[121] DW Tank, M Sugimori, J Conner, and R Llinás, Spatially resolved calcium dynamics of mammalian Purkinje cells in the cerebellar slice, *Science* 242 (1988) 773–777.

[122] L Tauc and GM Hughes, Modes of initiation and propagation of spikes in the branching axons of molluscan central neurons, *J. Gen. Physiol.* 46 (1963) 533–549.

[123] DW Thompson, *On Growth and Form,* abridged edition, Cambridge University Press, Cambridge, 1961.

[124] AM Thomson and J Deuchars, Temporal and spatial properties of local circuits in neocortex, *Trends Neurosci.* 17 (1994) 119–126.

[125] RD Traub, RKS Wong, R Miles, and H Michelson, A model of the CA3 hippocampal pyramidal neuron incorporating voltage-clamp data on intrinsic conductances, *J. Neurophysiol.* 66 (1991) 635–650.

[126] HC Tuckwell, *Introduction to Theoretical Neurobiology,* Cambridge University Press, Cambridge, 1988.

[127] DA Turner, XG Li, K Pyapali, A Ylinen, and G Buzaki, Morphometric and electrical properties of reconstructed hippocampal CA3 neurons recorded in vivo, *J. Comp. Neurol.* 356 (1995) 580–594.

[128] DA Turner, submitted.

[129] YW Turner, DER Meyers, and JL Barker, Localization of tetrodotoxin-sensitive field potentials of CA1 pyramidal cells in the rat hippocampus, *J. Neurophysiol.* 2 (1991) 1375–1387.

[130] AM Vallet and JC Coles, Is the membrane voltage amplifier of drone photoreceptors useful at physiological light intensities? *J. Comp. Physiol. A* 173 (1993) 163–168.

[131] AM Vallet, JC Coles, JC Eilbeck, and AC Scott, Membrane conductances involved in amplification of small signals by sodium channels of drone honeybee, *J. Physiol. (London)* 456 (1992) 303–324.

[132] P Vetter, A Roth, and M Häusser, Propagation of action potentials in dendrites depends on dendritic morphology, *J. Neurophysiol.* 85 (2001) 926–937.

[133] SG Waxman, Regional differentiation of the axon, a review with special reference to the concept of the multiplex neuron, *Brain Res.* 47 (1972) 269–288.

[134] N Wiener, *Nonlinear Problems in Random Theory,* Technology Press, Cambridge, MA, 1958.

[135] HR Wilson, *Spikes, Decisions, and Actions: The Dynamical Foundations of Neuroscience,* Oxford University Press, Oxford, 1999.

[136] RKS Wong, DA Prince, and A Busbaum, Intradendritic recordings from hippocampal neurons, *Proc. Natl. Acad. Sci. USA* 76 (1979) 986–990.

[137] JZ Young, *Doubt and Certainty in Science,* Oxford University Press, Oxford, 1951.

[138] Y Zhou and J Bell, Study of propagation along nonuniform excitable fibers, *Math. Biosci.* 119 (1994) 169–203.

10
Constructive Brain Theories

Given some modest appreciation for the dynamics of individual neurons, it is natural to ask how they might act in concert, which is a central question of neuroscience. In accord with the constructive perspectives of modern science, this problem can be phrased: How does a collection of interacting neurons manage to behave like a brain? The aim of this chapter is to sketch some answers to this question.

We begin with a brief review of the first McCulloch–Pitts paper, in which brain modeling was approached by stripping real neurons down to their essential features: all-or-nothing response and a threshold for firing (see Section 2.4.1) [36]. Interestingly, this simple "M–P neuron" survives to the present day as a workhorse of neural network modeling.

McCulloch and Pitts also suggested the division of neural network models into two broad classes: "nets with circles" and "nets without circles." Their quaint jargon distinguishes between networks possessing internal feedback loops and those simpler networks for which the information flows in one direction only, from input to output terminals, providing a basis for organizing this chapter.

A key property of biological brains is their ability to learn, which was modeled in 1958 by Rosenblatt through a significant modification of the M–P neuron [44]. In a class of neural networks called the "perceptron," information is constrained to flow only in one direction (a net without circles), but the input weightings of each model neuron are allowed to change during the course of a *training period* as required by an appropriate *learning algorithm.*

Nets *with* circles are of central interest in neuroscience because biological brains—even those of the most simple creatures—do indeed have many internal loops of positive feedback threading through their constituent neurons. As we have seen in previous chapters of this book, such closed causal loops (or "re-entry") lead to the emergence of new dynamic entities, the nerve impulse being an outstanding example. With the emergence of novel coherent states arises the need for describing their dynamics, compounding the difficulties of mathematical formulation and analysis. Such matters are addressed in the following two chapters.

If each model neuron in a network is allowed to compute the most general Boolean function of its inputs, as suggested in the previous chapter, it is straightforward to compute the number of nets with circles that can be created from a given number of neurons and to sketch the various types of behavior. The number of such systems grows very rapidly with the number of constituent neurons, however, soon becoming unmanageable; thus, some guiding perspectives are needed.

As a simple brain model that includes closed loops of causal implication (positive feedback), Hopfield's "spin-glass" model is presented in the context of previously noted concepts of phase-space analysis of nonlinear systems [27]. The number of stable stationary states in this model is considered as an estimate for the information storage capacity of real brains.

The chapter closes with a brief introduction to cortical field theories, the dynamics of which are in accord with observations of Gestalt psychology, suggesting means for communication among the emergent states of real brains.

10.1 Nets Without Circles

In this section, attention is restricted to nets *without* circles for two reasons. First, it is evident that such network models are easier to analyze and understand just because they do not give rise to emergent entities. (There is an adage, no less true for being ancient, that one should learn to walk before trying to run.) Second, from a mathematical perspective, there are several rather simple results on the geometric interpretation of the pattern-classification problem and on procedures for learning that are of general interest and may play supporting roles in the information-processing activities of real brains.

Although it was proposed back in the 1950s that the trainable properties of nets without circles offer a basis for understanding the human brain [4, 44, 45], this view has not been widely held since the demise of behaviorism as a credible psychological theory. Nonetheless, nets without circles do comprise a class of *learning machines* that have been of engineering interest

since the late 1950s for a variety of tasks, including automatic sorting of photographs, converting handwritten characters to digitally defined letters, recognizing speech, generating suggestions for medical diagnoses, making weather predictions directly from atmospheric data, analyzing aerial photographs for economic data, and so on [23, 35].

However such systems fare in the realms of engineering, the peculiar properties of nets without circles may be employed for special purposes in certain restricted regions of the human brain, such as processing information on the way from the retina to the primary areas of the visual cortex or from the ears to the temporal lobes. Thus it seems prudent for neuroscientists to be aware of what nets without circles can do.

10.1.1 McCulloch–Pitts (M–P) Networks

In their 1943 paper, McCulloch and Pitts began by assuming a class of neural networks with the following properties: the activity of any constituent "neuron" is an all-or-nothing process; a fixed number of synapses must be stimulated within the period of latent addition in order to ignite a "neuron," and this number is independent of previous activity; the only significant delay occurs at synapses; ignition of a "neuron" is prevented by activation of a single inhibitory synapse; and the network structure does not change with time [36]. The term "neuron" is used here with quotation marks to emphasize that real neurons are more intricate than the model. Although this indication will be dropped in subsequent discussions, the reader should keep the caveat in mind.

McCulloch and Pitts were under no illusion that their assumptions are physiologically correct; indeed, they specifically mention that *facilitation and extinction* ("in which antecedent activity temporarily alters responsiveness to subsequent stimulation") and *learning* have been ignored. They defended their approach, however, as a way to establish baseline estimates of what neural networks can do.

A key aspect of the M–P formulation was their recognition that the all-or-nothing property of a neuron (an impulse is either present or it is not on a certain nerve at a certain time) can be viewed as a logical proposition (this statement is either true or false), so Boolean algebra (the algebra of classes) can be invoked to describe their model networks [3]. Thus they obtained two main results.

First, M–P showed that their model neuron could represent the three fundamental circuit elements of the computer engineer—the AND, OR, and NOT gates—which we met in the preceding chapter. Second, they appealed to the algebra of classes to show that any Boolean function can be modeled by one or more of their networks, and each such network corresponds to one or more Boolean functions. What is a Boolean function?

Written in the two-element number system "1" and "0" (which indicates that a statement is true or false or that an all-or-nothing impulse is present

or absent), the three basic operations of Boolean arithmetic are:

$$\begin{bmatrix} 1\text{ AND }1 &=& 1 \\ 1\text{ AND }0 &=& 0 \\ 0\text{ AND }1 &=& 0 \\ 0\text{ AND }0 &=& 0 \end{bmatrix}, \quad \begin{bmatrix} 1\text{ OR }1 &=& 1 \\ 1\text{ OR }0 &=& 1 \\ 0\text{ OR }1 &=& 1 \\ 0\text{ OR }0 &=& 0 \end{bmatrix}, \text{ and } \begin{bmatrix} \text{NOT }1 &=& 0 \\ \text{NOT }0 &=& 1 \end{bmatrix}.$$

In the context of this arithmetic, a Boolean function specifies the output variable for each combination of input variables. Thus a particular Boolean function of three inputs A, B, and C might be denoted as $F(A, B, C)$ and defined as in the following table.

A	B	C	$F(A,B,C)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

A Boolean expression for this particular function is

$$\begin{aligned} F(A,B,C) &= (A\text{ AND }B\text{ AND }C)\text{ OR}(A\text{ AND NOT }B\text{ AND }C) \\ &= A\text{ AND }C \end{aligned} \tag{10.1}$$

indicating in ordinary English that an output impulse will appear if either of two input conditions occurs: there are impulses at A, B, and C, or there are impulses at A and at C but not at B. In this formulation, "at" refers to a location in space-time because the AND operation requires temporal coincidence.

Because a Boolean function of N inputs has 2^N input combinations for which the corresponding output is either 0 or 1, there are evidently

$$2^{2^N}$$

distinct functions of N inputs. Each of these Boolean functions can be defined as in the preceding table and expressed as in Equation (10.1).

In retrospect, the demonstration by McCulloch and Pitts that any possible dependence on the output of a neural network can be realized (as engineers like to say) through a suitable combination of model neurons may seem modest. These results are now well known to computer engineers, and techniques for designing networks with a minimum number of switching elements (AND, OR, and NOT functions) have been available for decades [25]. In the early 1940s, however, engineers were striving to construct telephone switching stations with networks of magnetomechanical relays, and the modern digital computer was but a dream. In its day, therefore, the M–P paper was strikingly original.

More to the point in evaluating McCulloch–Pitts networks is the recognition that each nerve cell is modeled by a *single switch* represented by the Heaviside step function $H(I)$ in Equation (2.10), an assumption with two implications.

- This is a *convenient* assumption to make because the linear summation of input variables to the jth neuron

$$I_j = \sum_{k=1}^{N} \alpha_{jk} V_k(t) - \theta_j \tag{10.2}$$

  in Equation (2.10) keeps the threads of causality distinct, facilitating analysis of the system [7].

- In the context of neuroscience, however, it is a *dangerous* assumption because causal relations among input signals to real neurons are far more intricate than is indicated in Equation (10.2).

10.1.2 Learning Networks

Although M–P networks can "in principle" be arranged to do whatever can be done without circles, their design is not straightforward and requires selection of the weighting parameters α_{jk} and θ_j in Equation (10.2) for all neurons in the net. How might a neuron manage to solve this problem?

In 1958, Rosenblatt suggested that the α_{jk} and θ_j could be changed incrementally if a particular neuron is not responding correctly [44, 45]. His *training algorithm* led to a class of learning networks composed of M–P neurons with adjustable weights, which he called the *perceptron* [4, 5, 37].

At about the same time, an identical idea arose within the engineering community [23, 50]. Here, the class of networks was dubbed "ADALINE" (for ADAptive LInear NEtworks), and the constituent element was called a "linear threshold unit" (LTU). In this stream of activity, the aim was not to understand brain dynamics but to design computing machines that could be trained to recognize patterns in data sets.

To be specific, let us suppose that the Boolean function of Equation (10.1) is to be used for predicting the weather, where $A = 1$ indicates that

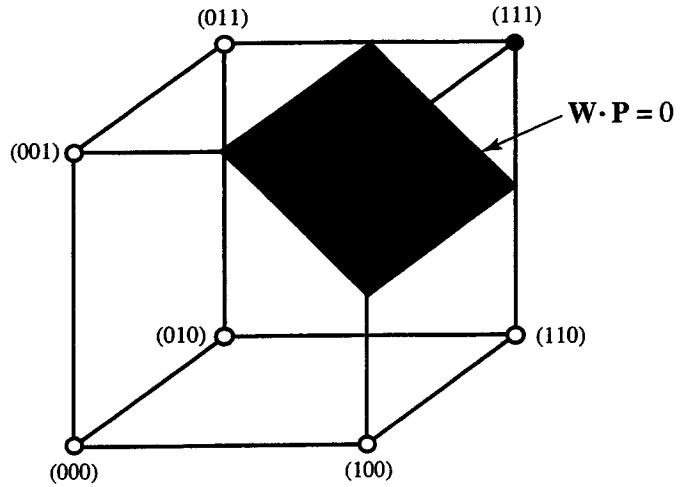


Figure 10.1. The geometrical interpretation of the pattern-recognition task indicated by Equations (10.1) and (10.3).

the barometer is rising and $A = 0$ that it is falling, $B = 1$ implies daytime and $B = 0$ night, and $C = 1$ indicates that it is clear and $C = 0$ indicates cloudiness. With $F(A, B, C)$ defined as in Equation (10.1), it is reasonable to expect that $F = 1$ implies that no rain is to be expected within the next few hours.

To understand how the training algorithm works, it helps to view pattern-recognition problems in a geometrical context. Thus, the eight values of these three input variables can be taken as vertices of a cube, as indicated in Figure 10.1, with the black dots indicating where $F = 1$ and the open dots where $F = 0$. The shaded area indicates a *linear discriminant plane* in pattern space on one side of which $F = 1$ and on the other $F = 0$.

Suppose that we wish to realize the logical function of Equation (10.1) with the M–P model neuron

$$\tilde{F} = H\left(\sum_{k=1}^{3} \alpha_k V_k(t) - \theta\right), \tag{10.3}$$

where $V_1 \equiv A$, $V_2 \equiv B$, and $V_3 \equiv C$. (Recall that $H(x)$ is the Heaviside step function, which equals 0 when x is negative and 1 otherwise.)

Two questions arise: (1) How do we choose α_1, α_2, α_3, and θ? (2) If these weighting parameters are incorrectly chosen, how can they be altered so that the functions computed from Equations (10.1) and (10.3) are the same?

To answer these questions, it is convenient to define a four-dimensional *weight vector* as

$$\mathbf{W} \equiv (\alpha_1, \alpha_2, \alpha_3, -\theta)$$

and a four-dimensional *augmented pattern vector* as

$$\mathbf{P} \equiv (V_1, V_2, V_3, 1)\,.$$

Then the inner product of the weight vector and the augmented pattern vector,[1]

$$\mathbf{W} \cdot \mathbf{P} = \sum_{k=1}^{3} \alpha_k V_k(t) - \theta\,,$$

is just the argument of the Heaviside step function in Equation (10.3). Thus to realize the Boolean expression of Equation (10.1) with the M–P neuron of Equation (10.3), it suffices to choose the three α_js and θ so that the condition

$$\mathbf{W} \cdot \mathbf{P} = 0$$

corresponds to a discriminant plane lying between the vertices where $F = 1$ (the dark circles) and those where $F = 0$ (the open circles), as shown in Figure 10.1. This answers question (1).

To answer question (2), suppose that we have mistakenly chosen the components of the weight vector ($\mathbf{W_1}$) such that

$$\mathbf{W_1} \cdot \mathbf{P} < 0$$

for (say)

$$\mathbf{P} = (1, 1, 1, 1)\,,$$

but all of the other vertices in Figure 10.1 lie on the correct side of the discriminant plane. Then Equation (10.3) tells us that $\tilde{F} = 0$ for $V_1 = V_2 = V_3 = 1$. In other words, if the barometer is rising, it is daytime, and the sky is not cloudy, we should expect rain. Clearly, this is not a correct prediction and the weight vector must be changed, but how?

If the weight vector were altered by adding an increment in a direction orthogonal (at right angles) to $\mathbf{P}$, the inner product $\mathbf{W} \cdot \mathbf{P}$ would not change; thus, it is necessary to alter the weight vector in the direction of $\mathbf{P}$. To accomplish this, assume

$$\mathbf{W_2} = \mathbf{W_1} + c\mathbf{P}\,, \tag{10.4}$$

where c is a positive real constant that must be determined. Taking the inner product of both sides of Equation (10.4) with $\mathbf{P}$ and requiring that $\mathbf{W_2} \cdot \mathbf{P} > 0$ shows that for

$$c > -\frac{\mathbf{W_1} \cdot \mathbf{P}}{\mathbf{P} \cdot \mathbf{P}} \tag{10.5}$$

the inner product $\mathbf{W_2} \cdot \mathbf{P} > 0$.

[1]The inner (or "dot") product of two vectors is the sum of the products of their components.

If $\mathbf{W_1} \cdot \mathbf{P} > 0$ gives an incorrect result for some $\mathbf{P}$, on the other hand, it is necessary to decrease $\mathbf{W} \cdot \mathbf{P}$, so making

$$c < -\frac{\mathbf{W_1} \cdot \mathbf{P}}{\mathbf{P} \cdot \mathbf{P}} \tag{10.6}$$

will give the correct response.

When the inequality in Equation (10.5) or (10.6) is barely satisfied, then the weight vector has been readjusted with a minimum of change. In our example of the weather predictor, this ensures that $\tilde{F} = F$ for $V_1 \equiv A = 1$, $V_2 \equiv B = 1$, and $V_3 \equiv C = 1$. Because these changes in the weight vector may have caused some of the other inputs to give erroneous results, it is necessary to check all of the other input conditions and make minimal corrections corresponding to Equations (10.4) and (10.5) wherever necessary.

In more general cases, it may be that no discriminant plane exists, for example, with a function defined as 1 at two diagonally opposite vertices—(000) and (111)—and 0 otherwise. If such cases are excluded, the Boolean function is said to be *linearly separable.*

In other words, an M–P (or LTU) representation for a linearly separable Boolean function exists by definition, leading to the following theorem.

> **Training theorem:** If a Boolean function (F) is linearly separable and the weight vectors of an M–P neuron ($\tilde{F}$) are successively modified as indicated in Equations (10.4) and (10.5) or (10.6), then the sequence
>
> $$\mathbf{W_1} \rightarrow \mathbf{W_2} \rightarrow \mathbf{W_3} \rightarrow \mathbf{W_4} \rightarrow \cdots$$
>
> converges in a finite number of steps to a weight vector for which $\tilde{F}$ is identical to F [37, 38].

This result is biologically interesting because the information needed to make such successive weight modifications is just what a neuron has available at the tips of its dendrites. From Equations (10.5) and (10.6), this information comprises the current values of the synaptic strengths and the threshold (given by $\mathbf{W_j}$) and the current values of the input signals (given by $\mathbf{P}$).

Put differently, if a neuron were informed that its response to a particular pattern is undesired, it could correct that behavior by increasing or decreasing its synaptic weights in amounts proportional to the current input signals, which suggests the following question for neuroscientists: Are there biologically credible means through which a real neuron might come to know that a certain response is unwanted by its organism?

Table 10.1. The number of Boolean networks ($\mathcal{N}$) for various numbers of switches (N).

N	$\mathcal{N} = 2^{N2^N}$
1	$2^2 = 4$
2	$4^4 = 256$
3	$8^8 \doteq 1.7 \times 10^7$
4	$16^{16} \doteq 1.8 \times 10^{19}$
5	$32^{32} \doteq 1.5 \times 10^{48}$
6	$64^{64} \doteq 3.9 \times 10^{115}$

10.2 Nets with Circles

Because the human brain is threaded through with myriad closed loops of causal implication, any serious study of its dynamics must deal with the many new entities that emerge. This section presents two constructive theories of such networks. The first indicates the degree of intricacy to be expected, and the second suggests ways in which methods of statistical physics may lead to understanding.

10.2.1 General Boolean Networks

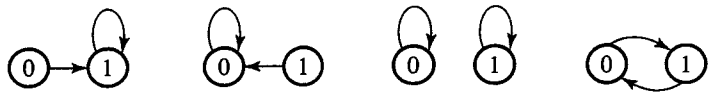
Let us begin by imagining the most general class of networks that can be constructed from N model neurons (or switches), each of which is allowed to compute an arbitrary Boolean function of its N inputs. Because there are

$$2^{2^N}$$

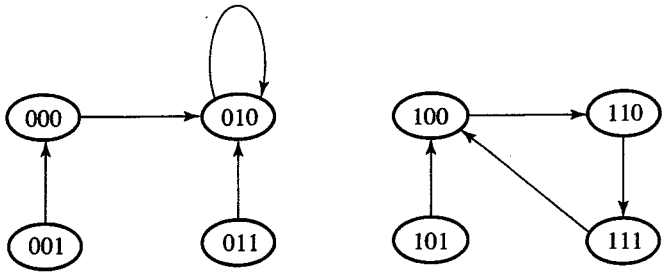
Boolean functions of N inputs and each of the N neurons is chosen to be one of these, there are

$$\mathcal{N} = \left(2^{2^N}\right)^N = 2^{N2^N}$$

different systems in this class of general Boolean networks. For modest numbers of neurons, the number of possible systems soon becomes very large, as is seen from Table 10.1. To deal with such large numbers, combinatoric mathematicians have whimsically defined the *googol* $\equiv 10^{100}$ as a finite number above which arithmetic becomes problematic [11]. To see why



(a)



(b)

Figure 10.2. (a) The four Boolean systems that can be constructed from a single switch. (b) One of the more than sixteen million systems that can be constructed from three switches.

they have introduced such a definition, let us go down the table, making references along the way to Figure 10.2.

In this figure are indicated *state diagrams* of certain Boolean systems. Each such diagram shows the 2^N states of some system, which advances from one state to the next in a fixed time interval. Each state, therefore, has one outgoing arrow, indicating which of the 2^N states the system will go to in the next increment of time. Thus, the 2^N arrows can be chosen in 2^N different ways, leading once again to

$$\mathcal{N} = \left(2^N\right)^{2^N} = 2^{N2^N}$$

different systems composed of N switches.

For $N = 1$, there are $2^2 = 4$ Boolean systems that can be constructed from a single switch. These four systems are shown in Figure 10.2(a), where "0" indicates that the switch is off and "1" implies on. Reading from left to right, the first system turns on if it is off and then stays on. The second system turns off if it is on and then stays off. The third system stays off if it is off and stays on if it is on, as expected for a light switch. Finally, the last system—which electrical engineers call a "free running multivibrator"—turns off if it is on and turns on if it is off, thereby generating a periodic signal.

For $N = 2$, implying two switches or neurons, there are $4^4 = 256$ possible Boolean systems, which could be worked out in a few hours. For $N = 3$, this number has increased to $8^8 = 16,777,216$ systems, one of which is shown in Figure 10.2(b). Presumably, a computer code could be written to generate diagrams for all of these systems. For $N = 4$ and 5, it would be imprudent to attempt such a code.

For $N \geq 6$, interestingly, it can be asserted with confidence that no computing system will ever be constructed that generates and records the diagrams for all possible systems. This is because the atomic weight (or the total number of protons and neutrons) of the universe is only about 10^{80}; thus there is not enough paper—or memory storage of any sort—for the task.

To emphasize the importance of this point in theoretical biology, physicist Walter Elsasser has proposed that the term *immense* be used to describe finite numbers larger than a googol [12]. Typically, the number of possible members of a biological species is immense, whereas the number of actual members—past, present, and future—is not.

Dealing with sets in which the number of possible members is much larger than the number of actual members, Elsasser suggests, helps to make the biological and social sciences fundamentally different from the physical sciences. Thus, a physicist studying (say) hydrogen can perform as many experiments as desired on identical atoms, leading to generalizations formulated as laws of physics. Similarly, the chemist can study as many (say) benzene molecules as are needed to formulate reliable laws of chemistry.

The biologist or psychologist studying (say) *Homo sapiens*, on the other hand, faces quite a different challenge. There are a great many more possible humans than will ever actually exist—past, present, or future. To see this, note first that the total number of actual human beings is certainly less than a googol (not immense). Then, consider what $\mathcal{N}$ would be in Table 10.1 if N were anything like the 10^{10} neurons in a human brain. It is clear from such a comparison that psychological observations are necessarily made on very limited subsets of the possible members of our species.[2]

In Elsasser's terminology, biologists, psychologists, and anthropologists study *heterogeneous sets*, the members of which exhibit substantial differences. (We have seen an example of biological heterogeneity in Table 4.1 showing the variability of membrane data on giant axons of the squid.) Physicists and chemists, on the other hand, deal with *homogeneous sets*, for which members are essentially identical. To emphasize this distinction, consider the case of alloys.

The conducting wires of electric circuits, for example, are often connected together by an alloy of tin and lead called solder, for which the melting point depends on the ratio of the components. Given two small

[2]It may be objected that the brain's neurons do not compute arbitrary Boolean functions of their inputs, so the estimate of $\mathcal{N}$ may be high. From work of Yajima et al. [54], the number of systems composed of N threshold (M–P) neurons with n inputs each is greater than $2^{n^2N/2}$, but this estimate is low. Recently, Poirazi and Mel [42] have studied memory capacity of systems of model neurons in which individual dendritic trees respond nonlinearly to their inputs. From both combinatoric and numerical calculations, these authors find that capacities of nonlinear neurons exceed those of linear models by "orders of magnitude."

blobs of solder with the same tin-to-lead ratio, the detailed arrangements of the tin and lead atoms will differ, and it is not difficult to show that the total number of possible arrangements is immense. Nonetheless, the average properties of solder (melting point, electrical resistivity, specific heat, ductility, and so on) are almost exactly the same for most arrangements; thus all possible solder blobs form a homogeneous set, falling comfortably within the purview of physical science.

In a heterogeneous set, on the other hand, small variations count, so members have very different global properties. Thinking about it, there are many heterogeneous sets in our ordinary experience, including the number of possible natural languages, protein molecules, musical compositions, English sonnets, chess games, people, and so on. Thus there will always be many different languages, useful proteins, beautiful melodies and poems, interesting chess games, and exciting human personalities that have not been realized and never will be.

10.2.2 Attractor Neural Networks

Although the large numbers of the previous section may discourage some who would develop a constructive theory of the brain's dynamics, one should not give up altogether. With reference to the state diagram of Figure 10.2(b), three qualitatively different sorts of behavior are observed.

(1) First, we note *transients* such as

$$(001) \rightarrow (000) \rightarrow (010)$$

in which the system passes through a sequence of states, never to return.

(2) Amid these transients emerge *stable attractors* such as (010) with the *basin of attraction* (000), (001), and (011).

(3) Finally, there are *limit cycles* such as

$$(100) \rightarrow (110) \rightarrow (111) \rightarrow (100) \rightarrow (110) \rightarrow (111) \rightarrow \cdots$$

having a period of three time units and the basin of attraction (101).

It comes as no surprise to find emergent entities in switching networks; such behavior arises directly from the positive feedback associated with closed causal loops, as we have often seen. The problem with nets comprising a realistic number of neurons is that there is no direct way of accounting for all of the emerging states. Some limit cycles may be as short as one time unit; others might approach 2^N time units, snaking through most states of the system, and one cannot proceed by looking at all possible systems.

In 1982, Hopfield made some progress in circumventing such combinatoric difficulties by assuming a modified version of the neural network we have previously considered [27]. Motivated by the McCulloch–Pitts formulation of Equation (2.10), his basic "neuron" obeys the dynamic

equation

$$s_j(t+\tilde{\tau}) = \text{sign}\left(\sum_{k=1}^{N} J_{jk}\, s_k(t)\right), \tag{10.7}$$

where $J_{kk} = 0$ and sign(·) is the "sign" function with properties

$$\text{sign}(y) = \begin{cases} +1 \quad \text{for} \quad y \geq 0 \quad \text{and} \\ -1 \quad \text{for} \quad y < 0\,, \end{cases}$$

and the "spin" variables s_j take the values +1 and −1 instead of the Boolean numbers 1 and 0.

In comparing Hopfield's *attractor neural network* with the McCulloch–Pitts (M–P) model, the following points of difference should be noted.[3]

- In M–P, the variables V_j (with $j = 1, 2, \ldots, N$) indicate the instantaneous voltages of the N neurons in the network, whereas the s_j in an attractor neural network represent the average firing rate of the jth neuron measured on a linear scale from *quiescent* (−1) to *fully active* (+1).
- In an attractor neural network, the states of individual switches are not changed all at once but successively altered in a randomly selected order. Thus $\tau = N\tilde{\tau}$ is the time required for updating average firing rates for the entire net, corresponding to about 1–2 s in the human brain.
- The attractor neural network threshold is assumed to be the same for all N model neurons, which are joined by a symmetric $N \times N$ interconnection matrix

  $$J = [J_{jk}]\,,$$

  where $J_{kk} = 0$ and $J_{jk} = J_{kj}$. (This symmetry condition means that the coupling from neuron k to neuron j is equal to the coupling from neuron j to neuron k, which is not neurologically realistic.)
- When numerically convenient, the "sign(·)" function in Equation (10.7) can be replaced by a *sigmoid* function, qualitatively like tanh(·), which rises in a monotone manner from −1 to +1 as x increases from $-\infty$ to $+\infty$ [28].

This attractor neural network is convenient for analysis because it has a *Lyapunov functional* (E) possessing the following pair of properties [29, 32]:

[3] Although attractor neural networks are sometimes referred to as "ANNs" [1], others use these initials for the broader class of "artificial neural networks"; thus the acronym is avoided here.

first, E must be bounded from below, and second, E must either decrease or remain constant with each time step.[4]

For the dynamics described by Equation (10.7), a Lyapunov functional is

$$E = -\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} J_{jk}\, s_j s_k\,, \tag{10.8}$$

with the proof as follows.

(1) Note first that for finite J_{jk}, E is evidently bounded from below.

(2) To see how E changes under the dynamics of Equation (10.7), suppose first that the ith switch changes from $s_i = +1$ to -1. The corresponding change in E is

$$\begin{aligned}\Delta E &= s_i \sum_{j=1}^{N} J_{ij}\, s_j + s_i \sum_{j=1}^{N} J_{ji}\, s_j \\ &= 2\sum_{j=1}^{N} J_{ij}\, s_j\end{aligned}$$

because $J_{ij} = J_{ji}$. Because the summation must be negative for s_i to decrease, ΔE is negative.

(3) Next suppose that the ith switch changes from $s_i = -1$ to $+1$. The corresponding change in E is

$$\Delta E = -2\sum_{j=1}^{N} J_{ij} s_j\,.$$

Because the summation must be positive for s_i to increase, ΔE is again negative.

(4) Finally, if s_i does not change, ΔE is zero.

Because E is bounded from below and ΔE is either negative or zero after each time step ($\tilde{\tau}$), E eventually ceases to decrease, and the system either moves around on a limit cycle at constant E or sits at a stable attractor, which may be found numerically. In Hopfield's formulation, each stable attractor is viewed as a *pattern* stored nonlocally by the net.

Each such pattern will have a basin of attraction into which the system can be forced by sensory inputs. In other words, if external stimulations nudge an attractor neural network into the basin of attraction for one of its patterns, the system will move to that attractor and remain there, providing

[4]Further motivation for Hopfield's model is the fact that Equation (10.8) is an expression for the total energy of interacting magnets (or atomic spins), which have been of interest to physicists for decades. Thus several standard results from condensed-matter physics translate directly into statements concerning the dynamics of nerve systems, drawing physical scientists into neuroscience [1, 26].

a model for the brain's ability to recall intricate memory patterns under the influence of sensory information.

Assume a pattern of the form $\mathbf{X_m} = (x_1^m, x_2^m, \ldots, x_N^m)$, where the components are either $+1$ or -1 with equal probability. To learn this pattern, the interconnection matrix can then be constructed by the rules

$$\begin{aligned}\Delta J_{ij} &\propto x_i^m x_j^m\,, \\ \Delta J_{ij} &= 0\,,\end{aligned}$$

which has the effect of increasing interconnection strengths where both neurons are in the same state and reducing them where the states are different. If p patterns are learned in this manner,

$$J_{ij} = \frac{1}{N}\sum_{m=1}^{p} x_i^m x_j^m\,,$$

where the normalization by N is chosen to keep the components of J at a uniform level as the number of patterns is increased.[5]

How many such patterns can be stored by this system?

Supposing that p patterns are randomly chosen, each will seem like noise to the others. Every time a new pattern is added to the store, in other words, the elements of J are adjusted, effectively introducing noise into the task of recovering formerly stored patterns. Because the elements of J are normalized to N, the total noise amplitude seen by each stored pattern will grow as $\sqrt{p/N}$.

If p remains constant while $N \to \infty$, the noise amplitude goes to zero and the storage system works well. With N held constant while p is increased, on the other hand, a maximum number of patterns (p_{m}) is eventually reached that is proportional to N [1, 26].

Allowing the network to have stable patterns close to those learned (where "close" means that no more than 1% of the neurons deviate), it appears from a combination of theoretical arguments and numerical evidence that

$$p_{\mathrm{m}} \approx 0.14\, N\,. \tag{10.9}$$

[5]For example, if $\mathbf{X} = (+1, -1, +1, -1)$, then

$$\Delta J = \begin{bmatrix} 0 & -1 & +1 & -1 \\ -1 & 0 & -1 & +1 \\ +1 & -1 & 0 & -1 \\ -1 & +1 & -1 & 0 \end{bmatrix},$$

and $\Delta J X \propto X$, which is a stationary point of Equation (10.7).

Interestingly, this is a sharp boundary. If one attempts to store more than the critical number of patterns, the probability of retrieval falls rapidly to zero.

Assuming this spin-glass model of the brain bears some relation to neurological reality, a human neocortex of 10^{10} to 10^{11} neurons might be expected to store something like 10^9 to 10^{10} intricate patterns. We will consider another derivation of this important number in the following chapter.

10.3 Field Theories for the Neocortex

Another way to deal with the immense number of possible systems into which the brain's neurons may become organized is to develop a *field theory* for neural activity. Such an approach was first proposed in 1956 by Beurle [2], who assumed that the neural mass of the neocortex can be locally described by the fraction $F(x,t)$ of cells at position x that are firing at time t, and the probability $p(x)$ of two cells being interconnected is an exponentially decreasing function of the distance between them [47]; thus,

$$p(x) \propto e^{-|x|/\sigma}.$$

In this theory, activity at a particular region of the cortex induces activity at neighboring regions, which leads to a *wave of information* propagating through the neural mass. Because Beurle supposed that all of the cortical neurons are excitatory, the waves described by his theory correspond roughly to the leading-edge formulations of Chapter 5, albeit with the activity averaged over many neurons rather than localized on a single fiber. Salient features of his study include the following.

- A wave of information may involve the activity of only a small fraction of the local neurons, allowing waves to pass through each other with little interference. Thus many different messages may propagate throughout the neocortex and carry information from one region to another.[6]
- If the neural interconnections (synapses) are supposed to increase in strength upon exposure to the activity of a particular wave (a learning mechanism), one can imagine a means for holographic-like recall [2, 6, 16, 17, 34, 43]. Thus, waves induced by external (sensory) stimulation would become coupled to subsequent internal waves, leading to the possibility that but a fragment of the original stimulation is required to trigger the related internal response.

[6]That cortical information waves pass through one another leads some to confuse them with *solitons*, but the two phenomena are quite different. Solitons conserve energy, whereas waves of information do not, being dynamically akin to waves of activity in the heart [46].

- Although neuroscientists often fret over the "binding problem" of relating activities in different parts of the cortex—combining, for example, the voice, image, and personality of a friend into a single perception—Beurle's information waves may provide a means of achieving such coupling.

Following Beurle's lead, Griffith (among others) developed a field theory of neural activity in which time and space dependencies are brought in through their lowest derivatives [19, 20, 21, 24, 39, 40, 41]. In this theory, the probability of a neuron firing in the next time interval (S) is a sigmoid function of the present firing rate (F), say

$$S(F) = \frac{F^2}{F^2 + \theta^2},$$

with S and F necessarily positive and θ a threshold parameter.

Thus, without spatial variations (i.e., "space-clamped"), the first time derivative can be approximated as

$$\frac{dF}{dt} \approx \frac{S(F) - F}{\tau},$$

implying

$$\frac{dF}{dt} \approx -f(F),$$

where

$$f(F) = -\frac{F^3 - F^2 + \theta^2 F}{\tau(F^2 + \theta^2)}$$

is a cubic function qualitatively like those sketched in Figures 5.2 and 5.3.

To introduce spatial dependence, Griffith reasoned that the connectivity would be the same in both the $+x$ and the $-x$ directions; thus, the lowest space derivative is the second, implying a nonlinear diffusion equation

$$D\frac{\partial^2 F}{\partial x^2} - \frac{\partial F}{\partial t} = f(F). \tag{10.10}$$

For a mean interconnection distance indicated by σ and a neural response time of τ, the diffusion constant is of order

$$D \sim \frac{\sigma^2}{\tau}.$$

Although Equation (10.10) is formally identical to Equation (5.5), its interpretation is quite different. Equation (5.5) describes the leading edge of a nerve impulse, traveling along an unmyelinated axon, whereas Equation (10.10) represents a wave of activity propagating through a neural medium such as the neocortex.

To bring *recovery* into the picture, Wilson and Cowan took advantage of the fact that some of the neocortical neurons are inhibitory; thus, they developed a theory in two dependent variables [51, 52]:

- $E(x,t)$: the fraction of excitatory neurons that are firing as a function of x and t, and
- $I(x,t)$: the corresponding fraction of inhibitory neurons,

which are assumed to interact as [53]

$$\tau\frac{dE}{dt} = S_E\left(\int[w_{EE}(x,x')E(x') - w_{IE}(x,x')I(x')]dx' + P(x,t)\right) - E\,, \tag{10.11}$$

$$\tau\frac{dI}{dt} = S_I\left(\int[w_{EI}(x,x')E(x') - w_{II}(x,x')I(x')]dx' + Q(x,t)\right) - I\,.$$

In these coupled integro-differential equations, the nonlinearity of neural response is introduced through the sigmoid functions

$$S_E(y) \equiv \frac{100y^2}{\theta_E^2 + y^2} \text{ and } S_I(y) \equiv \frac{100y^2}{\theta_I^2 + y^2}\,,$$

diffusion stems from the interconnection probabilities between neurons at x and x',

$$w_{ij}(x,x') = b_{ij}\exp\left(-\frac{|x-x'|}{\sigma_{ij}}\right)\,,$$

and external (sensory) inputs to the excitatory and inhibitory cells are represented by $P(x,t)$ and $Q(x.t)$, respectively.

It is, of course, difficult to fix the many parameters of such a model, but Wilson suggests the following values as reasonable "guesstimates" for the human neocortex: $\sigma_{EE} = 40\,\mu$m, $\sigma_{EI} = \sigma_{IE} = 60\,\mu$m, $\sigma_{II} = 30\,\mu$m, $\theta_E = 20$, and $\theta_I = 40$, and he has made available several MATLAB codes for exploring the resulting dynamics [53]. Those familiar with MATLAB are encouraged to play with these codes and explore the following spectrum of behaviors.

- *Stationary patterns of activity:* With $b_{EE} = 1.95$, $b_{IE} = b_{EI} = 1.4$, and $b_{II} = 2.2$, a short pulse of stimulation ($P = 1.0$ for 10 ms over a range of 100 μm, and $Q = 0$) induces a stationary pattern in which the longer-ranging inhibitory activity surrounds and contains the more localized excitatory activity. Qualitatively, this behavior is similar to *Turing patterns*, which are found in studies of nonlinear reaction-diffusion systems of more than one dimension [46].
- *Transient activity:* With $b_{EE} = 1.5$, $b_{IE} = b_{EI} = 1.3$, and $b_{II} = 1.5$, a brief stimulation ($P = 2.0$ over 5 μm for 5 ms, and $Q = 0$) causes a transient response, with E rising to a maximum value of about 28 in about 30 ms and then relaxing back to zero.
- *Localized oscillations:* With $b_{EE} = 1.9$, $b_{IE} = b_{EI} = 1.5$, and $b_{II} = 1.5$, a constant stimulation over a spatial range of 100 μm or more results in a variety of spatially localized oscillations.

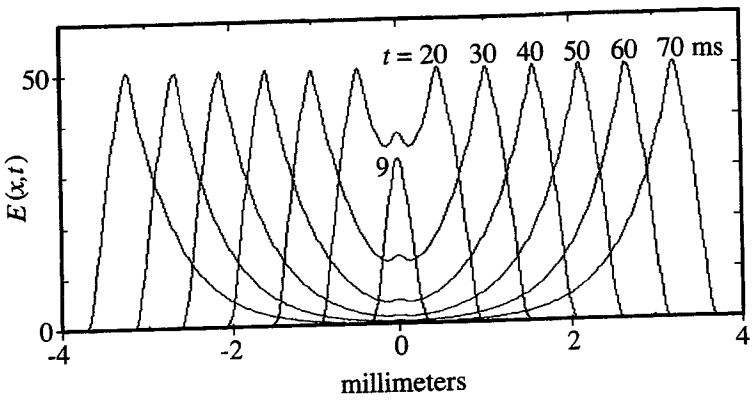


Figure 10.3. Outgoing wave solutions of the Wilson–Cowan equations (10.11) generated by a brief pulse of excitation near the origin.

- *Waves of activity:* With $b_{EE} = 1.9$, $b_{IE} = b_{EI} = 1.5$, and $b_{II} = 1.5$, and a strongly inhibitory input ($Q = -90$ while P is applied briefly over 100 μm), subsequent outgoing waves of activity are shown in Figure 10.3. Traveling at a speed of about 0.06 mm/ms, these waves are qualitatively similar to the impulses with recovery that were discussed in Chapter 6 for single fibers.

In his book, Wilson offers many more examples of such dynamics, discussing ways in which neural field theories similar to Equations (10.11) can model a variety of mental phenomena, including phase transitions, hallucinations, and epileptic seizures [8, 13, 15, 18, 22, 30, 31, 48, 53]. Recently, Ermentrout and Kleinfeld have developed a simple model of cortical wave motion through a network of weakly coupled oscillators in which only the phase of the oscillators is influenced by interactions [14].

On a much longer time scale, note that nonlinear field effects may also play a role in the development of mesoscopic cortical structure along lines suggested by Alan Turing in 1952 [49]. Such structure occurs in the visual cortices of mammals, where alternating bands of neurons, receiving inputs dominated by one or the other eye, are seen. Starting from this observation, Martha Constantine-Paton and Margaret Law have used experiments on "three-eyed frogs" to show that these cortical "stripes" are a form of "Turing pattern" [9, 10, 33]. Interestingly, the active nonlinearity driving the pattern-formation process stems not from biochemistry as was originally proposed by Turing. Rather, it is a positive feedback phenomenon with a closed causal loop having the structure

Increased spiking activity
$\downarrow \quad \uparrow$
Dendritic strength

Thus, the basic driving force is neural activity.

The idea that spiking activity causes dendritic strengths to increase is (somewhat incorrectly) associated with psychologist Donald Hebb, whose work is the subject of the following chapter.

10.4 Recapitulation

The aim of this chapter has been to introduce certain key threads of development in neural network theories beginning with the seminal work of McCulloch and Pitts and the subsequent development of learning machines. Although geometrical ideas help us to understand the fundamental learning theorem for such systems, the immense number of possible neural arrangements precludes exhaustive searches of particular networks.

To deal with this difficulty, physicists have proposed "spin-glass" models of the brain in which neural behavior is idealized in order to obtain a Lyapunov (or energy) functional governing global dynamics. Following this approach, it has been estimated that the number of complex concepts that the human brain can store is of the order of 10^9 to 10^{10}.

Finally, some nonlinear field theories for cortical dynamics were sketched that display a wide variety of qualitative behaviors, including stationary patterns, transients, localized oscillations, and waves of information. Although it is difficult to fix the parameters of these nonlinear diffusion models with precision from neurological data, they can be studied numerically with currently available tools.

References

[1] DJ Amit, *Modeling Brain Function: The World of Attractor Neural Networks,* Cambridge University Press, Cambridge, 1989.

[2] RL Beurle, Properties of a mass of cells capable of regenerating pulses, *Philos. Trans. R. Soc. London* A240 (1956) 55–94.

[3] G Birkhoff and S MacLane, *A Survey of Modern Algebra,* Macmillan, New York, 1953.

[4] HD Block, The Perceptron: A model for brain functioning, *Rev. Mod. Phys.* 34 (1962) 123–135.

[5] HD Block, BW Knight, Jr, and F Rosenblatt, Analysis of a four-layer series coupled Perceptron, *Rev. Mod. Phys.* 34 (1962) 135–142.

[6] A Borsellino and T Poggio, Holographic aspects of temporal memory and optomotor responses, *Kybernetik* 10 (1972) 58–60.

[7] M Bunge, *Causality and Modern Science,* third revised edition, Dover, New York, 1979.

[8] PS Churchland and TJ Sejnowski, *The Computational Brain,* MIT Press, Cambridge, MA, 1994.

[9] M Constantine-Paton and MI Law, Eye-specific termination bands in tecta of three-eyed frogs, *Science* 202 (1978) 639–641.

[10] M Constantine-Paton and MI Law, The development of maps and stripes in the brain, *Sci. Am.* December 1982.

[11] RE Crandall, The challenge of large numbers, *Sci. Am.* February 1997, 72–78.

[12] WM Elsasser, *Reflections on a Theory of Organisms: Holism in Biology,* The Johns Hopkins University Press, Baltimore, 1998 (first published in 1987).

[13] GB Ermentrout and JD Cowan, A mathematical theory of visual hallucination patterns, *Biol. Cybern.* 34 (1979) 137–150.

[14] GB Ermentrout and D Kleinfeld, Traveling electrical waves in cortex: Insights from phase dynamics and speculation on a computational role, *Neuron* 29 (2001) 33–44.

[15] A Fuchs, JAS Kelso, and H Haken, Phase transitions in the human brain: Spatial mode dynamics, *Int. J. Bifurcation Chaos* 2 (1992) 917–939.

[16] D Gabor, Holographic model of temporal recall, *Nature* 217 (1968) 584.

[17] D Gabor, Improved holographic model of temporal recall, *Nature* 217 (1968) 1288.

[18] J Glanz, Mastering the nonlinear brain, *Science* 277 (1997) 1758–1760.

[19] JS Griffith, A field theory of neural nets: I. Derivation of field equations, *Bull. Math. Biophys.* 25 (1963) 187–195.

[20] JS Griffith, A field theory of neural nets: II. Properties of field equations, *Bull. Math. Biophys.* 27 (1965) 111–120.

[21] JS Griffith, *Mathematical Neurobiology: An Introduction to the Mathematics of the Nervous System,* Academic Press, New York, 1971.

[22] H Haken, *Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior and Cognition,* Springer-Verlag, Berlin, 1996.

[23] J Hawkins, Self-organizing systems: A review and commentary, *Proc. IRE* 49 (1961) 31–48.

[24] CE Hendrix, Transmission of electric fields in cortical tissue: A model for the origin of the alpha rhythm, *Bull. Math. Biophys.* 27 (1965) 197–213.

[25] FC Hennie, *Finite-State Models for Logical Machines,* John Wiley & Sons, New York, 1968.

[26] J Hertz, A Krogh, and RG Palmer, *Introduction to Neural Computation,* Addison-Wesley, Reading, MA, 1991.

[27] JJ Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci. (USA)* 79 (1982) 2554–2558.

[28] JJ Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA* 81 (1984) 3088–3092.

[29] EA Jackson, *Perspectives of Nonlinear Dynamics,* Cambridge University Press, Cambridge, 1990.

[30] VK Jirsa, R Friedrich, H Haken, and JAS Kelso, A theoretical model of phase transitions in the human brain, *Biol. Cybern.* 71 (1994) 27–35.

[31] JAS Kelso, SL Bressler, S Buchanan, GC De Guzman, A Fuchs, and T Holroyd, A phase transition in human brain and behavior, *Phys. Lett. A* 169 (1992) 134–144.

[32] J La Salle and S Lefschetz, *Stability by Liapunov's Direct Method,* Academic Press, New York, 1961.

[33] MI Law and M Constantine-Paton, Right and left eye bands in frogs with unilateral tectal ablations. *Proc. Natl. Acad. Sci. USA* 77 (1980) 2314–2318.

[34] HC Longuet-Higgins, Holographic model of temporal recall, *Nature* 217 (1968) 104.

[35] K Mainzer, *Thinking in Complexity: The Complex Dynamics of Matter, Mind, and Mankind,* Springer-Verlag, Berlin, 1994.

[36] WS McCulloch and WH Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* 5 (1943) 115–133.

[37] M Minsky and S Papert, *Perceptrons,* MIT Press, Cambridge, MA, 1969.

[38] NJ Nilsson, *Learning Machines: Foundations of Trainable Pattern-classifying Systems,* second edition, Morgan Kaufmann, San Mateo, CA, 1990.

[39] PL Nuñez, The brain wave equation: A model for the EEG, *Math. Biosci.* 21 (1974) 279–297.

[40] PL Nuñez, Wave-like properties of the alpha rhythm, *Trans. IEEE Biomed. Eng.* BME-21 (1974) 473–482.

[41] PL Nuñez, *Electric Fields of the Brain: The Neurophysics of EEG,* Oxford University Press, New York, 1981.

[42] P Poirazi and BW Mel, Impact of active dendrites and structural plasticity on the memory capacity of neural tissue, *Neuron* 29 (2001) 779–796.

[43] KH Pribram, The neurophysiology of remembering, *Sci. Am.,* January 1969, 73–85.

[44] F Rosenblatt, The Perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.* 65 (1958) 386–408.

[45] F Rosenblatt, *Principles of Neurodynamics,* Spartan Books, New York, 1962.

[46] AC Scott, *Nonlinear Science: Emergence and Dynamics of Coherent Structures,* Oxford University Press, Oxford, 1999.

[47] DA Sholl, *The Organization of the Cerebral Cortex,* Methuen, London, 1956.

[48] P Tass, Cortical pattern formation during visual hallucinations, *J. Biol. Phys.* 21 (1995) 177–210.

[49] AM Turing, The chemical basis of morphogenesis, *Philos. Trans. R. Soc. London* B237 (1952) 37–72.

[50] B Widrow and JB Angell, Reliable, trainable networks for computing and control, *Aerospace Eng.* September, 1962, 78–123.

[51] HR Wilson and JD Cowan, Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.* 12 (1972) 1–24.

[52] HR Wilson and JD Cowan, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik* 13 (1973) 55–80.

[53] HR Wilson, *Spikes, Decisions, and Actions: The Dynamical Foundations of Neuroscience,* Oxford University Press, Oxford, 1999.

[54] S Yajima, T Ibaraki, and I Kawano, On autonomous logic nets of threshold computers, *Trans. IEEE on Comp.* 17 (1968) 385–391.

11
Neuronal Assemblies

Although the suggestion that neurons in the human brain may act in functional groups reaches back at least to the beginning of the twentieth century (when Charles Sherrington published his *The Integrative Action of the Nervous System* [85]), it was in Donald Hebb's classic *Organization of Behavior* that the cell-assembly concept was first carefully formulated. Largely neglected for several decades [13], Hebb's theory of neural assemblies has more recently begun to attract broad interest from the neuroscience community. Why, one wonders, was such a reasonable suggestion so long ignored? Several answers come to mind.

First, Hebb was far ahead of his time. As a psychologist, moreover, he was telling electrophysiologists and neurologists what they should be doing when these people had much on their collective plate. Throughout most of the twentieth century, electrophysiologists were facing numerous difficulties in recording from single neurons. Adequate impulse amplifiers needed to be designed and suitable microelectrodes fabricated before voltages could be measured from even a single cell. If mere hit-or-miss recordings were to be avoided, it was necessary to position accurately the tips of these electrodes, knowing what cells are located where. As the levels of the observed signals became smaller, means for shielding measurements from ambient electromagnetic noise were ever more in demand. With single-neuron recording being the primary experimental focus, therefore, it is not surprising that theoreticians refrained from embracing more complicated formulations that required simultaneous recordings from many neurons for which empirical support was not soon expected.

Second, as we have seen in Chapter 9, it is difficult enough to describe properly the dynamics of individual neurons; thus, a theory that assumed interacting assemblies of neurons would be venturing even further out onto the thin ice of speculation.

A third reason for the tendency to simplify the theoretical picture—in North America, at least—was the unfortunate domination of psychology by the beliefs of behaviorism, which focused attention on the conditioning of stimulus–response reflexes, thereby ignoring much that comprises mental reality. From the behaviorist perspective, the concept of internal cerebral states was rightly shouldered into the background because the simpler ideas of "connection theory" seemed adequate to explain acceptable psychological data.

With all of these strikes against it, how did Hebb's theory ever manage to see the light of day?

11.1 Birth of the Cell-Assembly Theory

During the 1940s, Hebb became impressed with several sorts of evidence that cast doubt on behaviorist assumptions and suggested that more subtle theoretical perspectives were needed to explain psychological facts [34]. Among such facts is the surprising robustness of the brain's dynamics, a well-known example of which was provided by railroad workman Phineas Gage, who survived having a piece of iron rod go through his brain [56]. With characteristic directness, Hebb put the matter thus: How is it that a person can register an IQ of 160 after the removal of a prefrontal lobe [32]?

His first publication on the cell assembly stemmed from observations of chimpanzees raised in a laboratory where, from birth, every stimulus was under experimental control. Such animals, Hebb noted, exhibited spontaneous fear upon seeing a clay model of a chimpanzee's head [33]. The chimps in question had never witnessed decapitation, yet some of them "screamed, defecated, fled from their outer cages to the inner rooms where they were not within sight of the clay model; those that remained within sight stood at the back of the cage, their gaze fixed on the model held in my hand" [35, 36, 38].

Such responses are clearly not reflexes; nor can they be explained as conditioned responses to stimuli, for there was no prior example in the animals' repertory of responses. Moreover, they earned no behavioral rewards by acting in such a manner. But the reactions of the chimps do make sense as disruptions of highly developed and meaningful internal configurations of neural activity according to which the chimps somehow recognized the clay head as a mutilated representation of beings like themselves.

Another contribution to the birth of his theory was Hebb's rereading of Marius von Senden's *Space and Sight* [84], which was originally published in Germany in 1932. In this work, von Senden gathered records on 65

patients who had been born blind due to cataracts up to the year 1912. At ages varying from 3 to 46 years, the cataracts were surgically removed, and a variety of reporters had observed the patients as they went about handling the sudden and often maddeningly novel influx of light.

One of the few generalizations over these cases, von Senden noted, was that the process of learning to see "is an enterprise fraught with innumerable difficulties, and that the common idea that the patient must necessarily be delighted with the gifts of light and colour bequeathed to him by the operation is wholly remote from the facts." Not every patient rejoiced upon being forced to make sense of incoming light that was all but incomprehensible, and many found the effort of learning to see to be so difficult that they simply gave up.

That such observations are not artifacts of the surgery or uniquely human was fortuitously established through observations on a pair of young chimpanzees that had been reared in the dark by a colleague of Hebb [81]. After being brought out into the light, these animals showed no emotional reactions to their new experiences. They seemed unaware of the stimulation of light and did not try to explore visual objects by touch. Hebb conjectured that the chimps showed no visual response because they had not yet formed the neural assemblies needed for perception.

Finally, Hebb pointed out that the learning curve for an individual subject in a behavioral experiment is not the smoothly rising curve shown in psychology textbooks. This is because the textbook curves are averages over many learning experiments, whereas the observations in a particular experiment are influenced by whether the subject is paying attention to the task. Thus the factor of *attention* (otherwise called attitude, expectancy, hypothesis, intention, vector, need, perseveration, or preoccupation), Hebb felt, must somehow be included in any satisfactory theory of learning.

As was noted in Chapter 1, these considerations led Hebb to propose that nerve cells do not necessarily act as individuals in the dynamics of the brain but often as functional groups, which he called cell assemblies, with the following properties.

- Each complex assembly comprises a "three-dimensional fishnet" of many thousands of interconnected cells sparsely distributed over much of the brain.

- The interconnections among the cells of a particular assembly grow slowly in numbers and strength as a person matures in response to both external stimuli and internal dynamics that are tailored to the particular experiences of the organism.

- One mechanism suggested for the growth of neuronal interconnections postulated the strengthening of dendritic contacts through use. (That this feature has become widely known among nerve network mavens as a "Hebbian synapse" amused Hebb because it was one of

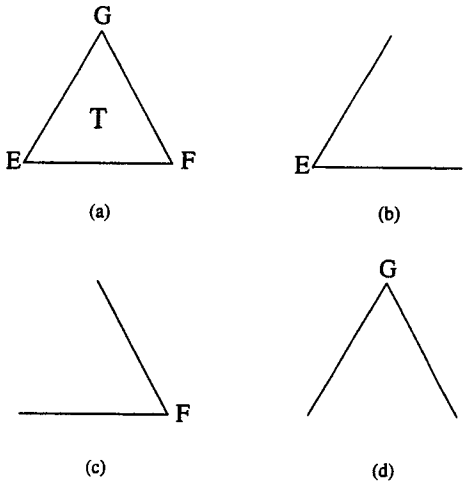


Figure 11.1. Diagrams related to the process of learning to see a triangle.

the few aspects of the theory that he did *not* consider to be original [64].) In Chapter 9, we saw that a real neuron has several means for altering its behavior, including changes in the geometry of dendritic spines or branching, variations in the distributions of ionic channels over the dendritic and axonal membranes, development of dendro-dendritic interactions, changes in amplification levels of decremental conduction, and so on.

- Upon ignition—effected through some combination of external stimuli and the partial activities of other assemblies—a particular assembly remains briefly active, yielding in a second or so to partial exhaustion of its constituent neurons.
- During the period of time that an assembly is active, the attention of the brain is focused on the concepts embodied in that assembly.
- As one assembly ceases its activity, another ignites, then another, and so on, in a temporal series of events called the *phase sequence*, which is experienced by each of us as a train of thought.

As a simple example of assembly formation, consider how an infant might learn to perceive the triangle T shown in Figure 11.1(a). The constituent sensations of the vertices are first supposed to be centered on the retina by eye movement and mapped onto the primary visual area (V1) of the optical lobes of the neocortex (located in the back of your head). Corresponding cell assemblies E, F, and G then develop in the secondary visual area through nontopological connections with area V1. The process of examining the triangle involves elementary phase sequences in which E, F, and G are sequentially ignited. Gradually, these subassemblies are supposed to fuse together into a common assembly for perception of the triangle T.

With further development of the assembly T—which reduces its threshold for ignition through the strengthening of the internal connections among E, F, and G—a glance at one corner, with a few peripheral cues, serves to ignite the entire assembly representing T. At this point in the learning process, T is established as a second-order cell assembly for perception of a triangle, including E, F, and G among its constituent subassemblies.

Is there empirical evidence supporting Hebb's theory?

11.2 Early Evidence for Cell Assemblies

Upon formulating the cell-assembly theory for brain dynamics, Hebb and other psycholigists began the process of empirical evaluation that is central to science. By the mid-1970s, these efforts had produced the following results.

Robustness

In Chapter 1, we considered a social analogy for the cell-assembly concept in which the brain is likened to a community and the neurons to its individual citizens. From this perspective, the remarkable robustness of the brain to physical damage can be understood. If a motorcycle club gets into a fight, losing several of its members, the strength of the club is not permanently reduced because new members can be added. Similarly, a damaged cell assembly can recruit additional neurons to participate in its activities. (Such recruitment of new assembly members may occur during rehabilitation from a stroke, a lobotomy, or other forms of neurological damage.)

Furthermore, because the cells of an assembly may be widely dispersed over much of the brain, partial destruction of the brain does not completely destroy any of the assemblies. Thus, the cell-assembly theory offers the same sort of robustness under physical damage as a hologram but is more credible because it does not require a regular structure that can reinforce scattered waves of neural activity.

Learning a New Language

As a graduate student in the "post-Sputnik" days of the late 1950s, I had the experience of learning to read Russian, having no prior knowledge of the language whatsoever. This effort proceeded in stages, commencing with the task of recognizing Cyrillic letters and associating these new shapes with novel sounds. Upon mastering the alphabet, it became possible to learn words comprising these letters, and with enough words, sentences and then paragraphs could eventually be understood. Thus it appears to me an empirical observation that language learning is a step-by-step process, during which a hierarchically organized memory is slowly constructed.

Interestingly, the full perception of a letter or word involves the melding of visual, auditory, and motor components, which underscores the concept

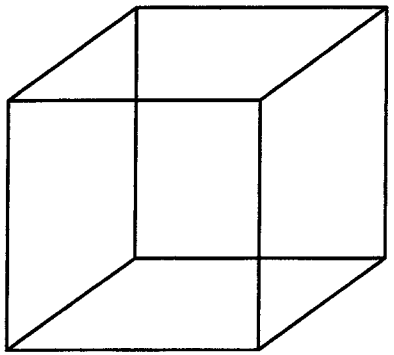


Figure 11.2. The Necker cube.

of subassemblies being distributed widely over the brain, a point to which we will return in the following chapter.

The general idea of hierarchical learning and memory has been rather carefully formulated by Braitenberg and Pulvermüller [8]. Although the acquisition of most of our basic skills lies buried in the forgotten past, most learning seems layered, with each stage necessarily mastered before it becomes possible to move on to the next. In the context of Hebb's theory, these stages involve the formation of subassemblies from which assemblies of higher order will subsequently emerge.

Ambiguous Perceptions

No discussion of the brain can neglect the mention of ambiguous figures, which have fascinated Gestalt psychologists for generations, and my favorite example—the Necker cube—is shown in Figure 11.2. Attempting to "bridge the long gap between the facts of neurology and those of psychology," Hebb's theory provides an explanation for the properties of such figures [34]. Gestalt phenomena are thus understood in a visceral manner by supposing that an assembly is associated with the perception of each orientation. Upon regarding Figure 11.2, I sense something switching inside my head every few seconds as the orientations change.

From the several cases of people learning to see that were cited by von Senden [84], it is clear that the ability to perceive an object in three spatial dimensions is itself learned, and the Necker cube is particularly interesting because perceptions of its two possible orientations would seem to be of equal likelihood. In the following section, we model the dynamics of switching between perceptions of two such orientations, where the overall symmetry of the situation suggests that the parameters of the two assemblies are identical, thereby simplifying analysis.

Stabilized Images

In Hebb's view, some of the strongest evidence in support of the cell-assembly theory was obtained from *stabilized-image* experiments, which

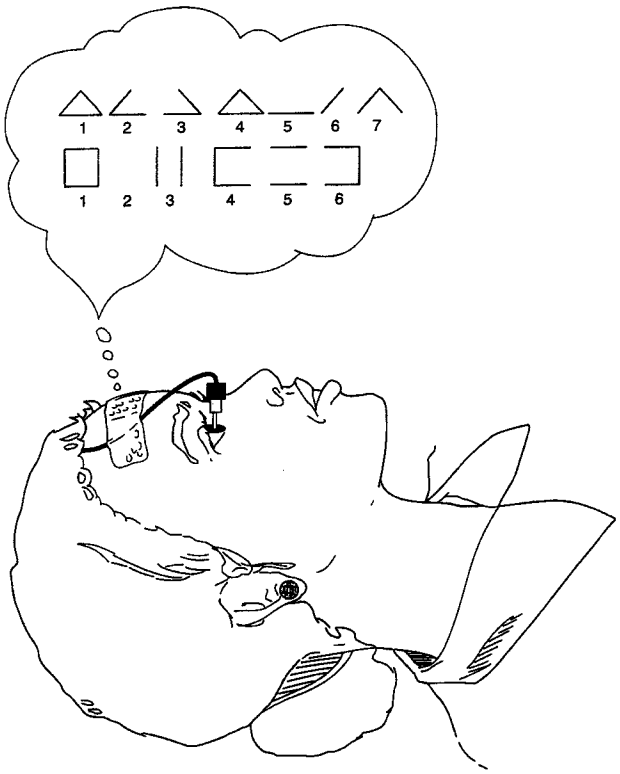


Figure 11.3. Sketch of contact lens and optical apparatus mounted on the eyeball of a reclining observer. The wire is connected to a small lamp that illuminates the target. The thought balloon shows sample sequences of patterns perceived by the subject with images that are stabilized on the retina by the apparatus. In the upper row a triangle is the target, and in the lower row, the target is a square (after a photograph in Pritchard [77]).

were carried out at McGill University in the early 1960s [35, 36, 64, 76, 77]. The experimental setup is sketched in Figure 11.3, where a simple geometric figure (e.g., a triangle or a square) is projected as a fixed image onto the retina. The subjects are asked to relax and simply report what they see, and because this is an introspective experiment, typical results are displayed in a thought balloon.

At first, subjects report seeing the entire figure, but after a few moments the figures change. Habituation effects (perhaps electrochemical changes in the stimulated retinal neurons) cause entire parts of the figures to disappear or to fall out of perception. It is the manner in which perceptions of the figures alter that is of particular interest. Subjects reported that the component lines or angles (i.e., subassemblies) of a triangle and a square would jump in and out of perception all at once. These observations are

as expected from Hebb's original formulation of the theory and the learning sequence for a triangle indicated in Figure 11.1; thus stabilized-image experiments confirm a prediction of the theory.

Learning Environments for Animals

According to Hebb's theory, adult thought processes involve continuous interactions among cell assemblies, which in turn are organized by sensory stimulation and internal interactions during the learning period of a young animal. How does adult behavior depend on opportunities for percept formation during development? Experiments show that rats reared in a rich perceptual environment—a "Coney Island for rats"—are notably more intelligent as adults than those raised in restricted environments, which provides yet another confirmation of the theory [64, 78]. As is anticipated from the cell-assembly theory, this positive influence of perceptual stimulation occurred only during youthful development; increased stimulation of adults is less effective in increasing rodent smarts.

Similar experiments with Scottish terriers showed even more striking differences, again as expected from the cell-assembly theory [89]. This is because the fraction of the neocortex that is not under the influence of sensory inputs—the *associative cortex*—is larger for a dog than a rat. Thus, the internal organization of the dog's brain should play a greater role in its behavior. Terriers reared in single cages, where they could not see or touch other dogs, had abnormal personalities and could neither be trained nor bred. Other studies showed that dogs reared in such restricted environments did not respond to pain, as if they were lobotomized [62].

Sensory Deprivation of Humans

In his original formulation of the cell-assembly theory [34], Hebb speculated that perceptual isolation would cause emotional problems because the phase sequence needs the guidance of meaningful sensory stimulation to remain organized in an intelligible manner. To test this aspect of the theory, experiments on perceptual isolation were performed by Heron and his colleagues in the 1950s [37, 64]. In these studies, the subjects were college students who were paid to do nothing. Each subject lay quietly on a comfortable bed wearing soft arm cuffs and translucent goggles, hearing only a constant buzzing sound for several days. During breaks for meals and the toilet, the subjects continued to wear their goggles, so they averaged about 22 hours a day in total isolation.

Many subjects took part in the experiment intending to plan future work or prepare for examinations. According to Hebb [35], the main results were that a subject's ability to solve problems in his or her head declined rapidly after the first day as it became increasingly difficult to maintain coherent thought, and for some it was difficult to daydream. After about the third day, hallucinations became increasingly complex. One student said that his mind seemed to be hovering over his body like a ball of cotton wool. Another

reported that he seemed to have two bodies but did not know which was really his. Such observations are in accord with a variety of anecdotal reports from truck drivers, shipwreck survivors, solitary sailors, long-distance drivers, and the like that extended periods of monotony breed hallucinations. (Reporting on his famous solo flight across the Atlantic Ocean, for example, U.S. aviator Charles Lindbergh noted "vapor-like shapes crowding the fuselage, speaking with human voices, giving me advice and important messages" [50].)

After the perceptual isolation experiments were concluded, subjects experienced difficulties with visual perception lasting for several hours and were found to have a significant slowing of their electroencephalograms or brain waves. They also seemed more vulnerable to propaganda. Although the specific results of these experiments were not predicted by the cell-assembly theory, the disorganizing effect of sensory deprivation on coherent thought had been anticipated.

Structure of the Neocortex

While presenting a plausible theory for the dynamics of a brain, Hebb's classic book contains but one lapse into mathematical notation: he discusses in some detail the ratio

$$\frac{A}{S} \equiv \frac{\text{total association cortex}}{\text{total sensory cortex}}$$

for various mammalian species [34]. This ratio relates the area of the neocortex that is not directly tied to sensory inputs—the *associative* (A) regions—to the area of the *sensory* (S) regions, which are under direct environmental control from eyes, ears, and senses of touch and smell. If this ratio is zero, all of the cortex is under sensory control, and necessary conditions for behaviorist psychology are satisfied. On the other hand, larger values of the ratio imply increasing opportunities for the cortex to construct abstract cell assemblies with dynamics beyond direct control of the senses.

In general, Hebb pointed out, this A/S ratio increases as one moves through mammalian species from rat to dog to primate to human, in general agreement with two aspects of brains' behaviors. First, as most would agree, the character of a human's inner life is significantly more intricate than that of a chimp, which in turn is more than for a dog or a rat. Second, the time required for *primary learning* (until adulthood is reached) increases with the A/S ratio. Human infants are essentially helpless and remain so for several years as they slowly build the myriad assemblies upon which the complexities of their lives will eventually be based.

11.3 Elementary Assembly Dynamics

In this section, some simple models of cell-assembly dynamics are presented that describe the average behavior of a relatively large number of interacting model neurons. Because these descriptions are restricted to very simple representations of the neurons—little like the more realistic picture that was developed in Chapter 9—they should be viewed as indicating lower bounds on the possible behaviors of real neural systems. The generalization of such analyses to more realistic neural models is a challenge for current neuroscience research, and some such attempts are described in Section 11.5.

11.3.1 Ignition of an Assembly

To model the dynamics of an individual neural assembly as it turns on (ignites) or turns off (becomes extinguished), we can imagine a large mass of randomly connected McCulloch–Pitts (M–P) neurons as described by Equation (2.10), a problem that goes back to the 1950s [3, 26, 28, 79, 86, 87, 90]. In developing a simple formulation, it is convenient to make the following assumptions and definitions of additional variables.

- Time (t) is defined on a discrete lattice, with the duration of each interval equal to the *synaptic delay* τ.
- $F(t)$ represents the fraction of neurons that are firing at time t.
- I is the number of input connections to each neuron. These are received randomly from outputs of other neurons in the assembly.
- The refractory times of the neurons are shorter than the synaptic delay.

With these definitions, we can write the probability of a neuron receiving exactly j input signals at time t as

$$\left(\frac{I!}{j!(I-j)!}\right) F^j(1-F)^{I-j},$$

an expression that can be understood as follows.[1]

(1) $I!/j!(I-j)!$ is the number of different ways that j input signals can be selected from among I input channels.

(2) F^j is the probability of having signals appear on j of the input channels.

[1]The alert reader will recall that we met the same expression in Equation (2.5) of Chapter 2 describing the probability for k synaptic vesicles to release their transmitter substance through n presynaptic sites.

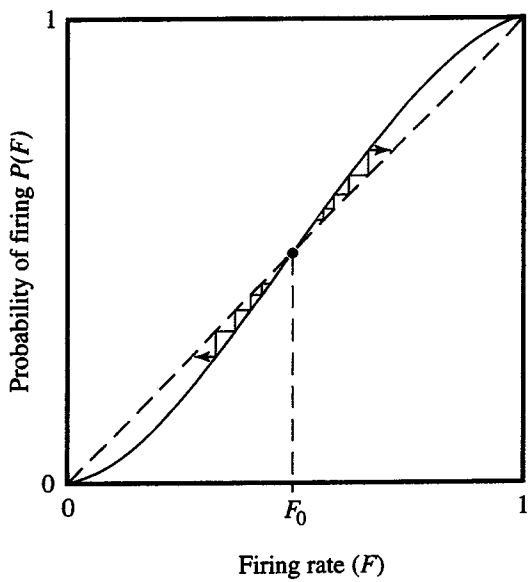


Figure 11.4. Qualitative behavior of the probability of a neuron firing in the next time increment $P(F)$ as a function of F, the current firing rate, assuming that $1 < \theta < I$.

(3) $(1-F)^{I-j}$ is the probability of not having signals on the other $I-j$ input channels.

Because the M–P model neuron gives an output signal when its inputs are equal to or greater than the threshold θ, the probability of a neuron firing in the next increment of time is given by the summation

$$P(F) = \sum_{j=\theta}^{I} \left(\frac{I!}{j!(I-j)!}\right) F^j(1-F)^{I-j}. \tag{11.1}$$

Although this expression appears unwieldy, its qualitative behavior is straightforward; thus for

$$1 < \theta < I,$$

$P(F)$ is the sigmoid function of F sketched in Figure 11.4.[2]

The condition

$$P(F) = F, \tag{11.2}$$

[2]To see this, note that $P(F) \sim B(I,\theta)F^\theta$ near $F = 0$, where $B(I,\theta) \equiv I!/\theta!(I-\theta)!$ is a *binomial coefficient*. Similarly $P(F) \sim 1 - B(I,\theta-1)(1-F)^{I-\theta+1}$ near $F = 1$. Because direct calculation shows that $P(F)$ is a monotone increasing function, it must have the shape indicated in Figure 11.4.

which is satisfied for three values of F, indicates *stationary solutions* of the system because these are the values of F for which the probability of firing in the next time increment is equal to the present firing rate. Let us consider these three stationary solutions in detail.

1. The minimum stationary condition $F = 0$ corresponds to none of the neurons firing. This is a stable solution because if F is increased slightly from 0, Figure 11.4 shows that the corresponding increase in $P(F)$ is less than that of F, implying that the activity will relax back to zero.

2. The maximum stationary condition $F = 1$ corresponds to all of the neurons firing at their maximum rates. This is also a stable solution because if F is decreased slightly from 1, the corresponding value of $P(F)$ is greater than F, implying that the activity will rise back to one.

3. The stationary condition at $F = F_0$ corresponds to an intermediate firing rate, where F_0 increases from 0 to 1 as θ increases from 1 to I. In contrast to $F = 0$ and $F = 1$, this intermediate stationary level is *unstable*. To see this, note from Figure 11.4 that if F is increased slightly above F_0, the increase in $P(F)$ is greater than that of F, causing F to rise even more in the subsequent time increment. If F is decreased slightly below F_0, on the other hand, the decrease of $P(F)$ is more than that of F, causing F to fall even more in the subsequent time increment.

In the context of nonlinear system theory, therefore, a cell assembly shares properties of the Hodgkin–Huxley nerve impulse that were discussed in Section 4.6. Thus the stationary state at $F = 1$ can be viewed as an *attractor*, as can the null state at $F = 0$. In these terms, the intermediate stationary state at $F = F_0$ defines a *separatrix* lying on the boundary between the basins of these two attractors.

In other words, cell-assembly activity emerges from a net of interconnected neurons, much as a nerve impulse emerges from the Hodgkin–Huxley equations for a squid axon. Both exhibit the interrelated properties of all-or-nothing response and threshold, providing a basis for the hierarchical structures of assemblies shown in Figure 11.1 and to be considered in the following chapter.

From the perspectives of Chapter 1, the ignition of an assembly can be represented by the following positive feedback diagram:

Firing rate: F
$\downarrow \quad \uparrow$
Probability of firing: $P(F)$

Above the level of ignition ($F = F_0$), positive feedback causes $P(F)$ to grow faster than F, so activity increases until the stable stationary state at $F = 1$ is reached. What is the time course of this growth?

Because the function $P(F)$ indicates the level of activity at time $t + \tau$, it was noted in the previous chapter that the discrete formulation of the dynamics is roughly equivalent to the ordinary differential equation

$$\frac{dF}{dt} = \frac{P(F) - F}{\tau}, \tag{11.3}$$

where t is now considered to be a continuous variable.[3] For $F_0 < F < 1$, it is evident from Figure 11.4 that the right-hand side of this ODE has the same qualitative features as the right-hand side of Equation (1.3), which was used to derive the Verhulst curve for population growth shown in Figure 1.3.

Thus, $F(t)$—the dependence of the firing rate on time during assembly ignition—is given implicitly by the integral relation

$$\int_{F_{\text{init}}}^{F(t)} \frac{dF'}{P(F') - F'} = \frac{t}{\tau}. \tag{11.4}$$

Here, $F_{\text{init}} > F_0$ is the initial value of F at $t = 0$, which may have been established by inputs from other assemblies, external sensory inputs, or some combination of the two. (Although one actually calculates t as a function of F, it can be seen from Figure 11.4 that $F(t) \to 1$ as $t \to \infty$.)

To model its qualitative features, Equation (11.3) can be written as

$$\frac{dF}{dt} \approx -\frac{1}{\tau} F(F - F_0)(F - 1), \tag{11.5}$$

an ODE that is interesting to compare with the representation of a space-clamped patch of nerve membrane developed in Chapter 5. In that case, the reader will recall, transmembrane voltage obeys an ODE of the form

$$\frac{dV}{dt} = -\left(\frac{G}{C}\right)\left[\frac{V(V - V_1)(V - V_2)}{V_2(V_2 - V_1)}\right], \tag{11.6}$$

where C/G is an active time constant for the membrane, and a cubic approximation is used to model the transmembrane current that is plotted in Figure 5.1. Thus, we see a mathematical relationship between the switching of a patch of membrane and the switching of an assembly, although they are at quite different levels of description. This correspondence is of central importance for the perspectives being developed in this book and will be further discussed in the following chapter.

[3] Beware the analytic sleights of hand here. Time was assumed to be a discrete variable in order to derive an expression for $P(F)$ in Equation (11.1), and now it is redefined as a continuous variable in order to use that expression in an ODE.

Once an assembly has been ignited, Equation (11.5) indicates that it remains firing forever, but this overlooks habituation effects, inhibitory inputs from other assemblies, and external sensory inputs, all of which may reduce the firing rate and increase the ignition threshold F_0. (Similarly, Equation (11.6) neglects the recovery effects on a nerve fiber stemming from potassium ion current, which are treated in Chapter 6.) The time course of the extinction dynamics is again given more precisely by Equation (11.4), but now F_{init} is *less* than F_0 at $t = 0$, and it is seen from Figure 11.4 that $F(t) \to 0$ as $t \to \infty$.

This analytic formulation is tidy, but can we believe it? Should real nerve networks be expected to behave at all like the variables in these equations? Because the candid answer is that I do not know, it seems appropriate to underscore some areas of present concern with the hope that they will be selected for further study.

First, I repeat that we do not yet know how to accurately model a single nerve cell, thus the McCulloch–Pitts representation may miss essential neural properties. In particular, the preceding formulation reduces the communication among neurons to passing information about their average firing rates, an assumption that overlooks important aspects of neural dynamics. Perhaps real neurons talk to each other in languages that are based on time codes, space codes, or some subtle combinations thereof. Perhaps they use chemical or ephaptic interactions as a sort of body language. Over longer distances, cell assemblies might communicate via the information waves that were considered in the previous chapter. Finally, it could be that assemblies engage in activities beyond our present ken.

However assemblies interact, an important aspect of neural behavior that has been neglected in the preceding analysis is the fact that synaptic influences can be inhibitory as well as excitatory. We will see in the following section that inhibition plays a key role in determining the ways in which two or more cell assemblies behave.

11.3.2 Inhibition among Assemblies

At the time of Hebb's original formulation of the cell-assembly theory, there was no experimental evidence for inhibition among cortical neurons, so he conservatively assumed only excitatory interactions. By 1957, however, cortical inhibition had been observed, so Peter Milner, a colleague of Hebb's at McGill University, developed a "Mark II" version of the theory [63]. The most striking feature of this revised theory is that it allows independent assemblies to develop from an undifferentiated mass of model neurons.

To evaluate the effect that synaptic inhibition among cortical neurons might have on cell-assembly dynamics, it is convenient to represent the behavior of an individual assembly as simply as possible. To this end, let us set $\theta = 1$ in Equation (11.1), whereupon $P(F) = 1 - (1 - F)^I$. For $I = 2$ (two inputs for each neuron), this expression becomes

$$P(F) = 2F - F^2 ,$$

with the same qualitative behavior for larger values of I.

Under these simplifying assumptions ($\theta = 1$, $I = 2$), Equation (11.3) reduces to

$$\frac{dF}{dt} = F(1 - F) ,$$

where time is measured in units of the synaptic delay (τ). This is just the Verhulst equation with solution

$$F(t) = \frac{F(0)e^t}{1 + F(0)\,(e^t - 1)} ,$$

which follows from integration of Equation (11.4) and is displayed in Figure 1.3 for several initial values. The same growth equation describes both the firing rate of a cell assembly and the population of Belgium. Again, we find that identical mathematical formulations are useful at widely different levels of description.

Thus motivated, let us model the dynamics of two identical neural assemblies with inhibitory interactions by the coupled ODE system

$$\begin{aligned} \frac{dF_1}{dt} &= F_1(1 - F_1) - \alpha F_2 , \\ \frac{dF_2}{dt} &= F_2(1 - F_2) - \alpha F_1 , \end{aligned} \tag{11.7}$$

where $0 \le F_1 \le 1$ and $0 \le F_2 \le 1$ because F_1 and F_2 represent the fraction of neurons in each assembly that are firing. When positive, the parameter α introduces an inhibitory interaction between the two assemblies because the $-\alpha F_2$ term in the first equation reduces dF_1/dt and similarly for the second equation.

To see how these equations model the role that inhibition plays in the formation of cell assemblies, let us recall a bit of history. As digital computers became available for scientific problems in the mid-1950s, Frankel reviewed several approaches to the numerical studies of brains, concluding that Hebb's cell-assembly theory was the most promising [17]. Rochester et al. [82] then began to study the growth of cell assemblies in a group of 99 McCulloch–Pitts style model neurons, allowing only excitatory interactions as had originally been proposed by Hebb [34]. Although they found a diffuse reverberation with a period on the order of the synaptic delay, assemblies did not develop.

This disappointing result follows directly from Equations (11.7). How? If we let α be negative, only excitatory interactions among the neurons are allowed. In this case, as is seen from Figure 11.5(a), all points on the (F_1, F_2)

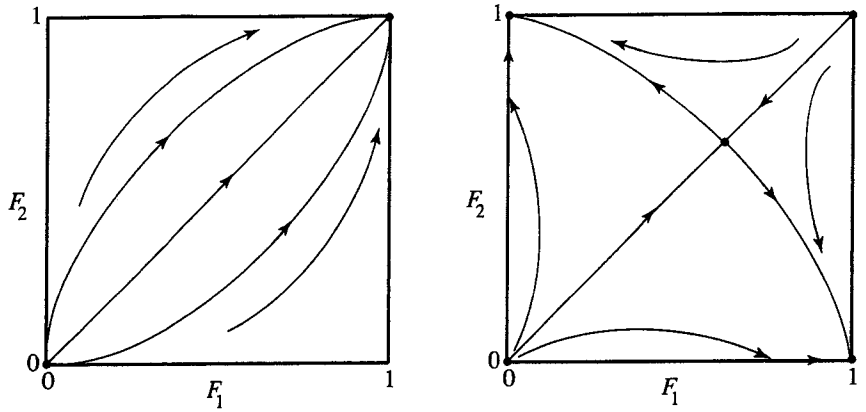


Figure 11.5. (a) A phase-plane plot from Equations (11.7) with $\alpha < 0$ (only excitatory interactions). (b) A similar plot for $\alpha > 1/3$ (excitatory and inhibitory interactions).

phase plane move to (1,1), and no individual assemblies are permitted to ignite. In other words, all neurons end up firing at their maximum rates.

Rochester et al. then talked with Milner, who was revising Hebb's theory to include inhibition [63]. Thus inspired, they modified their computer model to include the growth of both excitatory and inhibitory interactions among 512 M–P neurons, with six neurons being externally driven [82]. Cell assemblies were then observed to form with excitatory interactions developing among cells in the same assembly and inhibitory interactions among different assemblies. How can this be seen in the context of our model?

Upon introducing inhibition in Equations (11.7) by making $\alpha > 0$, one finds a singular point at

$$F_1 = F_2 = 1 - \alpha\,,$$

where the time derivatives are zero. For $0 < \alpha < 1/3$, this singular point is stable, but for $\alpha > 1/3$, it becomes *unstable,* as shown in Figure 11.5(b). Stable states of the system are then at either

$$(F_1, F_2) = (1,0) \text{ or } (0,1)\,.$$

Thus, with sufficiently large inhibition, Equations (11.7) suggest that assemblies can be individually ignited in accord with both the numerical observations of Rochester et al. [82] and the theoretical considerations of Milner's "Mark II" cell-assembly theory [28, 63].

At this point in the discussion, you should revisit Figure 11.2 and experience how your perception switches back and forth between the two orientations of the Necker cube. Although it is easy to see the cube in either orientation, note that you cannot perceive both orientations at the same time. (How rapidly can you switch between perceptions of the two orientations? Might the speed of these transitions be taken as a measure of how well your brain is working?)

Now, consider Equations (11.7) with $\alpha > 1/3$ and the corresponding phase-plane diagram shown in Figure 11.5(b). Evidently, these equations model the switching on and off of assemblies that correspond to the dynamics of those in your head as you regard the Necker cube.

From an engineering perspective, the interactive dynamics of two assemblies are like a "flip-flop" circuit widely used in the design of information storage and processing systems [27]. With a cell assembly, however, the bit of information being switched on or off is not the voltage level of a transistor but an intricate psychological perception embodied in the connections among thousands of neurons scattered about the brain that have developed in response to the lifelong experiences of the organism. Although this has been a "bottom-up" discussion of the brain's dynamics, it suggests the utility of "top-down" approaches. Regarding assembly firing rates as *order parameters* for higher level representations of the brain's dynamics, for example, Haken and his colleagues have been able to model a variety of psychological experiments [29, 30, 31].

To represent more than two assemblies, Equations (11.7) can be generalized to

$$\begin{aligned}
\frac{dF_1}{dt} &= +F_1(1-F_1) - \alpha F_2 - \alpha F_3 - \cdots - \alpha F_n\,,\\
\frac{dF_2}{dt} &= -\alpha F_1 + F_2(1-F_2) - \alpha F_3 - \cdots - \alpha F_n\,,\\
&\cdots\\
\frac{dF_n}{dt} &= -\alpha F_1 - \alpha F_2 - \alpha F_3 - \cdots + F_n(1-F_n)\,,
\end{aligned} \tag{11.8}$$

where $0 \le F_j \le 1$ for $j = 1, 2, \ldots, n$. In this n-assembly model, interestingly, all of the previous analysis (for $n = 2$) can be carried through. Thus, there is a singular point for positive α (the inhibitory case) at

$$F_1 = F_2 = \cdots = F_n = 1 - (n-1)\alpha\,,$$

which is stable for

$$\alpha < \alpha_c = 1/(2n-1)$$

and unstable for

$$\alpha > \alpha_c = 1/(2n-1)\,.$$

Below this critical value of inhibition (α_c), all of the assemblies can become simultaneously active. It turns out that the switching time (τ_{sw}) of this instability is

$$\tau_{sw} = \frac{1}{(2n-1)\alpha - 1}\,,$$

counterintuitively implying that the rate at which a neural system can change from one perception to another *increases* with inhibition (α). This result is in accord with Hebb's suggestion that we humans are more intelligent than our fellow mammals in part because we can switch our attention more quickly from one assembly to another [35, 36].

Another aspect of intelligence, however, is the total number of assemblies that can be remembered.

11.4 How Many Assemblies Can There Be?

Having considered some of the evidence for the existence of cell assemblies, it is interesting to ask how many of them can be stored in a human brain. This is a difficult question to answer because—as we have seen in Chapter 9—there is not yet a clear understanding of what the individual neurons are doing, but it is possible to make certain lower estimates. To this end, let us review three considerations.

First, it is presently necessary to use a McCulloch–Pitts style model in which each neuron is represented by a single switch. Evidently, conclusions based on this unrealistic assumption can provide only lower bounds on the possible number of assemblies.

Second, it is not correct to estimate the number of assemblies by dividing the number of neurons in the brain by the number of neurons in an assembly. Why not? Recall the social analog for cell assemblies, which was presented in Chapter 1. Just as a particular person in a city may be a member of more than one social assembly, so may a single neuron participate in several different cell assemblies.

Finally, any estimate of the maximum number of assemblies should account for the fact that the brain is hierarchically structured. Thus, complex assemblies comprise simpler assemblies, which in turn are composed of yet simpler ones, and so on.

In an important paper that appeared in the mid-1960s, Charles Legéndy assessed human brain capacity from a simple model [47]. Although the basic structure of his work is presented here, additional statistical details are in the original publications [48, 49].

To introduce hierarchical character, Legéndy assumed that the brain is already organized into subassemblies and modeled their organization into assemblies. In the spirit of Hebb's theory, an assembly and one of its subassemblies variously represent

> a setting and a person who is part of it, a word and one of its letters, an object and one of its details.

To avoid complications of spatial organization, interconnections among assemblies are taken to be evenly distributed over the brain. (Following a familiar caricature of a mathematician's approach to biology, this is the assumption of a "spherical brain".)

Like individual neurons, subassemblies and assemblies have excitation thresholds that must be exceeded for ignition. Whereas the threshold for a subassembly is assumed to be a certain number of active neurons, the threshold for an assembly is a certain number of active subassemblies. Legéndy considered the subassemblies to be already formed by *weak* contacts, whereas assemblies develop from subassemblies through the development of *latent* into *strong* contacts among neurons.

To proceed further, let us introduce the following notation.

- N is the number of neurons in the brain.
- A is the maximum number of assemblies that can form in the brain.
- n is the number of neurons in a subassembly.
- y is the number of subassemblies in an assembly.
- a is the number of strong (latent) contacts per neuron.
- m is the maximum number of strong contacts from an assembly to one of its subassemblies.

Assuming that half of the strong (latent) contacts make output (axonal) connections and the other half make input (dendritic) connections, the number of output contacts from an assembly is $nya/2$. Those outputs connecting to a particular subassembly reach a fraction n/N of the neurons in the brain; thus

$$m = \frac{n^2 ya}{2N} .$$

The maximum number of assemblies are stored in the model when about half of the latent connections have been converted into strong contacts. Why half? Think of a black and white photograph. If all of the pixels are all white or all black, the image conveys very little information. It is when about half of the pixels are black and the others are white that the most information is being stored, and so it is with the conversion from latent to strong contacts. Thus

$$A \sim \frac{Na}{2my} .$$

In ordinary English, this equation says that the maximum number of assemblies in the brain is given by half of the total number of strong (latent) connections in the brain ($Na/2$) divided by the number of strong (latent) contacts in a single assembly (my).

Combining the previous two equations yields an estimate for the maximum number of assemblies that can be stored in the brain:

$$A \sim \left(\frac{N}{ny}\right)^2 . \tag{11.9}$$

Table 11.1. The number of cell assemblies (A) in a brain versus the number of neurons in the brain (N) and the number of neurons in an assembly (ny). These values are estimated from Equation (11.9).

	$N = 10^{10}$	$N = 10^{11}$
$ny = 10^3$:	10^{14}	10^{16}
$ny = 10^4$:	10^{12}	10^{14}
$ny = 10^5$:	10^{10}	10^{12}

Some values of the maximum number of assemblies (A) implied by this estimate for different values of the number of neurons in a brain (N) and in a subassembly (n) are given in Table 11.1. Because the number of neurons in the brain is variously estimated as from ten to a hundred billion [9, 12, 39], these two values are selected in the upper row of the table. The values for ny are not empirically established and are expected to vary widely according to the intricacy of the concept perceived. (Palm has suggested that "a total assembly should have somewhere around 10^4 neurons with a working range from a few thousand to several tens of thousands" [73].) Lower values for ny would increase estimates of the number of assemblies that can be stored in a brain.

From these approximate values, it appears that

$$A > 10^9$$

is a comfortable lower bound on the maximum number of assemblies stored in the human brain. Equal to the number of seconds in 30 years, 10^9 is also in accord with estimates by Griffith based on the rate at which the brain is able to absorb information [28].

Finally, it is interesting to compare Equation (11.9) with the maximum number of patterns (p_m) that can be stored in an attractor network from Equation (10.9) of Section 10.2.2. Although 10^9 is again a rough lower bound on the number of attractors that emerge for a brain comprising 10^{10} to 10^{11} neurons, the bases for these two estimates differ; in particular, $A \propto N^2$, whereas $p_m \propto N$.

An explanation for this difference is that under the analysis of Section 10.2.2, every neuron is assumed to be firing 50% of the time. Thus, p codes of length N were found to introduce noise of amplitude $\sqrt{p/N}$ into the retrieval task, which limits the number of stored codes to $O(N)$. Under Legéndy's analysis, on the other hand, a particular neuron fires only when assemblies in which it participates are ignited, which leads to smaller average firing rates in closer accord with empirical observations or cortical activity.

11.5 Cell Assemblies and Associative Networks

As most have now seen, holograms use a well-defined reference beam (usually a laser source) to translate information from a distributed memory (the hologram) into a family of three-dimensional images. A small piece of the hologram is able to reproduce the entire image, albeit with reduced resolution. Inspired by the realization of holographic memories in laser laboratories of the 1960s, it was suggested that similar nonlocal storage principles might apply to memory in the neocortex [6, 22, 23, 52, 75].

Because several requirements of a holographic memory are not satisfied in the neocortex (e.g., well defined reference beam, stable wave medium), a memory principle was sought that would capture the distributed features of holographic storage in a realistic neural context [53, 95]. Thus, it emerged that the neocortex might operate as an *associative memory* [40, 53, 95].

The basic element of an associative memory is a connection matrix relating two sets of patterns. Feeding a portion of one pattern into the matrix and introducing threshold discrimination often allows aspects of the corresponding pattern to be recovered. Such a system can be useful for a variety of information-processing tasks, including feature extraction, pattern reconstruction, pattern identification, and sequential association [42].

To make contact with the previous section, think of the connection matrix as $N \times N$, with elements indicating interconnection ("synaptic") strengths between N neocortical neurons, and take the fraction of nonzero elements in the patterns to be $O(ny/N)$. Then, the fraction of matrix elements (or synapses) used up in the learning of a pattern pair is $O[(ny/N)^2]$. Because the number of unactivated synapses after the learning of r random code pairs will be of the order

$$\left[1 - \left(\frac{ny}{N}\right)^2\right]^r \approx 1 - r\left(\frac{ny}{N}\right)^2,$$

the maximum number of pattern pairs that can be learned is $O[(N/ny)^2]$ [43, 69].

It was recognized in the 1970s that Hebb's brain model can be regarded as an *autoassociative memory*, where the paired patterns can be the same [69, 93]. To see this, turn back to Figure 11.4—which shows the dynamics of a single assembly—and consider what the network is doing as the firing rate (F) increases from its threshold value of F_0. Outputs from a fraction F of the neurons are fed back as inputs to all neurons of the assembly, further increasing the firing fraction until the entire assembly is firing ($F = 1$). In this manner, it may be said, an ignited assembly has recognized itself.

Noting that the maximum number of assemblies (A) is equal to the maximum number of pattern pairs that can be related by the synaptic matrix $(N/ny)^2$, Legéndy's Equation (11.9) is confirmed.

Since the 1970s, the relationship between autoassociative memories (or *associative networks*, as they are coming to be called) and Hebb's cell assemblies has been an increasingly active area of neuroscience research, which comprises mathematical [69, 70, 72], neurological [7, 8, 9, 59, 73], and numerical components [13, 16, 41, 42, 69, 74, 92, 94].

Although much of this work supports the idea that Hebb's cell assemblies "provide an intermediate description of the brain between the psychological and the electrophysiological level" (as Günther Palm, a leader in associative nets research, has put it [71]), further tests of the theory depend on more realistic neural models.

11.6 More Realistic Assembly Models

As we have learned in Chapter 9, the dynamic behavior of a real neuron is far more intricate than that of an M–P model; thus, the "elementary assembly dynamics" formulated in Section 11.3 are suspect. To move in the direction of more realistic models of cell-assembly dynamics, descriptions of the basic neurons must be improved.

An early attempt in this direction modeled the basic units on *motor neurons* (MN), with disappointing results [55]. Interconnected populations of model neurons showed little tendency for activity to continue after their initiating inputs were turned off, at variance with Hebb's original concept of an assembly "acting briefly as a closed system." Since the 1980s, however, more biologically based models of assembly dynamics have been studied by investigators at the Royal Institute of Technology in Stockholm, leading to positive results.

The research group (called Studies of Artificial Neural Systems, or SANS, in the Department of Numerical Analysis and Computing Science, with a web site at www.nada.kth.se/sans) stems from the doctoral research of Anders Lansner, which was published in 1986 [42]. From the beginning, this work concentrated on the development of flexible models that could be incorporated into system studies with nuanced tradeoffs between neural realism and the numerical demands of large networks.

Currently, the best introduction to this effort is the doctoral thesis published in 1996 by Erik Fransén under the direction of Lansner [18]. A key feature of these investigations was the assumption of excitatory neurons based on cortical *pyramidal* (P) cells, with Hodgkin–Huxley style parameters differing from those of MN-cells as follows [44]:

- Less negative resting potential (-50 mV rather than -70 mV).
- Larger "depolarizing after potential" and smaller "after hyperpolarization." (In Section 4.7, these effects are referred to as "enhancement zones" and "refractory zones," respectively.)
- Smaller repolarizing voltage (V_K in Table 4.1).

- Spikes of smaller amplitude and duration.

In addition to either P-cells or MN-cells, the numerical representation included inhibitory fast-spiking (FS) cells, which are modeled after cortical interneurons. The earliest simulations consisted of 50 pairs of an excitatory cell and an FS-cell with 408 excitatory synapses to excitatory cells, 1538 excitatory synapses to inhibitory interneurons, and 50 inhibitory synapses to excitatory cells.

The numerical model was a neural simulator called SWIM, which is based on biologically plausible compartmental models for the neurons and synapses [16]. To reduce the overall computational task, excitatory neurons (MN-cells or P-cells) comprised four compartments each, whereas the inhibitory interneurons (via FS-cells) had only two.

The system of 50 cell pairs was taught eight different assemblies consisting of eight cells each. Thus some of the excitatory cells (MN or P) were necessarily members of more than one assembly. As is suggested by the analysis of Section 11.3.2, interconnections among cells of the same assembly were excitatory, whereas inhibitory interconnections (FS-cells) were established between different assemblies.

Differences between the behaviors of MN-cells and P-cells in such studies of Hebb's cell-assembly theory indicate that details of neural modeling are qualitatively important. The salient results are now discussed.

After Activity and Reaction Time

As Hebb assumed and the simple analysis of Figure 11.4 suggests, a cell assembly is expected to maintain its activity for a significant period of time after the stimulation is turned off. In studies with P-cells, such *after activity* was typically observed, with assemblies remaining active for periods of 350 to 400 ms after the termination of a 40-ms-long stimulation. Using MN-cells, on the other hand, after activity occurred only in exceptional cases, in accord with the previous work of MacGregor and McMullen [55]. During sustained firing of the P-cells, the frequency gradually decreased due to buildup of internal concentrations of Ca^{++} ions, which activate a Ca-dependent hyperpolarizing K^+ current. This is analogous to Hebb's cellular "partial exhaustion," and it eventually leads to extinction of the assembly activity.

Interestingly, these numerical experiments showed very short *reaction times* of about 50–70 ms, implying that each P-cell fires only about five times before an ignited assembly becomes fully active. This numerical observation blunts criticisms of the cell-assembly theory based on the suggestion that the turn-on process (indicated by the up-going arrows in Figure 11.4) might be significantly longer than typical perceptual response times.

Ignition Threshold and Pattern Completion

A critical firing level, denoted as F_0 in Figure 11.4, was readily observed

numerically. Typically, the stimulation of three cells was sufficient to ignite an assembly, whereas two was not, implying an *ignition threshold* in the range

$$\frac{1}{4} < F_0 < \frac{3}{8}.$$

The existence of a threshold for ignition is closely related to the phenomenon of *pattern completion*, under which stimulation of only part of a pattern is required for a correct response. In connection with the discussions of Figure 11.1, for example, it was noted that one need not examine every aspect of a geometric figure before its global form is perceived.

Competition and Noise Suppression

The strong lateral inhibition among neurons of different assemblies—parameterized by α in Equations (11.8)—implies that two or more mutually active assemblies will compete for dominance. Such competition (subjectively perceived for the Necker cube shown in Figure 11.2) was readily observed in Lansner and Fransén's numerical studies, with the winning assembly both activating its missing members and suppressing spurious activity (noise) of other cells [44].

Influence of Time Delay

Introduction of variable *time delays* in the firing of excitatory cells mimics the propagation of signals over extended axonal pathways of varying lengths. With P-cells, it was found that such axonal delays could be increased up to an average value of about 10 ms without significant changes in assembly behavior [19]. How far apart does this delay allow neurons of an assembly to be located?

Assuming that long cortical axons are myelinated (see Chapter 7) and of the order of a micron or more in diameter [80, 88], Equation (7.19) suggests an outside fiber diameter of at least 1.5 μm. The data on myelinated nerves of the cat summarized in Figure 7.3 then imply impulse speeds of more than 0.84 cm/ms. During 10 ms of axonal delay, therefore, a spike can travel at least 8.4 cm, or 3.3 inches, which is about the average distance between two randomly selected neurons in the human cortex. Thus, the SANS model seems to permit the extension of Hebb's "three-dimensional fishnet" throughout most of the brain.

Slow Firing Rates and the Role of Inhibition

One discrepancy between the foregoing results and the behavior of real brains involves the maximum firing rate of an ignited assembly. As is suggested by Figure 11.4 and observed numerically, neurons in an active assembly are expected to fire at their maximum rates, which can be as high as 300 Hz (every 3.3 ms) for typical pyramidal cells. Under normal physiological conditions, however, cortical neurons are observed to fire at about 20–60 Hz but seldom higher.

In response to this objection to Hebb's theory, Fransén and Lansner show that reduced firing rates are observed for fully active assemblies when synapses are assumed to be realistically slow and also *saturating*, implying an upper limit on its peak conductance [20]. From the discussion in Section 2.3.1, saturation of postsynaptic membrane conductance is a reasonable constraint because the density of channels in this membrane is limited [2].

During these simulations, inhibitory neurons were not included because only maximum firing rates were under investigation. Thus, the authors concluded [20]:

> Cortical inhibition may not be as critically involved in regulating firing rates of individual cells and producing oscillatory activity as has often previously been assumed. From the perspective of the cell-assembly theory, the role for inhibition in preventing spread of activity among overlapping assemblies and in the shaping of cellular response properties could be emphasized. In fact, in the neocortex a reduction of the inhibition by only 30% (Lindström, personal communication, 1994) leads to epileptiform seizures. This may be an example of activity spreading uncontrollably when inhibition no longer separates the partly overlapping assemblies.

These remarks are in accord with the preceding analysis of the system described by Equation (11.8), where a reduction in the inhibiting parameter (α) below a critical value (α_c) allows all of the assemblies to become simultaneously active. They are also relevant to an evaluation of the field theory models of epilepsy discussed in Section 10.3.

Modeling of Cortical Columns

In a more recent study, Fransén and Lansner have extended their numerical simulations to include *columns* of cells, which corresponds more closely to the structure of the neocortex [21]. In this model, there are 50 functional units (columns) comprising 12 pyramidal neurons and 3 fast-spiking (FS) inhibitory interneurons each (rather than individual cells) for a total of 750 neurons. Pyramidal cells were modeled with six compartments each and FS-cells with three for a total of 4050 compartments.

All of these properties (significant afteractivity, short reaction times, ignition thresholds, pattern completion, competition, and noise suppression) were observed in this more realistic context while rendering the neural interconnections more realistic. Thus, the interconnection probability between pyramidal cells from different columns is both sparse and asymmetric, as is observed in cortical tissue, whereas the interconnection between columns is symmetric, in closer correspondence with the attractor neural networks discussed in Section 10.2.2.

In response to certain qualitative objections raised by Malsburg [58], the numerical studies of Fransén and Lansner have established that Hebb's cell-assembly hypothesis is in approximate accord with both the elementary analysis of Section 11.3 and with present knowledge of cortical structure.

How has the theory fared in current electrophysiology laboratories?

11.7 Recent Evidence for Cell Assemblies

In the half century since Hebb's theory of cell assemblies was first proposed, the experimental techniques of electrophysiology have greatly improved. Classical methods have been refined and new techniques introduced, leading Nicolelis, Fanselow, and Ghazanfar to comment in 1997 [66]:

> What we are witnessing in modern neurophysiology is increasing empirical support for Hebb's views on the neural basis of behavior. While there is much more to be learned about the nature of distributed processing in the nervous system, it is safe to say that the observations made in the last 5 years are likely to change the focus of systems neuroscience from the single neuron to neural ensembles. Fundamental to this shift will be the development of powerful analytical tools that allow the characterization of encoding algorithms employed by distinct neural populations. Currently, this is an area of research that is rapidly evolving.

In assessing this optimistic perspective, it is important to remember that observing the dynamic behavior of a "three-dimensional fishnet" comprising several thousand neurons (each receiving several thousand synaptic inputs) and spread over much of the brain is a daunting task, yet not hopeless. Although there is presently no possibility of taking microelectrode readings from most of the neurons in an assembly, records from as few as two may offer interesting opportunities for research because the experimenter can ask whether the recorded voltages are *correlated* and observe how the degree of correlation depends upon the global behavior of the organism.

Suppose that voltages $V_1(t)$ and $V_2(t)$ are measured from two different neurons of a brain. To learn whether these signals are related to each other, one can compute their correlation as

$$C(\tau) = \int V_1(t+\tau) \times V_2(t) dt\,, \tag{11.10}$$

where the integration is over the greatest practical temporal range. If, for example, $V_1(t)$ is defined between 0 and T_1 and $V_2(t)$ is defined between 0 and T_2, with $T_1 \gg T_2$, then appropriate limits of integration would be from $t = 0$ to T_2. Thus, varying τ over the range

$$0 \le \tau \le T_1 - T_2$$

effectively slides V_1 past V_2.

With data sets of moderate length and currently available computing equipment, this is a straightforward numerical task that indicates how much alike the two signals look. If they are totally unrelated, $C(\tau)$ will be small and random, and one would judge the signals to be uncorrelated. If $C(\tau)$ exhibits some reproducible structure, the signals are partially correlated. Finally, a large peak at some value of τ suggests that part of V_1 looks much like a temporal translation of V_2.

To avoid unrealistic computing times in applying correlation analysis to neural data, it is important to choose discrete approximations of Equation (11.10) that place in evidence the features of interest. Thus an impulse train might be represented by a series of times $(t_1, t_2, \cdots, t_n)$ at which spikes are observed to occur. Dividing the time axis into B "bins" that are larger than the minimum interpulse intervals but much less than the total recording time, an impulse train can then be approximated as the B-dimensional vector $(v_1, v_2, \cdots, v_B)$, where v_j is the number of spikes appearing in the jth bin.

Two such vectors take the form

$$\begin{aligned} \mathbf{V_1} &= (v_{11}, v_{12}, \cdots, v_{1b}, \cdots, v_{1B_1}) \\ \mathbf{V_2} &= (v_{21}, v_{22}, \cdots, v_{2b}, \cdots, v_{2B_2})\,, \end{aligned}$$

where B_1 is not necessarily equal to B_2 because there may be a longer run of reliable data from one measurement than from another.

Assuming that $B_1 \gg B_2$, Equation (11.10) can be approximated as

$$C(\beta) = \sum_{b=1}^{B_2} v_{1(b+\beta)} \times v_{2b}\,. \tag{11.11}$$

Informally, this equation says to slide the longer vector $(\mathbf{V_1})$ along the shorter vector $(\mathbf{V_2})$ by β bins, where β is an integer lying within the range

$$0 \le \beta \le B_1 - B_2\,,$$

multiply the number of spikes in the (B_2) overlapping bins, and add the (B_2) products. A large peak of $C(\beta)$ at some value of β suggests that part of $\mathbf{V_1}$ looks much like a temporal translation of $\mathbf{V_2}$.

As noted in Chapter 1, it is now feasible to measure voltages from several dozen microelectrodes while the subject is undergoing behavioral tests [14, 60, 61, 67, 68, 96]. Because each electrode may indicate the dynamics of several neurons, up to 100 or more individual signals can be simultaneously recorded, analyzed, and compared with concomitant behavior.

Groups of neurons producing correlated signals might be acting as members of a common cell assembly, a possibility that can be checked by comparing correlation functions with behavioral observations. In maze experiments on rats, for example, the experimenter might notice whether such

correlated signals both turn on when a certain behavior begins and turn off when it ceases. If so, it would be reasonable to suspect that correlated neurons are acting as part of an assembly related to that behavior.

Using such techniques, multiple-electrode recordings from a variety of animal species tend to confirm the hypothesis that neurons act not "as single spies but in battalions." Although far from an exhaustive survey of this work, the examples that follow give the flavor of current activities. Many more such results are expected to appear in the next few years.[4]

Mollusk

Multineuronal optical studies of the abdominal ganglion of the mollusk (*Aplysia*) were carried out by Wu, Cohen, and Falk, who recorded from up to 30% of the 900 neurons involved and related this activity to global behavior (gill withdrawal reflex, respiratory pumping, and so on) [98]. Instead of finding neural circuitry developed for specific tasks, these researchers observed that

> different behaviors appear to be generated by altered activities of a single, large distributed network rather than by small dedicated circuits.

Locust

Laurent and colleagues have used several glass microelectrodes to record from projection neurons (PNs) in the antennal lobe (a structural and functional analog of the vertebrate olfactory bulb) of the locust (*Schistocerca americana*) [45, 91]. Focusing on 1 s bursts of stimulants to which the locust has been previously exposed, odor-specific oscillatory responses were observed at 20 Hz, which suggest that memories of different odors are encoded as stimulus-specific assemblies (or "ensembles") of coherently firing neurons [46]. However,

> each odor appears to be represented not simply by an ensemble of synchronized neurons but by a progressive and odor-specific transformation of that ensemble, so that each neuron synchronizes with several others only during one or more precise epochs of the ensemble response.

In a style of interdisciplinary research that one hopes to see more often in coming years, Bazhenov et al. have reported numerical modeling of Laurent's experiments on the locust olfactory system by a team of biologists, neuroscientists and applied mathematicians [4, 5]. Similar to the modeling of the SANS group (described previously in Section 11.5), the

[4]Although some question the ethics of the work, there have been several multiple-electrode experiments on monkeys that also draw conclusions that support the cell-assembly hypothesis [1, 10, 15, 24, 51, 83].

model comprised 90 PNs and 30 inhibitory local neurons (LNs) randomly interconnected with biologically realistic representations of the membrane conductances. Although only one compartment was included in each PN and LN, numerical simulations of about 0.5 s duration showed 20–30 Hz oscillations and pattern discrimination in accord with the foregoing biological measurements. Detailed analyses of the model dynamics led this group to speculate

> that a stimulus does not simply set the initial state of a fixed dynamical system, but instead that each stimulus creates a new and unique dynamical system. This dynamical system has a stimulus-specific global attractor that determines its spatiotemporal response patterns or trajectory.

How is this possible? The global nature of the oscillatory activity, it seems, is rather sensitive to the stimulus pattern delivered to the inhibitory LNs, with different LN patterns igniting different assemblies. Stimulations of the PNs by a certain odor are then translated into a specific PN oscillatory response pattern mediated by the particular assembly that has been ignited.

Moth

Evidence for neural assemblies in the antennal lobe of the moth (*Manduca sexta*) has also emerged from correlation studies of recordings on silicon microprobe arrays published by Christensen et al. [11]. To better represent the brevity of natural odors in the context of turbulent air flow, stimulating pulses were only 0.1 s in duration, yielding data implying that

> the patterns of synchrony among different members of an odor-encoding ensemble are not the same for different concentrations of the same odor. Furthermore, the responses to odor blends cannot necessarily be predicted from the responses to the individual odors in the blend. We therefore propose that ensembles of olfactory PNs must use multiple and overlapping coding strategies to process olfactory information, and that these strategies are matched to the particular circumstances surrounding odor presentation.

At variance with the results of Laurent et al. because of the shorter time scale (0.1 s vs. 1 s), these multiunit recordings again suggest the importance of cell-assembly codes in insect olfactory systems.

Rat

How does a rat get around in the dark? Much as you or I would use our fingers, this little fellow employs his whiskers to interrogate his surroundings. Thus, the cheek (or trigeminal) nerves are of particular interest in understanding a rat's perceptions. With this in mind, Nicolelis and his col-

leagues have been recording from up to 48 (and more recently up to 100) cortical, thalamic, and brainstem neurons of freely moving rats [25, 65, 67]. Widespread oscillations in the range from 7 to 12 Hz were observed, which began when the animals were still but alert and predicted the onset of whisker twitching. Starting as a traveling wave of activity in the cortex (see Figure 10.3), this action spread to the thalamus and the trigeminal brainstem complex. Correlation calculations between pairs of these signals indicate that

> the coding of sensory information in most cortical and subcortical relays of the trigeminal pathways occurs at the ensemble rather than at the single unit level and involves both spatial and temporal domains.

Deadwyler et al. have studied the relationships among recordings from ten different locations in the rat's hippocampus while the animal was undergoing a behavioral learning task [14]. Because the hippocampus is regarded as essential for storage and readout of cerebral information, these experiments were expected to shed light on the nature of neural dynamics during learning. From observations on seven animals, these authors show that

> ensemble encoding and retrieval of "functionally relevant" information are represented as distinct firing patterns in hippocampal networks.

If neuronal assemblies exist in the neocortex, one would hope to be able to switch them on and off—as suggested by Figure 11.5(b) and the dynamic properties of Equations (11.7)—and this is what Maldonado and Gerstein have managed by inserting ten tungsten microelectrodes into the rat's auditory cortex [57]. Both intracortical microstimulation (ICMS) through the electrodes and acoustic stimuli were used as probes during 15 different experiments. Based on correlation analyses of their data, these researchers conclude as follows.

> We have identified neuronal assemblies in two ways, defined through similarity of receptive field properties and defined through correlated firing. Close anatomical spacing between neurons was conducive to, but not sufficient for membership in, the same assembly with either definition. ICMS changed cortical organization by altering assembly membership. Our data showed that neuronal assemblies in the rat's auditory cortex can be established transiently in time and that their membership is dynamic.

Finally, it is interesting to note recent evidence in support of Hebb's phase sequence in which a series of assemblies are ignited one after another to comprise a train of thought [54].

To this end, Louie and Wilson used implanted multielectrodes to record from hippocampal CA1 pyramidal cells of rats (see Figure 9.1), which are known to be "place cells" that tend to fire when the animal is in a particular location [96]. The rats were trained to run around a circular track in search of food, and recordings were made during the actual awake activity (RUN) and also during shorter periods of "rapid eye movement sleep" (REM) [97].

Only those cells judged to be "active" (with firing rates greater than 0.2 Hz) were included in the analysis, leading to impulse train recordings from between 8 and 13 electrodes for a particular experiment. With bin sizes of 1 s and RUN recording times up to 4 minutes, the RUN-REM correlation was computed for each electrode as in Equation (11.11) and then averaged over the electrodes.

Such computations of RUN-REM correlation showed no similarity between the two measurements, but this fails to account for the possibility that the time scale of the REM signal could differ from that of awake activity (RUN). Stretching out (or slowing down) the REM data by a factor of about 2, on the other hand, gave sharply defined correlation peaks that could not be ascribed to happenstance. The authors claim that these results demonstrate that "long temporal sequences of patterned multineuronal activity suggestive of episodic memory traces are reactivated during REM sleep."

11.8 Recapitulation

This chapter opened with a survey of Donald Hebb's seminal formulation of the cell-assembly hypothesis for the robust storage and retrieval of information in the human brain and emphasized key aspects of the theory. Early evidence in support of Hebb's theory was reviewed, including the hierarchical nature of learning, perceptions of ambiguous figures, stabilized image experiments, sensory deprivation experiments, and anatomical data from the structure of the neocortex.

A simple mathematical model for interacting cell assemblies was then developed that describes ambiguous perceptions and suggests the importance of inhibitory interactions among cortical neurons for assembly formation and switching.

This model implies that cell assemblies emerge from intricate closed causal loops (subnetworks) of positive feedback threading sparsely through the neural system. Assemblies exhibit all-or-nothing response and threshold properties (just like the Hodgkin–Huxley impulse or an individual neuron); thus, an assembly is also an attractor. Interestingly, speed of switching from one assembly to another is found to increase with the level of interassembly inhibition. Under simple assumptions, a generous lower bound on the number of complex assemblies that can be stored in a human brain is estimated as about one thousand million—the number of seconds in 30 years.

Conclusions drawn from the simple analytic model are in accord with numerical studies on more realistic neural representations, which predict several hundred milliseconds of significant afteractivity (Hebb's "acting briefly as a closed system"), psychologically reasonable reaction times (less than 100 ms), and pattern recognition (or completion) from imperfect data.

Finally, the concept of correlation was defined and some experimental observations were cited that appear to confirm Hebb's cell-assembly theory in neuronal activities of a mollusk, locust, moth, and rat.

References

[1] M Abeles, H Bergman, I Gat, I Meilijson, E Seidemann, N Tishby, and E Vaadia, Cortical activity flips among quasi-stationary states, *Proc. Nat. Acad. Sci. USA* 92 (1995) 8616–8620.

[2] P Andersen, Factors influencing the efficiency of dendritic synapses on hippocampal pyramidal cells, *Neurosci. Res.* 3 (1986) 521–530.

[3] WR Ashby, H von Foerster, and CC Walker, Instability of pulse activity in a net with threshold, *Nature* 196 (1962) 561–562.

[4] M Bazhenov, M Stopfer, M Rabinovich, HDI Abarbanel, TJ Sejnowski, and G Laurent, Model of transient oscillatory synchronization in the locust antenna lobe, *Neuron* 30 (2001) 553–567.

[5] M Bazhenov, M Stopfer, M Rabinovich, HDI Abarbanel, TJ Sejnowski, and G Laurent, Model of cellular and network mechanisms for odor-evoked temporal patterning in the locust antennal lobe, *Neuron* 30 (2001) 569–581.

[6] A Borsellino and T Poggio, Holographic aspects of temporal memory and optomotor responses, *Kybernetik* 10 (1972) 58–60.

[7] V Braitenberg, Cell assemblies in the visual cortex. In *Theoretical Approaches to Complex Systems,* Springer-Verlag, Berlin, 1978.

[8] V Braitenberg and F Pulvermüller, Entwurf einer neurologischen Theorie der Sprache, *Naturwissenschaften* 79 (1992) 102–117.

[9] V Braitenberg and A Schüz, *Anatomy of the Cortex,* Springer-Verlag, Heidelberg, 1991.

[10] M Castelo-Branco, R Goebel, S Neuenschwander, and W Singer, Neural synchrony correlates with surface segregation rules, *Nature* 405 (2000) 685–689.

[11] TA Christensen, VM Pawlowski, H Lei, and JG Hildebrand, Multi-unit recordings reveal context-dependent modulation of synchrony in odor-specific neural ensembles, *Nat. Neurosci.* 3 (2000) 927–931.

[12] PS Churchland and TJ Sejnowski, *The Computational Brain,* MIT Press, Cambridge, MA, 1992.

[13] GJ Dalanoort, In search of the conditions for the genesis of cell assemblies: A study in self-organization, *J. Soc. Biol. Struct.* 5 (1982) 161–187.

[14] SA Deadwyler, T Bunn, and RE Hampson, Hippocampal ensemble activity during spatial delayed-nonmatch-to-sample performance in rats, *J. Neurosci.* 16 (1996) 354–372.

[15] G Deco, K Laskey, M Diamond, W Freiwald, and E Vaadia, Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies, *Neural Comput.* 12 (2000) 2621–2653.

[16] Ö Ekeberg, P Wallén, A Lansner, H Travén, L Brodin, and S Grillner, A computer based model for realistic simulations of neural networks, I: The single neuron and synaptic interaction, *Biol. Cybern.* 65 (1991) 81–90.

[17] S Frankel, On the design of automata and the interpretation of cerebral behavior, *Psychometrika* 20 (1955) 149–162.

[18] E Fransén, Biophysical simulation of cortical associative memory, Doctoral thesis, Royal Institute of Technology, Stockholm, 1996.

[19] E Fransén, A Lansner, and H Liljenström, A model of cortical memory based on Hebbian cell assemblies. In *Computation and Neural Systems,* FH Eeckman and JM Bower (eds), Kluwer, Boston, 1993.

[20] E Fransén and A Lansner, Low spiking rates in a population of mutually exciting pyramidal cells, *Network* 6 (1995) 271–288.

[21] E Fransén and A Lansner, A model of cortical associative memory based on a horizontal network of connected columns, *Network* 9 (1998) 235–264.

[22] D Gabor, Holographic model of temporal recall, *Nature* 217 (1968) 584.

[23] D Gabor, Improved holographic model of temporal recall, *Nature* 217 (1968) 1288.

[24] AP Georgopoulos, AB Schwartz, and RE Kettner, Neuronal population coding of movement direction, *Science* 233 (1986) 1416–1419.

[25] AA Ghazanfar and MAL Nicolelis, Nonlinear processing of tactile information in the thalamocortical loop, *J. Neurophysiol.* 78 (1997) 506–510.

[26] JS Griffith, On the stability of brain-like structures, *Biophys. J.* 3 (1963) 299–308.

[27] JS Griffith, *A View of the Brain,* Oxford University Press, Oxford, 1967.

[28] JS Griffith, *Mathematical Neurobiology,* Academic Press, New York, 1971.

[29] H Haken, *Synergetics,* third edition, Springer-Verlag, Berlin, 1983.

[30] H Haken, *Advanced Synergetics,* Springer-Verlag, Berlin, 1983.

[31] H Haken, *Principles of Brain Functioning: A Synergetic Approach to Brain Activity,* Springer-Verlag, Berlin, 1996.

[32] DO Hebb, Intelligence in man after large removals of cerebral tissue: Report of four left frontal lobe cases, *J. Gen. Psychol.* 21 (1939) 73–87.

[33] DO Hebb, On the nature of fear, *Physiol. Rev.* 53 (1946) 259–276.

[34] DO Hebb, *Organization of Behavior: A Neuropsychological Theory,* John Wiley & Sons, New York, 1949.

[35] DO Hebb, The structure of thought. In *The Nature of Thought,* PW Jusczyk and RM Klein, (eds), Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[36] DO Hebb, *Essay on Mind,* Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[37] W Heron, The pathology of boredom, *Sci. Am.* January 1957.

[38] RM Klein, D.O. Hebb: An appreciation. In *The Nature of Thought,* PW Jusczyk and RM Klein, (eds), Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[39] C Koch, *Biophysics of Computation,* Oxford University Press, New York, 1999.

[40] T Kohonen, *Associative Memory,* Springer-Verlag, Berlin, 1977.

[41] T Kohonen, P Lehtiö, J Rovamo, J Hyvärinen, K Bry, and L Vainio, A principle of neural associative memory, *Neuroscience* 2 (1977) 1065–1076.

[42] A Lansner, Investigations into the pattern processing capabilities of associative nets, Doctoral thesis, Royal Institute of Technology, Stockholm, 1986.

[43] A Lansner and Ö Ekeberg, Reliability and recall in an associative network, *Trans. IEEE Pattern Anal. Mach. Intell.* PAMI-7 (1985) 490–498.

[44] A Lansner and E Fransén, Modelling Hebbian cell assemblies comprised of cortical neurons, *Network* 3 (1992) 105–119.

[45] G Laurent, Dynamical representation of odors by oscillating and evolving neural assemblies, *Trends Neurosci.* 19 (1996) 489–496.

[46] G Laurent, M Wehr, and H Davidowitz, Temporal representations of odors in an olfactory network, *J. Neurosci.* 16 (1996) 3837–3847.

[47] CR Legéndy, On the scheme by which the human brain stores information, *Math. Biosci.* 1 (1967) 555–597.

[48] CR Legéndy, The brain and its information trapping device. In *Progress in Cybernetics* 1, J Rose (ed), Gordon and Breach, New York, 1969.

[49] CR Legéndy, Three principles of brain structure and function, *Int. J. Neurosci.* 6 (1975) 237–254.

[50] CA Lindbergh, *The Saturday Evening Post,* June 6, 1953.

[51] C Lee, WH Rohrer, and DL Sparks, Population coding of saccadic eye movements by neurons in the superior colliculus, *Nature* 332 (1988) 357–360.

[52] HC Longuet-Higgins, Holographic model of temporal recall, *Nature* 217 (1968) 104.

[53] HC Longuet-Higgins, DJ Willshaw, and OP Buneman, Theories of associative recall, *Q. Rev. Biophys.* 3 (1970) 223–244.

[54] K Louie and MA Wilson, Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep, *Neuron* 29 (2001) 145–156.

[55] R MacGregor and T McMullen, Computer simulation of diffusely connected neuronal populations, *Biol. Cybern.* 28 (1978) 121–127.

[56] M MacMillan, *An Odd Kind of Fame: Stories of Phineas Gage,* MIT Press, Cambridge, MA, 2000.

[57] PE Maldonado and GL Gerstein, Neuronal assembly dynamics in the rat auditory cortex during reorganization induced by intracortical microstimulation, *Exp. Brain Res.* 112 (1996) 431–441.

[58] C von der Malsburg, Synaptic plasticity as basis of brain organization. In *The Neural and Molecular Basis of Learning,* JP Changeux and M Konishi (eds) John Wiley & Sons, New York, 1987.

[59] D Marr, Simple memory, *Philos. Trans. R. Soc. London* 262 (1971) 23–82.

[60] EM Maynard, CT Nordhausen, and RA Normann, The Utah intracortical electrode array: A recording structure for potential brain-computer interfaces, *Electroencephalogr. Clin. Neurophysiol.* 102 (1997) 228–239.

[61] TJ McHugh, KI Blum, JZ Tsien, S Tonegawa, and MA Wilson, Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice, *Cell* 87 (1996) 1339–1349.

[62] R Melzak and TH Scott, The effects of early experience on the response to pain, *J. Comp. Phys. Psychol.* 50 (1957) 155–161.

[63] PM Milner, The cell assembly: Mark II, *Psychol. Rev.* 64 (1957) 242–252.

[64] PM Milner, The mind and Donald O. Hebb, *Sci. Am.* January 1993, 124–129.

[65] MA Nicolelis, LA Baccala, RCS Lin, and JK Chapin, Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system, *Science* 268 (1995) 1353–1358.

[66] MAL Nicolelis, EE Fanselow, and AA Ghazanfar, Hebb's dream: The resurgence of cell assemblies, *Neuron* 19 (1997) 219–221.

[67] MAL Nicolelis, AA Ghazanfar, BM Faggin, S Votaw, and LMO Oliveira, Reconstructing the engram: Simultaneous, multisite, many single neuron recordings, *Neuron* 18 (1997) 529–537.

[68] CT Nordhausen, EM Maynard, and RA Normann, Single unit recording capabilities of a 100 microelectrode array, *Brain Res.* 726 (1996) 129–140.

[69] G Palm, On associative memory, *Biol. Cybern.* 36 (1980) 19–31.

[70] G Palm, On the storage capacity of an associative memory with randomly distributed storage elements, *Biol. Cybern.* 39 (1981) 125–127.

[71] G Palm, Toward a theory of cell assemblies, *Biol. Cybern.* 39 (1981) 181–194.

[72] G Palm, *Neural Assemblies: An Alternative Approach to Artificial Intelligence,* Springer-Verlag, Berlin, 1982.

[73] G Palm, Cell assemblies, coherence, and corticohippocampal interplay, *Hippocampus* 3 (1993) 219–226.

[74] G Palm and T Bonhoeffer, Parallel processing for associative and neural networks, *Biol. Cybern.* 51 (1984) 201–204.

[75] KH Pribram, The neurophysiology of remembering, *Sci. Am.* January 1969, 73–85.

[76] RM Pritchard, W Heron, and DO Hebb, Visual perception approached by the method of stabilized images, *Can. J. Psychol.* 14 (1960) 67–77.

[77] RM Pritchard, Stabilized images on the retina, *Sci. Am.* June 1961, 72–79.

[78] MS Rabinovitch and HE Rosvold, A closed field intelligence test for rats, *Can. J. Psychol.* 4 (1951) 122–128.

[79] A Rapoport, "Ignition" phenomena in random nets, *Bull. Math. Biophys.* 14 (1952) 35–44.

[80] JM Ritchie, On the relation between fiber diameter and conduction velocity in myelinated nerve fibres, *Proc. R. Soc. London* B217 (1982) 29–35.

[81] AH Riesen, The development of visual perception in man and chimpanzee, *Science* 106 (1947) 107–108.

[82] N Rochester, JH Holland, LH Haibt, and WL Duda, Tests on a cell assembly theory of the action of a brain using a large digital computer, *Trans. IRE Inf. Theory* IT–2 (1956) 80–93.

[83] E Seidemann, I Meilijson, M Abeles, H Bergman, and E Vaadia, Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task, *J. Neurosci.* 16 (1996) 752–768.

[84] M von Senden, *Space and Sight: The Perception of Space and Shape in the Congenitally Blind Before and After Operation,* Methuen & Co., London, 1960 (a republication of *Raum-und Gestaltauffassung bei Operierten vor und nach der Operation*, Barth, Leipzig, 1932).

[85] CS Sherrington, *The Integrative Action of the Nervous System,* Yale University Press, New Haven, 1906.

[86] A Shimbel and A Rapoport, A statistical approach to the theory of the central nervous system, *Bull. Math. Biophys.* 10 (1948) 41–45.

[87] DR Smith and CH Davidson, Maintained activity in neural nets, *J. Assoc. Comput. Mach.* 9 (1962) 268–279.

[88] A Surkis, B Taylor, CS Peskin, and CS Leonard, Quantitative morphology of physiologically identified and intracellularly labeled neurons from the guinea-pig laterodorsal segmental nucleus *in vitro, Neuroscience* 74 (1996) 375–392.

[89] WR Thompson and W Heron, The effects of restricting experience on the problem-solving capacity of dogs, *Can. J. Psychol.* 8 (1954) 17–31.

[90] E Trucco, The smallest value of the axon density for which "ignition" can occur in a random net, *Bull. Math. Biophys.* 14 (1952) 365–374.

[91] M Wehr and G Laurent, Odor encoding by temporal sequences of firing in oscillating neural assemblies, *Nature* 384 (1996) 162–166.

[92] H Wigström, Associative recall and formation of stable modes of activity in neural network models, *J. Neurosci. Res.* 1 (1975) 287–313.

[93] G Willwacher, Fähigkeiten eines assoziativen Speichersystems im Vergleich zu Gehirnfunktion, *Biol. Cybern.* 24 (1976) 181–198.

[94] G Willwacher, Storage of a temporal pattern sequence in a network, *Biol. Cybern.* 43 (1982) 115–126.

[95] DJ Willshaw, OP Buneman, and HC Longuet-Higgins, Non-holographic associative memory, *Nature* 222 (1969) 960.

[96] MA Wilson and BL McNaughton, Dynamics of the hippocampal ensemble code for space, *Science* 261 (1993) 1055–1058.

[97] MA Wilson and BL McNaughton, Reactivation of hippocampal ensemble memories during sleep, *Science* 265 (1994) 676–679.

[98] JY Wu, LB Cohen, and CX Falk, Neuronal activity during different behaviors in *Aplysia*: A distributed organization? *Science* 263 (1994) 820–823.

12

The Hierarchical Nature of Brain Dynamics

In previous pages of this book, we have considered mathematical formulations at several levels of neuroscience, from the Newtonian dynamics of individual membrane proteins, through the switching of isolated patches of membrane and the interactions among propagating nerve impulses, to the intricate dynamics of cell assemblies, extending over much of the brain. The picture of the brain that arises from this survey is a *cognitive hierarchy* of distinct dynamic levels in which each level of description is built upon—or emerges from—those below. Because the brain's hierarchical structure is a matter of observation, little is debatable about the preceding statement, but the implications of this perspective for the social sciences are not yet fully appreciated.

Since the demise of behaviorism as a credible theory of the human brain, a variety of alternative formulations have been advanced and are currently the subject of intense discussions among neuroscientists, psychologists, philosophers, and humanists [31]. This final chapter briefly surveys aspects of these debates, paying particular attention to the claims of reductive materialism and closing with a few modest suggestions for future research on the brain's dynamics.

12.1 The Biological Hierarchy

Before taking up the cognitive hierarchy, let us fix ideas by considering a related structure, the *biological hierarchy,*