A STOCHASTIC MODEL OF NONLINEARLY INTERACTING NEURONS

Terrence Joseph Sejnowski

A DISSERTATION PRESENTED TO THE FACULTY OF PRINCETON UNIVERSITY IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY THE DEPARTMENT OF PHYSICS

June, 1978
A Stochastic Model of Nonlinearly Interacting Neurons

Terrence Joseph Sejnowski

Neurons in the vertebrate nervous system respond to inputs nonlinearly, are highly and precisely interconnected, and have a randomly fluctuating component in their membrane potentials. The Hartline-Ratliff model of the Limulus retina is generalized to incorporate these three fundamental properties. Because not all neurons produce all-or-none action potentials, the membrane potential rather than the firing rate is taken as the primary variable in the model. However, the average firing rate, which is an important measurement in many experiments, plays a key role in coupling the equation which governs the average membrane potentials with the equation which governs the correlations between membrane potentials.

Many neurons in visual cortex respond best to special patterns of light which are called "trigger features". The properties of such "trigger features" are examined in a variational analysis of the steady-state equation for the average membrane potentials. This equation always has at least one solution and an example is analyzed which demonstrates bifurcations to multiple stable solutions and hysteresis between them. A recent cooperative model of binocular depth perception (Marr and Poggio, 1976) is
identical with the steady-state equation analyzed here; a time-dependent generalization of the model is given which should have better performance with dynamic input.

If the membrane potentials of a group of neurons have Gaussian distributions, which is a reasonable assumption in a highly interconnected area such as cerebral cortex, then the equation for the covariances between membrane potentials unexpectedly becomes linear and represents a linear filter. If this statistical assumption is experimentally confirmed, then control and communication theory in systems engineering may have direct applications in some parts of the brain.

The model is applied to motor learning in the cerebellum and a prediction is made concerning the underlying plasticity: the change in synaptic strength between a parallel fiber and a Purkinje cell should be proportional to the covariance between impulses in the parallel fiber and the Purkinje cell's climbing fiber. Unlike previous proposals for synaptic plasticity in the cerebellum, this prediction requires both facilitation and depression to occur (under different conditions) at the same synapse.
Acknowledgements

I am indebted to John Wheeler who made the initiation of this work possible. I want to thank my parents and all my friends for their support and encouragement during its completion, and I am especially grateful to David Bender, Alan Gelperin, Charles Gross, John Hopfield, Bruce Knight, Jr., and Murray Lampert for their generous help.
Table of Contents

Abstract iii
Acknowledgements v
Table of Contents vi
List of Illustrations viii

Introduction 1

Neurons and Synapses
Slow Potentials and Stochastic Variability
The *Limulus* Retina
Generalizing the *Limulus* Model

I. The Nonlinear Model 39

Neurons and "Microcircuits"
Quantitative Model

II. Average Membrane Potentials in Steady State 48

Stochastic Model
Steady-State Solutions
Feature Detection
Stereopsis
III. Correlations between Membrane Potentials 82

Correlations
Covariance Equation
Stationary Case
Neuronal Filters and Memory

IV. Motor Learning in the Cerebellum 110

Cerebellar Cortex
Storing Covariance
Synaptic Plasticity
Long-term Memory

Discussion 133

Appendix: Point-Process Model 138

References 144
List of Illustrations

<table>
<thead>
<tr>
<th>Page</th>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.</td>
<td>Golgi-stained neurons</td>
</tr>
<tr>
<td>5</td>
<td>2.</td>
<td>Representative vertebrate neuron</td>
</tr>
<tr>
<td>7</td>
<td>3.</td>
<td>EEG recordings</td>
</tr>
<tr>
<td>11</td>
<td>4.</td>
<td>Action potential</td>
</tr>
<tr>
<td>13</td>
<td>5.</td>
<td>Chemical synapses</td>
</tr>
<tr>
<td>15</td>
<td>6.</td>
<td>Slow potential processing</td>
</tr>
<tr>
<td>16</td>
<td>7.</td>
<td>Vertebrate retina: anatomy</td>
</tr>
<tr>
<td>17</td>
<td>8.</td>
<td>Vertebrate retina: physiology</td>
</tr>
<tr>
<td>19</td>
<td>9.</td>
<td>Single neuron in cat visual cortex</td>
</tr>
<tr>
<td>23</td>
<td>10.</td>
<td>Prog neuromuscular junction</td>
</tr>
<tr>
<td>25</td>
<td>11.</td>
<td>Limulus ommatidium</td>
</tr>
<tr>
<td>26</td>
<td>12.</td>
<td>Limulus retina response to light</td>
</tr>
<tr>
<td>27</td>
<td>13.</td>
<td>Limulus retina lateral inhibition</td>
</tr>
<tr>
<td>29</td>
<td>14.</td>
<td>Limulus retina response to step illumination</td>
</tr>
<tr>
<td>32</td>
<td>15.</td>
<td>Limulus retina response to modulated input</td>
</tr>
<tr>
<td>33</td>
<td>16.</td>
<td>Limulus retina transfer functions</td>
</tr>
<tr>
<td>43</td>
<td>17.</td>
<td>Membrane potential as a function of time</td>
</tr>
<tr>
<td>43</td>
<td>18.</td>
<td>Firing rate as a function of membrane potential</td>
</tr>
<tr>
<td>45</td>
<td>19.</td>
<td>Schematic drawing of two interacting neurons</td>
</tr>
<tr>
<td>53</td>
<td>20.</td>
<td>Average firing rate and its derivative</td>
</tr>
<tr>
<td>59</td>
<td>21.</td>
<td>General two-neuron model</td>
</tr>
<tr>
<td>60</td>
<td>22.</td>
<td>Solutions of the two-neuron model</td>
</tr>
<tr>
<td>61</td>
<td>23.</td>
<td>Hysteresis in the two-neuron model</td>
</tr>
<tr>
<td>71</td>
<td>24.</td>
<td>Random-dot stereogram</td>
</tr>
<tr>
<td>73</td>
<td>25.</td>
<td>Depth localization</td>
</tr>
<tr>
<td>75</td>
<td>26.</td>
<td>Geometry of the disparity algorithm</td>
</tr>
<tr>
<td>92</td>
<td>28.</td>
<td>Summary of the physical and statistical variables</td>
</tr>
<tr>
<td>102</td>
<td>29.</td>
<td>Block diagram of a neuronal filter</td>
</tr>
<tr>
<td>105</td>
<td>30.</td>
<td>Single neurons in cat lateral geniculate nucleus</td>
</tr>
<tr>
<td>111</td>
<td>31.</td>
<td>Gross cerebellum</td>
</tr>
<tr>
<td>113</td>
<td>32.</td>
<td>Basic elements in the cerebellar cortex</td>
</tr>
<tr>
<td>115</td>
<td>33.</td>
<td>Patch of Purkinje cell dendrite</td>
</tr>
<tr>
<td>117</td>
<td>34.</td>
<td>Correlations in two climbing fibers</td>
</tr>
<tr>
<td>141</td>
<td>35.</td>
<td>&quot;Instantaneous rate&quot;</td>
</tr>
<tr>
<td>141</td>
<td>36.</td>
<td>Afterpotential following an action potential</td>
</tr>
</tbody>
</table>
Introduction

The neuronal basis of animal behavior is largely a mystery despite the significant progress that has been made in our knowledge of the brain's structure. Each component of the nervous system has been painstakingly studied, but only recently has it been possible to explore integrated behavior during which the components of the nervous system work together in concert. The two questions addressed by this study are first, what parallel computations can the nervous system make based on our present knowledge of its structure, and second, how can these computations be investigated with existing experimental techniques?

The complexity of the nervous system is unprecedented and our understanding of large-scale parallel computation is limited. One way to develop an intuition for parallel processing is to study the properties of simple nonlinear networks — the conceptual and mathematical insights gained could serve as a basis for more realistic models. The simplicity of a model is by no means a drawback if its behavior resembles the regularities exhibited by a complex system. An example from statistical mechanics is the two-dimensional Ising model for ferromagnetism. A two-dimensional lattice of spins with nearest neighbor interaction is a caricature of three-dimensional atomic
structure. However, the Ising model has the twin virtues of an exact analytic solution and a phase transition remarkably similar to the experimentally observed ferromagnetic transition. Biological systems, and especially the nervous system, are considerably more complex than this example, but the same strategy could be useful, and certainly the simplest models should be thoroughly studied before attempting to study more complicated ones.

General agreement has not yet been reached about what among the rich variety of chemical and electrical detail in the nervous system is most important for its function. Each approach, experimental as well as theoretical, makes drastic assumptions to limit the scope of enquiry. A brief review of nerve cells and their electrical properties is given here to provide a background for this study.

Neurons and Synapses

In 1906 the Nobel Prize in Physiology and Medicine was shared by Santiago Ramón y Cajal and Camillo Golgi for their contributions to neurohistology. Golgi had introduced a silver method for staining whole nerve cells in 1873, but his work was ignored for 14 years until Cajal, among others, improved Golgi's method and systematically applied it to tissues from a wide range of the animal kingdom. Cajal's
Fig. 1. Drawings of mammalian neurons stained by the Golgi method:  

a) Purkinje cell from the human cerebellum.  
b) Pyramidal cell from the rabbit cerebral cortex.  
c) Motor neuron from the cat spinal cord (from Cajal, 1911).
careful descriptions and extensive illustrations laid the foundation for modern neuroanatomy, and his bold interpretations of what he saw have proved remarkably accurate.

In the "black reaction", as the Golgi method is sometimes called, silver ions taken up by a neuron are reduced to a black precipitate which dramatically outlines the neuron. However, the stain is capricious: by some unknown mechanism only a few neurons are stained, but these are stained in their entirety. Although many specialized stains have been introduced, the Golgi method has not been superseded in the clarity and completeness with which the morphology of neurons can be visualized, as illustrated in Fig. 1.

The awarding of the Nobel Prize in 1906 heated a vigorous debate which had begun 25 years earlier and would not be settled until 50 years later. Ironically, Golgi and Cajal were leading spokesmen for the opposing groups. Golgi favored the idea that nerve cells were joined in a reticular network by continuous protoplasm. The main evidence for this view was the existence of fibrils within nerve cells which appeared to cross between cells. Cajal supported the opposing view that nerve cells are contiguous but not continuous. According to the "neuron doctrine", nerve cells were functional units and their zones of contact were not
Fig. 2. a) Schematic illustration of a representative vertebrate neuron. b) Functional idealization of a neuron according to the classical model. The arrows indicate typical electrical responses: top — excitatory (EPSP) and inhibitory (IPSP) postsynaptic potentials; middle — formation of the action potential at the spike-initiating zone; bottom — propagation of the action potential along the axon (from Kandel, 1976).
continuously joined. The length and bitterness of the 75 year debate are difficult to imagine now that direct electron-micrographic evidence exists proving that, with only special exceptions, the "neuron doctrine" is correct. In the last 25 years the zones of contact between neurons, called synapses, have been the focus of intensive study.

Neurons are the basic units in the brain and are greater in variety than any other cell type in the body. The diversity of shapes and sizes is matched by the complex but specific pattern of their synaptic interconnections. Certain features of the nervous system, however, appear to be universal in the animal kingdom, such as the forms of electrical signaling used for communication within and between neurons, which will be briefly summarized here. More detailed introductory accounts are given by Katz (1966), Stevens (1966), Aidley (1971), Kuffler and Nicholls (1976), Kandel (1976), and Bullock (1977).

Although no single type of neuron is typical of all the neurons in the vertebrate nervous system, the idealized neuron in Fig. 2 combines characteristics common to sensory neurons, which transduce physical stimuli into electrical signals, motor neurons, which innervate muscles, and many types of interneurons, which mediate between neurons. According to the principle of "dynamic polarization", introduced by Cajal (1911), information within a neuron
Fig. 3. EEG recordings from several vertebrates. The calibration bar in all recordings represents one second. The human recordings are from scalp electrodes; all other recordings are from electrodes on the surface of cerebral cortex. The bottom recording shows the blockage of the alpha rhythm when the subject opened both eyes (from Morrell, 1967).
flows from receptive zones, usually called dendrites, toward the transmitting element, called the axon. The cell bodies of neurons are typically between 10\(\mu\) and 50\(\mu\) in diameter, but the branching dendrites often extend several millimeters from the cell body and some axons are over a meter long. The electrical signals collected by dendrites are spatially and temporally summed by graded electrical potentials within the neuron. In contrast with electrically passive dendrites, the axon produces an active, all-or-none, regenerative event, called the action potential, which typically propagates at speeds between 0.1 and 100 m/sec depending on the size and properties of the axon. Synonyms for the action potential are impulse and spike.

The first electrical recording from an exposed brain was made by Richard Caton in 1875. Hans Berger in 1929 showed that similar microvolt potentials could be recorded on the surface of the skull. The marked rhythms observed in these recordings, as shown in Fig. 3, vary throughout the brain, are correlated with behavioral states, and are routinely used to diagnose brain injury. Because the sources of the rhythms in the electroencephalogram (EEG) are not known and because the EEG represents the pooled activity of many neurons, it is not as useful as electrical recordings from individual neurons.

The surface of a neuron, like that of any cell, is a
lipid bilayer membrane which can support a potential difference, called the membrane potential. In the absence of disturbance a resting potential of approximately 60 mV, with the inside negative, is maintained across the membrane of most neurons. Membranes are selectively permeable to particular ions, such as sodium, potassium, and chlorine; the resting potential is produced by a metabolic pump which keeps the concentration of sodium within a neuron low, and the concentration of potassium high, relative to the extracellular medium.

The membrane potential of a single neuron can be recorded with a micropipette. A finely-drawn glass capillary with a tip of about 0.1 μ is filled with electrolyte and used to impale a neuron. The membrane usually reseals around the micropipette and the intracellular membrane potential can be reliably monitored for many hours. Examples of intracellularly recorded membrane potentials are shown in Figs. 5 and 8. The weak extracellular potentials that are found in the vicinity of a neuron can be recorded with an extracellular metal microelectrode. A fine tungsten wire with its tip etched to about 1 μ was used to record the impulses shown in Fig. 9.

Most of our detailed knowledge about the electrical properties of membranes is based on the large nerve fibers (up to 1 mm in diameter) that the squid uses to initiate
rapid escape movement. The pioneering work on the giant axon of the squid in the 1950s, and in particular the investigations of Alan Hodgkin and Andrew Huxley on the action potential, has since proven relevant to many different tissues, such as vertebrate nerve fibers and heart muscle. An action potential is a sudden spike in the membrane potential which occurs in some membranes when a threshold potential is reached and which lasts for a few milliseconds, as shown in Fig. 4a. The action potential is triggered by voltage-sensitive channels in the membrane: the permeability for sodium ions is transiently increased during the rising phase (a positive shift of the membrane potential called a depolarization), and the permeability for potassium ions is increased during the falling phase (a negative shift of the membrane potential called a hyperpolarization), as indicated in Fig. 4b.

At most synapses a chemical neurotransmitter is used to communicate across the 200 Å separation, called the synaptic cleft, between the presynaptic membrane and the postsynaptic membrane, as shown in Fig. 5a. When an action potential or depolarization invades the presynaptic terminal, calcium enters the terminal, and shortly thereafter neurotransmitter is released into the synaptic cleft. The neurotransmitter diffuses to the postsynaptic membrane, where it binds with a specialized protein receptor and causes a temporary
Fig. 4. Theoretical changes of the a) membrane potential and b) membrane conductances during an action potential. $E_{Na}$ and $E_K$ are respectively the sodium and potassium equilibrium potentials, and $g_{Na}$ and $g_K$ are the sodium and potassium conductances. The equilibrium potential for an ionic species is the potential across a semipermeable membrane caused by an ionic concentration difference (from Aidley, 1971).
permeability increase for particular ions. The magnitude and sign of the resulting postsynaptic potential (PSP) depends on the postsynaptic membrane potential before the synapse is activated; in fact, at the reversal potential the PSP is zero. The synapse is called excitatory if the reversal potential is above threshold and inhibitory if the reversal potential is below threshold. In most normal circumstances the activation of an excitatory synapse moves the membrane potential closer to threshold and the activation of an inhibitory synapse drives the membrane potential away from threshold. An example of excitatory and inhibitory postsynaptic potentials (EPSP's and IPSP's) in an intracellular recording is shown in Fig. 5b. Some electric synapses have been found which are direct, low-resistance connections between neurons and for which there is no appreciable time delay between the presynaptic activation and the postsynaptic response.

Slow Potentials and Stochastic Variability

The massive bundles of nerve fibers in the vertebrate brain and the striking all-or-none character of the action potential led early investigators to compare the nervous system with digital computers and logical automata (McCulloch and Pitts, 1943; Shannon and McCarty, 1956).
Fig. 5. a) Schematic drawings of typical chemical synapses. The presynaptic terminal is filled with vesicles containing a neurotransmitter. Synapses often occur not on the dendrite itself but on thorn-like protuberances called spines, as shown on the right. b) Intracellular recording from an Aplysia neuron which displays both excitatory and inhibitory postsynaptic potentials (from Aidley, 1971).
However, as the techniques for recording intracellular potentials were developed and applied to nerve cells in a variety of nervous systems, the importance of subthreshold graded potentials for the integration of information within the neuron was soon appreciated (Grundfest, 1957; Bullock, 1959).

Graded membrane potentials, such as those produced by sensory receptors and synapses, are called slow potentials to distinguish them from action potentials. One of the earliest investigators to emphasize the role of slow potentials for signal processing was Bishop (1956), who summarized the evidence for graded responses and concluded that the action potentials in a sensory nerve transduce the graded sensory stimulation into a corresponding slow potential at the next neuron. A similar view was expressed by Fuortes (1959) in discussing reflex responses:

"It turns out therefore that, even in the simplest conditions applying to excitatory monosynaptic reflexes, a whole succession of frequency-intensity transformations is involved in the production of normal actions: intensity to frequency in receptors, frequency to intensity at synaptic regions, intensity to frequency at the pacemaker regions [spike-initiating zones] of motoneurones, and finally, frequency to intensity in muscles."

The evidence for processing by graded potentials is now overwhelming and the view expressed by Bishop and Fuortes, sometimes called the "slow potential theory", is generally
Fig. 6. Schematic illustration of the "slow potential theory" of information processing in neurons. The graded potentials in the two neurons on the left are converted into firing rates which are decoded by the postsynaptic neuron on the right. The membrane potential of the postsynaptic neuron temporally and spatially sums the excitatory (bottom) and inhibitory (top) influences. According to the "slow potential theory", the primary computations in the nervous system are performed by the graded membrane potentials within neurons, and action potentials are only a secondary information carrier (from Stevens, 1966).
Fig. 7. Schematic summary diagram of the five principal cell types and typical synaptic contacts found in vertebrate retinas. Light is transduced into electrical signals by the photoreceptor cells whose receptor terminals (RT) are shown at the top of the diagram. Electrical signals from the receptors are conveyed by the bipolar cells (B) to the ganglion cells (G) whose long axons make the optic nerve. Lateral interaction within the retina is provided by the horizontal cells (H) near the receptor layer and the amacrine cells (A) near the ganglion cell layer. The direction of information flow can be inferred from the round vesicles found only in presynaptic terminals. Some terminals, called reciprocal synapses, are both presynaptic and postsynaptic (from Dowling, 1968).
Fig. 8. Intracellular recordings from the five principal cell types in the mudpuppy retina. The responses were elicited with a flash of light focussed on the left photoreceptor. The recordings are placed in a schematic wiring diagram to suggest how the responses could be produced by synaptic interactions. The receptors (R), the horizontal cells (H), and the bipolar cells (B) respond with slow, graded potentials. The amacrine cells (A) respond more transiently and some ganglion cells (G₂) respond with a burst of action potentials at the onset or offset of the light, or both. Other ganglion cells (G₁) respond with sustained firing or sustained inhibition (from Dowling, 1970).
accepted by most investigators. Stevens (1964) has given a particularly clear statement of the "slow potential theory", and Fig. 6 is an instructive illustration from his book (1966). Signal processing by graded potentials is particularly important in the vertebrate retina where most neurons do not produce action potentials (Tomita, 1965; Werblin and Dowling, 1969; Kaneko, 1970). A schematic diagram of the five principal cell types in the vertebrate retina is shown in Fig. 7, and intracellular recordings from the mudpuppy retina are shown in Fig. 8. Graded interaction between dendrites has also been found in the olfactory bulb and elsewhere in the central nervous system (Rakic, 1975; Pearson, 1976; Schmitt, Dev & Smith, 1976; Shepherd, 1978).

A typical pyramidal cell in cerebral cortex may have 1,000 to 10,000 synapses and receive 10,000 to 100,000 postsynaptic potentials per second, most of which are spatially and temporally summed in the dendritic tree. Normal electrical conduction in a dendrite is governed by an equation identical to that studied by Lord Kelvin for electrical transmission in submarine cables. In the steady state the potential from a battery at one end of a uniform cable decrements with distance $X$ along the cable according to

$$V = V_0 e^{-X/\lambda},$$

where $\lambda$ is the space constant which depends on the
Fig. 9. Extracellular recordings from a single neuron in the cat visual cortex. This neuron responded best to a slit of light obliquely oriented in a particular part of the visual field. The first 12 successive responses of the neuron to 50 ms exposures of the light are shown above, and the average response for 20 trials is shown below. Although the pattern of firing varied from trial to trial (and some parts of the response drop out entirely, such as in trials 5, 10, and 11), the average over the ensemble of trials, called the poststimulus time histogram (PST), is a repeatable measurement (from Morrell, 1972).
electrical properties of the membrane. For most nerve fibers the space constant is typically between 100 \( \mu \) and 1 mm.

Electrical recordings from single neurons in the vertebrate nervous system have a surprising degree of variability. For example, extracellular recordings of a neuron in the cat visual cortex are shown in Fig. 9. A neuron in visual cortex will respond to light only in a restricted region of the visual field, called its receptive field, and many neurons respond best to slits of light with a preferred orientation and direction of movement (Hubel and Wiesel, 1962). The successive trials shown in Fig. 9 illustrate the variability of the response under identical conditions. Although the details of the response are different from trial to trial — the action potentials appear to occur at irregular intervals — the average firing rate from pooled trials, shown at the bottom of Fig. 9, is a more consistent measurement. Stochastic variability is found at every level of the nervous system, from sensory receptors to cerebral cortex.

One of the chief sources of noise in the nervous system is variation in the amount of neurotransmitter released at chemical synapses. Neurotransmitter is stored in small packets called vesicles, shown in Fig. 5a, and is released from the presynaptic terminal in discrete units called
quantal fluctuations and threshold fluctuations occur on a millivolt scale. Microvolt variations are also observed owing to shot noise from the opening and closing of ionic channels in the membrane. For example, fluctuations in intracellular recordings, as shown in Fig. 10b, have been used to estimate the kinetics of a single acetylcholine receptor during activation of the frog neuromuscular junction: when acetylcholine binds to the receptor, a channel is opened for about 1 ms, about $5 \times 10^4$ univalent ions flow through the membrane (mainly potassium and sodium ions), and a postsynaptic potential of about $0.3 \mu V$ is
Fig. 1). Recordings made from the motor nerve synapse of the frog. a) The lower trace on the left panel shows the postsynaptic membrane potential in response to nerve stimulation (scale: 50 mv and 2 ms). The upper traces show spontaneous activity in the absence of stimulation (scale: 3.6 mv and 47 ms). The miniature synaptic potentials are a result of the random release of acetylcholine-filled vesicles from the presynaptic terminal. The right panel presents the same measurements made 2 mm from the synapse. b) Low-gain DC recordings (narrow traces) and high-gain AC recordings (noisy traces) of the presynaptic membrane potential (top). When acetylcholine is applied to the junction (bottom), the membrane potential increases and the membrane noise from the opening and closing of ionic channels increases. Spontaneous miniature synaptic potentials are also shown. c) Recording of the current through a small patch of postsynaptic membrane bathed with suberyldicholine, which is known to open the acetylcholine receptor for a longer period than acetylcholine. The rapidity with which the channel opens and closes indicates that the receptor has only two states (from DeFelice, 1977).
produced. Similar measurements have been made on the sodium and potassium channels which are activated during an action potential (Neher and Stevens, 1977; DeFelice, 1977).

The *Limulus* Retina

One of the most carefully studied and best understood pieces of nervous tissue is the lateral eye of *Limulus polyphemus*. Sometimes called the horseshoe crab but in fact an arachnid, *Limulus* is today not much different from its ancestors 500 million years ago. The compound lateral eye of *Limulus* contains about 600 ommatidia, each of which is a functional photoreceptor unit and resembles a segmented orange, as shown in Fig. 11. There are about a dozen segments and all but one are retinular cells containing a rhodopsin-like visual pigment. The one odd segment is an eccentric cell whose axon generates impulses and leaves the eye through the optic nerve. Lying just beneath the ommatidia is a layer of finely branching axons and synaptic contacts which provides lateral interaction between the eccentric cells.

The action potentials from a single ommatidium can be recorded by carefully dissecting the optic nerve and teasing out a single fiber. The response of an ommatidium to a sudden increase in the illumination is a transient burst of
Fig. 11. Longitudinal cross-section of a *Limulus* photoreceptor ommatidium, which is about the size of a pencil lead. Light enters through the lens at the top. The retinular cells (R) contain light-sensitive rhabdoms (r) and are arranged like the segments of an orange around the dendrite (DP) of an eccentric cell (E). Illumination of the ommatidium gives rise to impulses in the axon (ax) of the eccentric cell (from Ratliff, Hartline, & Miller, 1963).
Fig. 12. Electrical recordings from a single optic nerve fiber of the Limulus lateral eye. The retina was illuminated by light with the relative intensity given on the left. The time marks at the bottom of each trace represent 200 ms. The white band above the time marks is blackened during the period of illumination. Several features of the photoreceptor response are illustrated: first, the rate of firing in steady state depends monotonically on the intensity over a wide range; second, a burst of firing occurs at the onset of illumination; and third, the delay of the response decreases with increasing illumination. Like our own eyes, the Limulus retina has a wide dynamic range while remaining sensitive to small changes in the intensity (from Hartline, 1941).
Fig. 13. Demonstration of lateral inhibition in the *Limulus* retina. The schematic drawing of the *Limulus* retina on the left shows two dissected nerve fibers. The illumination on each ommatidium was independently controlled. The graphs on the right show the decrease in the firing rate of one fiber as a function of the firing rate in the other fiber. In both cases the data above the threshold for lateral inhibition are fit by a single straight line. The mutual inhibition arises in the lateral layer beneath the photoreceptor ommatidia (from Ratliff, 1965).
impulses which subsides to a regular discharge, as shown in Fig. 12. In 1932 Haldan Keffer Hartline found that the steady-state firing rate of a dissected ommatidium varies in proportion to the logarithm of the incident light intensity over at least four decades. Since then it has been possible to study every step in the process of sensory transduction and neural integration.

In 1956 Hartline and his colleagues discovered that the firing rate of an ommatidium is diminished by the illumination of neighboring ommatidia, as shown in Fig. 13. The inhibition between neighboring ommatidia depends linearly on their firing rates (Hartline and Ratliff, 1957). For steady-state illumination above the threshold for lateral inhibition, the firing rate of the a-th ommatidium \( \gamma_a \) is determined by

\[
\gamma_a = e_a - \sum_b K_{ab}^I (\gamma_b - \gamma_b^0),
\]

where \( e_a \) is the input, \( K_{ab}^I \) are the inhibitory coefficients, and \( \gamma_b^0 \) is the threshold for inhibition. The validity of this linear model has been established by extensive tests that typically involved the measurement of two or more ommatidia, as shown in Fig. 13. The statistical distribution of the inhibitory coefficients has the shape of a volcanic crater around each ommatidium (Barlow, 1969).

The response of the Limulus retina illuminated with a spatial step function is shown in Fig. 14. Lateral
Fig. 14. The firing rate of a single ommatidium in response to a step function in illumination. The density of the photographic plate used to project the stimulus is shown in the upper right corner. When all the receptors except one were covered by a mask, the receptor responded directly to the light intensity, as shown in the top graph. The lower graph shows the firing rate of the same receptor when the mask was removed. The response at the border is enhanced by lateral inhibition: a brightly-illuminated receptor inhibits a less-brightly-illuminated neighboring receptor more strongly than it is inhibited in return (from Ratliff and Hartline, 1959).
inhibition enhances the response contrast at borders between regions with differing light intensity. Humans perceive a similar contrast enhancement, known as Mach bands, which may arise from lateral inhibitory interaction in the human retina (Ratliff, 1965).

The steady-state firing rate of an eccentric cell is proportional to its depolarization whether the depolarization is produced by phototransduction or by current injected intracellularly (Puortes, 1959b). Thus, the logarithmic relationship between light intensity and firing rate occurs in the first step that produces the generator potential. The transient burst of impulses in response to sudden increase in illumination (Fig. 12) is the result of self-inhibitory feedback (Stevens, 1964; Purple, 1964; Purple and Dodge, 1965). The inhibitory postsynaptic potential has a decay time constant of about 0.5 sec.

The linearity and translation invariance of the Limulus retina have been exploited in a series of experiments which generalize the Hartline-Ratliff model for arbitrary temporal and spatial patterns of illumination. Bruce Knight, Jun-ichi Toyoda, and Frederick Dodge (1970) have used linear systems theory to examine each element of signal processing in the Limulus retina: (i) the transduction of light to a generator potential in the eccentric cell; (ii) the transduction of the graded potential in the eccentric cell
into action potentials; and (iii) the inhibitory influence of neighboring ommatidia.

Since each transformation of the signal past the generator potential is approximately linear, the response of the *Limulus* retina should be predictable from a transfer function. Knight, Toyoda, and Dodge experimentally measured transfer functions for all three elements of signal processing in the *Limulus* retina and demonstrated self-consistency between the model and the measurements. They illuminated a single ommatidium in the *Limulus* retina with light containing a sinusoidally-modulated component and measured the response of the generator potential and the firing rate of the eccentric cell (top and bottom frames of Fig. 15). Because the system is linear, the responses also had a sinusoidally-modulated component, but with different amplitudes and phases. The experiment was repeated with the light modulation replaced by modulated current directly injected into the eccentric cell (center frame of Fig. 15). If the system were truly linear, then the product of the measured gain from the first transduction (light to generator potential) with the measured gain from the second transduction (generator potential to firing rate) should equal the measured gain for the overall transduction (light to firing rate), and furthermore, the respective phases should add. Good agreement was found at every modulation
Fig. 15. Responses of a *Limulus ommatidium* to modulated inputs. The top and bottom frames show respectively the generator potential and firing rate in response to sinusoidally-modulated illumination. The middle frame shows the firing rate in response to steady background illumination with sinusoidally-modulated current injected directly into the eccentric cell. Least-squares fits were made to the data between the vertical lines. The transfer functions in Fig. 16 were based on the results of this experiment and similar experiments at different modulation frequencies (from Knight, Toyoda, and Dodge, 1970).
Fig. 16. The amplitudes (above) and phases (below) of the transfer functions for two elements of signal processing in the *Limulus* retina. In the left panel, the measured transduction of light to generator potential (open circles) and current to firing rate (solid circles) are shown with best-fit curves through the data points. In the right panel, the overall transduction of light to firing rate is shown based on direct measurements (open circles) and predicted from the data in the left panel (solid curves) (from Knight, Toyoda, and Dodge, 1970).
frequency tested between 0.1 and 10 Hz, as shown in Fig. 16. The response of the ommatidium to an arbitrary temporal pattern of illumination can be predicted from the measured transfer function.

Knight, Toyoda, and Dodge also found the transfer function for the lateral inhibitory pathway and were able to successfully generalize the Hartline-Ratliff equation to the frequency domain. If light modulation of intensity $i_a$ and frequency $f$ produces a generator potential

$$ G(f) = G(f) i_a, $$

where $G(f)$ is the transfer function, then the modulated firing rate $\gamma_a(f)$ is determined by

$$ \gamma_a = e_a - K T_S(f) \gamma_a - T_L(f) \sum_{b \neq a} k_{ab} \gamma_b, $$

where $T_S(f)$ and $T_L(f)$ are respectively the transfer functions for self-inhibition and lateral inhibition, and $K$ and $k_{ab}$ are real coefficients. The quantities in this equation are generally complex and of the form $A e^{i \phi}$, where $A$ is the amplitude and $\phi$ is the phase. The transfer function for lateral inhibition depends on the separation between ommatidia only through the amplitude, which scales by a real coefficient. This model allow the response of the *Limulus* retina to be predicted for arbitrary spatial as well as temporal patterns of illumination.
The Hartline-Ratliff model of the Limulus retina, "battle tested" and refined by several generations of investigators, represents a standard against which future models will be judged. The present study attempts to build on the Limulus model by generalizing it to the vertebrate nervous system. Further details about the Limulus model can be found in several reviews (Dodge, 1969; Dodge, Shapley, and Knight, 1970; Hartline and Ratliff, 1972; Knight, 1973) and a volume of collected papers (Ratliff, 1974).

Generalizing the Limulus Model

The lateral inhibitory pathways in the Limulus retina operate in a way which is consistent with the "slow potential theory" of neuronal interaction discussed on page 14. Stevens (1964) has shown that the "slow potential theory" also explains the transient response of the Limulus retina to temporal patterns of light. A sudden increase in the illumination of an ommatidia causes a sudden rise in the firing rate of the eccentric cell, followed by decay to a new resting level. Stevens presented evidence that a self-inhibitory pathway exists which obeys the "slow potential theory". Thus, in at least one piece of nervous system, the "slow potential theory" provides an adequate description and a quantitative model.
The *Limulus* retina is an ideal preparation because of its regular structure and easy accessibility. Is the *Limulus* retina special in other respects as well, and might the "slow potential theory" be less successful when applied to other nervous systems? The linearity of the *Limulus* retina is an aspect that may not apply elsewhere. Below threshold the firing rate of a neuron which is not autoactive falls to zero, and for large inputs the firing rate cannot exceed some maximum rate owing to a minimum recovery time. Although some neurons, like those in the *Limulus* retina, operate in the approximately linear region between threshold and saturation, many neurons in the vertebrate nervous system operate in their nonlinear region. Nonlinearity introduces qualitatively new features into neuronal interaction and greatly complicates quantitative models.

Another qualitative difference between the *Limulus* retina and the vertebrate nervous system is the degree of variability in electrical recordings. An eccentric cell in the *Limulus* retina fires regularly in response to a flash of light (Fig. 12), but a cell in the cat visual cortex responds irregularly (Fig. 9). Stochastic models are commonly used when sources of noise are present. Given the degree of variability found in most parts of the vertebrate nervous system, a stochastic model of interacting neurons
would be more appropriate than a deterministic one. In systems engineering the type of model developed in this study is called a stochastic state-variable model.

The *Limulus* retina has a particularly simple and regular pattern of lateral inhibitory interconnections. Although the details of synaptic interaction in the neuropil beneath the ommatidia are in fact quite complicated, the lattice-like structure of the *Limulus* retina produces a statistically averaged interaction that is quite regular. In other nervous systems the pattern is more complicated and detailed connections are known in only a few special cases, such as the vertebrate cerebellum. To early anatomists the branching patterns of dendrites and axons in cerebral cortex had a random appearance, but evidence is accumulating that connections between cortical neurons are specific and far from random (Mountcastle, 1978). Without knowledge of the precise connectivity only qualitative properties of a model can be studied.

Although the firing rate is the primary variable in the Hartline-Hatliiff model of the *Limulus* retina, the graded membrane potential is an important integrative variable within a neuron and has a central role in the "slow potential theory". A majority of neurons in the vertebrate retina do not even produce action potentials, and parts of neurons in other areas of the nervous system seem to be
specialized for locally processing graded potentials. Thus, the membrane potential is a more fundamental variable than the firing rate, although for those neurons which produce action potentials the firing rate remains an important secondary variable.

A simple model of interacting neurons is motivated in Part I which incorporates the nonlinearity, the electrical variability, and the precise connectivity observed in vertebrate nervous systems. Probability theory is used to study the statistical properties of the model in a two-tier analysis: the average membrane potentials and average firing rates are examined in Part II, and the variances and correlations between membrane potentials are analyzed in Part III. In Part IV the model is applied to motor learning in the cerebellum.

All the results in the main body of this study are derived from a model whose primary variable is the continuous membrane potential. In the appendix a point-process model is motivated which is a more realistic model for impulse-producing neurons and which takes into account refinements such as propagation delays in axons and electrotonic conduction in dendrites. All the main results of the continuous model also hold for the more general point-process model.
I. The Nonlinear Model

In most of the present study a neuron is considered a point with a set of properties, such as an electrical potential and an input-output function, and connections between neurons are represented by a set of coupling strengths, arrayed for convenience in a square matrix. A reasonable question to ask is whether such a "cartoon" neuronal model oversimplifies essential properties of real neurons. The answer lies ultimately in the value of the model for organizing experimental data, but the question can be partially answered by testing whether results based on the model are sensitive to particular simplifying assumptions. For example, the main results of the model are shown in the appendix to hold even when more realistic dendritic inputs are included.

Neurons and "Microcircuits"

According to the principle of "dynamic polarization" (Cajal, 1911), information within neurons is transmitted one way: passive dendrites receive electrical signals through synapses; the integrated input at the spike-initiating zone is coded into impulses; the impulses are then sent out through the axon to influence other neurons. Although this
classical picture, as shown in Fig. 2, is a good approximation for many neurons, some recent findings are inconsistent with it: synapses have been found from dendrite to dendrite and from axon to axon, and some synaptic interactions take place that are not mediated by impulses (Shepherd, 1974; Rakic, 1975; Schmitt, Dev, & Smith, 1976).

A new concept that has emerged is the "microcircuit" (Shepherd, 1978), which is a localized anatomical region that processes information through continuously graded membrane potentials. In the model developed here the membrane potential is the primary physical variable so that both the classical impulse-producing neuron and the local "microcircuits" can be included.

Because electrotonic potentials decrement exponentially in passive dendrites, a neuron which transmits signals over a long distance must use regenerative action potentials, or impulses, to provide fast and reliable communication. These projection neurons most nearly resemble the classical model shown in Fig. 2, and have been studied in greater detail than the "local circuit neurons" (Rakic, 1975) which often form the "microcircuits". The generally accepted explanation for how information is coded into impulses and transmitted by projection neurons will be summarized here and used to motivate a quantitative nonlinear version.

Graded membrane potentials are called slow potentials
to distinguish them from the much more rapidly varying action potential. Above the threshold for producing action potentials, the firing rate increases monotonically with the neuron's membrane potential. The impulses transmitted by the axon to the synaptic connections with other neurons then produce postsynaptic potentials. According to the "slow potential theory" of neuronal interaction (Stevens, 1964, 1966), the temporally summed slow potential in the postsynaptic neuron is an approximate reproduction of the membrane potential in the presynaptic neuron, diminished in amplitude but unaltered in shape. In short, this approach assumes that information about the membrane potential in one neuron is linearly transduced to other neurons by the firing rate. However, below threshold no impulses are produced, and for large membrane potentials the firing rate saturates at some maximum. These nonlinearities are taken into account in the quantitative model given here and developed in more detail in the appendix.

Quantitative Model

One of the simplest linear models for neurons which do not produce action potentials is given by

\[ \tau \frac{d}{dt} V_a + V_a = \sum_b K_{ab} V_b + R_a I_a, \]  

(1)

where \( V_a \) are the somatic membrane potentials, \( \tau \) is
the membrane time constant, \( I_a \) are the synaptic input currents, \( R_a \) are the effective load resistances, and \( K_{ab} \) are dimensionless coupling strengths. The left side of Eq. (1) provides temporal integration like that of a leaky capacitor, and the right side takes into account the inputs and interconnections.

Action potentials introduce a strong nonlinearity into neuronal interaction. The response of an idealized impulse-producing neuron to a constant input current is shown in Fig. 17. Imagine for the moment that action potentials are prevented so that the membrane potential is allowed to vary smoothly above the threshold for discharge, as indicated by the dashed line in Fig. 17. Define the effective membrane potential \( \phi \) as the membrane potential which would be present in a neuron if the action potential were absent. The firing rate \( \rho(\phi) \) of an idealized neuron, which is a function of the effective membrane potential as shown in Fig. 18, increases monotonically above threshold and, because of the absolute refractory period, approaches an upper bound.

Action potentials cannot be eliminated from a neuron when the integrity of an interacting system of neurons is important. Hence, indirect methods must be used to estimate the effective membrane potential above the threshold for discharge. The regenerative part of the action potential
Fig. 17. a) The membrane potential of an idealized neuron in response to a constant input current as a function of time. The dashed line represents the effective membrane potential which would be present in the absence of action potentials. b) The membrane potential of a second idealized neuron which receives synaptic input from the first, as illustrated in Fig. 19 (from Sejnowski, 1977a).

Fig. 18. The firing rate $\rho(\phi)$ as a function of effective membrane potential $\phi$ for an idealized neuron with firing threshold $\Theta$ (from Sejnowski, 1977a).
lasts for only a few milliseconds and can be easily recognized and removed from the recording. Afterpotentials following an action potential reset the membrane potential below threshold, but the effect of the reset decays exponentially (with the membrane time constant) after each impulse (see appendix). An intracellular recording with action potentials removed and the reset compensated should be a good approximation to the effective membrane potential defined above. In a neuron which does not produce an action potential, the difficulty above threshold does not arise and the effective membrane potential is the membrane potential.

For the coupled pair of neurons represented in Fig. 19, repetitive firing of the presynaptic neuron produces a repetitive postsynaptic potential, as shown in Fig. 17b. Under steady-state conditions the average postsynaptic potential is constant and, for an idealized neuron, proportional to the firing rate. A model for a collection of idealized interacting neurons is given by

$$\tau \frac{d}{dt} \phi_a + \phi_a = \sum_b K_{ab} \rho_b(t) + \sum_b B_{ab} \eta_b,$$  \hspace{1cm} (2)

where $\tau$ is the membrane time constant, $\eta_b(t)$ is the input firing rate, $\rho_b(t)$ is the firing rate of the nonlinear interaction defined above, and $K_{ab}$ are the input coupling strengths and $B_{ab}$ are the internal coupling strengths, with units of potential/firing rate. Because the effective membrane potentials $\phi_a(t)$ are continuous, this
Fig. 19. Schematic illustration of two neurons with a synaptic connection $K_{21}$ from the first to the second and with inputs $\eta_1$ and $\eta_2$ respectively. Micropipettes record the intracellular membrane potentials $V_1$ and $V_2$ (from Sejnowski, 1977a).
nonlinear model applies equally well to neurons which do not produce an action potential and parts of neurons which interact through graded synapses. Hence, "microcircuits" can be modeled as well as classical neurons.

Above the threshold for lateral inhibition, the response of the Limulus retina to a steady-state pattern of light is given, to a good approximation, by the Hartline-Ratliff model (1957)

\[ \gamma_a = e_a - \sum_b^I K_{ab} (\gamma_b - \gamma_b^0), \]  

where \( \gamma_a \) is the rate of firing of an ommatidium, \( \gamma_b^0 \) is the inhibitory threshold, \( e_a \) represents the input, and \( K_{ab} \) are the inhibitory coefficients. Under normal experimental conditions a steady background illumination produces a steady background firing rate against which variations in the rate are measured.

If in the nonlinear model the effective membrane potentials are constant, then they can be eliminated in favor of the firing rates

\[ \gamma_a = \rho_a (\sum_b B_{ab} \eta_b + \sum_b K_{ab} \gamma_b). \]  

The Hartline-Ratliff equation is equivalent to this equation when \( \rho(\phi) \), shown in Fig. 18, is restricted to the approximately linear region above threshold.

Although many details of real neurons are not included in the continuous model motivated here, the main results
based on it also hold in a more general model, given in the appendix, which takes into account axonal latency and dendritic electrotonus.

In summary, the highly nonlinear action potential was eliminated by first redefining the membrane potential above the threshold for discharge, and secondly by smoothing the postsynaptic potentials. The nonlinear model (2) based on this effective membrane potential differs from the linear model (1) by an effective nonlinear interaction, as represented in Fig. 18.
II. Average Membrane Potentials in Steady State

In electrical recordings from neurons in various parts of the brain, impulses often occur irregularly and membrane potentials have a randomly fluctuating component. Some examples are given in Figs. 8, 9, and 10. As discussed in the introduction, the sources of this random component include membrane noise from the opening and closing of ionic channels, fluctuations in the threshold for producing action potentials, and fluctuations in the number of transmitter packets released during synaptic transmission (Verveen and DeFelice, 1974; Holden, 1976; Neher and Stevens, 1977; Stevens, 1977; DeFelice, 1977).

Because of the variability in electrical recordings, the electrical response of the brain to a sensory stimulus is often measured on a number of trials under conditions as uniform as possible from trial to trial. If the trials are reasonably independent and conditions are reasonably stationary, then the uncontrolled part of the variability cancels and the average response does not depend on which block of trials were chosen. When spike trains are measured the average response is called the poststimulus time histogram (PST), as shown in Fig. 9, and when gross potentials from groups of neurons are measured, the average is called the average evoked response (AER).
In a PST each response is a function of time after the stimulus and the average is computed from a collection or ensemble of trials. This form of average, called an ensemble average, is used throughout this study. If conditions are stationary (the stimulus is constant and the system is in equilibrium), then a time average can be computed from a single trial which (under further conditions) is equal to the ensemble average. For example, in statistical mechanics the time-averaged macroscopic variables, such as temperature and pressure, can be computed from a Gibbs ensemble of identically prepared systems.

Let \( f(t) \) represent a stationary signal, such as an intracellular recording of a membrane potential or an extracellular recording of a spike train. The time-average mean of \( f(t) \) over time \( T \) is defined as

\[
\langle f \rangle_T = \frac{1}{T} \int_0^T dt \, f(t).
\]

Now consider an ensemble of recordings \( \{ f_a(t) \} \) from \( N \) independent trials under uniform conditions. The ensemble-average mean is defined as

\[
\hat{f}_N(t) = \frac{1}{N} \sum_{a=1}^N f_a(t).
\]

The ensemble-average mean can be a function of time \( t \), as in the case of the PST. If, however, conditions are stationary, then \( \lim_{N \to \infty} \hat{f}_N(t) \) is independent of time, and if certain ergodic conditions are also satisfied (Halmos,
1956), then the appropriate limits of the time-average mean and the ensemble-average mean are equal:

$$\lim_{T \to \infty} \langle f \rangle_T = \lim_{N \to \infty} \hat{f}_N.$$ 

**Stochastic Model**

Engineering models of physical systems usually take random influences into account by adding noise terms to deterministic equations. For example, the input $\eta_a(t)$ to the nonlinear model (2) might be modeled by a deterministic part $\hat{\eta}_a(t)$ and a white noise component $\omega_a(t)$:

$$\eta_a(t) = \hat{\eta}_a(t) + \omega_a(t).$$

Then the mean and covariance of the input are

$$E\eta_a(t) = \hat{\eta}_a(t),$$

$$\text{Cov}(\eta_a(t), \eta_b(s)) = E(\eta_a(t) - \hat{\eta}_a(t))(\eta_b(s) - \hat{\eta}_b(s))$$

$$= Q_{ab} \delta(s - t).$$

where $E$ is the expectation or ensemble average and $Q_{ab}$ is the covariance of the white noise process. More generally, the input can be considered an arbitrary Markov process with finite second-order moments

$$\text{Cov}(\eta_a(t), \eta_b(s)) = Q_{ab}(t, s).$$

The physical significance of the covariance is discussed in the next part.
The response of the nonlinear model to an ensemble of randomly-varying inputs is an ensemble of randomly-varying membrane potentials. Much can be learned about the stochastic model by examining the lowest-order moments of the membrane potentials. From the expectation of Eq. (2), the mean or average effective membrane potentials

\[ \hat{\phi}_q(t) = E \phi_a(t) \]  

(5)
satisfy

\[ \tau \frac{d}{dt} \hat{\phi}_a + \hat{\phi}_a = \sum \alpha_{ab} R_b(\hat{\phi}_b) + \sum \beta_{ab} \hat{\eta}_b, \]  

(6)
where the average firing rates are

\[ R_b(\hat{\phi}_b) = E \rho_b(\phi_b). \]

Since the right side of Eq. (7) depends on all the moments of \( \phi_b \), the average firing rate \( R_b(\hat{\phi}_b) \), written explicitly as a function of the average membrane potential \( \hat{\phi}_b \), depends implicitly on all the higher-order moments as well. Alternatively, the average firing rate can be considered a functional over an ensemble of membrane potentials \( R_b[\phi_b] \), but this notation might be confusing and will be avoided. Thus, Eq. (7) is not a strictly closed set of relations for \( \hat{\phi}_b \), but is coupled to equations for the higher-order moments of \( \hat{\phi}_b \).

Because of membrane potential fluctuations, the average rate of firing \( R_b(\hat{\phi}_b) \) as a function of \( \hat{\phi}_b \), holding all higher-order moments fixed, is smoother than \( \rho_b(\phi_b) \) as a
function of $\phi_b$. For example, Fig. 20 shows $R(\hat{\phi})$ for $p(\phi)$ a step function at threshold $\Theta$ and with $\phi$ having a Gaussian distribution. In general, Eq. (6) for the mean effective membrane potentials is coupled, through the terms with $R_b(\hat{\phi}_b)$, to equations for the higher moments.

Nonlinear equations similar to Eq. (6) have been investigated by White (1961), Wilson and Cowan (1968), and Grossberg (1973), who base their models on populations of neurons. The primary variable in their equations is the fraction of neurons in an "excited state". In the present case the membrane potentials of individual neurons are studied and the averages are over ensembles of identically prepared experimental trials rather than over physical populations. The expectation corresponds to the average response from the ensemble.

In the stationary case the mean effective membrane potentials are constant and satisfy

$$\hat{\phi}_a = \sum_b K_{ab} \hat{R}_b(\hat{\phi}_b) + \hat{\mathcal{U}}_a,$$

(8)

with the mean inputs

$$\hat{\mathcal{U}}_a = \sum_b B_{ab} \hat{\eta}_b.$$

(9)

Eliminating $\hat{\phi}_a$ from Eq. (8) in favor of the average firing rate

$$\hat{\gamma}_a = R_a(\hat{\phi}_a),$$

(10)

an alternative form of the steady-state equation is

$$\hat{\gamma}_a = R_a(\sum_b K_{ab} \hat{\gamma}_b + \hat{\mathcal{U}}_a).$$

(11)
Fig. 20. The average firing rate $R(\hat{\phi})$ and its derivative $R'(\hat{\phi})$ as a function of the mean effective membrane potential $\hat{\phi}$. The threshold for firing is $\Theta$ (from Sejnowski, 1977a).
Steady-State Solutions

Although the Hartline-Ratliff model (3) of the *Limulus* retina is linear above threshold, the cutoff below threshold makes the model nonlinear. Hadeler (1974) has used a fixed point theorem to study the existence of solutions to the *Limulus* model with thresholds. His approach can be generalized to the nonlinear steady-state equation (8) for the average membrane potentials (Sejnowski, 1976a).

Because the steady-state equation (8) for $\hat{\phi}_a$ depends implicitly on the higher-order moments of $\hat{\phi}_a$, it cannot be treated independently of the higher-order equations. The question asked here is whether the steady-state equation for $\hat{\phi}_a$ has a solution when all higher-order moments are held fixed, which is a necessary but not a sufficient condition for there to exist a stationary solution to the coupled set of equations. The answer is in the affirmative for all values of the higher-order moments.

A fixed point of a function is a point which the function maps into itself. Topological fixed-point theorems give very general conditions under which a fixed-point solution must exist. In operator notation the steady-state equation (8)

$$\hat{\phi} = KR(\hat{\phi}) + \hat{\omega}$$
can be written as the fixed point

\[ x = F(x) \]

of

\[ F(x) = KR(x + \hat{u}) \]

with

\[ x = \phi - \hat{u}. \]

Theorem. (Brouwer). Every continuous mapping of a closed \( n \)-ball into itself has a fixed point.

The map \( F \) is continuous since \( R \) is continuous; \( F \) is bounded by

\[ B = \| K \| \text{max}_{x} R(x); \]

and \( F \) maps the ball

\[ \mathbb{B} = \{ x \mid \| x \| \leq B \} \]

into itself

\[ F: \mathbb{B} \rightarrow \mathbb{B}. \]

Hence by the Brouwer fixed-point theorem the steady-state equation always has at least one solution. However, the solution may not be stable and periodic solutions, as well as less regular behavior, may also occur. This case requires treatment of the coupled equations for all the moments of the membrane potentials.

A function \( F \) on a metric space is a contraction map if, starting with two arbitrary points \( x_1 \) and \( x_2 \), \( F(x_1) \)
and \( F(x_2) \) are always closer together than the original two points. The contraction mapping theorem states that a contraction map always has a unique fixed-point solution and that the solution can be constructed by iteration starting from an arbitrary point.

The contraction mapping theorem can be used to show that if the nonlinearity is sufficiently weak then the solution of the steady-state equation (8) for \( \Phi_a \), holding all higher-order moments fixed, is unique. Take two arbitrary vectors \( x_1 \) and \( x_2 \) in an \( N \)-dimensional Euclidean space and consider

\[
\| F x_1 - F x_2 \| \leq \| K \| \cdot \| R(x_1 + \hat{u}) - R(x_2 + \hat{u}) \|
\]

\[
\leq ( \| K \| \cdot \max_{x} R'(x) ) \| x_1 - x_2 \| ,
\]

where

\[
R'(x) = \frac{d}{dx} R(x).
\]

If the above quantity in parenthesis is less than one then the map contracts and by the contraction mapping theorem there is a unique steady-state solution. For the sigmoid \( R(x) \) shown in Fig. 20, the width of the transition region \( W \) from the minimum to maximum firing rate is at least as great as the width defined by the steepest tangent

\[
W \geq R_{\text{M}} / \max_{x} R'(x),
\]

where \( R_{\text{M}} \) is the maximum firing rate. Hence the contraction condition is

\[
1 \geq \| K \| \cdot \max_{x} R'(x) \geq \| K \| \frac{R_{\text{M}}}{W} ,
\]
Thus, if no single connection is capable of moving any neuron across the transition region (weak coupling condition), then there is a unique steady-state solution.

The graphical method used by White (1961) and by Wilson and Cowan (1968) allows the steady-state solutions for two interacting neurons to be visualized. The steady-state equation (8) for $\hat{r}_\alpha$ can be inverted

$$\sum_b K_{ab} \hat{r}_b + \hat{u}_\alpha = R_{\alpha}^{-1}(\hat{r}_\alpha).$$

For two neurons (Fig. 21) the equations are

$$K_{12} \hat{r}_2 = R_{1}^{-1}(\hat{r}_1) - K_{11} \hat{r}_1 - \hat{u}_1,$$

$$K_{21} \hat{r}_1 = R_{2}^{-1}(\hat{r}_2) - K_{22} \hat{r}_2 - \hat{u}_2,$$

which are of the form $\hat{r}_2(\hat{r}_1)$ and $\hat{r}_1(\hat{r}_2)$. The intersection of these two equations is a graphical solution, as illustrated in Fig. 22.

When the transition regions of $R_{\alpha}^{-1}(\hat{r}_\alpha)$ are sufficiently wide (weak coupling), only one solution exists, as proven above. When the transition regions are narrow (strong coupling), multiple solutions can exist (Fig. 22) and hysteresis between these solutions may occur as the inputs vary (Fig. 23). Bifurcations to new solution branches can occur even when the inputs are held fixed and
higher-order moments, such as the variances of $\phi_a$, which contribute to the transition widths of $R_a(\hat{\phi}_a)$, are allowed to vary parametrically (Sejnowski, 1976a). At bifurcation points, where new solutions appear or multiple solutions merge, the behavior of the system is probably sensitive to the higher-order moments of $\phi_a$, which were assumed to be constant in the analysis given here.

The stability of a steady-state solution can be studied by examining small perturbations around the solution. The results of the perturbation analysis given below are used in the next section to study feature detection.

Let $\hat{\phi}_a(x)$ be a family of solutions to the steady-state equation (8) for inputs $\hat{u}_a(x)$ smoothly parameterized by a real variable $x$. From the Taylor series expansions

$$\hat{\phi}_a(x) = \hat{\phi}_a(0) + x \hat{\phi}'_a(0) + \frac{x^2}{2} \hat{\phi}''_a(0) + \ldots,$$

$$\hat{u}_a(x) = \hat{u}_a(0) + x \hat{u}'_a(0) + \frac{x^2}{2} \hat{u}''_a(0) + \ldots,$$

the first-order variations $\hat{\phi}'_a$ of the steady-state equation (8) satisfy

$$\hat{\phi}'_a = \hat{u}'_a + \sum_b K'_{ab} \hat{\phi}'_b$$

(12)

where

$$K'_{ab} = K_{ab} R'_b$$

(13)

$$R'_b = \frac{\partial R_b(\hat{\phi}_b)}{\partial \hat{\phi}_b}$$
Fig. 21. Schematic diagram of the general two-neuron model with inputs $\hat{u}_1$ and $\hat{u}_2$, reciprocal couplings $K_{12}$ and $K_{21}$, and recurrent collaterals $K_{11}$ and $K_{22}$. 
Fig. 22. Solutions of the two-neuron model are given by the intersections of $\hat{r}_1(\hat{r}_2)$ (broken line) with $\hat{r}_2(\hat{r}_1)$ (solid lines), shown here for the case of reciprocal excitation ($K_{12} > 0, K_{21} > 0$) without recurrent collaterals ($K_{11} = K_{22} = 0$). A variety of steady-state solutions is exhibited by the four different values of the input chosen (from Sejnowski, 1976a).
Fig. 23. Solutions of the two-neuron model are given by the intersections of the two curves, as in Fig. 22 but with \( K_{\alpha \nu} > 0 \). Of the three solutions shown, the center one is unstable and the outer two are stable. As the input \( \hat{U}_I \) is varied (translating the vertical curve vertically) the stable solution on one branch may disappear and the state may jump (arrow) to the other solution branch. The closed circuit describes a hysteresis curve.
Because of the factor $R_b'$, only those neurons whose firing rates are very sensitive to changes in their membrane potentials, such as those neurons near threshold, contribute significantly to $K_{ab}'$, as shown in Fig. 20. The solution for the first-order variation, if it exists, is given by

$$\hat{\Phi}_a' = \sum_b R_{ab} \hat{u}_b'$$

(14)

where

$$R_{ab} = (\delta_{ab} - K_{ab}')^{-1}$$

(15)

This resolvent exists if and only if none of the eigenvalues $\lambda_n$ of $K_{ab}'$ is equal to one. The Neumann series for the resolvent is

$$R_{ab} = \delta_{ab} + K_{ab}' + \sum_c K_{ac}' K_{cb}' + \ldots$$

(16)

which converges if and only if $|\lambda_n| < 1$ for all the eigenvalues. Each term in the Neumann series represents a chain of neurons, the n-th term corresponding to a pathway with n links. The pathways which contribute most significantly to $R_{ab}$ are the ones through neurons whose average membrane potentials are near threshold.

The stability of the steady-state solution is determined by the linearized time-dependent equation

$$\tau \frac{d}{dt} \hat{\Phi}_a' + \sum_b (\delta_{ab} - K_{ab}') \hat{\Phi}_b' = \hat{u}_a'(t)$$

(17)

which is asymptotically stable if and only if $Re \lambda_n < 1$ for all the eigenvalues. Bifurcations to new solution branches (Berger, 1977) can occur only when $Re \lambda_n = 1$. 
The second-order variations from steady state satisfy
\[ \hat{\phi}'' = \left( \hat{\mathcal{U}}'' + \sum_b K''_{ab} \hat{\phi}'_b^2 \right) + \sum_b K'_{ab} \hat{\phi}''_b, \] (18)
where
\[ K''_{ab} = K_{ab} R''_b \] (19)
\[ R''_b = \frac{\partial^2 R_b(\hat{\phi}_b)}{\partial \hat{\phi}_b^2}, \]
the solution of which is
\[ \hat{\phi}''_a = \sum_b R_{ab} \left( \hat{\mathcal{U}}'' + \sum_c K''_{bc} \hat{\phi}'_c^2 \right). \] (20)

Feature Detection

Vision is the best-studied vertebrate sensory system. The transformation of the visual field from the retina to cerebral cortex has been extensively studied with extracellular recordings from single neurons and with increasingly sophisticated anatomical techniques (Cooke and Lipkin, 1972; Hubel and Wiesel, 1977).

It was not at first clear that significant progress could be made in understanding the visual system by recording from one neuron at a time. The key to success was, first, the use of simple visual stimuli which mimic those occurring under natural conditions, and second, the search for the most effective stimulus for each cell. The "trigger feature" of a neuron is defined as the stimulus
which produces the highest average firing rate (Barlow and Levick, 1965). Many surprising and remarkable results have been found by applying this approach to each stage of visual processing.

In the vertebrate retina there are already several layers of visual processing between the photoreceptor mosaic, which transduces photons into electrical signals, and the ganglion cells, whose axons compose the optic nerve. A ganglion cell responds to light over a restricted region of the visual field called its receptive field (Adrian, 1928). In cats and monkeys the receptive fields of ganglion cells are circular and the cells respond best to spots of light. In the retinas of other vertebrates, such as frogs and rabbits, some "trigger features" are more complex, such as edges or movement in a particular direction.

A more radical transformation of the visual field occurs in visual cortex (Hubel and Wiesel, 1962). Neurons there are only weakly influenced by diffuse illumination and spots of light. However, a vigorous response can usually be elicited by using such stimuli as edges or slits of light in a particular orientation, and perhaps moving in a particular direction. Important parameters which distinguish the "trigger features" in visual cortex include position, size, orientation, direction of movement, ocularity, binocular disparity, and color contrast. Moreover, at least two of
these parameters, orientation and occularity, vary in a regular pattern over the cortical surface (Hubel and Wiesel, 1977).

The "trigger features" become more complicated at each successive stage of visual processing, but in another respect the response becomes less specific: the size of the receptive field increases so that the position of the "trigger feature" becomes less important. In the inferior temporal cortex, an area of the cerebral cortex which is important for visual learning, some neurons have extremely specific "trigger features" together with receptive fields that cover much of the visual field (Gross, Rocha-Miranda, & Bender, 1972). These findings have intriguing implications for visual perception (Barlow, 1972), but the sole concern here is with the origin of the observed specificity and invariance in the response of single neurons.

Although the concept of a "trigger feature" has been useful in organizing experimental data, some of its usefulness may be due more to the organizing ability of the experimenter than to the design of the nervous system. In the visual system, for example, an infinite number of spatial and temporal patterns of light could be used as stimuli, but in practice the response of a neuron is tested with only a small number of stimuli chosen by tradition. At most, an experimenter can claim that a neuron responded best
to a "trigger feature" from a restricted class of possible stimuli. Thus, the response may be maximum on only a subspace of all possible variations around the "trigger feature", and even if the maximum were a true local maximum, it might not be a global maximum, which requires that all possible stimuli be tested.

In practice the "trigger feature" of a neuron is found by varying a few parameters of the stimulus to optimize the neuron's firing rate. If the response of a neuron is at a local maximum then small variations around the optimal stimulus should produce no variation in the average firing rate to first order and only negative variations to second order. These variational conditions will be applied to the nonlinear model.

Define the local "trigger feature" as the sensory input \( \hat{U}_b(0) \) for which the response of a neuron is maximum

\[
\hat{\phi}_\alpha(0) \geq \hat{\phi}_\alpha(x),
\]

where the real variable \( x \) parameterizes a one-dimensional subset of stimuli in the space of all possible stimuli. The average membrane potential, and hence the average firing rate, is at an extremum if and only if the first-order variations vanish \( \hat{\phi}_\alpha'(0) = 0 \). From Eq. (14) this extremum condition implies that

\[
R_{ab} = 0 \quad b \in \mathcal{B}_\alpha,
\]
where \( \mathcal{B}_a \) is the subset of inputs in the receptive field of the \( a \)-th neuron.

The type of extremum depends on the sign of the second-order variation. From Eq (20) there is a maximum if and only if

\[
\phi''_a = \sum_b R_{ab} (\hat{u}''_b + \sum_c K''_{bc} \hat{\phi}'_c) \leq 0 ,
\]

where \( \hat{u}'_b \) and \( \hat{u}''_b \) are allowed to vary arbitrarily in the receptive field of the \( a \)-th neuron, \( b \in \mathcal{B}_a \). The first term on the right side vanishes by virtue of the condition on first-order variations (22). The second term on the right depends only on \( \hat{u}'_c \) and only indirectly through at least one neuron outside the receptive field \( \mathcal{B}_a \). From Eq. (14) the second-order condition can be written as a quadratic form

\[
\sum_k \sum_l F_{kl} \hat{u}'_k \hat{u}'_l \leq 0,
\]

where

\[
F_{kl} = \sum_b \sum_c R_{ab} K''_{bc} R_{ck} R_{cl} .
\]

Local maxima are not possible in linear models since

\[
R''_a = 0 \Rightarrow K''_{ab} = 0 \Rightarrow F_{kl} = 0 .
\]

In terms of the eigenvectors \( \alpha^{(b)} \) and the eigenvalues \( \nu_b \) of \( F_{kl} \), the second-order condition can be written

\[
\sum_b \nu_b \left( \sum_a \hat{u}'_a \alpha^{(b)}_a \right)^2 \leq 0
\]

for all \( \hat{u}'_a \) in the receptive field, from which it follows that

\[
\sum_a \hat{u}'_a \alpha^{(b)}_a \neq 0 \Rightarrow \nu_b \leq 0 .
\]
Hence, the spectrum of \( F_{kl} \) must be nonpositive in the receptive field for there to be a local maximum. The eigenvectors with the largest negative eigenvalues are those directions along which the response of the neuron is most sensitive to variations from the "trigger feature." For example, the orientation of an edge parameterizes such a direction for most neurons in visual cortex. Similarly, the eigenvectors with eigenvalues near zero give the directions which leave the response of the neuron invariant.

The existing experimental techniques for tracing functional couplings between neurons are difficult to use and are often ambiguous (Gerstein, 1970). Even in the best-studied parts of the brain, such as the cerebellum where the basic wiring diagram is already known (Eccles, Ito, and Szentágothai, 1967), progress in constructing realistic models is slow (Pellionisz, Llinás, and Perkel, 1977). Simple models have been put forward where less structural information is known, such as the Hubel and Wiesel hierarchical model (1962) of cell types in visual cortex. Although such oversimplified models are unlikely to survive, they at least summarize the data and often stimulate further advances, such as the recent evidence for parallel visual processing in response to the hierarchical model (Dow, 1976). The variational analysis of "trigger features" given above might be useful in analyzing these
hypothetical models.

There is an alternative to modeling incomplete structural data when the function of a brain area is known or partially understood: if a computationally successful model can be found for that function, then connections between neurons might be predicted from the model. Such a testable model is examined in the next section.

Stereopsis

The perception of depth is of great biological importance for many animals, so it is not too surprising that several independent mechanisms can be found in the human visual system for computing depth (Gulick and Lawson, 1976). Artists have discovered many monocular cues to distance, such as perspective, shadowing, and variations in texture. Another monocular depth cue, discovered by Leonardo, is movement perspective: as an observer moves his head, nearby objects pass more rapidly across the retina than more distant objects.

In binocular depth perception, or stereopsis, depth is computed from the parallax between the two eyes. Stereopsis is independent of monocular cues, as shown by Julesz (1960) who constructed the first random-dot stereogram. In a random-dot stereogram each eye sees a random field of dots,
but the dots shown to the right eye are horizontally shifted with respect to the dots shown to the left eye: the amount of shift, or binocular disparity, is interpreted by the visual system as depth variation in a single "cyclopean" field of dots. When fused, the central dots in the random-dot stereogram shown in Fig. 24 appear to hover above the plane of the paper even though each eye only sees a slightly different pattern of random dots.

Since 1958 a machine has been commercially available, the Wild-Raytheon B8 stereomat automated plotter, which can draw a contour map from two overlapping aerial photographs. The machine correlates scan paths from the two images and attempts to assign a depth to each point. Unfortunately, serious problems occur owing to local minima, and even special techniques do not completely eliminate false assignments (Mori, Kidode, & Asada, 1973).

Viewing the random-dot stereogram in Fig. 24, most people, after a delay, experience a sharp transition from the disordered dot pattern to an ordered image. Careful experiments show that fusion of the image exhibits hysteresis: although the binocular disparity must be less than about 6' in humans for fusion to occur, after fusion the disparity can be increased to nearly 2° before the perceived image is lost (Fender and Julesz, 1967). Ambiguous random-dot stereograms have been constructed which
Fig. 24. When this random-dot stereogram is stereoscopically fused (by crossing or everting the eyes or by using a prism over one eye), a square is seen hovering over the random background (from Julesz, 1971).
can be perceived as two different surfaces; once locked onto one image, most people have difficulty shifting their perception to the alternative image (Julesz and Johnson, 1968).

Julesz (1974) has emphasized that the properties of stereopsis — such as a sharp transition to an ordered state, hysteresis, and multiple stable states — are characteristic of cooperative phenomena. The term "cooperative" refers to the way in which local operations nonlinearly cooperate in forming a global order; examples can be found in many physical and biological systems (Haken, 1974). Julesz (1971, 1976) has proposed a metaphorical model of stereopsis composed of springs and magnets which qualitatively reproduces his stereopsis data and which exhibits all the general characteristics of a cooperative phenomenon.

Another more biologically-oriented model has emerged through the combined efforts of several workers (Julesz, 1960, 1963; Sperling, 1970; Dev, 1975; Nelson, 1975; Marr and Poggio, 1976). Given binocular images as input, the model attempts to find the best global three-dimensional reconstruction by performing many local computations in parallel. The input to the model is a pair of random-dot fields where each point in a discrete field \((x, y)\) is either a one (white) or a zero (black), as shown in Fig. 24.
Fig. 25. Depth localization by the identification of corresponding dots in the right and left eyes. Each eye views four dots from which sixteen correspondences are possible. There are six false depth planes above and below the true depth plane (solid squares). The depth plane can be moved up or down by increasing or decreasing the relative shift of the dot positions, called the disparity, between the two retinas (from Julesz, 1971).
The first step of the model is the computation of a three-dimensional disparity matrix, \( D_d(x,y) \), composed of two-dimensional layers parameterized by \( d \). Each layer is the pointwise product of the left visual field and the right visual field after a horizontal shift of \( d \) positions.

\[
D_d(x,y) = R(x+d,y) \cdot L(x,y)
\]

The geometric relationship between horizontal shift and disparity is illustrated in Fig. 25.

Within the three-dimensional disparity matrix, which represents an internal representation of space in the field of view, a two-dimensional surface must be found representing the boundaries of objects in that space. The model proposed by Dev (1975) and Nelson (1975) for performing this computation consists of stacked two-dimensional layers, similar to the disparity matrix, with excitatory connections between nearby units within the same layer, and inhibitory interconnections between units in nearby layers, as schematically shown in Fig. 26. Marr and Poggio (1976) implemented this computational scheme with an iterative algorithm:

\[
C^{(n+1)}_d(x,y) = \sigma \left\{ \sum_{y'} C^{(n)}_d(x,y') - \epsilon \sum_{y'} C^{(n)}_d(x',y') + D_d(x,y) \right\}, \tag{25}
\]

where \( \sigma \) is a sigmoid function, \( D_d(x,y) \) is the input disparity matrix defined above, \( C^{(n)}_d(x,y) \) is the state of
Fig. 26. a) One-dimensional images in the right retina $R_x$ and left retina $L_x$ are plotted as a two-dimensional graph. Lines of constant disparity run diagonally (dashed lines). b) The expanded-scale inset shows the local computation scheme at a single point on the graph: nearby points of the same disparity (dashed line) have excitatory (+) interconnections, while nearby points with different disparity (solid lines) have inhibitory (-) interconnections (from Marr and Poggio, 1976).
the unit at position \((x, \gamma)\) and disparity \(d\) at iteration \(n\), \(\xi\) is the inhibition constant, and \(S'\) and \(O'\) are the subsets of \((d', x', \gamma')\) near \((d, x, \gamma)\) represented respectively by the dashed \((d=d')\) and thick \((d\neq d')\) neighborhoods appearing in the inset of Fig. 26.

Because of the excitatory coupling within each layer, neighboring units will "cooperate" with each other. However, because of inhibitory "competition" between layers, the dominant layer will suppress neighboring layers. Thus, the disparity algorithm attempts to convert the raw disparity input into a consistent depth assignment — defined as the maximally excited layer — at each point.

Julesz (1960) was the first to use disparity layers in a model of stereopsis, but he suggested taking pointwise differences rather than pointwise products in the computation of the disparity matrix. The threshold of the sigmoid in the disparity algorithm (25) suppresses weak background "noise", a technique which Julesz (1963) exploited in an early model. Dev (1975) and Nelson (1975) contributed the excitatory and inhibitory computational scheme. Marr and Poggio (1976) were the first to introduce the nonlinearity required for successful convergence, as demonstrated in Fig. 27.

Marr and Poggio did not give a physiological interpretation for their algorithm, although implicit in
Fig. 27. Starting with the random-dot stereogram shown at the top, the disparity algorithm of Marr and Poggio converges to the image perceived by humans: a rectangle standing in front of the page. The size of each dot is proportional to the disparity or depth assigned to each point (from Marr and Poggio, 1977).
their work and in that of their predecessors is the assumption that "units" could represent neurons. Several important questions remain open: Are the terms in the disparity algorithm related to directly measurable physical quantities or some population average? How is the discrete-time computation related to the continuous-time computation in the nervous system?

There is a close relationship between the disparity algorithm (25) and the physiologically-motivated model being studied here. Let us identify solutions of the algorithm with average firing rates

$$\hat{R}_d(x,y) = \lim_{n \to \infty} C^{(n)}_d(x,y).$$

Then a fixed-point solution of the disparity algorithm can be written

$$\hat{R}_d(x,y) = R\left(\sum_{d',x',y'} K_{d,x,y}^{d',x',y'} \hat{R}_d'(x',y') + D_d(x,y)\right).$$

This equation is identical to the steady-state equation (11) for the average firing rates, with the subscript $\alpha$ indexing $(d,x,y)$, $D_d(x,y)$ identified with the inputs $\hat{U}_\alpha$, the sigmoid $\sigma$ identified with the average firing rate $R_\alpha(\hat{\phi}_\alpha)$ as a function of the average membrane potential (Fig. 20), and the excitatory and inhibitory geometry identified with the coupling strengths $K_{d,x,y}^{d',x',y'}$.

If the "units" in the disparity algorithm (25) are identified with neurons and the state of a unit is
identified with a neuron's average firing rate, then the model of stereopsis predicts that (i) disparity-sensitive neurons should be organized in layers or "columns" of constant disparity, (ii) neighboring neurons within the same disparity layer should have excitatory interconnections, and (iii) neighboring neurons in different disparity layers should have inhibitory interconnections.

Three consequences follow from identifying the neuronal network model with the model of stereopsis: first, the mathematical analysis of the steady-state equation given here applies equally well to fixed-point solutions of the disparity algorithm; second, the nonphysiological discrete-time algorithm can be naturally generalized to a physiological continuous-time model; and third, testable predictions can be made about the computation of stereopsis in the vertebrate nervous system. Each of these consequences will be discussed in turn.

The properties of the steady-state equation are precisely those that are characteristic of cooperative phenomena. A proof is given in the above section on steady-state solutions that the equation always has at least one solution and a sufficient condition is given for there to exist a unique solution. A simple model is analyzed which demonstrates multiple stable states and hysteresis between them.
Although the discrete-time disparity algorithm works well on static random-dot stereograms, the algorithm is less successful on dynamic random-dot stereograms, which humans perceive as moving surfaces (Julesz, personal communication). The problem probably arises because the algorithm does not take into account temporal integration.

A physiologically natural continuous-time generalization of the discrete-time algorithm is suggested by the present model: the steady-state equation (11), which is identical with fixed-point solutions of the disparity algorithm, is a special case of the time-dependent equation (6) for the average membrane potentials

$$\tau \frac{d}{dt} \hat{\phi}_a + \hat{\phi}_a = \sum_b K_{ab} R_b(\hat{\phi}_b) + \hat{U}_a.$$ 

This time-dependent model has the same inputs and coupling strengths as the steady-state model, but the average membrane potential rather than the average firing rate is the primary variable and temporal integration is provided by the term with the membrane time constant $\tau$. The average firing rate is related to the average membrane potential by $\hat{r}_a = R_a(\hat{\phi}_a)$, as shown in Fig. 20. Although these are equivalent variables, the physiologically natural continuous-time generalization of the disparity algorithm favors the average membrane potential as the primary variable.

Information from the two eyes converges for the first
time in visual cortex area 17, and single neurons sensitive to binocular disparity have been found in area 17 of the cat (Bishop, 1970) and areas 17 and 18 in the monkey (Hubel and Wiesel, 1970; Poggio and Fischer, 1977). The model of stereopsis with the physiological interpretation given here is experimentally testable. The functional couplings between disparity-sensitive neurons could be measured by recording simultaneously from pairs of neighboring neurons and cross-correlating their spike trains (Gerstein, 1970).

Disparity-sensitive neurons have thus far been studied with spots and slits of light; in view of Julesz's work and the evidence for cooperativity, more complex spatial patterns, including random-dot stereograms, might be valuable stimuli to use while recording from disparity-sensitive areas of visual cortex. Random-dot stereopsis may not occur in an anesthetized monkey, but monkeys can be trained to perceive depth in a random-dot stereogram (Bough, 1970) and recordings can be made from awake monkey trained to fixate on the stereogram (Poggio and Fischer, 1977).
III. Correlations Between Membrane Potentials

Most sensory neurons in the vertebrate nervous system maintain continual impulse firing in the absence of sensory stimulation, a feature which is usually called spontaneous activity (Granit, 1955). In the visual system, for example, retinal ganglion cells in complete darkness have spontaneous activity which may increase, decrease, or remain unchanged when the retina is exposed to a steady background level of illumination (Kuffler, FitzHugh, & Barlow, 1957). Numerous investigators have studied spontaneous activity and have attempted to determine its origin (Levick, 1973). Some investigators regard spontaneous activity as inherent noise in the nervous system; others view it as a carrier which allows for inhibitory as well as excitatory modulation. Regardless of speculation, spontaneous activity is clearly an important phenomenon and may provide a clue to the operation of the nervous system.

Information processing by average firing rates was discussed in the previous part. Further information could be coded in the temporal pattern of the impulses, such as in spike-train correlations (Perkel and Bullock, 1968; Gerstein, 1970). However, an impulse-producing neuron is not an ideal device for processing temporally-coded information because this type of neuron is maximally
sensitive to the timing of an incoming impulse only when its membrane potential is just below threshold. One might therefore expect that a neuron which processes correlations would operate near threshold most of the time and, as a consequence, maintain a moderate rate of firing. Thus, spontaneous activity in a neuron may reflect a state of sensitivity to temporally-coded information and a readiness to respond to input correlations.

Correlations

As an experimental technique, autocorrelation can be used to detect serial dependence in a single spike train, such as a tendency for spikes to occur at preferred intervals, and cross-correlation between a pair of spike trains can be used to detect mutual dependence, such as interaction or a common influence (Moore, Perkel, & Segundo, 1966; Gerstein, 1970). The correlation is called a second-order property because it depends on two events, in contrast to the mean which is a first-order property.

Let $f(t)$ and $g(t)$ represent two simultaneously measured stationary signals, such as intracellular recordings from a pair of neurons or extracellular recordings of a pair of spike trains. The time-average correlation between $f(t)$ and $g(t)$ over time $T$ is
The correlation depends on a delay parameter $u$ which is the shift between the two signals in the product. If $f = g$ then $C_T(u)$ is called the autocorrelation and $C_T(0)$ is called the variance.

Now consider an ensemble of simultaneously measured pairs of signals, $\{f_a(t)\}$ and $\{g_a(s)\}$, from $N$ independent trials under uniform conditions. The ensemble-average correlation is defined as

$$C_N(t, s) = \frac{1}{N} \sum_{a=1}^{N} f_a(t) g_a(s).$$

In general the ensemble-average correlation depends on both times $t$ and $s$, but if the signals are stationary then $C_N(t, s)$ depends only on the difference $u = s - t$.

Under certain ergodic conditions (Halmos, 1956) the appropriate limits of the two types of correlation are equal:

$$\lim_{T \to \infty} C_T(u) = \lim_{N \to \infty} C_N(u).$$

The correlation depends on the means which are first-order properties. By subtracting off the dependence of the means, a normalized correlation can be defined, called the covariance, which depends only on second-order properties:

$$\text{Cov}(f(t), g(s)) = \frac{1}{N} \sum_{a=1}^{N} (f_a(t) - \hat{f}(t))(g_a(s) - \hat{g}(s)).$$
where $\hat{c}(t)$ and $\hat{g}(s)$ are the ensemble-average means defined in the previous part. Under stationary conditions the covariance differs from the correlation only by an additive constant, but for time-varying signals the correlation may be significantly influenced by the time-varying means.

The covariance is a measure of the joint relationship between two events relative to their chance or accidental rate of coincidence. A positive covariance means that the events occur together more often than by chance, and a negative covariance means that the events occur together less often than by chance. If the two events are independent, so that they occur together entirely by chance, then their covariance is zero. Although covariance is a measure of the temporal pattern in a signal, it represents only a small fraction of all the information that could be contained in the detailed timing of each impulse or in the fine structure of the membrane potential variations. However, the presence of covariance in a signal indicates that the system at least has available to it temporally-coded information beyond the temporally-varying mean value of the signal. As summarized below, there is in fact some indirect evidence for the sensory coding and the use of temporally-coded information in the auditory, somatosensory, and visual systems.
In the auditory system, phase information in the phase-locked action potentials from the two ears is used for spatial localization at frequencies below 1 KHz (Jeffress, 1975; Konishi, 1977). Neurons have been found in the auditory system of the owl with a spatial resolution of 5°, which corresponds to a time delay between the owl's ears of less than 10 microseconds (Knudsen, Konishi, & Pettigrew, 1977).

In the somatosensory system information coded as interspike intervals can apparently be used for perceiving the frequency of flutter-vibration (Mountcastle, 1967), and some central neurons fire with preferred interspike intervals even under normal stimulation of the touch receptors (Amassian and Giblin, 1974).

In the visual system, random-dot stereograms, which have a random texture when viewed monocularly, are perceived in depth if the dots are binocularly correlated, either spatially (Julesz, 1971) or temporally (Ross, 1974). Color can be perceived in a temporally varying spot of white light; the apparent color depends on the temporal pattern of the intensity modulation (Festinger, Allyn, and White, 1971).

Simultaneous recordings of spike trains have been obtained from pairs of neurons in the retina (Rodiek, 1967), the lateral geniculate nucleus (Stevens and Gerstein, 1976),
the cerebellum (Bell and Kawasaki, 1972), the auditory cortex (Dickson and Gerstein, 1972), and elsewhere. In many cases the spike trains were significantly correlated. From the present perspective, the question of whether the correlations were produced by direct interaction or a common input is secondary to the question of whether large-scale correlations exist and are related to sensory processing. When the membrane potential of a neuron is below the threshold for producing impulses, correlations in its membrane potential cannot contribute to correlations in its spike train; hence, correlations between membrane potentials should be at least as prominent as correlations measured between spike trains.

Although the ensemble average of the firing rate, called the PST, is a well-established experimental technique, ensemble-average correlation has not yet been exploited. Time-average correlation is restricted to stationary conditions, which does not allow the stimulus to vary with time. More interesting results should be found in nonstationary responses to stimuli for which the ensemble-average correlation is the natural technique. One aim of this part is to examine how the nervous system might utilize correlations in processing sensory information.
Covariance Equation

The covariance between the effective membrane potentials is defined as

$$\text{Cov}(\phi_a(t), \phi_b(s)) = E \left( \phi_a(t) - \hat{\phi}_a(t) \right) \left( \phi_b(s) - \hat{\phi}_b(s) \right),$$  \hspace{1cm} (27)

and by virtue of the nonlinear model in Eq. (2) the covariance satisfies

$$\tau \text{Cov}(\frac{d}{dt} \phi_a, \phi_b) + \text{Cov}(\phi_a, \phi_b) =$$

$$\sum_C K_{ac} \text{Cov}(\rho_c(\phi_c), \phi_b) + \sum_C B_{ac} \text{Cov}(\eta_c, \phi_b).$$  \hspace{1cm} (28)

In the monkey's cerebral cortex there are approximately \(10^{10}\) neurons and each neuron typically receives connections from \(10^4\) others. The large number of neurons and their high degree of connectivity will prove helpful in analyzing these nonlinear equations.

An unexpected simplification occurs if a physically reasonable assumption is made concerning the probability distribution of the membrane potentials (Sejnowski, 1976b). By the central limit theorem the sum of a large number of independent random inputs has, under quite general conditions, a Gaussian distribution. The membrane potential, which depends on an extremely large number of synaptic events along many inputs, might be approximately Gaussian. With this assumption, the nonlinear terms in Eq. (28) significantly simplify by the following:
Theorem. (Bussgang, 1952). If \( X \) and \( Y \) are jointly Gaussian random variables, and \( \rho \) is any function such that \( \rho(Y) \) is a well-defined random variable with finite second-order moments, then

\[
\text{Cov}(x, \rho(y)) = \text{Cov}(x, y) R'(\hat{\gamma}),
\]

where

\[
R'(\hat{\gamma}) = \frac{\partial R(\gamma)}{\partial \gamma},
\]

\[
R(\hat{\gamma}) = E \rho(\gamma).
\]

If the function \( \rho(Y) \) is considered a nonlinear transformation of an input \( Y \), then Bussgang's theorem states that the output covariance \( \text{Cov}(X, \rho(Y)) \) with an arbitrary Gaussian random variable \( X \) is proportional to the input covariance \( \text{Cov}(x, y) \). Bussgang's theorem is useful for estimating the characteristics of a nonlinear system (Gelb, 1974; Haddad, 1975) and has been applied to the nonlinearity in the goldfish retina (Spekreijse, 1969; Spekreijse and Costing, 1970).

Assuming that \( \phi_a(t) \) are Gaussian and applying this theorem to Eq. (28), we find that the differences \( \phi_a(t) - \hat{\phi}_a(t) \) have the same joint distribution as \( \phi'_a(t) \) defined by

\[
\tau \frac{d}{dt} \phi'_a = \sum_b A_{ab} \phi'_b + \sum_b B_{ab} \eta'_b, \tag{29}
\]

where \( \eta'_b(t) \) are Gaussian processes having zero mean and the same covariance as \( \eta_b(t) \), and \( A_{ab} = K'_{ab} - \delta_{ab} \).
where $\delta_{ab}$ is the Kronecker delta and

$$K'_{ab} = K_{ab} R'_{b}(\hat{\phi}_b), \quad (30)$$

with $R'_{b}(\hat{\phi}_b)$ as defined in the above theorem. Equation (29), which determines the covariance of $\phi'_a(t)$ and will be called the covariance equation, resembles the linear model for graded electrical coupling in Eq. (1), with the interaction matrix $K'_{ab}$ playing the role of the linear coupling coefficients. The multiplicative weights $R'_{b}(\hat{\phi}_b)$ appearing in these effective coupling strengths depend on the membrane potentials, as shown in Fig. 20. Those neurons with average membrane potentials near threshold and those connections between such critical neurons contribute most effectively to the covariance equation.

Despite the linear form of the covariance equation, the coupled equations for the means and covariances are, of course, nonlinear. Because only the mean and covariance of a Gaussian process are independent, the equations for higher-order moments contribute no new information. If the membrane potentials are indeed Gaussian, then only a small part of all the detailed timing information in input spike trains is available for processing by the membrane potentials. Hence, membrane potential covariance could be an important form of temporal coding in the central nervous system.

The assumption that the effective membrane potentials
are Gaussian can be tested experimentally and enters at the same level as the assumptions which led to the nonlinear model. Membrane potentials are most likely to be Gaussian in a highly interconnected area such as cerebral cortex. Departures from a Gaussian distribution may occur in a neuron which is dominated by a few inputs. However, even in a neuron with only one input the membrane potential might still be approximately Gaussian if the input spike train were sufficiently random. A more general model is given in the appendix which takes into account the randomness observed in spike trains and from which it follows as a theorem that the membrane potentials are Gaussian.

The covariance equation is identical with the linearized time-dependent model in Eq. (17), as it must be to remain consistent in the limit of small departures from steady-state. However, for Gaussian membrane potentials the covariance equation is exactly linear and, because the probability distribution enters Bussgang's theorem in a well-behaved manner, small departures from normality introduce correspondingly small departures from strict linearity.

Since all the information about the covariance of $\phi_a(t)$, and only covariance information, is contained in $\phi'_a(t)$, this variable will be called the covariance even though, strictly speaking, the covariance is a second-order
Fig. 28. Summary diagram of the variables which enter in the analysis of neuronal interaction. a) The membrane potential $V_a(t)$ includes action potentials as well as graded membrane potentials. b) $\Phi_a(t)$ is the effective membrane potential which would be present if the action potential were absent. c) $\hat{\Phi}_a(t)$ is the mean effective membrane potential. d) $\Phi'_a(t)$ is equivalent to the difference $\Phi_a(t)-\hat{\Phi}_a(t)$ and is quadratically related to the covariance of $\Phi_a(t)$ (from Sejnowski, 1977a).
property. Similarly, a prime will denote the covariance in other variables.

The nonlinear model of neuronal interaction, summarized in Fig. 28, becomes more linear at each successive stage of probabilistic analysis: a) The membrane potential, including the highly nonlinear action potential, is the primary variable. b) A smoother effective membrane potential is introduced which leads to a nonlinear model (2). c) At the next level the equation (5) for the mean effective membrane potential is significantly less nonlinear. d) In the last stage of analysis the covariance equation (29) has a linear form. The last two levels are coupled by the average firing rates (7).

Stationary Case

The equations for the means and covariances of the membrane potentials are implicitly coupled through the variances, which affect $R_a(\hat{\phi}_a)$ in Eq. (6) and $R'_a(\hat{\phi}_a)$ in Eq. (29). If the coupled equations have a stationary solution, then the interaction matrix is constant and the covariance equation can be solved analytically.

In terms of the fundamental homogeneous solutions of Eq. (29), which satisfy

$$\tau \frac{d}{dt} T_{ab}(t) - \sum_c A_{ac} T_{cb}(t) = 0$$

(31)
with
\[ T_{ab}(t) = \frac{1}{t} \delta_{ab}, \] (31)

the solution of the stationary covariance equation is
\[ \phi_a'(t) = \int_{-\infty}^{t} ds \sum_b T_{ab}(t-s) \sum_c B_{bc} \eta_c'(s), \] (32)

where \( T_{ab} \) is called the transition matrix and represents the impulse response.

The solution of the covariance equation represents a multidimensional linear filter (Wiener, 1949), and the covariance equation is the so-called state-variable form (Kalman & Bucy, 1961). Identical equations are used in communication theory to extract signals from noise and in systems theory to model and control physical systems (Bucy & Joseph, 1968; Åström, 1970).

The solution of the covariance equation is particularly simple in a basis where the interaction matrix \( \mathbb{K}' \) is in Jordan form (Firkhoff and Rota, 1969). It is always possible to find a similarity transform
\[ \mathbf{\bar{K}}' = S^{-1} \mathbb{K}' S \]
such that \( \mathbf{\bar{K}}' \) only has entries on the diagonal or just above it. Furthermore, \( \mathbf{\bar{K}}' \) can be written as a direct sum of blocks with an eigenvalue \( \lambda_n \) of \( \mathbb{K}' \) on the diagonal of each block:
\[
\begin{pmatrix}
\lambda_n & 1 & 0 \\
\lambda_n & 1 & \ddots & 0 \\
0 & \ddots & \ddots & \ddots \\
0 & & \ddots & \ddots & \ddots \\
\lambda_n & & & \ddots & \ddots \\
\end{pmatrix}
\]
Since several blocks may have the same eigenvalue, let the \( l \)-th block of \( \lambda_n \) have dimension \( m_{nl} \). Define \( \psi_m^{nl} \) as the \( m \)-th column of the \((n,l)\)-block of the similarity transform \( S \) which transforms the interaction matrix to Jordan form. Similarly, let \( \psi_n^{*nl} \) be the \( m \)-th row of \( S^{-1} \). By virtue of their definition, these two sets of vectors are biorthonormal
\[
(\psi_m^{*nl}, \psi_m^{nl'}) = \delta_{nn'} \delta_{ll'} \delta_{mm'}
\]
where the left side is the inner product of the two vectors. Each set of vectors forms the basis for the linear vector space, though not necessarily an orthonormal one.

The two sets of vectors \( \{\psi_m^{nl}\} \) and \( \{\psi_m^{*nl}\} \) satisfy
\[
(\lambda_n I - K') \psi_m^{nl} = 0
\]
\[
\psi_m^{*nl} (\lambda_n I - K')^{m_{nl} - m + 1} = 0
\]
and are sometimes called generalized eigenvectors (Schmeidler, 1965). The conventional eigenvectors of \( K' \) are the \( \psi_m^{nl} \) with \( m = 1 \), and the conventional eigenvectors of the adjoint of \( K'_{ab} \) are the \( \psi_m^{*nl} \) with \( m = m_{nl} \). Generalized eigenvectors arise because a nonsymmetric matrix cannot always be diagonalized. However, the generic nonsymmetric matrix is diagonalizable and only special cases require off-diagonal elements. An example is given later in this section which illustrates the difference.
In operator notation the transition matrix is

\[ T(t) = \frac{1}{c} e^{At} \]

and takes the form

\[ T(t) = \frac{1}{c} \sum_{n=1}^{\infty} e^{(-\alpha_n - 1) t} \sum_{k=1}^{k-1} \left( \frac{t}{c} \right)^{k-1} \]

\[ \cdot \left\{ \cos(\beta_n t/c) Re E^n_k - \sin(\beta_n t/c) Im E^n_k \right\}, \]  
\[ \text{with} \quad \alpha_n = Re \lambda_n , \quad \beta_n = Im \lambda_n , \]

and

\[ E^n_k = \sum_{l}^{m} \sum_{m-k}^{m+n} \psi_{m-k+l}^{n+l} \otimes \psi_{m}^{n+l} \]  
\[ \text{The transition matrix represents the response of the filter to an impulse input. If the solution is stable \( \alpha_n < 1 \) for all the eigenvalues), then each eigenvalue of \( K' \) contributes a term which, in the general case, is exponentially damped and sinusoidally modulated. The envelope of the k-th term for the eigenvalue \( \lambda_n \), owing to the factor \( t^{k-1} \), has a peak occurring at \((k-1)c/(1-\alpha_n)\) after the impulse. Attached to each \( \lambda_n \) is a sequence of nested subspaces spanned by \( E^n_k \) in Eq. (34) for the transition matrix: \( E^n_k : \mathcal{H}_k \rightarrow \mathcal{H}_k \).}
The solution of the covariance equation to an impulse input can be understood geometrically: the input is projected onto the subspaces $\mathcal{E}_k^n$ and the output from each subspace unfolds in time; each term of $T_{ab}(t)$ in Eq. (34) represents a spiral (first outgoing, then ingoing) in the plane spanned by the real and imaginary parts of $\psi_{m}^{nl}$; for example, if the input were $\psi_{m}^{nl} \delta(t)$, then the output would follow the sequence $\psi_{m}^{nl}, \psi_{m-1}^{nl}, \ldots \psi_{1}^{nl}$.

A deeper insight into the analytic structure of the covariance equation can be gained by examining the Laplace transform of the transition matrix in Eq. (31)

$$\mathcal{L}[T](s) = (c s I - A)^{-1}$$

which has the form of a resolvent. In terms of Laplace transforms, the solution of the covariance equation is

$$\mathcal{L}[\phi_a'] = \sum_b \mathcal{L}[T_{ab}] \sum_c b_{bc} \mathcal{L}[\eta_c'] .$$

Thus, the Laplace-transformed output $\mathcal{L}[\phi_a']$ is just a linear algebraic transformation of the Laplace-transformed input $\mathcal{L}[\eta_c']$ with an interaction part $\mathcal{L}[T_{ab}]$ and an input branching part $b_{bc}$.

An explicit form of the interaction part is given by the Laplace transform of Eq. (34):

$$\mathcal{L}[T](s) = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \frac{E_{nk}}{(c s - \lambda_n + 1)^k} .$$

Each pole of the resolvent has associated with it an
invariant subspace \( \mathcal{F}_K^n \) as defined above. An input vector within \( \mathcal{F}_K^n \) is transformed by \( E_K^n \) into an output vector within the same subspace. As the order of the pole increases, which is only possible when there are off-diagonal elements in the Jordan form of \( K' \), the dimension of the invariant subspace attached to the pole decreases in a nested fashion

\[
\mathcal{F}_K^{n+1} \subset \mathcal{F}_K^n .
\]

The Laplace transform of the transition matrix in Eq. (37) is an analytic function of small perturbations to \( A \) even though the eigenvectors and eigenfunctions which appear in Eq. (39) are generally not analytic (Kato, 1966). The stability of the covariance equation to small perturbations of the coupling strengths between neurons is examined in the next part.

The damped harmonic oscillator is a simple example which illustrates some of the properties of the general case. The second-order equation

\[
\frac{d^2}{dt^2} x + 2D x + \omega_0^2 x = 0 ,
\]

where \( \omega_0 \) is the frequency of the undamped oscillator and \( D \) is the damping coefficient, can be written as the equivalent first-order equation

\[
\frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & \omega_0 \\ -\omega_0 & -2D \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ,
\]
which has the same form as the two-dimensional covariance equation. The characteristic equation for this problem is

$$\lambda^2 + 2D\lambda + \omega_0^2 = 0,$$

which has the solutions

$$\lambda_\pm = -D \pm \sqrt{D^2 - \omega_0^2}.$$

If \(D > \omega_0\) then the solution is \{overdamped, critically damped, underdamped\}.

The underdamped solution is

$$x(t) = e^{-Dt}(A\cos \omega_1 t + B\sin \omega_1 t),$$

where

$$\omega_1^2 = \omega_0^2 - D^2,$$

which can be compared to Eq. (34) with \(D = (1 - \alpha_n)/\tau\) and \(\omega_1 = \beta_n/\tau\). Only in the critically damped case can the interaction matrix fail to diagonalize, but this circumstance only occurs in two and three dimensions. For there to exist two equal complex eigenvalues (which would allow an underdamped off-diagonal Jordan form) requires at least four dimensions because complex eigenvalues always come in conjugate pairs.

The "quality factor" or \(Q\) of a damped oscillation is a measure of the sharpness of its resonance. For the oscillator represented by Eq. (37) the \(Q\) is defined as

$$Q = \frac{1}{2} \frac{\omega_1}{D}.$$
If the oscillator is excited by a sudden displacement, then the number of cycles during decay of the envelope to $1/e$ of its original amplitude is approximately $Q/\pi$. Let us generalize the $Q$ to the damped oscillations represented in Eq. (34) by assigning to each eigenvalue

$$Q_n = \frac{1}{2} \frac{1/|\beta_n|}{(1 - \alpha_n)}.$$  

(40)

Overdamped modes arise from eigenvalues which lie on the real axis ($Q_n = 0$).

The damping time constant for a mode is

$$\tau_n = \tau/(1 - \alpha_n).$$  

(41)

For there to occur underdamped oscillations lasting longer than $N$ seconds with a typical membrane time constant $\tau \sim 10$ msec requires that $(1 - \alpha_n) < 0.01/N$. Hence, only those eigenvalues whose real parts are nearly equal to one contribute to correlations on a long time scale.

Although no experimentally determined interaction matrix is available in the vertebrate nervous system, that some modes in the brain are underdamped might be inferred from the widespread rhythms observed in EEG recordings over human cerebral cortex (Morrell, 1967) and from recordings of intracellular membrane potentials in cortical neurons (Elul, 1968; Offenbach, 1975). With a typical membrane time constant of $\tau \sim 10$ msec, the alpha rhythm at $\nu_\alpha \sim 10$/sec corresponds to

$$|\beta_n| \sim \left(\frac{2\pi}{\nu_\alpha}\right) \tau \sim 0.63.$$
Other EEG rhythms occur between 5/sec and 40/sec.

**Neuronal Filters and Memory**

In the stationary case treated above, the membrane potential covariances are particularly sensitive to input covariances with the spatial and temporal patterns associated with each eigenvalue: these will be called the covariance modes of the neuronal filter in analogy with the normal modes of mechanical systems. However, the normal modes of an undamped mechanical system are derived from a symmetric matrix and are always orthogonal, but the interaction matrix is generally not symmetric and its eigenvectors are generally not orthogonal. As a consequence, coupled covariance modes can appear which emerge and decay in temporal sequence.

Because the interaction matrix in the covariance equation depends on \( R_b'(\hat{\Phi}_b) \), the covariance modes of a neuronal filter are adjustable and depend on the average firing rates, as shown in Fig. 29. Thus, by controlling the background firing rates in a "lower" processing center, a "higher" processing center could control the filtering of incoming information. Such an adjustable filter may underlie the ability of an organism to selectively direct its attention and to extract special features from sensory
Fig. 29. Block diagram representing a neuronal filter: the input covariances between spike trains $\eta_b(t')$, linearly filtered, drive the output covariances between membrane potentials $\phi_a(t')$. The impulse-response $T_{ab}(t-t')$ of the filter depends on the background firing rates $R_a$. 
information.

Some inputs to an area may contribute little to the background firing rates of postsynaptic neurons, but could have a significant effect on membrane potential covariances. An experiment concerned solely with average firing rates might not detect such an input even if it were a major source of information to the area. Cases of major anatomical pathways without corresponding influence as judged by average firing rates are not uncommon. For example, most ascending projections from thalamic nuclei to cerebral cortex are accompanied by reciprocal corticothalamalic tracts (Carpenter, 1976), but the electrophysiological influence of these descending projections to the thalamus is weak compared to the influence of other inputs, as measured by the firing rates of thalamic neurons. As a specific example, in the visual system of the cat the response of cells in the thalamic dorsal lateral geniculate nucleus to spots of light is not affected by cooling visual cortex (Kalil and Chase, 1970). However, for some neurons the cooling had a significant reversible effect on the pattern of impulses in response to moving slits of light, as shown in Fig. 30. If the primary purpose of descending influence to thalamic relay nuclei is to provide such timing information, then cortical feedback could play an important role in covariance processing.
Fig. 30. Average firing rates of two neurons in the cat lateral geniculate nucleus (LGN) in response to a moving slit of light. The LGN receives inputs from both the retina and visual cortex. The part of cortex that projects to the LGN was cooled (thereby abolishing the output from the cooled area) in order to study its effect on the response of the LGN to light. At the start of each histogram the slit began moving from a position 10° to the right of the receptive-field center. After passing to a position 10° to the left of center the slit reversed direction. The top pair of histograms shows the response before and after the cortex was cooled, averaged over 20 trials each. Their similarity indicates that the effect of cooling was reversible. a) The neuron was inhibited when the slit was in the receptive field. The average response of the neuron while the cortex was cooled, shown at the bottom, was significantly different from the normal response: the average background activity was reduced, the response to the slit moving from left to right was increased, and a new burst appeared when the slit moved in the opposite direction. b) The neuron gave a burst of firing when the slit was in the receptive field. Secondary bursts (arrows) appeared when the cortex was cooled (from Kalil and Chase, 1970).
Fig. 30
Remarkably little is known about the physical basis of memory. It is generally believed that short-term memory depends on transient electrical activity, and that long-term memory is the result of relatively permanent structural changes. Early neuroanatomical evidence for feedback loops between neurons (Lorente de Nó, 1938) led to the idea that reverberation and reverberatory circuits were a possible physical basis for short-term memory (Hebb, 1949; Kimble, 1965). Although circulatory chains of neurons do exist, some synapses between neurons are known to be inhibitory; in addition, a single action potential is usually insufficient by itself to cause an action potential to occur in another neuron. Hence, the computer-like circulation of impulses in closed neural circuits is unlikely.

The analysis of correlation processing given in this part shows how repetitive patterns can circulate as a "statistical reverberation" even though individual impulses need not. The information is distributed in a large collection of neurons and is represented as temporal sequences of covariance modes. If some of the brain's covariance modes had a sufficiently long coherence time (a sufficiently high Q), then membrane potential covariance could serve as the physical basis for short-term memory. A neuronal filter has the additional advantage that different covariance modes can be selected by adjusting the background
firing rates of the neurons. In this way the filter characteristics of an area can be selectively altered in order to selectively store a particular type of information. This is the same mechanism proposed for selective attention, but in the case of selective storage a "lower" processing center controls the background firing rates in a "higher" processing center.

The cerebral cortex receives inputs from thalamic nuclei, association fibers from other cortical areas, commisural fibers from the corpus collosum, and a variety of other afferents from the brain stem, basal ganglia, and elsewhere (Carpenter, 1976). The background firing rates in a cortical area could be controlled by several of these inputs, one part arising from general arousal and another part from a specific sensory modality, while other inputs may be primarily concerned with processing covariance. Inputs from areas which provide temporally-coded information should synapse near the spike-initiating region of a neuron to reduce the delay and decrement from electrotonic conduction in dendrites. Inputs from areas which control the background firing rates could as well, or with advantage, synapse on the distal parts of dendrites.

Our understanding of the physical basis for long-term memory is not much better than that of Aristotle, who compared memory with the impressions left in wax. The
advanced technologies of each age have provided metaphors for memory — from hydraulics in the 17th century and telephone switchboards in the 19th century, to digital computers in the 20th century. The most recent metaphor is holography: as in long-term memory, information in a hologram is stored distributively and can be recalled associatively (Julesz and Pennington, 1965; Longuet-Higgins, 1968; Gabor, 1968). In attempting to make this metaphor more precise, some workers have sought to identify parts of the holographic process with physical components of neurons, such as the computation of sinusoidal interference fringes at synapses (Westlake, 1970; Pribram, Nwara, & Baron, 1974). Another group of workers, recognizing that the mathematical basis for the storage and recall of holographic information is correlation storage and linear filtering (Longuet-Higgins, Willshaw, & Buneman, 1970; Pfaffelhuber, 1975), have studied the properties of simplified linear filters (Steinbuch, 1961; Anderson, 1972; Wigström, 1973; Kohonen 1977).

An entirely different starting point — the stochastic modeling of nonlinearly interacting neurons — has led to equations for a linear filter at the level of covariance between membrane potentials. The stochastic model includes additional features, such as the selective storage and retrieval of information, which arise from the essential
nonlinearity of the model.

The gap between the mathematical study of memory in simple models and the experimental study of the physical basis for memory could be bridged by refining existing techniques for intracellular recording from cortical neurons and by measuring ensemble-average membrane potential correlations during controlled sensory stimulation (initially by measuring autocorrelation in single neurons and later by measuring the cross-correlation between neighboring pairs of neurons). A testable theoretical prediction for the physical basis of long-term memory is made in the next part.
IV. Motor Learning in the Cerebellum

Although the mechanisms which underlie learning are unknown, experimental evidence indicates that stored information in the nervous system is distributed in large collections of neurons rather than localized as in the case of digital computers (Rosenzweig and Bennett, 1976). The probabilistic analysis of interacting neurons in Parts II and III suggests that distributed information could be processed and temporarily stored as correlations between membrane potentials. The main purpose of Part IV is to examine how these correlations could be permanently stored and subsequently retrieved (Sejnowski, 1977a).

Cerebellar Cortex

The region of the brain whose local circuits are best known is the cerebellum, a grapefruit-sized structure which lies at the back of the skull, as shown in Fig. 31 (Llinás, 1975). Animals deprived of a cerebellum suffer severe disturbances of coordination and equilibrium. Because the cerebellum receives inputs from both the motor system and the sensory monitor system, there is little doubt that it is intimately involved in motor coordination, but the precise function of the cerebellum in organizing movement is still
Fig. 31. Side views of the human brain with the front of the head to the left. a) The cerebellum lies at the back of the brain beneath the convoluted cerebrum. b) A midline cross-section of the brain stem and cerebellum is shown below (from Llinas, 1975).
unknown. The basic elements of the cerebellum are the same in all vertebrates, indicating that its evolution is highly conservative. The cerebellum has enlarged between threefold and fourfold in the past million years of human evolution — whatever the cerebellum contributes to motor coordination is obviously of great adaptive value. Surprisingly, in some electric fish which can sense very weak electric fields, the cerebellum is enormous, filling most of the skull.

Like the cerebrum, the cerebellum is elaborately folded and has a thin cortical layer of grey matter. Unlike the cerebral cortex, however, cerebellar cortex contains a remarkably regular, repetitive structure. The fundamentals of cerebellar structure which were established by Ramón y Cajal in 1888 have been repeatedly confirmed in a wide range of vertebrates (Llinás, 1969). The cerebellar cortex has seven basic elements: two carry inputs into the cerebellum, the mossy fibers and the climbing fibers, and only one element serves as output, the Purkinje cell axons; the four remaining elements are neurons intrinsic to the cerebellar cortex — the granule cells, the Golgi cells, the stellate cells, and the basket cells. Each of these elements has a distinct geometrical relationship in the cortex, as shown in Fig. 32.

The Purkinje cell is one of the most distinctive and
Fig. 32. Three-dimensional schematic diagram of the basic elements in the cerebellar cortex. The mossy fibers and climbing fibers are the two main inputs and the Purkinje cell axons are the only outputs. The axons of the granule cells rise to the surface of the cortex where they bifurcate to form long parallel fibers. Each Purkinje cell dendritic tree receives many excitatory synapses from passing parallel fibers and multiple excitatory synapses from a single climbing fiber. The stellate cells, the Golgi cells, and the basket cells receive excitatory input from the parallel fibers and have exclusively inhibitory effects on the other neurons in the cortex (from Bullock, 1977).
dramatic neurons in the nervous system. Each Purkinje cell has a large dendritic tree with intricately bifurcating branches that resembles an espaliered pear tree, as shown in Fig. 1. The flat dendritic trees are densely stacked, without contact, perpendicular to the surface of the cortex. Purkinje cells receive two types of excitatory input: in the human cerebellum each Purkinje cell receives approximately 300 synapses from a single climbing fiber and approximately 100,000 synapses from passing parallel fibers. Granule cell axons, called parallel fibers, make synaptic contacts as they run perpendicularly through the plane of the dendritic trees and parallel to the cortical surface. In contrast, a single climbing fiber entwines each dendritic tree like a climbing vine and makes multiple synaptic contacts on every major branch. A schematic view of a patch of Purkinje cell dendrite is shown in Fig. 33.

The granule cell bodies and dendrites form a dense layer immediately beneath the Purkinje cells. The mossy fibers enter the granule cell layer from below and form nests of synaptic contacts with the granule cell dendrites. The remaining interneurons — the basket cells, the stellate cells, and the Golgi cells — are all inhibitory in their action and mediate between the parallel fibers and other cerebellar neurons.

The cerebellar cortex receives inputs, through the
Fig. 33. Schematic illustration of a cerebellar Purkinje cell dendrite Pd (with a dendritic branchlet), a climbing fiber CF (entwining the dendritic trunk), and a parallel fiber pf (passing through the dendritic tree), based on Palay and Chan-Palay (1974). Climbing fiber varicosities make numerous synaptic contacts with spines on the dendritic trunk (from Sejnowski, 1977a).
mossy fibers and climbing fibers, from a bewildering variety of sources: there are several nerve bundles in the spinal cord destined for the cerebellum which carry information about the condition of muscles, tendons, joints, and spinal interneurons; motor neurons in cerebral cortex that send motor commands down the spinal cord have branches which reach the cerebellum through relay nuclei; the vestibular organs, which sense head velocity and contribute to our balance, have direct projections to the cerebellum. Inputs to the cerebellum generally send branches to the cerebellar nuclei which lie beneath the cerebellar cortex, as shown in Fig. 32. These integration centers also receive the output from the Purkinje cells, which is entirely inhibitory. Thus, the cerebellar cortex works in parallel with the cerebellar nuclei.

The regularity of the cerebellar structure has made it an ideal material for electrophysiological recording (Eccles, Ito, and Szentágothai, 1967). The electrical response of each cell type in the cerebellum has been carefully studied and all the synapses between the seven basic elements have been characterized. Purkinje cells generally have a high spontaneous firing rate of 20-100/second, but climbing fibers on average discharge at only 1/second. Even though their average firing rates are low, neighboring climbing fibers sometimes fire
Fig. 34. a) Cross-correlation between two nearby climbing fibers in the cerebellar cortex and b) their autocorrelations. The correlation is given by the solid and dotted columns added together; the solid columns alone represent cross-interval histograms based only on nearest pairs of impulses. Two Purkinje cells were recorded simultaneously on two extracellular microelectrodes and the discharges of their climbing fibers were identified by the characteristic complex-spike pattern. The autocorrelations indicate that most complex spikes are separated by at least 100 ms. The cross-correlation indicates a tendency for the two climbing fibers to discharge either in synchrony or at a 130 ms interval (from Bell and Kawasaki, 1972).
synchronously. Simultaneous recordings from pairs of climbing fibers show significant correlations lasting for several hundred milliseconds, as shown in Fig. 34 (Bell and Kawasaki, 1972). Thus, climbing fibers may be primarily concerned with temporally-coded information.

Among its many functions, the cerebellum is involved in "fine-tuning" the vestibulo-ocular reflex, a compensatory eye movement induced by head rotation (Ito, 1975; Robinson, 1976; Miles, 1977). Because synaptic delays would be greater than the response time of the reflex, no closed-loop feedback from the retina to the vestibular nuclei is possible. However, the cerebellum does receive visual feedback through the climbing fiber system, and the accuracy of the open-loop vestibulo-ocular reflex is improved by visual experience over a time scale of days. Lesion of the vestibulo-cerebellum entirely abolishes this behavioral plasticity. Thus, the cerebellum is required for vestibulo-ocular learning; however, the evidence does not yet prove that the cerebellum is indeed the site of this plasticity.

The one-to-one match between climbing fibers and Purkinje cells and the complete compartmentalization of each Purkinje cell dendritic tree is unique in the nervous system. Brindley (1964) and Szentágothai (1968) have pointed out that a climbing fiber is ideally positioned to
influence all the synapses from parallel fibers to a Purkinje cell, and have suggested that the purpose of the climbing fiber could be to control the plasticity of the parallel fiber synapses. Following their suggestion, Marr (1969) and Albus (1971) proposed detailed theories for how the cerebellum could coordinate motor acts and learn to perform motor skills. The aim here is more limited in scope: rather than propose another all-encompassing theory of the cerebellum, attention is focussed on the Purkinje cell dendritic tree, and the original suggestion made by Brindley and Szentagothai is reformulated within the framework of covariance processing.

Storing Covariance

Highly skilled movements, such as those required to play a musical instrument, are shaped by repetitive practice, suggesting that each repetition of a skilled movement slightly modifies the motor program. In this section motor learning is treated as a gradual perturbation of the coupling strengths between neurons in the cerebellum. A learning algorithm is derived which selectively modifies the impulse firing patterns of the output Purkinje cells.

Let $\phi_a(t)$ be the effective membrane potentials of all the neurons in the cerebellum, including the intrinsic
neurons, and let $\eta_b(t)$ and $\xi_a(t)$ represent the mossy fiber and climbing fiber input firing rates with input coupling strengths $B_{ab}$ and $C_a$ respectively. Since each Purkinje cell receives approximately 100,000 synaptic connections from parallel fibers, we can reasonably assume that the membrane potentials of Purkinje cells are Gaussian. Then, following the conventions in Part II, the covariances satisfy

$$T \frac{d}{dt} \phi'_a = \sum_b A_{ab} \phi'_b + \sum_b B_{ab} \eta'_b + C_a \xi'_a,$$

and the part of the covariances $\psi'_a(t)$ arising from climbing fiber inputs alone satisfy

$$T \frac{d}{dt} \psi'_a = \sum_b A_{ab} \psi'_b + C_a \xi'_a.$$

How should synaptic strengths $K_{ab}$ be altered so that the original membrane potential covariances $\phi'_a(t)$ are augmented by a small amount $\xi'_a(t)$ after learning has occurred? If the synaptic strengths were altered by a small amount $\xi_{K_{ab}}$, then Eqs. (42) and (43) require, to first order in $\xi$ , that

$$C_a \xi'_a(t) = \sum_b K_{ab} R'_b \phi'_b.$$

If the plastic synapses are the ones between parallel fibers and Purkinje cells, then the only contributions to the right side of Eq. (44) are the output covariances from the granule cells. If the membrane potentials of the granule cells have Gaussian distributions, then in terms of
the firing rates of the granule cells
\[ \xi_b(t) = \rho_b(\phi_b(t)), \] (45)
the covariances in the parallel fibers can be written
\[ \xi'_b(t) = R'_b \phi'_b(t). \]
The condition on $K_{ab}$ in Eq. (44) becomes
\[ C_a \xi'_a(t) = \sum_b K_{ab} \xi'_b(t), \] (46)
where $\xi'_a(t)$ and $\xi'_b(t)$ are the input covariances to Purkinje cells arising respectively from the climbing fibers and the parallel fibers. Thus, the climbing-fiber input covariances can be associatively stored by matching them to a linear combination of the parallel-fiber input covariances. However, the alterations in the coupling strengths are fixed modifications, but the right and left sides of Eq. (46) are arbitrary temporal processes. Thus, only an approximate $K_{ab}$ can be found.

In the stationary case, a particularly simple optimal choice of $K_{ab}$ is that which minimizes the mean square error between the right and left sides of Eq. (46) at a fixed time
\[ E(K_{ab}) = E \sum_a (C_a \xi'_a - \sum_b K_{ab} \xi'_b)^2. \] (47)
The minimum, which occurs when the variation of $K_{ab}$ vanishes,
\[ K^*_{ab} = C_a \sum_c [\text{Cov}(\xi'_a, \xi'_c)] [\text{Cov}(\xi'_c, \xi'_b)]^{-1}, \] (48)
is independent of time. In the nonstationary case, a similar optimal solution can be found based on the integrated mean square error (Sejnowski, 1977a).

The plasticity required by Eq. (48) depends on a priori knowledge of the covariances and cannot be practically implemented with neurons. Stochastic approximation is a constructive method for estimating the optimal solution when only raw data are given (Tsypkin, 1973). The adaptive learning algorithm for the present problem is

$$\frac{d}{dt} K_{ab} = \gamma \nabla_{K_{ab}} \mathcal{E}(K_{ab}), \quad (49)$$

where $\gamma$ is a positive constant (or a negative constant if the associative storage is complementary) corresponding to the strength of the plasticity. The result from Eqs. (47) and (49) is

$$\frac{d}{dt} K_{ab} = \gamma \left[ C_a \text{Cov}(E'_a, S'_b) - \sum_c K_{ac} \text{Cov}(S'_c, S'_b) \right]. \quad (50)$$

The present problem differs from most applications of this method in that the nonlinear background as well as the linear system being optimized depend on the coupling strengths. This difficulty will not be dealt with here; the present treatment is valid only for small changes to the synaptic strengths, in which case the second term on the right side of Eq. (50) is much smaller than the first term and can be neglected. Thus, the approximate learning algorithm is

$$\frac{d}{dt} K_{ab} = \gamma C_a \text{Cov}(E'_a, S'_b). \quad (51)$$
Notice that the integrated form of Eq. (51) is proportional to the first factor of the optimal solution $K_{ab}^*$ in Eq. (48) (Pfaffelhuber, 1975). When only raw data are given the learning algorithm is

$$\frac{d}{dt} K_{ab} = \gamma C_a \left[ \bar{E}_a(t) \bar{S}_b(t) - \hat{E}_a(t) \hat{S}_b(t) \right].$$  \hspace{1cm} (52)

The covariance on the right side is the temporal correlation between $\bar{E}_a(t)$ and $\bar{S}_b(t)$ relative to the product of their means — that part of the correlation owing to chance. The learning algorithm requires that a single climbing fiber influence all the modifiable synapses on an entire Purkinje cell. This requirement could be satisfied since a climbing fiber does make multiple synaptic contacts on all the major dendritic branches of a Purkinje cell and is therefore in a position to influence all the parallel fiber synapses.

Because the covariance in the learning algorithm can be either positive or negative, the plastic synapse should be capable of both long-term facilitation and depression. In comparison, Marr's theory (1969) only predicts facilitation when there is a "conjunction of presynaptic and climbing fiber (or post-synaptic) activity"; such a plastic synapse eventually reaches maximum strength through chance coincidences, or else loses information by nonspecific decay. Other theories for cerebellar motor learning (Albus, 1971; Gilbert, 1975) predict only weakening of the plastic
synapses; such a synapse eventually reaches its minimum strength through chance coincidences. The average strength of the synapse predicted here remains constant when the inputs are uncorrelated and can be flexibly adjusted anywhere within its dynamic range when the inputs are appropriately correlated (Sejnowski, 1977b).

**Synaptic Plasticity**

The form of the learning algorithm in Eq. (52) suggests a mechanism for motor learning in the cerebellum based on coincidences and anticoincidences between impulses in the climbing fiber and parallel fiber inputs. Suppose that the synaptic strength of the plastic synapse increases by a small amount $\alpha$ whenever a coincidence occurs between a discharge of the climbing fiber and a presynaptic activation of the plastic synapse by the parallel fiber within a time interval $T_\alpha/2$. Neglecting multiple coincidences, the accidental or chance coincidence rate for independently occurring events is $R_C R_P T_\alpha$, where $R_C$ and $R_P$ are respectively the average firing rate of the climbing fiber and average activation rate of the plastic synapse by the parallel fiber. (The subscripts "c" and "p" will always refer respectively to the climbing fiber and the parallel fiber.)
Chance strengthening must be balanced by a mechanism which weakens the synapse. For example, Stent (1973) has postulated that neurotransmitter receptors are removed from a plastic synapse when the postsynaptic membrane repeatedly discharges in the absence of a presynaptic activation. Stent further suggested that the mechanism might account for developmental plasticity in the cat visual cortex.

Assume that the plastic synapse diminishes in strength by $\beta$ whenever the climbing fiber discharges without an activation of the plastic synapse by the parallel fiber within $T_p/\alpha$. If multiple coincidences within this time interval are negligible, then the condition for balance between strengthening and weakening of the synaptic strength when the inputs are uncorrelated is

$$\beta(1 - R_pT_p) = \alpha R_p T \alpha.$$  

With appropriately correlated inputs the synaptic strength can be adjusted anywhere between its maximum and minimum.

If the balance condition is satisfied, then the plastic synapse is modified whenever the climbing fiber discharges; whether the synapse is strengthened or weakened depends on whether or not it is activated in coincidence with the climbing fiber. Consider the case when $T_p$ is sufficiently small so that $R_p T_p \ll 1$. Then the ratio of weakening to strengthening is, by Eq. (53), proportional to the firing rate of the plastic synapse: $\beta/\alpha \approx R_p T_p$. A
substance which was present in the synaptic cleft in proportion to the average activation rate of the synapse (such as a degradation product of the neurotransmitter) could serve as agent in satisfying this condition for balance.

An alternative way to balance the strengthening and weakening is to have the synaptic strength diminish by $\beta$ whenever the plastic synapse is activated without a discharge of the climbing fiber within $\tau_c/2$. The corresponding balance condition for uncorrelated inputs is

$$\beta (1 - R_c \tau_c) = \alpha R_c \tau_c \alpha.$$

The number of coincidences and anticoincidences is subject to random fluctuation. For stationary independent inputs over a time $T$, the synaptic strength, within its range, will undergo a random walk with standard deviation

$$\sigma(T) = \left[ (\alpha^2 + \alpha \beta) R_c R_p \tau \tau_c T \right]^{1/2}.$$

around its mean strength. Random fluctuation can be minimized by limiting the time during which the plastic synapse is sensitive to modification. A slow and diffuse chemical system — such as the recently discovered noradrenergic innervation of the cerebellar cortex by projections from the locus coeruleus (Olson and Fuxe, 1971; Hoffer, Siggins, Oliver, and Bloom, 1973) — could serve to regulate plasticity and minimize random fluctuation.
The magnitude of random fluctuation should be compared with the maximum possible nonrandom plastic change of the synaptic strength. If every climbing-fiber impulse were to occur in coincidence with an activation of the plastic synapse by the parallel fiber, then the maximum strengthening of the synaptic strength over a time $T$ would be

$$\mu(T) = \alpha R_c T.$$ 

Thus, the ratio of random fluctuation to maximum change over time $T$ is

$$\frac{\sigma(T)}{\mu(T)} = \left(\frac{T^*}{T}\right)^{\frac{1}{2}},$$

where the natural time scale is set by

$$T^* = (1 + \frac{\beta}{\alpha}) \left(\frac{R_p}{R_c}\right) \tau_\alpha.$$ 

Let us apply these general statistical constraints specifically to the cerebellum. Under typical conditions the firing rate of a parallel fiber is $R_p \sim 50$/sec and the firing rate of a climbing fiber is $R_c \sim 1$/sec. If we assume that the coincidence "window" for strengthening is $\tau_\alpha \sim 2$ msec (comparable to the time course of an action potential), then the accidental coincidence rate is $R_c R_p \tau_\alpha \sim 0.1$/sec; that is, about once in every 10 sec. If $\tau_p \ll 20$ msec (so that $R_p \tau_p \ll 1$), then the balance condition (53) becomes $\beta/\alpha \approx 0.1$; that is, the size of a
decrement from a single anticoincidence must be approximately one-tenth the size of an increment from a single coincidence. Under the same steady-state conditions the natural time scale is $T^* = 0.1$ sec, which is the minimum time required for the nonrandom change to equal the random fluctuation. The time scale for learning is typically between 10 sec and 1,000 sec, during which the random fluctuation is between 10% and 1% of the maximum nonrandom change. If a plastic synapse were always sensitive to modification, then the random fluctuation would be a cumulative, long-term noise problem.

If one of the above balance conditions is satisfied, then the average strength of the plastic synapse cannot be altered by simply increasing the firing rate of the climbing fiber or the parallel fiber: the strength of the synapse can be systematically altered only through correlated discharges, and the synapse is then either strengthened or weakened depending on the sign of the covariance. A direct test of the predicted synaptic plasticity in vertebrates is at present a difficult experiment; however, the general statistical constraints considered above apply equally well to invertebrates, and similar plasticity may occur in a more favorable preparation.

An example of the type of plasticity predicted in the cerebellum has been found in Aplysia (Kandel and Tauc, 1965;
Shimahara and Tauc, 1976; Kandel, 1976). The strength of an identified synapse has been shown to facilitate and to remain elevated for many minutes after an identified conditioning neuron has been activated. However, the demonstrated heterosynaptic facilitation is not contingent upon pairing of the synaptic activation with the conditioning stimulus: activation of the conditioning neuron alone produced the same facilitation. In this respect the facilitation demonstrated in *Aplysia* differs from that predicted in the cerebellum, which depends on whether or not the climbing fiber discharge is paired with a parallel fiber activation of the plastic synapse.

**Long-term Memory**

Motor learning in the cerebellum was treated as a gradual adaptation of the cortical filtering properties: input covariance among the climbing fibers was stored in association with the input covariance among the mossy fibers, a process which will be called covariance storage. A similar process may occur elsewhere in the nervous system and may underlie other forms of learning and memory.

Fibers similar in appearance to climbing fibers have been found in cerebral cortex by Cajal (1911), and more recently Szentágothai (1969) has written that "in the Golgi
picture it is quite common to see a number of fine terminal axons running for considerable distances closely associated into bundles which in many cases can be seen to contain a dendritic shaft in their axes." Thus, the learning algorithm (52) could also be used to store covariance in cerebral cortex.

There are, however, major differences between the cerebellar cortex and the cerebral cortex which would have to be taken into account by a detailed model. In the cerebellum, the membrane-potential covariances of granule cells depend mainly on the mossy fibers and very little on the climbing fibers. As a consequence, the learning algorithm simply associates the covariances along these two inputs. In a more highly interconnected area, such as cerebral cortex, the climbing fiber input could influence neurons which themselves appear on the right side of Eq. (46).

Although the presumed plastic synapses between parallel fibers and Purkinje cells in the cerebellum are all excitatory, plastic synapses elsewhere might be inhibitory. For inhibitory synapses the direction of the plastic change predicted by the learning algorithm is reversed: positive covariance would weaken the synapse and negative covariance would strengthen it.

Another difference between cerebellar cortex and
cerebral cortex is reflected in their processing time scales. The cerebellum participates in ballistic movements such as saccadic eye movements. Long reverberations would not be helpful and might even interfere with the precise timing required for fine control. Parts of the cerebral cortex are concerned with coordination on a longer time scale and can be expected to have a correspondingly longer coherence time for correlations between membrane potentials.

The long-term storage of membrane potential covariance was treated linearly, but the model is essentially nonlinear; one consequence is that under some conditions small changes in the coupling strengths might result in a large change in the background firing rates (for example, if the steady-state equation studied in Part II were to bifurcate to a new solution branch).

The stability of the stationary covariance equation to small changes in the coupling strengths can be treated analytically. The solution given by Eq. (32) depends on the transition matrix whose Laplace transform is

$$\mathcal{L} [T_\epsilon] (s) = (s \tau I - A_\epsilon)^{-1},$$

where $A_\epsilon = K_\epsilon - I$ and the interaction matrix is assumed to depend analytically on $\epsilon$. The transition matrix can be recovered by the inverse Laplace transform

$$T_\epsilon (t) = \frac{1}{2\pi i} \int_C (s \tau I - A_\epsilon)^{-1} e^{st} ds, \quad (57)$$
where the contour $\mathcal{C}$ is parallel to the imaginary axis and lies to the right of all singularities. For a sufficiently small $\epsilon$ the integrand of Eq. (57) is analytic everywhere on the contour $\mathcal{C}$ (Kato, 1966). Hence, around every solution there is an open neighborhood containing other solutions.

If the resolvent is nonsingular in the right half of the complex $s$-plane, then the solution is asymptotically stable. However, if the resolvent is nearly singular on the imaginary axis, then the solution is nearly unstable and small but finite changes in the coupling strengths could lead to a very large change in the behavior of the system. This condition corresponds to the necessary condition given in Part II for the nonlinear steady-state equation (8) to be near a bifurcation point. Hence, inherent in the nonlinear model in addition to the slowly-adapting continuous change treated in this part is the possibility of discontinuous change.
Discussion

No model of interacting neurons has yet succeeded in relating the brain structure of a vertebrate to its behavior. Because the present work was inspired primarily by current experimental problems and was based on concepts derived from current experimental procedures, there is at least the hope that theory and practice will benefit from one another as they have in other areas of science.

Part II, which was mainly concerned with average firing rates, makes the closest contact with current experimental work on single-neuron recording in vertebrates. The model of stereopsis examined in this part can be tested with available experimental techniques. As emphasis shifts toward relating complex sequences of sensory and motor patterns to neuron firing patterns, temporal and spatial correlations should become more common experimental measurements. The analysis of correlations between membrane potentials in simple models, such as that in Part III, may be helpful in interpreting experimental measurements. The next step toward describing biological reality is the detailed modeling of specific brain areas, such as the attempt in Part IV to study motor learning in the cerebellum.

The central result of this research program, of
interest in itself, arose in the analysis of covariance processing. Although the model is strongly nonlinear, the covariance equation has an exactly linear form when membrane potentials are Gaussian. If some part of the nervous system takes advantage of statistical linearity, then the membrane potentials of neurons in that area should have a Gaussian distribution. This possibility can be tested with intracellular recordings.

Linearity has far-reaching consequences which are nowhere more clearly seen than in a linear theory like quantum mechanics. For example, because the Schroedinger equation for a quantum system is linear, its solutions can be classified by its symmetry group (Wigner, 1959). Similarly, the solutions of the linear covariance equation can be classified by the symmetry group of the interaction matrix. Such a classification is of great help in mathematically analyzing the model; but more importantly, the classification could be used by the brain itself in detecting and analyzing complex patterns.

Physiologically realistic models of interacting neurons are difficult to analyze because of intrinsic nonlinearities. More progress has been made in analyzing solvable linear models (Kohonen, 1977), but this work is relatively remote from experiment. The present stochastic model of nonlinearly interacting neurons combines
cooperative nonlinearity, which governs the average membrane potentials, with linear processing, which occurs at the level of membrane potential covariances. Because of this novel synthesis many different linear systems can be embedded in a single collection of nonlinearly interacting neurons. Within such a "nonlinear linear filter" information can be selectively stored and retrieved relative to a background context. The prediction that information is processed as correlations between membrane potentials can be experimentally tested with intracellular recordings from neighboring pairs of neurons. In the cerebral cortex, for example, strongly interacting cells are located in vertical "columns" with relatively weak horizontal interaction between neighboring "columns" (Mountcastle, 1978). The results of this study suggest that the membrane potentials of neurons within a cortical "column" are highly correlated with each other and with the subcortical neurons to which they are reciprocally connected.

A neuron whose rate of firing is either very low or very high is insensitive to input correlations. In a highly interconnected area, such as cerebral cortex, the subset of neurons with intermediate firing rates forms a "skeleton network" of linearly-interacting neurons. A neuron in visual cortex, for example, firing briskly in response to its "trigger feature" in the visual field, would probably
contribute to the "skeleton network" if its firing rate were not near its maximum. Under natural conditions each visual scene picks out a different subset of "trigger features", and hence a different "skeleton network". The role of "trigger features" in a "skeleton network" processing information in parallel is an alternative to the role of "trigger features" in a serial hierarchy of increasingly specific "feature detectors" (Barlow, 1968).

Although the serial computation in a digital computer is very different from the parallel computation in the nervous system, the distinction made between hardware and software can also be applied to the brain. The present study was aimed at the intermediate level between neural hardware and behavioral software which in a digital computer is called the internal "machine language". One result is a candidate for the brain's "machine language": linear systems theory, which has widespread application in systems engineering, plays an unexpected role in the nonlinear model and may also have a significant role in the operation of the brain.

Despite the formidable complexity of the nervous system, its design principles need not be complicated. Just as the design principle of reproduction, the replication of DNA, was discovered in spite of complex details not yet fully understood, the basic design principles of the brain
might also be discoverable. The nervous system is a solution to numerous biological problems in the communication of information and the control of physical systems — problems which confronted the distant ancestors of existing animals. Systems engineers confronted with similar problems have also found some solutions. It is perhaps coincidental that the mathematical analysis of a simple neural model should result in equations which are formally identical to ones with which systems engineers are familiar. This parallel could nevertheless prove useful both in analyzing experimental data and in understanding the design principles of the nervous system.
Appendix

Point-Process Model

The results so far are based on a nonlinear model with continuous variables. The continuous model is derived here from a point-process model, and the variables in both models are given precise interpretations.

A spike train, regarded as a sequence of points in time, can be modeled by a stochastic point process on the real line (Lewis, 1972). Let \( N(t) \) represent the number of spikes on time interval \([0, T)\). Assume that the mean number of spikes \( \mathbb{E}N(t) \) is differentiable, and define the "instantaneous average rate"

\[
\eta(t) = \frac{d}{dt} \mathbb{E}N(t).
\]

(58)

For example, consider the case when the point process is Poisson. Then the probability that there are \( n \) spikes on the interval \([0, T)\) is

\[
P(N(t) = n) = \Lambda(t)^n e^{-\Lambda(t)}/n!,
\]

with

\[
\Lambda(t) = \int_0^t \lambda(s) ds.
\]

Since the mean number of spikes on the interval is

\[
\mathbb{E}N(t) = \Lambda(t),
\]

the "instantaneous average rate" is \( \eta(t) = \lambda(t) \).
The "instantaneous average rate" is related to the "instantaneous rate", an experimental variable which has been useful in measuring the dynamic response of the Limulus retina (Knight, Toyoda & Dodge, 1970). Given a spike train with spikes at times \{t_i\}, the "instantaneous rate" is defined as

\[ \sigma(t) = (t_{i+1} - t_i)^{-1} \quad t_i < t \leq t_{i+1} \quad (59) \]

It is apparent from Fig. 35 that

\[ \int_0^t \sigma(s) \, ds = N(t) + \epsilon(t), \]

where the error is bounded \(|\epsilon(t)| < 1\) and \(E\epsilon(t) = 0\). Consequently, the average of the "instantaneous rate" over an ensemble of independent but identically prepared trials gives an estimate for the "instantaneous average rate"

\[ E\sigma(t) = \frac{d}{dt} E N(t) = \eta(t). \quad (60) \]

If the trails in the ensemble are not independent, then this result may not be valid. For example, when a single eccentric cell in the Limulus retina is illuminated by light with a sinusoidally modulated component, the response histogram of the cell averaged over the modulation cycle is different from the average of the "instantaneous rate", and particularly so when the modulation frequency is near the average firing rate (Knight, 1972a,b). However, the ensemble from which the response histogram derives is not composed of independent trials since the responses over
successive cycles are highly dependent.

The "effective membrane potential" which was motivated in Part I can now be more precisely defined. Let \( \psi(t) \) be the membrane potential of a neuron which does not produce an action potential but which receives a spike train as input. Then the membrane potential should satisfy the stochastic equation

\[
\tau d\psi(t) + \psi(t)dt = B dN(t),
\]

where \( B \) is the jump in magnitude of the postsynaptic potential from a single synaptic event. The derivative of the counting function \( dN(t) \) is defined by

\[
\int_0^T F(t) dN(t) = \sum_i F(t_i),
\]

where \( \{ t_i \} \) are the times when impulses occur on \( [0, T) \). If \( F(t) \) is a stochastic function then the integral must be interpreted according to the Ito calculus (Snyder, 1975).

The effect of spike production on the membrane potential is shown in Fig. 36. The difference between the membrane potential with and without reset damps exponentially after each impulse. The average of Eq. (61) over an ensemble of point processes is

\[
\tau \frac{d}{dt} \phi(t) + \phi(t) = B \eta(t),
\]

where the average membrane potential

\[
\phi(t) = E \psi(t)
\]
Fig. 35. A typical spike train $S(t)$ as a function of time, the cumulative number of spikes $N(t)$, and the "instantaneous rate" $\sigma(t)$, defined by Eq. (59) (from Sejnowski, 1977a).

Fig. 36. Idealized intracellular recording of the membrane potential $\mathcal{V}(t)$ as a function of time. Upon reaching threshold, an impulse is released and the membrane potential is reset to a lower potential. In contrast, the effective membrane potential $\psi(t)$ is continuous. The difference $\psi(t) - \mathcal{V}(t)$ suffers a step discontinuity which damps exponentially (from Sejnowski, 1977a).
corresponds to the "effective membrane potential" motivated in Part I. The resemblance between Eq. (62) and the continuous model (2) further suggests that \( \rho(\phi) \), previously defined as the firing rate to a constant input current, should be interpreted as an "instantaneous average rate." Thus, a natural generalization of the continuous model is given by the stochastic integral equation

\[
\begin{align*}
\Psi_a(t) &= \sum_b \int_0^t K_{ab}(t,s) \, dB_b(s) + \sum_b \int_0^t B_{ab}(t,s) \, dN_b(s), \\
&= \sum_b \int_0^t K_{ab}(t,s) \, dB_b(s) + \sum_b \int_0^t B_{ab}(t,s) \, dN_b(s),
\end{align*}
\]

(63)

where the point processes \( dB_b \) and \( dN_b \) satisfy

\[
\begin{align*}
\frac{d}{dt} E M_b(t) &= \rho_b(\Psi_b(t)) \\
\frac{d}{dt} E N_b(t) &= \eta_b(t).
\end{align*}
\]

The kernel \( K_{ab}(t,s) \) is the temporal response of \( \Psi_a(t) \) to an impulse at time \( s \) from a neighboring neuron, and \( B_{ab}(t,s) \) is the response from an external input. Since \( \Psi_a(t) \) and \( \eta_b(t) \) are themselves stochastic processes, the point processes in Eq. (63) are doubly stochastic (Snyder, 1975). If a neuron receives many independent impulse inputs, then by a central limit theorem its membrane potential is approximately Gaussian. For example, the point-process model in Eq. (63) with Poisson impulse generators satisfies the conditions of the central limit theorem and the membrane potentials \( \Psi_a(t) \) satisfy a linear covariance equation similar to the one for the
The time-dependent coupling kernels in this point-process model take into account the electrical properties of the intervening axons, synapses, and dendrites, including delays and decremental decay. The special case

\[ K_{ab}(t,s) = \frac{1}{\tau} K_{ab} e^{-\frac{(t-s)}{\tau}} \]

corresponds to the simple model of exponential decay considered in part I.
References


Bishop, G. H.: Natural history of the nerve impulse. Physiol. Rev. 36, 376-399 (1956)

Bishop, P. O.: Beginning of form vision and binocular depth


Gross, C. G., Rocha-Miranda, C. E., Bender, D. B.: Visual properties of neurons in inferior temporal cortex of


Grundfest, H.: Electrical inexcitability and some consequences in the central nervous system. Physiol. Rev. 37, 337-361 (1957)


Halmos, P. R.: Lectures on ergodic theory. Tokyo: Mathematical Society 1956

Hartline, H. K.: The neural mechanisms of vision. The Harvey Lectures. Series XXXVII 39 (1941)


Kalman, R. E., Bucy, R. S.: New results in linear filtering and prediction theory. J. Basic Engineering 95-108 (March 1961)


Knight, B. W.: Dynamics of encoding in a population of neurons. J. Gen. Physiol. 59, 734-766 (1972)


Knight, B. W.: The horseshoe crab eye: a little nervous system whose dynamics are solvable. Lectures on Math. in the Life Sci. 5, 111-144 (1973)


Knudsen, E. I., Konishi, M., Pettigrew, J. D.: Receptive fields of auditory neurons in the owl. Science 198,
1278-1280 (1977)


Madrid: Consejo Superior de Investigaciones Científicas 1911


Ratliff, F., Hartline, H. K.: The responses of Limulus optic nerve fibers to patterns of illumination on the receptor mosaic. J. Gen. Physiol. 6, 1241-1255 (1959)


1969


continuous

\psi, \phi_a

viii, Fig 27