and can easily be verified. However it is virtually impossible to express the distribution of P_n in a workable form. We can calculate moments of P_n using Equation B3. For example the expected value of P_n is

$$\begin{aligned} E(P_n) &= \eta \sum_{j=1}^{n} g^{n-j}E(I_j) = \eta\pi \sum_{j=0}^{n-1} g^j \\ &= \eta\pi \frac{1 - g^n}{1 - g}, \end{aligned} \tag{B4}$$

while the second moment is

$$E(P_n^2) = \eta^2 E[(\sum_{j=1}^{n} g^{n-j} I_j)(\sum_{k=1}^{n} g^{n-k} I_k)] = \eta^2 \sum_{j,k=1}^{n} g^{n-j+n-k} E(I_j I_k)$$

$$= \eta^2 [\pi \sum_{j=1}^{n} g^{2(n-j)} + \sum_{j \neq k} g^{2n-j-k} E(I_j) E(I_k)]$$

$$= \eta^2 \left[\pi \frac{1-g^{2n}}{1-g^2} + \sum_{j=1}^{n} g^{n-j} E(I_j) \sum_{k=1}^{n} g^{n-k} E(I_k) - g^{n-j} E(I_j) \right]$$

$$= \eta^2 \left[\pi \frac{1-g^{2n}}{1-g^2} + \pi^2 \left(\frac{1-g^n}{1-g} \right)^2 - \pi^2 \frac{1-g^{2n}}{1-g^2} \right]$$

$$= \eta^2 \pi \left[(1-\pi) \frac{1-g^{2n}}{1-g^2} + \pi \left(\frac{1-g^n}{1-g} \right)^2 \right] \tag{B5}$$

The variance, then, of P_n is

$$\text{var}(P_n) = E(P_n^2) - E(P)^2 = \eta^2 \pi (1-\pi) \frac{1-g^{2n}}{1-g^2}. \tag{B6}$$

The higher moments can be found in the same way, although the calculations become more tedious.

As n approaches infinity, P_n cannot converge to a real number or to another random variable, since at each step we are adding a non-attenuating random element. However the distribution functions $F_n(x) \equiv P(P_n \leq x)$ converge to the distribution function of the random variable

$$P = \eta \sum_{j=0}^{\infty} g^j I_j, \tag{B7}$$

where the I_js are defined as before. Since the distributions of P_n are defined on the interval between 0 and $\eta/(1-g)$, the moments of P_n converge to the moments of P. The moments of P are easy to get by using an iterative scheme. We can write P as

$$P = gP' + \eta I, \tag{B8}$$

where $P' = \sum_{j=0}^{\infty} g^j I_{j+1}$, and $I = I_1$. Since P and P' have the same distribution, and since P' and I are independent we have for all $n \geq 1$

$$E(P^n) = E(gP' + \eta I)^n = g^n E(P^n) + \pi \sum_{j=1}^{n} \binom{n}{j} g^{n-j} \eta^j E(P^{n-j}), \tag{B9}$$

where (n, r) are the binomial coefficients [that is, $(n, r) = n!/r!(n-r)!$]. Therefore we have

$$E(P^n) = \frac{1}{1-g^n} \left[\pi \sum_{j=1}^{n} \binom{n}{j} g^{n-j} \eta^j E(P^{n-j}) \right]. \tag{B10}$$

It is easy to see that

$$\lambda_B(n) = 1 + \eta \sum_{j=1}^{n} g^{n-j}(1 - I_j) = 1 + \frac{\eta(1-g^n)}{1-g} - P_n, \tag{B11}$$

so that

$$\lambda(n) = \frac{\lambda_A(n)}{\lambda_B(n)} = \frac{1+P_n}{1 + \frac{\eta(1-g^n)}{1-g} - P_n} = \frac{1+a_n}{a_n - P_n} - 1, \tag{B12}$$

where $a_n = 1 + [(1-g^n)/(1-g)]$. However, the random variable $1/(a_n - P_n)$ is difficult to deal with even in the limiting case. All moments of $\lambda(n)$ must be derived from infinite series of moments of P_n. It is computationally feasible to get approximations, but we made no attempt to do so.

Fortunately, we have a remarkably different situation for the probability of selection. We write

$$\text{Prob}[\lambda(n)] = \frac{\lambda(n)[3\lambda(n) + \lambda(n)^2]}{[1+\lambda(n)]^3} = 1 - \frac{1+3\lambda(n)}{[1+\lambda(n)]^3}, \tag{B13}$$

and it can easily be verified that

$$\frac{1+3\lambda(n)}{[1+\lambda(n)]^3} = \frac{1}{(1+a_n)^3} \times [a_n^3 + 3a_n^2 - 6a_n P(n) + 3(1-a_n)P(n)^2 + 2P^3(n)], \tag{B14}$$

so that

$$\text{Prob}[\lambda(n)] = \frac{1+3a_n}{(1+a_n)^3} - \frac{1}{(1+a_n)^3}$$

$$\times [2P^3(n) + 3(1 - a_n)P^2(n) - 6a_nP(n)]. \quad \text{(B15)}$$

Therefore in order to determine the expected value of Prob[$\lambda(n)$] we need only to compute the first three moments of $P(n)$, while the variance of Prob[$\lambda(n)$] requires $P(n)$'s first six moments.

As n approaches infinity the distributions of Prob[$\lambda(n)$] approach the distribution of the random variable Prob(λ), where λ is given by

$$\lambda = \frac{1 + a}{a - P} - 1, \quad a = 1 + \frac{\eta}{1 - g}. \quad \text{(B16)}$$

Equation B15 holds true for λ with $P(n)$ replaced by P, so that that expected value and variance of Prob(λ) can be derived as in the finite case.

Appendix C

We would like to see if we must be restricted to the values given in Equation 5 for the feedback system to have its desirable properties. Suppose we let the main diagonal be zero. This means a cell influences only its neighbors, and not itself. We see that the value of the diagonal element given by Equation 5 is related to the mean square of the activities $f(i)$ over all the traces. If we make the assumption—in harmony with our general approach, but presently untestable—that, on the average, cells are equally active across the total stimulus set, that is, they are all equally "important" in some sense, then the diagonal elements will be almost the same.

Let us consider the matrix we shall call **D**, which contains only diagonal elements. Then,

$$\mathbf{D} \cong c\mathbf{I},$$

where c is a positive constant. If the covariance matrix given by Equation 5 is denoted **V** and we require the feedback matrix, **A**, to have zeroes along the diagonal, then

$$\mathbf{A} \cong \mathbf{V} - \mathbf{D} \cong \mathbf{V} - c\mathbf{I}. \quad \text{(C1)}$$

Suppose $\mathbf{e}_i$ is an eigenvector of the covariance matrix with eigenvalue λ_i and with all the important discriminative properties we have discussed. Let us assume that the equality in Equation C1 holds. Then if

$$\mathbf{A} = \mathbf{V} - c\mathbf{I}$$

$$\mathbf{A}\mathbf{e}_i = \mathbf{V}\mathbf{e}_i - c\mathbf{e}_i$$

$$= \lambda_i\mathbf{e}_i - c\mathbf{e}_i$$

$$= (\lambda_i - c)\mathbf{e}_i. \quad \text{(C2)}$$

Thus, $\mathbf{e}_i$ is an eigenvector of **A** and its eigenvalue is $(\lambda_i - c)$. We see that all the eigenvalues are reduced by an amount c. Thus some eigenvalues can be negative. If the equality does not hold, then for small deviations of **D** from a constant times the identity matrix, the eigenvectors of **A** will be close to the eigenvectors of the covariance matrix and will have about the same properties.

As a special case of interest, assume the eigenvectors pointed toward the corners of an N-dimensional hypercube. Suppose the system only learns these corners. Since the cube has equal sides, then all the $f_k^2(i)$ will be the same (i.e., the square of the saturation limit) and we see that the exact equality will hold). Thus, we can avoid the awkward necessity to have large positive elements along the main diagonal and lose little, if any, of the information processing power of the system.

Note that negative eigenvalues, as long as they are greater than -1, may cause the final state of the system to converge to zero, as can be seen from Equation 8. This will occur if the initial input is a linear combination of eigenvectors with eigenvalues in the interval $(-1, 0)$.

Received September 30, 1976 ■