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

7 Skeleton Filters in the Brain

Terrence J. Sejnowski
Department of Neurobiology
Harvard Medical School

7.1. LOOKING AT THE BRAIN

The human brain is a dull gray and glistening white tissue having the texture of stiff pudding. Under high magnification the brain looks like an intricate three-dimensional maze, as shown in Fig. 7.1(a). Each component is being studied in painstaking detail, but despite our increasing knowledge of the brain's structure our ignorance of how the brain works remains almost complete. However, neuroscientists have an advantage that workers in Artificial Intelligence do not yet have: a working model and an existence proof that problems in perception and cognition have at least one solution. If we knew how to look and what to look for, we might be able to see in Fig. 7.1(a), for example, a part of an algorithm for some problem in visual perception.

The fundamental design principles of a machine must be understood before its function can be deduced from its structure. For example, the piece of integrated circuit in Fig. 7.1(b) is a meaningless abstract design without knowing the principles of digital logic. Neuroanatomy is similarly meaningless without knowing how the signals that carry meaningful information are transformed by each component. Quite possibly we do not yet know the signals in the brain that encode thought, thus making the physiological study of cognition nearly impossible. We do have some understanding of how sensory information and motor commands are encoded, and a similar form of coding is probably exploited for central functions as well.

Peripheral sensory codes depend on labeled lines: The central nervous system "knows" where each input originates just as a central telephone exchange "knows" the origin of each telephone line. Are neurons in the central nervous

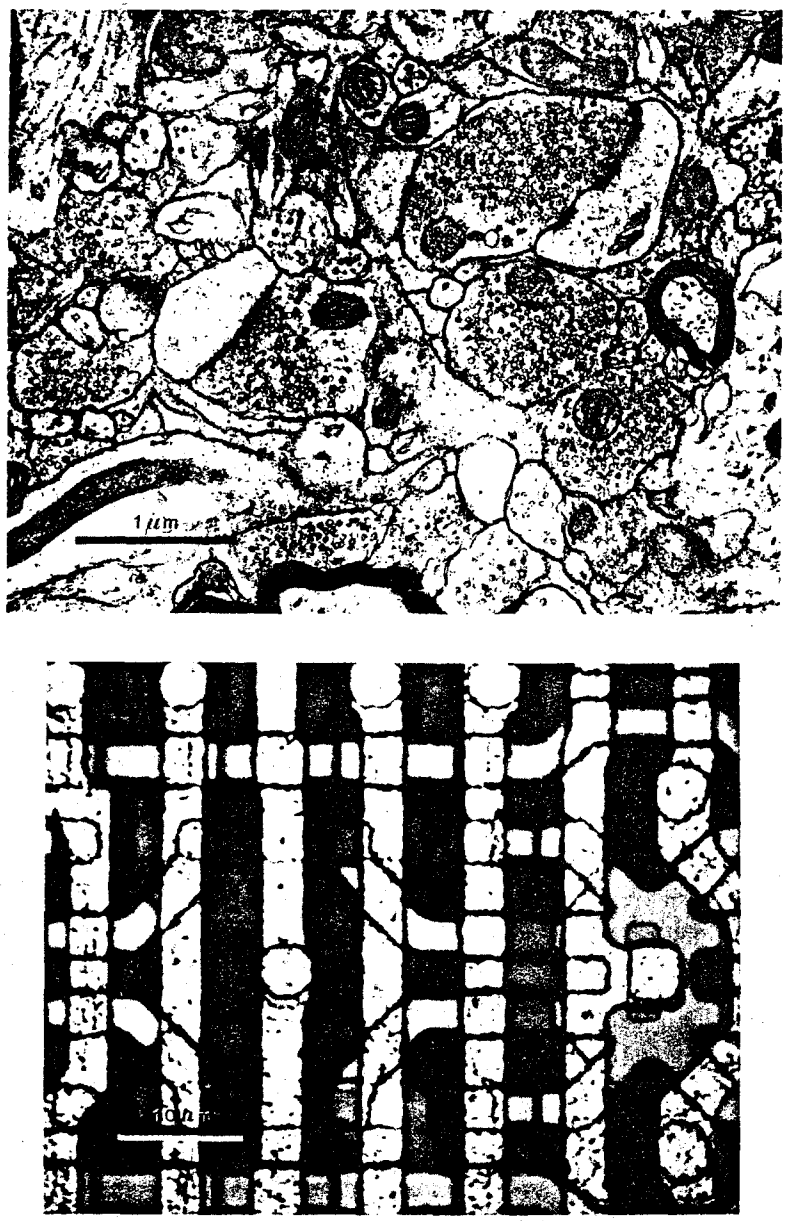

FIG. 7.1. (a) Highly magnified view of a cross section through the visual cortex of a rat using an electron microscope. The vesicle-filled profiles are presynaptic terminals. Several synapses, characterized by a presynaptic accumulation of vesicles and a postsynaptic thickening, are visible (courtesy of Simon LeVay). (b) A 16K Random Access Memory manufactured by MOSTEK. The magnification is about ten times less than in (a). Because silicon chips are essentially two-dimensional, the number of wires that can interconnect logical units in a large-scale device is severely limited.

system similarly labeled; that is, does the response of a neuron "mean" the same thing and represent a fixed address for a particular piece of information? Primary sensory areas of the brain appear to respond to sensory input in this way. For example, a visual scene is represented in the primary visual cortex by the subset of neurons that respond to particular features in the scene (Hubel & Wiesel, 1977). At higher levels of the nervous system, information may be represented by neurons that respond to a different set of primitive features—ones that are perhaps closer to the primitive components of perception. Are different perceptual states at some high level represented by the activation of different populations of neurons? The extreme possibility that small nonoverlapping populations represent percepts is called a localized representation, or sometimes a "grandmother cell" or "pontifical cell" theory (Barlow, 1972; Feldman, Chapter 2, this volume). The other extreme possibility that only large completely overlapping populations of neurons represent percepts is called a distributed representation. Both extremes are parallel models, but the essential information in one case is spatially separated and in the other case is spatially mixed.

These possibilities could be tested by mapping the electrical activity of a large number of neurons during different perceptual states. Although this type of experiment is not feasible with current physiological techniques, an anatomical technique using a radioactively labeled sugar analog, [^{14}C]-2-deoxyglucose, has been used for qualitatively mapping functional activity in the brain with a resolution of about 50 μ (Des Rosiers, Sakurada, Jehle, Shinohara, Kennedy, & Sokoloff, 1978; Hubel, Wiesel & Stryker, 1978; Sokoloff, Reivich, Kennedy, Des Rosiers, Potlak, Pettigrew, Sakurada, & Shinohara, 1977). Recent improvements in the technique now make it possible to measure the functional activity of single neurons with 1 μ resolution (Sejnowski, Reingold, Kelley, & Gelperin, 1980).

The implications of a distributed representation are explored in this chapter. Although different perceptual states are initially represented by overlapping populations of neurons, a new type of representation emerges, called a skeleton filter, which is intermediate between a localized representation and a distributed representation.

7.2. LISTENING TO THE BRAIN

Take a fine tungsten wire, etch its tip to about 1 μ, slowly lower it through a small hole made in the back of a cat's skull, and amplify the microvolt potentials from the microelectrode. Many neurons in the brain produce a brief signal that sounds like a pop when played through a loudspeaker. If the microelectrode is properly positioned in the cat's visual cortex, a burst of firing occurs whenever a bar of light moves in a particular direction at a particular position in the cat's visual field. The specificity of the response was a surprise to David Hubel and

Torsten Wiesel, who first performed this experiment, and it is surprising today how much they subsequently learned about the architecture of the visual cortex by recording from one cell—out of billions—at a time.

Although the average response of a neuron in the visual cortex from a dozen trials is a reasonably repeatable measurement, the firing pattern varies from trial to trial, as shown in Fig. 7.2. Stochastic variability is found not only in the cerebral cortex, but as well at every level of the nervous system, including the sensory receptors. One of the chief sources of noise in the brain occurs at synapses where a chemical neurotransmitter is used to signal between neurons.

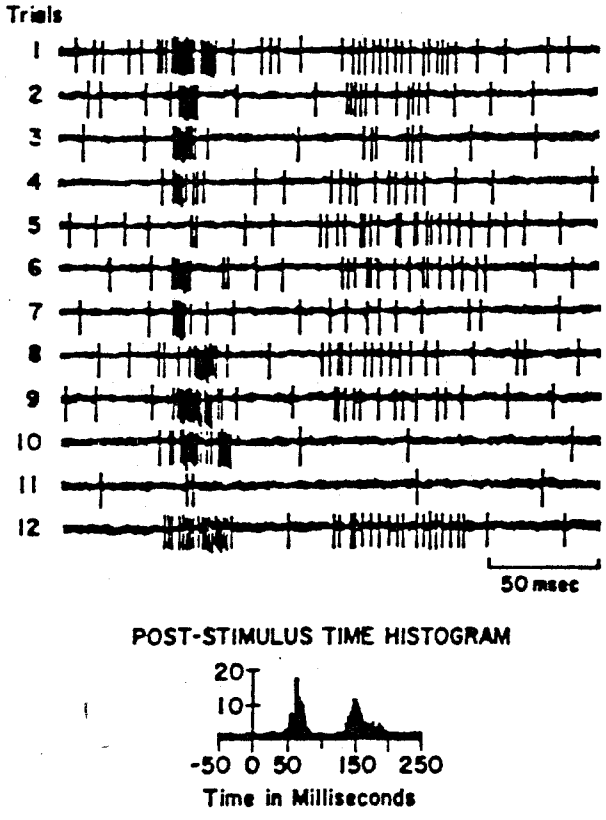

FIG. 7.2. Extracellular recordings from a single neuron in cat visual cortex. This neuron responded best to a slit of light obliquely oriented in a particular part of the visual field. Twelve successive responses of the neuron to 50 msec exposures of light are shown above, and the average response for 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, is a repeatable measurement (Morrell, 1972).

The neurotransmitter is stored in small packets called vesicles, visible in Fig. 7.1(a), and released from the presynaptic terminal in discrete units. For example, at the junction between a motor neuron and muscle in the frog, approximately 300 vesicles are released with each synaptic activation. A statistical variation of about 17 vesicles, or 6 percent of the total, can be expected during a normal activation of the synapse. Many central synapses are believed to release fewer vesicles and consequently have greater variation. Even in the absence of activation vesicles are released spontaneously, producing miniature synaptic potentials.

Quantal fluctuations at synapses is just one of several sources of noise in the nervous system. How is the brain able to function reliably with so much intrinsic variability? Perhaps the variability is not as serious as it appears (Bullock, 1970), or perhaps redundancy allows a reliable response from a population of neurons (Cowan, 1973); perhaps too we are asking the wrong question, being misled by the digital computer as a model of reliability. Could the apparent variability in the response of single neurons provide us with a clue to a basic design principle of the nervous system?

Let us take a closer look at the data in Fig. 7.2. By concentrating on the response of the neuron to the stimulus, we have overlooked another interesting feature. The neuron is active even before the stimulus and maintains an apparently random background firing. This so-called spontaneous activity is common in the nervous system although the average background firing rate varies from neuron to neuron. In the retina, for example, ganglion cells, which send signals from the eye to the brain, have spontaneous activity in complete darkness, which may increase, decrease, or remain unchanged when the retina is exposed to a steady background illumination.

Spontaneous activity is generally regarded as a bias against which inhibitory as well as excitatory signals can be imposed. Because an impulse-producing neuron has a threshold below which it can transmit no signal or information, a neuron is most sensitive to input changes when maintained near threshold. If a neuron is too far above threshold, then the signal gets swamped by the background. Threshold is, of course, an unstable region, so the price of high sensitivity is high susceptibility to noise. The high levels of spontaneous activity in the brain and the apparent variability in the response of single neurons are indications that many neurons operate near threshold much of the time.

Although the large-scale electrical activity of the brain was explored long before single-cell recording was perfected, relatively little has been established about brain mechanisms from gross recordings. One qualitative feature of EEG recordings, however, is so common that its implications are sometimes overlooked: Widespread rhythms occur throughout the cerebral cortex and subcortical structures with frequencies between 5–100 Hz. The fact that any signal survives averaging over millions of sources, is coherent over large areas of the brain, and changes with the behavioral state of the animal strongly suggests significant

temporal synchronization and spatial correlation among the sources of the EEG, one of which is believed to be the potentials generated at synapses. The possibility that synaptic events are correlated and synchronized is at present beyond the limits of experimental verification, but its consequences are worth exploring.

These three features—stochastic variability, spontaneous activity, and correlated electrical events—lead to a view of the brain that is probabilistic rather than deterministic, inherently distributed rather than local, and dynamic rather than static. Unfortunately, our experience with probablistic, distributed, dynamic systems is limited. Even simple examples and models would help us grasp the brain's complexity.

7.3. SIMPLIFYING THE BRAIN

A successful model in physics is often a caricature, extracting only a few essential features from a complex phenomenon but allowing these to be studied with clarity and precision. For example, the two-dimensional Ising model of the ferromagnetic phase transition, although unrealistic, is nonetheless important because it has an exact analytic solution and demonstrates a phase transition qualitatively similar to experimental measurements. Could a similar approach be useful in studying the brain? A simple but effective model of the brain does not yet exist, in part because its essential design features have not yet been identified. Nevertheless the strengths and limitations of simple models based on our present knowledge should be carefully examined. New ideas are more easily evaluated in comparison with already well-understood if inadequate models.

Consider a neuron, or some part of it, as a processing unit with several inputs and an output. In some models the processing is assumed to be linear: The output of each unit is proportional to the sum of its inputs. However, if a processing unit has a threshold or any other departure from proportionality, then the model is nonlinear. The class of all linear models is mathematically well understood, but each nonlinear model requires a difficult individual analysis. Linear models, such as those discussed by James Anderson and Geoffrey Hinton (Chapter 1, this volume) and Teuvo Kohonen, Pekka Lehtiö and Erkki Oja (Chapter 4, this volume) are useful for analyzing distributed properties of general networks. In nonlinear models localized computations must be studied in specific networks, such as the model of stereopsis by David Marr and Tomaso Poggio (1976) and the model of visual cortex by George Ermentrout and Jack Cowan (1979). Geoffrey Hinton (Chapter 6, this volume) demonstrates a nonlinear model of associative memory.

None of the models mentioned thus far explicitly takes into account the variability and randomness observed in the nervous system. A new approach is required based on probabilistic rather than deterministic mathematics. Fortunately, powerful tools from probability theory are available and have been

applied to a wide variety of problems in control and communication by electrical engineers. A simple probabilistic model of interacting neurons is presented in this section that provides an unexpected unification of the linear and nonlinear models.

A Simple Nonlinear Model

A cell maintains ionic gradients across its surface, which produce a potential difference between the outside and the inside of the cell. An incoming signal at a synapse, by altering the ionic conductance of the membrane, can change the membrane potential. A simple model for a single passive neuron, which to a first approximation behaves like a leaky capacitor, is given by

$$\tau \frac{d}{dt} \phi(t) + \phi(t) = B\eta(t), \tag{7-1}$$

where $\phi(t)$ is the membrane potential at time t and $\eta(t)$ is a single input with coupling strength B. The left side of Eq. (7-1) provides temporal integration of the input with a time constant τ. An excitatory input produces a sudden increase in the membrane potential, which then exponentially decays. The general solution of Eq. (7-1) for an arbitrary time-varying input is

$$\phi(t) = \int_{-\infty}^{t} e^{-(t-t')/\tau} B\eta(t')\, dt'. \tag{7-2}$$

A generalization of this linear model to a linearly interacting population of N neurons with membrane potentials $\phi_1, \phi_2, \ldots, \phi_N$ and M inputs $\eta_1, \eta_2, \ldots, \eta_M$ is given by

$$\tau \frac{d}{dt} \phi_a + \phi_a = \sum_{b=1}^{N} K_{ab}\phi_b + \sum_{c=1}^{M} B_{ac}\eta_c, \tag{7-3}$$

where K_{ab} is the strength of coupling from the bth neuron to the ath neuron, and B_{ac} is the strength of coupling between the cth input and the ath neuron. The general solution of this model is:

$$\phi_a(t) = \int_{-\infty}^{t} \sum_{b=1}^{N} T_{ab}(t - t') \sum_{c=1}^{M} B_{bc}\, \eta_c\, (t')\, dt', \tag{7-4}$$

where $\mathbf{T}(t - t')$ is the impulse response and depends only on $\mathbf{K}$. The network of neurons behaves like a multidimensional linear filter of the type used by electrical engineers to filter signals from noise. The network is especially sensitive to inputs with particular frequencies, given by the eigenvalues of $\mathbf{K}$, and to particular input patterns, given by the eigenvectors of $\mathbf{K}$.

The completeness with which the linear model can be analyzed is of great advantage when applying it to concrete cases. For example, Halden Hartline and Floyd Ratliff in 1957 using a linear model with lateral inhibition were able to

successfully predict the response of the *Limulus* lateral eye to steady-state patterns of light. More recently, Bruce Knight Jr., Fredrick Dodge Jr., and their colleagues have extended the model to predict the response of the *Limulus* retina to arbitrary time-varying illumination, thus making it the best-understood piece of nervous tissue. Details of the *Limulus* model can be found in a review (Knight, 1975) and a volume of collected papers (Ratliff, 1974).

A Simple Nonlinear Model

Signals propagated down the thin dendrites of neurons exponentially decrement with a typical length constant of 250μ. Long-distance communication is accomplished with active regenerative channels in the membrane that produce a brief all-or-none impulse, the action potential. Above threshold the rate of impulse firing increases monotonically with input and saturates at some maximum, as shown in Fig. 7.3. Some general qualitative properties of nonlinear models are already found in a single neuron which synapses onto itself. If the average effect of impulses at the synapse is assumed to be proportional to the firing rate, then the steady-state membrane potential of the neuron should satisfy

$$\phi = K\,\rho\,(\phi) + B\,\eta, \tag{7-5}$$

where η is an external input, B is the coupling strength of the input, K is the coupling strength of the neuron with itself, and $\rho\,(\phi)$ is the firing rate of the neuron, as shown in Fig. 7.3. The dependent variable in Eq. (7-5) is ϕ, the effective membrane potential, defined as the membrane potential that the neuron would have in the absence of impulses. Unlike the linear model, for which there is a unique solution for any input, the nonlinear model can have more than one steady-state solution to a single input, as shown in Fig. 7.4. The past history of the neuron determines which of the multiple states is obtained. As the input slowly changes, new solutions may appear and old ones disappear: The critical input at which a transition between solution branches occurs is called a bifurcation. (The nonlinear one-neuron model is by coincidence formally identical to the Curie-Weiss mean-field theory of magnetism, with ϕ playing the role of magnetization and η identified with the externally applied magnetic field.)

The number of multiple states and the complexity of transitions between them increases with the number of interacting neurons. A nonlinear model for N neurons with membrane potentials $\phi_1, \phi_2, \ldots, \phi_N$ and M inputs $\eta_1, \eta_2, \ldots, \eta_M$ is given by

$$\tau \frac{d}{dt}\,\phi_a + \phi_a = \sum_{b=1}^{N} K_{ab}\rho_b(\phi_b) + \sum_{c=1}^{M} B_{ac}\eta_c, \tag{7-6}$$

where B_{ac} is the coupling strength from the cth input to the ath neuron and K_{ab} is the coupling strength between the bth neuron and the ath neuron. Note that

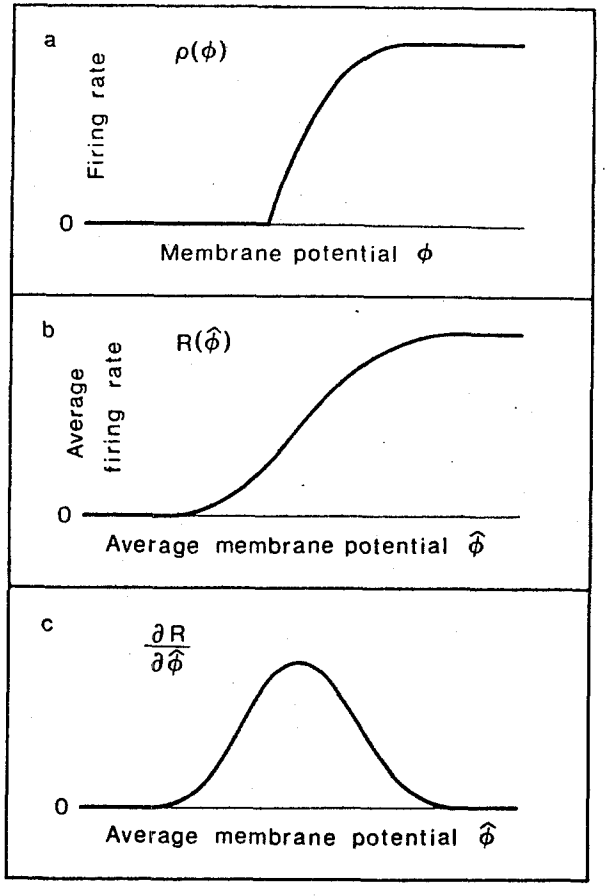

FIG. 7.3. (a) The firing rate, $\rho\,(\phi)$, as a function of the effective membrane potential ϕ for a typical neuron. (b) The average firing rate $R\,(\hat{\phi})$, defined in Eq. (7-8), as a function of the average membrane potential $\hat{\phi}$ holding all higher-order moments of ϕ fixed. (c) The partial derivative of $R\,(\hat{\phi})$ with respect to $\hat{\phi}$, which appears in Eq. (7-10).

the only difference between the linear model in Eq. (7-3) and the nonlinear model in Eq. (7-6) occurs in the nonlinear transduction, $\rho\,(\phi)$. For an impulse-producing neuron, the transduction between its input and output has a sigmoidal shape, as in Fig. 7.3, but in general the transduction can be an arbitrary nonlinear function. The model then applies equally well to neurons that do not produce impulses and to parts of neurons that are functionally independent processers (Shepherd, 1978).

Multiple perceptual states evoked by a single stimulus are common in the visual system, particularly in the perception of depth. A simple example is an

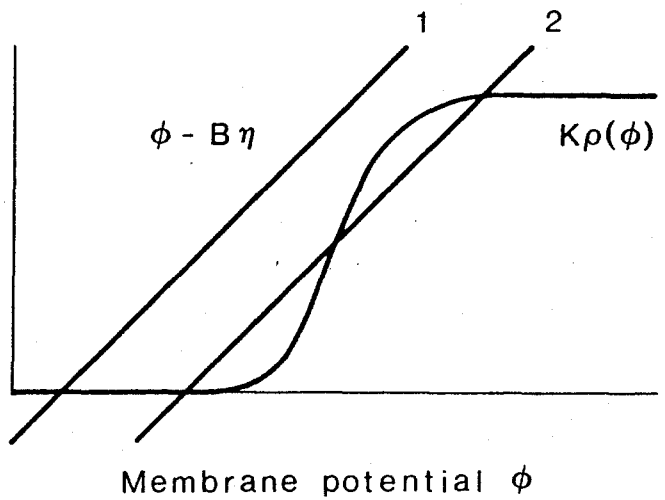

FIG. 7.4. Graphic solution of the nonlinear one-neuron model, Eq. (7-5). The intersection of the straight line $\phi - B\,\eta$ (shown for two different values of the input η) and the sigmoidal $\rho\,(\phi)$ are solutions of the model. For input η_1 there is a unique solution, but for the input η_2 there are three solutions of which the middle is unstable.

outline of a box, called the Necker cube, which can be seen in two stable three-dimensional configurations. Bela Julesz (1974) has emphasized that the properties of binocular depth perception, such as a sharp transitions and hysteresis between stable states, are characteristic of some nonlinear systems. David Marr and Tomaso Poggio (1976) have demonstrated a nonlinear model similar in form to Eq (7-6) that can detect depth in random dot stereograms. A nonlinear model of the visual cortex has been studied by George Ermentrout and Jack Cowan (1979) who found that the symmetries of solutions near a bifurcation point resemble the visual patterns reported by subjects during drug-induced visual hallucinations. The nonlinear model in Eq. (7-6) has a rich mathematical structure that we are only beginning to understand.

A Simple Probabilistic Model

The response of a neuron is the visual cortex to a pattern of light on the retina varies from trial to trial despite efforts to control experimental conditions strictly. By averaging the response over a number of trials, the variability in the response not related to the stimulus is reduced. The counterpart of experimental averaging in probability theory is called the ensemble average. Rather than model an input, for example, as a single function of time, an ensemble of inputs is chosen, the members of the ensemble differing from one another by random variations. The

solution of an equation for an ensemble of inputs is a corresponding ensemble of solutions. A great deal of information can be obtained from the ensemble in addition to the average solution, such as the average square variation or variance from the average. The average over an ensemble should not be confused with the average over time although under certain conditions the two may agree.

The nonlinear model in Eq. (7-6) can be made probabilistic by including noise with the inputs on the right side. Corresponding to an ensemble of inputs, each with a different random noise component, there is an ensemble of membrane potential responses derived from Eq. (7-6). Define the ensemble average of the membrane potential as

$$\hat{\phi}_a\,(t) = E\,\phi_a\,(t), \tag{7-7}$$

where E, the ensemble average or expectation operator, takes all the membrane potentials in the ensemble at a particular time and produces a single function, the average membrane potential $\hat{\phi}_a\,(t)$. Similarly, the ensemble average firing rate for an impulse-producing neuron is defined as

$$R_b\,(t) = E\,\rho_b\,(\phi_b\,(t)). \tag{7-8}$$

The ensemble average firing rate corresponds technically to the limit of the experimental poststimulus time histogram for an infinite number of trials.

A probabilistic analysis of the nonlinear model has been given elsewhere (Sejnowski, 1976b, 1977b). The strategy in the analysis is to set up a hierarchy of equations governing the statistical moments and to make reasonable simplifying assumptions to study each tier in the hierarchy. Although the equations in the hierarchy are coupled, each tier can, to some extent, be analyzed separately.

The lowest tier deals with first-order moments: the averages of single variables. The average membrane potentials satisfy an equation similar in form to the model itself, with $\rho_b\,(\phi_b)$ replaced by $R_b\,(\hat{\phi}_b)$, a somewhat smoother nonlinear function as shown in Fig. 7.3. The properties of the deterministic model in Eq. (7-6), which have already been discussed, hold as well for the equation that governs the average membrane potentials.

The second tier of the statistical hierarchy concerns second-order moments: the averages of squared variables and products of two variables. The average firing rate on the first tier is known to carry sensory information to the central nervous system and motor commands to muscles. Relatively little experimental effort has been devoted to measuring second-order moments, such as the variance of the firing rate, or correlations between membrane potentials, so it is not clear what information, if any, is carried on the second tier. An analysis of the second tier is nonetheless important for two reasons: First, the variances of the membrane potentials feed back to affect the first-order equations; and second, it is worth knowing what to look for if the nervous system does make use of higher-order moments.

Correlations between the spike trains of nearby neurons have been measured throughout the brain (e.g., retina: Rodieck, 1967; Mastronarde (in press); lateral geniculate nucleus: Stevens & Gerstein, 1976; cerebellum: Bell & Kawasaki, 1972; auditory cortex: Dickson & Gerstein, 1974). Relatively few experiments have been designed to measure changes in correlations in response to sensory stimulation. One intriguing example of stimulus-dependent correlations between two neurons in visual cortex is shown in Fig. 7.5. Because the membrane potential in a neuron is often below the spiking threshold, correlations between

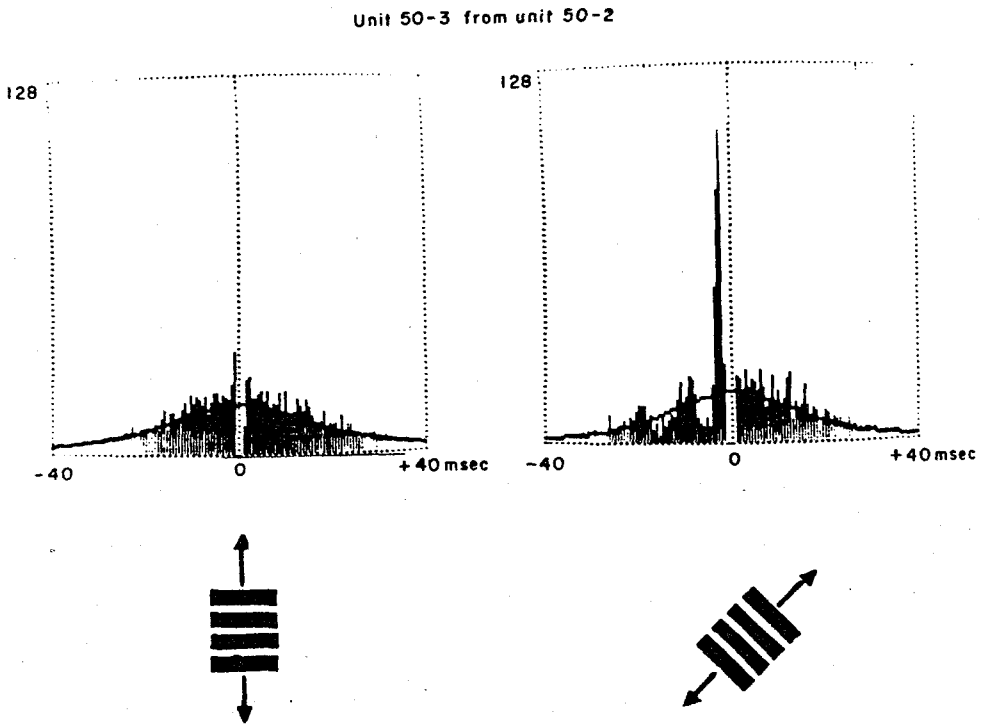

FIG. 7.5. Cross-interval histograms of impulse firing from two simultaneously recorded neurons in visual cortex of a cat. Each stimulus bar was 2° wide at the cat's eye and the pattern of bars (shown below each histogram) was moved sinusoidally through 20° every 6 sec in the direction indicated by the arrows. The histogram was computed by measuring the time difference from every impulse in one train to the nearest preceding and succeeding impulse of the second train. If the impulses in the two neurons were occurring independently, the histogram should agree with the solid line. The left histogram agrees well with the control calculation, but the right histogram shows a series of peaks at approximately 3, 11, and 22 msec, indicating that one neuron tended to follow the other with those intervals. Thus the correlations between the two neurons depended on the stimulus. (Gerstein, 1970)

membrane potentials should be at least as prominent as correlations between spike trains.

A second-order correlation that has been normalized to remove the influence of first-order averages is called a covariance and is defined as

$$\text{cov}\,[\phi_a(s), \phi_b(t)] = E\,[\phi_a(s)\,\phi_b(t)] - E\,\phi_a(s)\,E\,\phi_b(t). \qquad (7\text{-}9)$$

The covariance is positive when the two membrane potentials fluctuate together more often than by chance, negative when they fluctuate oppositely more often than by chance, and zero when they are completely independent.

A neuron in cerebral cortex, such as the pyramidal cell in Fig. 7.6, continually receives an extremely large number of synaptic events along thousands of inputs. By the central limit theorem in probability theory, the sum of a large number of independent random signals has an approximately Gaussian distribution. Hence

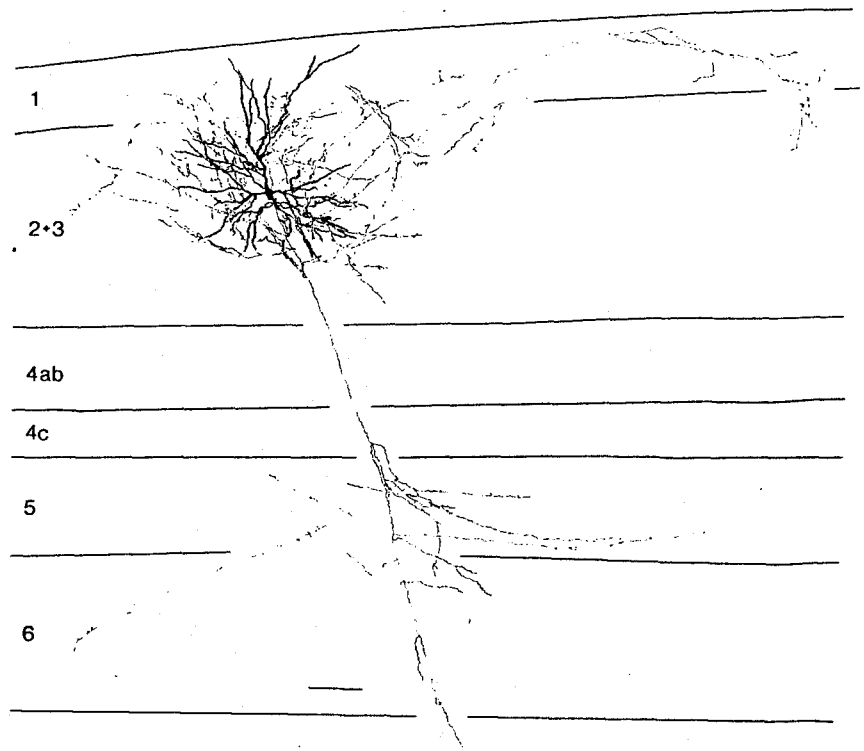

FIG. 7.6. Pyramidal cell in visual cortex of the cat. The cell was impaled with an intracellular electrode, and after its response to visual stimuli was determined, an enzyme, horseradish peroxidase, was injected into the cell. The neuron's axonal tree (thin processes) was as extensive as its dendritic tree (thick processes). (Gilbert & Wiesel, 1979)

it is reasonable to assume that the membrane potential of a typical neuron in cerebral cortex is Gaussian, an assumption that can be checked experimentally. This assumption can be rigorously proved as a limit theorem in some related dynamical systems (Stuart Geman, unpublished). An unexpected simplification occurs in the analysis of the covariance if the membrane potentials are Gaussian: Although the model is highly nonlinear, the covariance between membrane potentials satisfies a linear equation identical in form to Eq. (7-3). However, the coupling strength in the equation for the covariance is not K_{ab} but

$$K'_{ab} = K_{ab} \frac{\delta R_b(\hat{\phi}_b)}{\delta \hat{\phi}_b} \tag{7-10}$$

The multiplicative factor in the effective coupling strength K'_{ab} has a simple intuitive explanation. As shown in Fig. 7.3, this factor is small when the average membrane potential is far below or far above threshold and is largest when the average membrane potential is near threshold. The covariance is a signal contained in the fluctuations of the membrane potential: The neuron will not transmit information in the fluctuations if the membrane potential is below threshold or if the neuron is saturated above threshold but will transmit information in the fluctuations if the neuron is poised near threshold. As a consequence, only a skeleton network of neurons near threshold significantly affects the covariance. The processing of covariance is linear even for large fluctuations as long as the Gaussian assumption remains valid.

The probabilistic analysis summarized here unifies two classes of models with very different character. On the first tier, the average membrane potentials are governed by a nonlinear equation identical in form to those in the nonlinear class, Eq. (7-6). On the second tier, the covariance between membrane potentials satisfies a linear equation, Eq. (7-4). The coupling between the nonlinear and linear equations suggests a novel and flexible way to control the processing, storage, and retrieval of distributed information.

7.4. MODELING MEMORY

Models of memory are uncomfortably abstract: first, because cognitive processing is several stages removed from the physical representation of sensory information; and second, because no one knows where or how thinking takes place. What then is the value of modeling an unidentified brain area that processes undetermined information? There is, perhaps, something to be learned about the adequacy of the model for studying qualitative properties of the functioning brain.

The best-studied model of associative memory is the linear matrix model (Anderson, 1970; Kohonen, 1972; Steinbuch, 1961). In the simplest example,

the interactions between neurons in Eq. (7-3) are ignored and only static inputs are studied. The output of a neuron, ϕ_a, then satisfies

$$\phi_a = \sum_{b=1}^{M} B_{ab}\eta_b, \tag{7-11}$$

where the B_{ab} are the coupling strengths of the branching inputs η_b. How should the B_{ab} be chosen so that a specific input pattern will produce a desired output pattern? What happens as we increase the number of paired associations? Is there a simple algorithm for computing the optimal B_{ab}? All of these questions have precise answers, as discussed by Kohonen et. al. (Chapter 4, this volume). The same techniques can be applied to the static solutions of the linear model in Eq. (7-3), which includes interactions,

$$\phi_a = \sum_{b=1}^{N} K_{ab}\phi_b + \sum_{c=1}^{M} B_{ac}\eta_c, \tag{7-12}$$

where the coupling strengths between neurons, K_{ab}, are altered to store input-output associations rather than B_{ab}. A review of this model, including useful demonstrations, is given by Kohonen (1977).

The matrix model resembles memory in the same way that a toy glider resembles a bird. It does fly, in a rigid sort of way, but it lacks dynamics and grace. The input and output vectors in the matrix model are purely spatial, but as we know from common experience, associations have a temporal flow. Furthermore, associations depend overwhelmingly on context, which is entirely missing from the model. Can the matrix model be suitably generalized to overcome these shortcomings? In the case of dynamics the answer is yes, as shown shortly. No linear model, however, can ever be constructed to include context or contingencies: Like a toy glider a linear model always "flies" in a straight line.

Time

Most of us have a reasonably good memory for temporal sequences. Given the first few bars of a familiar tune, we can usually identify, if not reproduce, the rest of the tune. Christopher Longuet-Higgins (1968) proposed a model of temporal memory that he called the holophone in analogy with the distributed storage of spatial information in the holograph. The holophone can record temporal associations to a given input and respond with the associated signal whenever the input reoccurs. The original model of the holophone suggested by Longuet-Higgins involved banks of filters and variable amplifiers, that is, a realization in the frequency domain. Because the holophone is a linear filter whose output is of the form given by Eq. (7-4), the linear model in Eq. (7-3) is an equivalent state-variable realization of the holophone if only a single input and a single output are

condidered. David Willshaw (Chapter 3, this volume) discusses extensions of the original holophone model and some of its limitations.

The time-dependent linear model, by virtue of the first term in Eq. (7-3), adds a rich temporal dimension to the static model in Eq. (7-11). Moreover, the analytic solution is explicitly known: The filter matrix in Eq. (7-4) has the form

$$\mathbf{T}_{ab}(t) \sim \sum_n \sum_k t^k e^{-\lambda_n t} \sin(\omega_n t) P_{ab}(n, k), \tag{7-13}$$

where λ_n and ω_n are derived, respectively, from the real and imaginary parts of the eigenvalues for the coupling matrix $\mathbf{K}$, and $\mathbf{P}$ (n, k) are a set of matrices doubly indexed by (n, k) and derived from the eigenvectors of $\mathbf{K}$.

The filter matrix $\mathbf{T}$ (t) is the response of the model to a sudden burst of action potentials along the inputs. If for convenience we assume that the inputs do not branch ($\mathbf{B}$ is diagonal), then by Eq. (7-4) the response of the time-dependent model is

$$\phi_a(t) = \sum_b T_{ab}(t)\eta_b, \tag{7-14}$$

where η_b is proportional to the number of action potentials along the bth input. Thus the membrane potential of each neuron is the sum of many exponentially damped, sinusoidally varying components indexed by n. However, the envolope of the kth term in the second summation has a peak that appears later as k increases. Each term has a separate matrix $\mathbf{P}$ (n, k), that transforms the input into a different spatial output, and these successively unfold in time. Rather than give rise to a single output as in the case of the static model, a single input in the dynamic model produces a doubly indexed set of associated outputs, one set indexing the frequency spectrum and the second set indexing the sequence in time. Further details about the output pattern are given elsewhere (Sejnowski, 1976b).

Synaptic Plasticity

The physical basis of learning and memory is unknown. Alteration in the strengths of synapses between neurons has been shown to underlie habituation of a simple reflex in *Aplysia*, a marine mollusk, and similar mechanisms may underlie more complex forms of learning (Kandel, 1976). The matrix model and filter model of memory predict the conditions under which synaptic strengths should change in order to store new associations optimally (Kohonen, 1978; Sejnowski, 1977a). New experimental techniques are needed to test these predictions in the vertebrate central nervous system.

One of the best-studied areas of the brain is the cerebellum, an area that receives inputs from both motor and sensory systems and is intimately involved in motor coordination. Experiments on the vestibulo-ocular reflex indicate that

the cerebellum may be involved in motor learning (Ito, 1975; Robinson, 1976). Following the suggestions of Brindley (1964) and Szentágothai (1968), Marr (1969) and Albus (1971) have proposed detailed theories for associative motor learning in the cerebellum that predict plasticity for synapses between parallel fibers and Purkinje cells (Fig. 7.7). According to Marr (1969) the synapses should be "facilitated by the conjunction of presynaptic and climbing fiber (or postsynaptic) activity". Some conjunctions, however, take place purely by chance; because accidental coincidences are unrelated to an animals's experience, they can have little or no adaptive value. Moreover, unless means exist for weakening the plastic synapse, continual random coincidences inexorably push it to maximum strength. A plastic synapse whose strength can be flexibly adjusted within its range should therefore be capable of long-term depression as well as

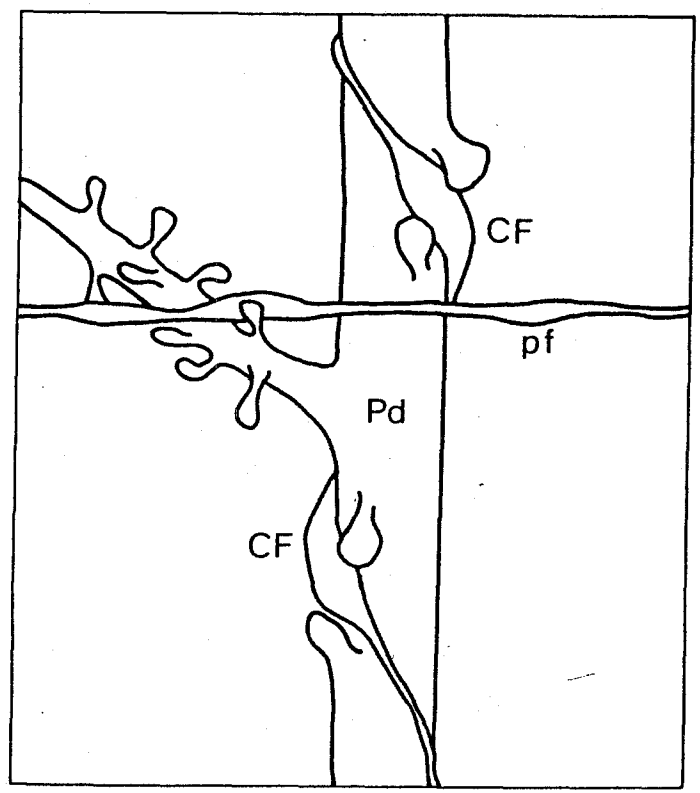

FIG. 7.7. Schematic illustration of a cerebellar Purkinje cell *Pd* (with a dendritic branchlet), a climbing fiber *CF* (entwining the dendritic trunk), and a parallel fiber *pf* (passing through the dendritic tree), based on Palay and Chan-Palay (1974). Climbing fiber varicosities make numerous synaptic contacts with spines on the dendritic trunk. (Sejnowski, 1977*a*)

long-term facilitation, and the condition for weakening the synapse should be as specific as that for strengthening it—otherwise the information stored as the synaptic strength is lost.

Without proposing an all-encompassing theory for the cerebellum, the probabilistic model in the previous section can be applied to the specific problem of plasticity in the cerebellar cortex (Sejnowski, 1977a, 1977b). The result overcomes some of the shortcomings of previous predictions. If K is the strength of a plastic synapse, then the learning algorithm derived from the dynamic filter model of memory is

$$\frac{d}{dt}K = \gamma[p(t)c(t) - \hat{p}(t)\hat{c}(t)], \tag{7-15}$$

where the constant γ determines the rate of change of the synaptic strength, $p(t)$ is the presynaptic input (a parallel fiber in the cerebellum), $c(t)$ is the "teaching" input (a climbing fiber in the cerebellum), and $\hat{p}(t)$ and $\hat{c}(t)$ are their average values. Thus the algorithm predicts that the strength of the synapse should increase whenever the parallel fiber and climbing fiber are activated together more often than by chance, decrease in strength whenever they are activated together less often than by chance, and maintain a constant average strength when the two inputs are uncorrelated. This covariance storage algorithm has two advantages: First, the problem of saturation from chance coincidences is overcome; and second, the entire dynamic range of synaptic strength is always available. One problem that the algorithm does not solve is deviation from the average strength owing to random fluctuations. However, this problem can be minimized by limiting the time during which a synapse is sensitive to modification.

A similar algorithm was independently proposed by Leon Cooper, Fishel Liberman, and Erkki Oja (1979), who used it to model the acquisition and loss of neuron specificity in the visual cortex during development. The convergence and stability of a wide class of learning algorithms has been studied by Élie Bienenstock (1980).

Skeleton Filters

The linear filter model for memory viewed as a processing unit has two types of terminals—inputs and outputs. Recall from memory, however, depends not only on sensory inputs but on expectation and context as well. How can these influences be accounted for in the model? Adding a second input for context does not help: The new output is simply a linear superposition. A new type of input is needed, one that can change the processing of other inputs.

The probabilistic model discussed in the previous section provides the required flexibility. The linear filtering of the input on the second tier depends on the skeleton network determined by the first tier. Two types of input can therefore be distinguished: (1) inputs that affect the average membrane potentials, and hence by Eq. (7-10) the skeleton network for the filter; and (2) inputs that affect the correlations between membrane potentials. (Both types of inputs may, of course, be carried by a single set of input fibers.) Contextual information, represented by the average firing rates of neurons in the filter, could completely alter the correlated output associations evoked by correlated inputs. Consequently, the probabilistic model allows many different skeleton filters to be embedded in the same population of neurons. For example, each different visual pattern excites a different subset of neurons in the visual cortex, which could in turn serve as a different skeleton filter for processing correlations.

Our sensitivity to a particular sensory signal can be greatly enhanced when our attention is properly focused. The cue can be physical, grammatical, meaningful, or any other perceivable dimension. A skeleton filter with internally generated inputs rather than sensory inputs driving the background firing rates is a candidate model for selective attention. The type of information to which the skeleton filter could be made sensitive depends on where the filter is placed in the processing stream. Evidence exists for filtering on all levels, from early sensory selection to late conceptual selection (Norman, 1976).

Items in human memory which are associated with each other can be related in many different ways. A table and chair can be related by color, style, function, or any other conceivable dimension. In a semantic network, relationships are graphically summarized as a set of items joined by relational arrows, as Scott Fahlman discusses in chapter 5 of this book. How is the discrete representation of knowledge in a semantic network related to the analog representation of information in distributed filters? The simplest unit of knowledge in a semantic network is a triple of two items and a relation between them. A relation cannot be included in a linear filter because, as we have seen, there is no way for a linear filter to account for a contingency in the association between two items. In a skeleton filter, however, contingency is represented by the background firing rates, as illustrated in Fig. 7.8. If the firing rates in an area were to represent a relation, then the skeleton filter could generate output associations to input items relative to that relation.

The skeleton filter viewed as a processing unit has three types of terminals, one of which can be used to represent contextual and relational information. Skeleton filters could, in principle, be used selectively to store and retrieve associations in long-term memory, to attend sensory information, and to manipulate relational knowledge. The selectivity of a skeleton filter does not depend on the details of the particular network model analyzed here. Any model composed of nonlinear threshold devices will exhibit transitions between different processing states. If the nonlinearity is strong, such as the step functions that Geoffrey Hinton uses in the model he discusses (Chapter 6, this volume), then the transitions between states is sharp and the control of the "skeleton" will be "tight". Weaker nonlinearities, such as a linear device with a threshold, allow more

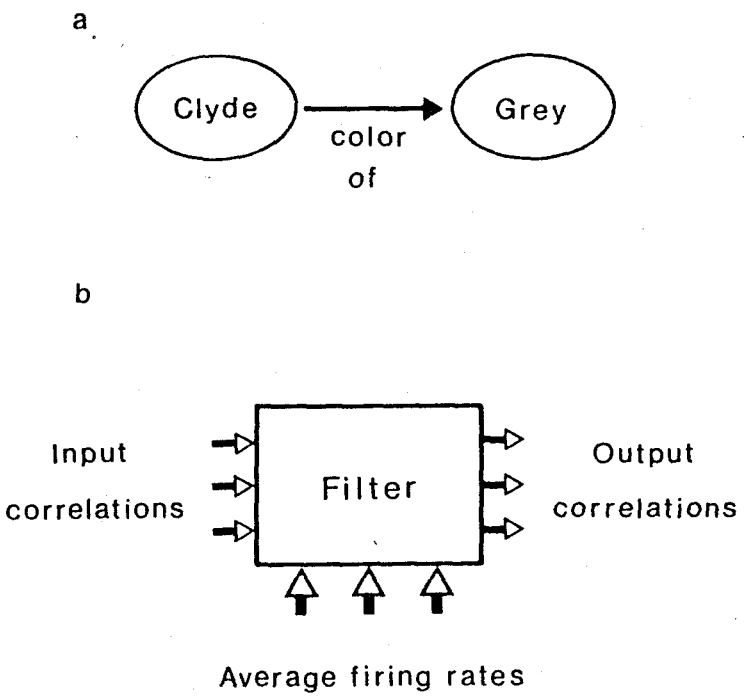

FIG. 7.8. (a) Example of an elementary triple from a semantic network. The item "Clyde" (an elephant) is linked with the item "grey" by the relation "color of." (b) Schematic diagram of inputs and outputs for a skeleton filter. Input correlations are transformed by the filter matrix T(t) given by Eq. (7-13). The filter matrix and hence the output correlations depend on the average firing rates R_b through the effective coupling matrix in Eq. (7-10).

gradual transitions and "softer" control. The probabilistic model in this chapter starts with an arbitrary sigmoidal nonlinearity, which includes step functions and linear threshold devices, and derives an exactly linear skeleton network embedded in the full nonlinear model. Thus, control of the skeleton can be "tight" in one area and "soft" in another, and the processing is linear in both cases.

7.5. THEORY AND PRACTICE

Three theoretical traditions have independently contributed to our present understanding of distributed information processing. Workers in two traditions were inspired by distributed processing in the brain: those interested in associative memory concentrated on linear models (Anderson, 1970; Kohonen, 1972; Longuet-Higgins, 1968) while those who emphasized cooperative properties developed nonlinear models (Ermentrout & Cowan, 1979; Julesz, 1974; Marr & Poggio, 1976). A third tradition, inspired by machines rather than man, deals with the control of complex physical systems and the communication of information. The probabilistic tools developed by systems engineers were applied in this chapter to a nonlinear model that previously had only been treated by deterministic techniques. A surprise occurred during the analysis of the model: If a reasonable assumption is made then the linear and nonlinear models become unified in a single probabilistic one. In addition to being theoretically attractive the unification has experimental implications that are directly testable.

The primary variable in most network models is the average firing rate, which is known to code sensory and motor information in the central nervous system. In the probabilistic model the membrane potential is taken as the primary variable, and the average firing rate appears as a derived statistical variable. In most linear network models the average firing rate of a neuron is assumed to vary linearly with total input, but this is only a valid approximation over a small part of a typical neuron's operating range, as shown in Fig. 7.3. Linearity appears in the probabilistic model not at the level of the average membrane potential or average firing rate but at the level of correlations between membrane potentials, a higher-order statistical variable that is just beginning to be explored experimentally.

Correlations are signals contained in the fluctuations of the membrane potentials from their average values, a component that is usually ignored in most experiments. Whether large-scale correlations exist and are related to sensory processing can be directly tested by intracellular recording from neurons in cerebral cortex. Charles Gilbert and Torsten Wiesel (1979), for example, have used intracellular recording in the visual cortex to identify the class of a neuron from its average response and to determine its morphology following injection of a marker (Fig. 7.6). An ensemble of responses contains information beyond the average response, such as the ensemble correlation, which may also depend on the stimulus. An ensemble of intracellular recordings from a pair of neurons responding to a controlled sensory stimulus could be used to determine the ensemble correlation between the membrane potentials and to test the key assumption that membrane potentials have a Gaussian distribution. These experiments are difficult and might seem unpromising: Only in a few areas of the cortex, such as the primary visual cortex, is enough known about the first-order average response to justify looking at second-order signals. However, the importance of higher-order processing cannot be properly assessed until data are available from carefully controlled experiments. A probabilistic model may be useful in suggesting worthwhile measurements and in analyzing the data.

The unification of the linear and nonlinear models by a single probabilistic model provides a rigorous basis for a new device, the skeleton filter, which combines the advantages of linear filters from systems engineering with the flexibility of nonlinear control. A skeleton filter is a skeleton network of neurons

in an area that linearly filters correlations along incoming spike trains. The subset of neurons in the skeleton network, and hence the filtering characteristics of the network, can be adjusted by changing the average firing rates of the neurons. In principle a skeleton filter could be used to implement selective attention, to provide for the selective storage and retrieval of information from associative memory, and to manipulate relational knowledge, which is not possible in a strictly linear model.

The aim of the probabilistic model summarized in this chapter is to provide a bridge between neural "hardware" and behavioral "software." If the model is a good first approximation to information processing in the central nervous system, then the parallels with communications and control engineering could prove useful not only in interpreting experimental data but in understanding the brain's design principles as well. In particular, linear systems theory, which has important applications in signal detection and spacecraft guidance, can be considered the machine language for a "chip" of highly interconnected neurons in a skeleton filter. Many chips, perhaps many millions, may be required for each sensory system and each level of cognitive processing. The "columns" (Mountcastle, 1979) and "dendritic bundles" (Roney, Scheibel, & Shaw, 1979) that have been found in the cerebral cortex are candidates for such skeleton filter chips.

The brain contains the solutions to numerous problems in control and communication that faced our distant ancestors. Despite the brain's formidable complexity, its design principles need not be complicated, just as the biochemical complexity of the cell does not obscure the simplicity of life's design principle, the replication of DNA. Future workers may uncover some of the brain's design principles.

REFERENCES

Albus, J. S. A theory of cerebellar function. *Mathematical Biosciences*, 1971, *10*, 25-61.

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

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

Bell, C. C., & Kawasaki, T. Relation among climbing fiber responses of nearby Purkinje cells. *Journal of Neurophysiology*, 1972, *35*, 155-169.

Bienenstock, É. L. *A theory of development of neuronal selectivity*. Unpublished doctoral dissertation, Brown University, 1980.

Bullock, T. H. The reliability of neurons. *Journal of General Physiology*, 1970, *55*, 565-584.

Brindley, G. S. The use made by the cerebellum of the information that it receives from sense organs. *International Brain Research Organization Bulletin*, 1964, *3*, 80.

Cooper, L. N., Liberman, F., & Oja, E. A theory for the acquisition and loss of neuron specificity in visual cortex. *Biological Cybernetics*, 1979, *33*, 9-28.

Cowan, J. D. The design of reliable systems. In M. A. B. Brazier, D. O. Walter, & D. Schneider (Eds.), *Neural Modeling*. Los Angeles, Calif.: Brain Information Service, 1973.

Des Rosiers, M. H. Sakurada, O., Jehle, J., Shinohara, M., Kennedy, C., & Sokoloff, L. Functional plasticity in the immature striate cortex of the monkey shown by the [^{14}C]deoxyglucose method. *Science*, 1978, *200*, 447-449.

Dickson, J. W., & Gerstein, G. L. Interactions between neurons in auditory cortex of the cat. *Journal of Neurophysiology*, 1974, *37*, 1239-1261.

Ermentrout, G., & Cowan, J. D. A mathematical theory of visual hallucination patterns. *Biological Cybernetics*, 1979, *34*, 137-150.

Gerstein, G. L. Functional associations of neurons: Detection and interpretation. In F. O. Schmitt (Ed.), *The Neurosciences: Second Study Program*. New York: The Rockefeller University Press, 1970.

Gilbert, C. D., & Wiesel, T. N. Morphology and intracortical projections of functionally characterized neurones in cat visual cortex. *Nature*, 1979, *280*, 120-125.

Hubel, D. H., & Wiesel, T. N. Functional architecture of macaque visual cortex. *Proceedings of the Royal Society (London)*, 1977, *B198*, 1-59.

Hubel, D. H., Wiesel, T. N., & Stryker, M. P. Orientation columns in macaque monkey visual cortex demonstrated by the 2-deoxyglucose autoradiographic technique. *Science*, 1978, *269*, 368-330.

Ito, M. Learning control mechanisms by the cerebellum flocculo-vestibulo-ocular system. In D. H. Tower (Ed.), *The Nervous System* (Vol. 1). New York: Raven Press, 1975.

Julesz, B. Cooperative phenomena in binocular depth perception. *American Scientist*, 1974, *62*, 32-43.

Kandel, E. R. *A cell-biological approach to learning*. Bethesda, Md.: Society for Neuroscience, 1978.

Knight, B. W. The horseshoe crab eye: A little nervous system whose dynamics are solvable. *Lectures on Mathematics in the Life Sciences*, 1975, *5*, 111-144.

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

Kohonen, T. *Associative memory*. Berlin-Heidelberg-New York: Springer-Verlag, 1977.

Longuet-Higgins, H. C. Holographic model of temporal recall. *Nature*, 1968, *217*, 104.

Marr, D. A theory of cerebellar cortex. *Journal of Physiology*, 1969, *202*, 437-470.

Marr, D., & Poggio, T. Cooperative computation of stereo disparity. *Science*, 1976, *194*, 283-287.

Mastronade, D. N. Correlated firing of cat retinal ganglion cells. *Journal of Neurophysiology*, in press.

Morrell, F. Integrative properties of parastriate neurons. In A. G. Karczmar & J. C. Eccles (Eds.), *Brain and Human Behavior*. Berlin-Heidelberg-New York: Springer-Verlag, 1972.

Mountcastle, V. B. An organizing principle for cerebral function: The unit module and the distributed system. In F. O. Schmitt (Ed.), *The Neurosciences: Fourth Study Program*. Cambridge, Mass.: MIT Press, 1979.

Norman, D. A. *Memory and attention*. New York: Wiley 1976.

Palay, S. L. & Chan-Palay, V. *Cerebellar cortex: Cytology and organization*. Berlin-Heidelberg-New York: Springer-Verlag, 1974.

Ratliff, F. (Ed.) *Studies in excitation and inhibition in the retina*. New York: The Rockefeller University Press, 1974.

Robinson, D. A. Adaptive gain control of vestibulo-ocular reflex by the cerbellum. *Journal of Neurophysiology*, 1976, *39*, 954-969.

Rodieck, R.W. Maintained activity of cat retinal ganglion cells. *Journal of Neurophysiology*, 1967, *30*, 1043-1071.

Roney, K. J., Scheibel, A. B. & Shaw, G. L. Dendritic bundles: Survey of anatomical experiments and physiological theories. *Brain Research Reviews*, 1979, *1*, 225-271.

Sejnowski, T. J. On global properties of neuronal interaction. *Biological Cybernetics*, 1976, *22*, 85-95. (a)

Sejnowksi, T. J. On the stochastic dynamics of neuronal interaction. *Biological Cybernetics*, 1976, *22*, 203–211. (b)

Sejnowski, T. J. Statistical constraints on synaptic plasticity. *Journal of Theoretical Biology*, 1977, *69*, 385–389. (a)

Sejnowski, T. J. Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, 1977, *4*, 303–321. (b)

Sejnowski, T. J., Reingold, S. C., Kelley, D. B., Gelperin, A. Localization of [^{3}H]-2-deoxyglucose in single molluscan neurons. *Nature*, 1980, *287*, 449–451.

Shepherd, G. M. Microcircuits in the nervous system. *Scientific American*, 1978, (2), 93–103.

Sokoloff, L., Reivich, M., Kennedy, C., Des Rosiers, M. H., Potlak, C. S., Pettigrew, K. D., Sakurada, O., Shinohara, M. The [^{14}C]deoxyglucose method for the measurement of local glucose utilization: Theory, procedure, and normal values in the conscious and anesthetized albino rat. *Journal of Neurochemistry*, 1977, *28*, 897–916.

Steinbuch, K. Die Lernmatrix. *Kybernetik*, 1961, *1*, 36–45.

Stevens, J. K., & Gerstein, G. Interactions between cat lateral geniculate neurons. *Journal of Neurophysiology*, 1976, *39*, 239–256.

Szentágothai, J. Structural–functional considerations of the cerebellar neuron network. *Proceedings of the IEEE*, 1968, 56, 960–968.