

Building Network Learning Algorithms from Hebbian Synapses

TERRENCE J. SEJNOWSKI
GERALD TESAURO

In 1949 Donald Hebb published *The Organization of Behavior*, in which he introduced several hypotheses about the neural substrate of learning and memory, including the Hebb learning rule, or Hebb synapse. We now have solid physiological evidence, verified in several laboratories, that long-term potentiation (LTP) in some parts of the mammalian hippocampus follows the Hebb rule (Brown, Ganong, Kariss, Keenan, & Kelso, 1989; Kelso, Ganong, & Brown, 1986; Levy, Brassel, & Moore, 1983; McNaughton, Douglas, & Goddard, 1978; McNaughton & Morris, 1987; Wigstrom and Gustafsson, 1985). The Hebb rule and variations on it have also served as the starting point for the study of information storage in simplified "neural network" models (Hopfield & Tank, 1986; Kohonen, 1984; Rumelhart & McClelland, 1986; Sejnowski, 1981). Many types of networks have been studied—networks with random connectivity, networks with layers, networks with feedback between layers, and a wide variety of local patterns of connectivity. Even the simplest network model has complexities that are difficult to analyze.

In this chapter we will provide a framework within which the Hebb rule serves as an important link between the implementation level of analysis, the level at which experimental work on neural mechanisms takes place, and the algorithmic level, on which much of the work on learning in network models is being pursued.

LEARNING MECHANISMS, LEARNING RULES, AND LEARNING ALGORITHMS

Long-term potentiation has been found in a variety of preparations. The common denominator is a long-lasting change in synaptic efficacy or spike coupling following afferent stimulation with a high-frequency tetanus. There are probably

several different molecular mechanisms underlying LTP in different preparations. For example, LTP in the CA1 region of the rat hippocampus can be blocked with 2-amino-5-phosphonovaleric acid (APV), but the same application of APV does not block LTP from the mossy fibers in the CA3 region of the hippocampus (Chattarji, Stanton, & Sejnowski, 1988). How LTP is implemented in these different regions is a question at the level of the learning mechanism.

The Hebb synapse, in contrast, is a rule rather than a mechanism. That is, many mechanisms can be used to implement a Hebb synapse, as we will show in a later section. A learning rule specifies only the general conditions under which plasticity should occur, such as the temporal and spatial relationships between the presynaptic and postsynaptic signals, but not the locus of plasticity, or even the neuronal geometry. As an extreme example, we show how the Hebb rule can be implemented without synaptic plasticity.

A learning algorithm is more general than a learning rule since an algorithm must also specify how the learning rule is to be used to perform a task, such as storing information or wiring up a neural system. Thus a description of the task to be performed and the type of information involved are essential ingredients of a learning algorithm. Several examples of learning algorithms will be discussed in later sections.

IMPLEMENTATIONS OF THE HEBB RULE

The Hebb Rule

Before considering the various possible ways of implementing the Hebb rule, one should examine what Hebb actually proposed:

When an axon of cell *A* is near enough to excite cell *B* or repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that *A*'s efficiency, as one of the cells firing *B*, is increased. (p. 62)

This statement can be translated into a precise quantitative expression as follows. We consider the situation in which neuron *A*, with average firing rate V_A , projects to neuron *B*, with average firing rate V_B . The synaptic connection from *A* to *B* has a strength value T_{BA} , which determines the degree to which activity in *A* is capable of exciting *B*. (The postsynaptic depolarization of *B* due to *A* is usually taken to be the product of the firing rate V_A times the synaptic strength value T_{BA} .) Now the preceding statement by Hebb states that the strength of the synapse T_{BA} should be modified in some way that is dependent on both activity in *A* and activity in *B*. The most general expression that captures this notion is

$$\Delta T_{BA} = F(V_A, V_B) \quad (1)$$

which states that the change in the synaptic strength at any given time is some as yet unspecified function F of both the presynaptic firing rate and the postsyn-

aptic firing rate. Strictly speaking, we should say that $F(V_A, V_B)$ is a functional, since the plasticity may depend on the firing rates at previous times as well as at the current time. Given this general form of the assumed learning rule, it is then necessary to choose a particular form for the function $F(V_A, V_B)$. The most straightforward interpretation of what Hebb said is a simple product:

$$\Delta T_{BA} = \epsilon V_A V_B \quad (2)$$

where ϵ is a numerical constant usually taken to be small. There are many other choices possible for the function $F(V_A, V_B)$; the choice depends on the particular task at hand. Equation 2 might be appropriate for a simple associative memory task, but for other tasks one would need different forms of the function $F(V_A, V_B)$ in Eq. 1. For example, in classical conditioning, as we shall see in the following section, the precise timing relationships of the presynaptic and postsynaptic signals are important, and the plasticity must then depend on the rate of change of firing, or on the "trace" of the firing rate (i.e., a weighted average over previous times), rather than simply depending on the current instantaneous firing rate. Once the particular form of the learning algorithm is established, the next step is to decide how the algorithm is to be implemented. We shall describe here three possible implementation schemes. This is meant to illustrate the variety of schemes possible.

Three Implementations

The first implementation scheme, seen in Figure 17.1a, is the simplest way to implement the proposed plasticity rule. The circuit consists solely of neurons A and B and a conventional axodendritic or axosomatic synapse from A to B . One postulates that there is some molecular mechanism operating on the postsynaptic side of the synapse which is capable of sensing the rate of firing of both cells and which changes the strength of synaptic transmission from cell B to cell A according to the product of the two firing rates. This is in fact similar to the recently discovered mechanism of associative LTP that has been studied in rat hippocampus (Brown et al., 1989). (Strictly speaking, the plasticity in LTP depends not on the postsynaptic firing rate, but instead on the postsynaptic depolarization. However, in practice these two are usually closely related [Kelso et al., 1986].) Even here, many different molecular mechanisms are possible. For example, even though the induction of plasticity occurs at a postsynaptic site, the long-term structural change may well be presynaptic (Dolphin, Errington, & Bliss, 1982).

A second possible implementation scheme for the Hebb rule is seen in Figure 17.1b. In this circuit there is now a feedback projection from the postsynaptic neuron, which forms an axoaxonic synapse on the projection from A to B . The plasticity mechanism involves presynaptic facilitation: one assumes that the strength of the synapse from A to B is increased in proportion to the product of the presynaptic firing rate times the facilitator firing rate (i.e., the postsynaptic

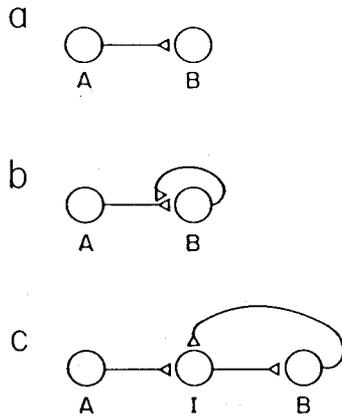


FIGURE 17.1. Three implementations of the Hebb rule for synaptic plasticity. The strength of the coupling between cell A and cell B is strengthened when they are both active at the same time. (a) Postsynaptic site for coincidence detection. (b) Presynaptic site for coincidence detection. (c) Interneuron detects coincidence.

firing rate). This type of mechanism also exists and has been extensively studied in *Aplysia* (Carew, Hawkins, & Kandel, 1983; Kandel et al., 1987). Several authors have pointed out that this circuit is a functionally equivalent way of implementing the Hebb rule (Gelperin, Hopfield, & Tank, 1985; Hawkins & Kandel, 1984; Tesauro, 1986).

A third scheme for implementing the Hebb rule—one that does not specifically require plasticity in individual synapses (Tesauro, 1988)—is seen in Figure 17.1c. In this scheme the modifiable synapse from *A* to *B* is replaced by an interneuron *I* with a modifiable threshold for initiation of action potentials. The Hebb rule is satisfied if the threshold of *I* decreases according to the product of the firing rate in the projection from *A* times the firing rate in the projection from *B*. This is quite similar, although not strictly equivalent, to the literal Hebb rule, because the effect of changing the interneuron threshold is not identical to the effect of changing the strength of a direct synaptic connection. A plasticity mechanism similar to the one proposed here has been studied in *Hermisenda* (Alkon, 1987; Farley & Alkon, 1985).

The three methods for implementing the Hebb rule seen in Figure 17.1 are by no means exhaustive. There is no doubt that nature is more clever than we are at designing mechanisms for plasticity, especially since we are not aware of most evolutionary constraints. These three circuits can be considered equivalent circuits since they effectively perform the same function even though they differ in the way that they accomplish it. There also are many ways that each circuit could be instantiated at the cellular and molecular levels. Despite major differences between them, we can nonetheless say that they all implement the Hebb rule.

Most synapses in cerebral cortex occur on dendrites where complex spatial interactions are possible. For example, the activation of a synapse might depolarize the dendrite sufficiently to serve as the postsynaptic signal for modifying an adjacent synapse. Such cooperativity between synapses is a generalization of the Hebb rule in which a section of dendrite rather than the entire neuron is considered the functional unit (Finkel & Edelman, 1987). Dendritic compart-

ments with voltage-dependent channels have all the properties needed for non-linear processing units (Shepherd et al., 1985).

USES OF THE HEBB RULE

Conditioning

The Hebb rule can be used to form associations between one stimulus and another. Such associations can either be static, in which case the resulting neural circuit functions as an associative memory (Anderson, 1970; Kohonen, 1970; Longuet-Higgins, 1968; Steinbuch, 1961); or they can be temporal, in which case the network learns to predict that one stimulus pattern will be followed at a later time by another. The latter case has been extensively studied in classical conditioning experiments, in which repeated temporally paired presentations of a conditioned stimulus (CS) followed by an unconditioned stimulus (US) cause the animal to respond to the CS in a way that is similar to its response to the US. The animal has learned that the presence of the CS predicts the subsequent presence of the US. A simple neural circuit model of the classical conditioning process that uses the Hebb rule is illustrated in Figure 17.2. This circuit contains three neurons: a sensory neuron for the CS, a sensory neuron for the US, and a motor neuron, R, which generates the unconditioned response. There is a strong, unmodifiable synapse from US to R, so that the presence of the US automatically evokes the response. There is also a modifiable synapse from CS to R, which in the naive untrained animal is initially weak.

One might think that the straightforward application of the literal interpretation of the Hebb rule, as expressed in Eq. 2, would suffice to generate the desired conditioning effects in the circuit of Figure 17.2. However, there are a number of serious problems with this learning algorithm. One of the most serious is the lack of timing sensitivity in Eq. 2. Learning would occur regardless of the order in which the neurons came to be activated. However, in conditioning we know that the temporal order of stimuli is important: if the US follows the CS, then learning occurs, whereas if the US appears before the CS, then no learning occurs. Hence Eq. 2 must be modified in some way to include this timing sen-

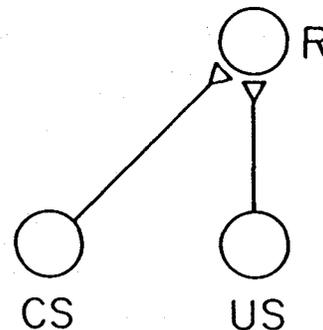


FIGURE 17.2. Model of classical conditioning using a modified Hebb synapse. The US elicits a response in the postsynaptic cell (R). Coincidence of the response with the CS leads to strengthening of the synapse between CS and R.

sitivity. Another serious problem is a sort of "runaway instability" that occurs when the CS-R synapse is strengthened to the point where activity in the CS neuron by itself causes the R neuron to fire. In that case, Eq. 2 would cause the synapse to be strengthened upon presentation of the CS alone, without being followed by the US. However, in real animals we know that presentation of CS alone causes a learned association to be extinguished, that is, the synaptic strength should decrease, not increase. The basic problem is that algorithm 2 is capable of generating positive learning only, and has no way to generate zero or negative learning.

It is clear then that the literal Hebb rule needs to be modified to produce desired conditioning phenomena (Klopf, 1987; Tesauro, 1986). One of the most popular ways to overcome the problems of the literal Hebb rule is by using algorithms such as the following:

$$\Delta T_{BA} = \epsilon \dot{V}_B \bar{V}_A \quad (3)$$

Here \bar{V}_A represents the stimulus trace of V_A , or the weighted average of V_A over previous times, and \dot{V}_B represents the time derivative of V_B . The stimulus trace provides the required timing sensitivity so that learning occurs only in forward conditioning and not in backward conditioning. The use of the time derivative of the postsynaptic firing rate, rather than the postsynaptic firing rate, is a way of changing the sign of learning and thus avoiding the runaway instability problem. With this algorithm, extinction would occur because upon onset of the CS, no positive learning takes place due to the presynaptic trace, and negative learning takes place upon offset of the CS. Many other variations and elaborations of Eq. 3 behave differently, taking into account other conditioning behaviors such as second-order conditioning and blocking. For details we refer the reader to Gelperin et al. (1985), Gluck and Thompson (1987), Klopf (1987), Sutton (1987), Sutton and Barto (1981), and Tesauro (1986). However, all of these other algorithms are built on the same basic notion of modifying the literal Hebb rule to incorporate a mechanism of timing sensitivity and a mechanism for changing the sign of learning.

Associative Memory

Memory and learning are behavioral phenomena, but there are correlates at many structural levels, from the molecular to the systems levels. Hebb (1949) went beyond the synaptic and cellular levels to speculate about information processing in networks of neurons, or assemblies as he called them, which he considered the fundamental unit of processing in the cerebral cortex. Information from the environment is encoded in an assembly by changing the synaptic strengths at many synapses simultaneously. Even though the Hebb rule is local, in the sense that only local information is needed to make a decision about how to change the strength of the synapse, the global behavior of the network of neu-

rons may be affected. This raises the possibility that new principles might emerge when assemblies of neurons are studied.

Probably the most important and most thoroughly explored use of the Hebb rule is in the formation of associations between one stimulus or pattern of activity and another. The Hebb rule is appealing for this use because it provides a way of forming global associations between macroscopic patterns of activity in assemblies of neurons using only the local information available at individual synapses.

The earliest models of associative memory were based on network models in which the output of a model neuron was assumed to be proportional to a linear sum of its inputs, each weighted by a synaptic strength. Thus

$$V_B = \sum_{A=1}^N T_{BA} V_A \quad (4)$$

where V_B are the firing rates of a group of M output cells, V_A are the firing rates of a group of N input cells, and T_{BA} is the synaptic strength between input cell A and output cell B . Note that A and B are being used here as arbitrary indices to represent one out of a group of cells.

The transformation between patterns of activity on the input vectors to patterns of activity on the output vectors is determined by the synaptic weight matrix, T_{BA} . How should this matrix be chosen if the goal of the network is to associate a particular output vector with a particular input vector? The earliest suggestions were all based on the Hebb rule in Eq. 2 (Anderson, 1970; Kohonen, 1970; Longuet-Higgins, 1968; Steinbuch, 1961). It is easy to verify by direct substitution of Eq. 2 into Eq. 4 that the increment in the output is proportional to the desired vector and the strength of the learning ϵ can be adjusted to scale the outputs to the desired values.

More than one association can be stored in the same matrix, so long as the input vectors are not too similar to one another. This is accomplished by using Eq. 2 for each input-output pair. This model of associative storage is simple and has several attractive features: (1) the learning occurs in only one trial; (2) the information is distributed over many synapses, so that recall is relatively immune to noise or damage; and (3) input patterns similar to stored inputs will give output similar to the stored outputs, a form of generalization. This model also has some strong limitations. First, stored items with input vectors that are similar (i.e., that have a significant overlap) will produce outputs that are mixtures of the stored outputs. However, discriminations must often be made among similar inputs, such as the phonetic distinction between the labial stops in the words "bet" and "pet." Second, the linear model cannot respond contingently to pairs of inputs (i.e., those that have an output that is different from the sum of the individual outputs). Some deficiencies can be remedied by making the learning algorithm and the architecture of the network more complex, as shown in the next section.

The Covariance Rule

Numerous variations have been proposed on the conditions for Hebbian plasticity (Levy, Anderson, & Lehmkuhle, 1984). One problem with any synaptic modification rule that can only increase the strength of a synapse is that the synaptic strength will eventually saturate at its maximum value. Nonspecific decay can reduce the sizes of the weights, but the stored information will also decay and be lost at the same rate. Another approach is to renormalize the total synaptic weight of the entire terminal field from a single neuron to a constant value (von der Malsburg, 1973). Sejnowski (1977a, 1977b) emphasized the need for a learning rule that decreases the strength of a plastic synapse as specifically as the Hebb rule increases it and proposed a covariance learning rule. According to this rule, the change in strength of a plastic synapse should be proportional to the covariance between the presynaptic firing and postsynaptic firing:

$$\Delta T_{BA} = \epsilon (V_B - \langle V_B \rangle) (V_A - \langle V_A \rangle) \quad (5)$$

where $\langle V_B \rangle$ is the average firing rates of the output neurons and $\langle V_A \rangle$ is the average firing rates of the input neurons (see also Chauvet, 1986). Thus the strength of the synapse should increase if the firings of the presynaptic and postsynaptic elements are positively correlated, decrease if they are negatively correlated, and remain unchanged if they are uncorrelated. Evidence for long-term depression has been found in the hippocampus (Levy et al., 1983; Stanton, Jester, Chatterji, & Sejnowski, 1988) and in visual cortex during development (Reiter & Stryker, 1987; Fregnac, Shulz, Thorpe, & Bienenstock, 1988).

The covariance rule is a special case of the general form of the Hebb rule in Eq. 1. It does go beyond the simple Hebb rule in Eq. 2; however, it is easy to show that traditional Hebb synapses can be used to implement Eq. 5, which can be rewritten as

$$\Delta T_{BA} = \epsilon (\langle V_B V_A \rangle - \langle V_B \rangle \langle V_A \rangle) \quad (6)$$

Both terms on the right-hand side have the same form as the simple Hebb synapse in Eq. 2. Thus the covariance learning algorithm can be realized by applying the Hebb rule relative to a "threshold" that varies with the product of the time-averaged presynaptic and postsynaptic activity levels. The effect of the threshold is to ensure that no change in synaptic strength should occur if the average correlation between the presynaptic and postsynaptic activities is at chance level.

Error-Correction Learning

One of the consequences of the linear associative matrix model with Hebbian synapses is that similar input vectors necessarily produce similar output vectors. Error-correction procedures can be used to reduce this interference. The weights

are changed to minimize the difference between the actual and correct output vectors:

$$\Delta T_{BA} = \epsilon (V_B^{(a)} - V_B^{(c)}) V_A \quad (7)$$

where $V_B^{(a)}$ is the actual output produced by the network by the current set of weights and $V_B^{(c)}$ is the correct output vector supplied by the teacher. Unlike the previous learning algorithms, which learn in one shot, error-correction procedures such as this are incremental and require several presentations of the same set of input vectors. It can be shown that the weights will evolve to minimize the average mean square error over the set of input vectors (Kohonen, 1984).

As with the covariance learning algorithm, the error-correction learning algorithm in Eq. 7 can be rewritten in a form that can be implemented with Hebb synapses:

$$\Delta T_{BA} = \epsilon (V_B^{(a)} V_A - V_B^{(c)} V_A) \quad (8)$$

Both terms on the right-hand side have the same form as the simple Hebb synapse in Eq. 2. Thus the error-correction procedure can be realized by applying the Hebb rule twice, first to the actual output produced by the network and then to the correct output supplied by a teacher, but with a negative rather than a positive increment. Alternatively, the difference can be computed by another neuron and used to control the Hebbian learning.

Further improvements have also been made to associative matrix models by introducing feedback connections, so that the networks are autoassociative, and by making the processing units nonlinear (Anderson & Mozer, 1981; Hopfield, 1984; Kohonen, 1984; Sejnowski, 1981; Toulouse, Dehaene, & Changeux, 1986). Associative memory models like this have been proposed for the CA3 region of the hippocampus (Lynch & Baudry, 1988; McNaughton & Morris, 1987; Rolls, 1987).

The learning algorithms for associative memory introduced in this section are supervised in the sense that information about the desired output vectors must be supplied along with the input vectors. One way to provide information about the desired output to a group of neurons is to have a separate "teaching" input that "clamps" the output firing rates to the desired values while the input corresponding to the desired output is simultaneously active. The climbing fibers in the cerebellar cortex, which make strong excitatory synapses on individual Purkinje cells, could have such a teaching role, as first suggested by Brindley (1964) and later developed by Albus (1971) and Marr (1969) in models of the cerebellar cortex as an adaptive filter. Evidence for plasticity in the cerebellar cortex has been found by Ito and his co-workers (Ito, 1982). However, evidence for Hebbian plasticity in the cerebellum does not necessarily imply that its function is associative storage, for there are many other possible functions. Evidence

for plasticity in the deep cerebellar nuclei has been found as well (Miles & Lisberger, 1981; Thompson, 1986).

Learning Internal Representations

The class of network models of associative memory discussed in the last section has a severe computational limitation in that all the processing units in the network are constrained by either the inputs or the outputs, so that there are no free units that could be used to form new internal features. What features should be used for the input units and output units if the network is deeply buried in the brain? New learning algorithms have been devised for multilayer networks with nonlinear processing units that overcome some of the limitations of single-layer networks (Hinton & Sejnowski, 1983; Rumelhart & McClelland, 1986). In particular, these algorithms use interneurons, or "hidden units," which become sensitive to the features that are appropriate for solving a specified problem and for performing context-sensitive computation. We will review one of these learning algorithms, based on the Boltzmann machine architecture, which can be implemented by Hebb synapses.

The Boltzmann machine is a network of stochastic processing units that solves optimization problems (Hinton & Sejnowski, 1983, 1986). The processing units in a Boltzmann machine have outputs that are binary valued and are updated probabilistically from summed synaptic inputs, which are graded. As a consequence, the state of the units in a Boltzmann machine fluctuate even for a constant input. The amount of fluctuation is controlled by a parameter that is analogous to the temperature of a thermodynamic system. Fluctuations allow the system to escape from local traps into which it would get stuck if there were no noise in the system. All the units in a Boltzmann machine are symmetrically connected: this allows an "energy" to be defined for the network and ensures that the network will relax to an equilibrium state which minimizes the energy (Hopfield, 1982).

The Boltzmann machine has been applied to a number of constraint satisfaction problems in vision, such as figure-ground separation in image analysis (Kienker, Sejnowski, Hinton, & Schumacher, 1986; Sejnowski & Hinton, 1987), and generalizations have been applied to image restoration (Geman & Geman, 1984), binocular depth perception (Divko & Schulten, 1986), and optical flow (Hutchinson, Koch, Luo, & Mead, 1988). These are problems in which many small pieces of evidence must be combined to arrive at the best overall interpretation of sensory inputs (Ballard, Hinton, & Sejnowski, 1983; Hopfield & Tank, 1986).

Boltzmann machines have an interesting learning algorithm that allows "energy landscapes" to be formed within the hidden units between the input and output layers. Learning in a Boltzmann machine has two phases. In the training phase a binary input pattern is imposed on the input group as well as the correct binary output pattern. The system is allowed to relax to equilibrium at a fixed

temperature while the inputs and outputs are held fixed. In equilibrium, the average fraction of time a pair of units is on together, the co-occurrence probability, is computed for each connection:

$$P_{BA}^+ = \langle S_B S_A \rangle |_{\text{clamped}} \quad (9)$$

where S_B is the output value of the B th unit, which can take on the values 0 or 1 only. In the test phase the same procedure is followed with only the input units clamped, and the average co-occurrence probabilities are again computed:

$$P_{BA}^- = \langle S_B S_A \rangle |_{\text{free}} \quad (10)$$

The weights are then updated according to

$$\Delta T_{BA} = \epsilon (P_{BA}^+ - P_{BA}^-) \quad (11)$$

where the parameter ϵ controls the rate of learning. A co-occurrence probability is related to the correlation between the firing or activation of the presynaptic and postsynaptic units and can therefore be implemented by a Hebb synapse. In the second phase, however, the change in the synaptic strengths is anti-Hebbian since it must decrease with increasing correlation. Notice that this procedure is also error-correcting, since no change will be made to the weight if the two probabilities are the same.

The Boltzmann machine demonstrates that the Hebb learning rule can be used to mold the response properties of interneurons within a network and adapt them for the efficient solution of difficult computational problems. The Boltzmann machine also shows that noise can play an effective role in improving performance of a network, and that the presence of noise in the nervous system does not necessarily imply a lack of precision. Other stochastic network models have also been studied (Barto, 1985). Although the Boltzmann machine is not meant to be a realistic brain model, it does serve as an existence proof that difficult computational problems can be solved with relatively simple processing units and biologically plausible learning mechanisms.

Development

The Hebb synapse has also been used by Linsker (1986) to model the formation of receptive fields in the early stages of visual processing. The model is a layered network having limited connectivity between layers, and it uses the covariance generalization of the Hebb rule given in Eq. 5. As the learning proceeds, the units in the lower layers of the network develop on-center and off-center receptive fields that resemble the receptive fields of ganglion cells in the retina, and elongated receptive fields develop in the upper layers of the network that resemble simple receptive fields found in visual cortex. This model demonstrates that some of the properties of sensory neurons could arise spontaneously during

development by specifying the general pattern of connectivity and a few parameters to control the synaptic plasticity. One surprising aspect of the model is that regular receptive fields develop even though only spontaneous activity is present at the sensory receptors. Similar models that require patterned inputs have also been proposed (Barrow, 1987; Bienenstock, Cooper, & Munro, 1982; von der Malsburg, 1973).

The visual response properties of neurons in the visual cortex of cats and monkeys are plastic during the first few months of postnatal life and can be permanently modified by visual experience (Hubel & Wiesel, 1962; Sherman & Spear, 1982). Normally, most cortical neurons respond to visual stimuli from either eye. Following visual deprivation of one eye by eyelid suture during the critical period, the ocular preference of neurons in primary visual cortex shifts toward the nondeprived eye. In another type of experiment, a misalignment of the two eyes during the critical period produces neurons that respond to only one eye, and, as a consequence, binocular depth perception is impaired. These and many other experiments have led to testable hypotheses for the mechanisms underlying synaptic plasticity during the critical period (Bear, Cooper, & Ebner, 1987).

Stent (1973) suggested that the effects of monocular deprivation could be explained if the synaptic weight were to decrease when the synapse is inactive and the postsynaptic cell is active. An alternative mechanism that incorporates another Hebbian form of plasticity was proposed by Bienenstock et al. (1982). The Bienenstock–Cooper–Munro algorithm for synaptic modification is a special case of the general Hebb rule in Eq. 1:

$$\Delta T_{BA} = \phi(V_B, \langle V_B \rangle) V_A \quad (12)$$

where the function $\phi(V_B, \langle V_B \rangle)$ is shown in Figure 17.3. The synapse is strengthened when the average postsynaptic activity exceeds a threshold and is weakened when the activity falls below the threshold level. Furthermore, the threshold varies according to the average postsynaptic activity:

$$\theta = \langle V_B \rangle^2. \quad (13)$$

Bienenstock et al. (1982) showed that this choice has desirable stability properties and allows neurons to become selectively sensitive to common features in input patterns.

Singer (1987) suggested that the voltage-dependent entry of calcium into spines and the dendrites of postsynaptic cells may trigger the molecular changes required for synaptic modification. This hypothesis is being tested at a molecular level using combined pharmacological and physiological techniques. *N*-methyl-D-aspartic acid (NMDA) receptor antagonists infused into visual cortex block the shift in ocular dominance normally associated with monocular deprivation (Kleinschmidt, Bear, & Singer, 1986). The NMDA receptor is a candidate mechanism for triggering synaptic modification because it allows calcium to enter a

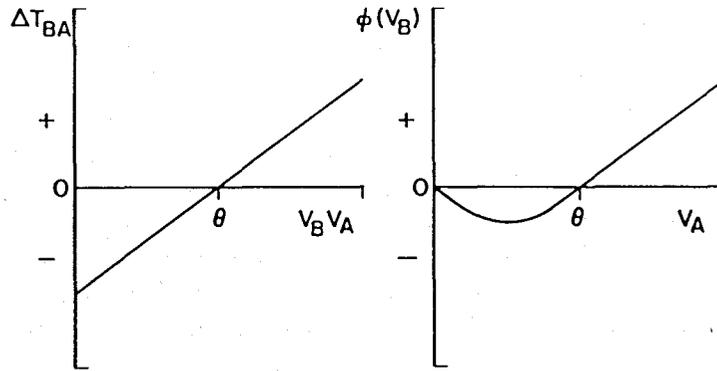


FIGURE 17.3. (left) Change in coupling strength ΔT_{BA} as a function of the correlation $\langle V_B V_A \rangle$ between the presynaptic and postsynaptic activity levels, as indicated in Eq. 5. The threshold θ is given by $\langle V_B \rangle \langle V_A \rangle$. (right) The postsynaptic factor $\phi(V_B, \langle V_B \rangle)$ in the Bienenstock–Cooper–Munro learning algorithm in Eq. 1, where the threshold θ is given by $\theta = \langle V_B \rangle^2$.

cell only if the neurotransmitter binds to the receptor and the postsynaptic membrane is depolarized. In a sense, the NMDA receptor is a “Hebb molecule,” since it is activated only when there is a conjunction of presynaptic and postsynaptic activity. The NMDA receptor also is involved in LTP in the hippocampus (Collingridge, Kehl, & McLennan, 1983).

The mechanisms for plasticity in the cerebral cortex during development may be related to the mechanisms that are responsible for synaptic plasticity in the adult. The evidence so far favors the general form of Hebbian plasticity in Eq. 1. However, the details of how this plasticity is regulated at short and long time scales may be quite different during development and in the adult.

CONCLUSIONS

The algorithmic level is a fruitful one for pursuing network models at the present time for two reasons. First, working top-down from functional considerations is difficult, since our intuitions about the functional level in the brain may be wrong or misleading. Knowing more about the computational capabilities of simple neural networks may help us gain a better intuition. Second, working from the bottom up can be treacherous, since we may not yet know the relevant signals in the nervous system that support information processing. The study of learning in model networks can help guide the search for neural mechanisms underlying learning and memory. Thus network models at the algorithmic level are a unifying framework within which to explore neural information processing.

Three principles have emerged from our studies of learning in neural networks. The first is the principle of locality. The Hebb algorithm depends only

on information that is present or can be extracted from highly localized regions of space and time. This is of practical importance for any physical system since nonlocal algorithms have a high overhead for the communications needed to bring together the relevant information. In spite of the limitation of locality, networks can nonetheless achieve a global organization during both development and long-term information storage. The second principle is gradient descent. A global energy or cost function can usually be found whenever large-scale network organization emerges from local interactions. That is, global organization is the result of local changes which optimize a function of the entire system (Sejnowski, 1987). This raises the important possibility that such global functions may be exploited in the nervous system and may be discoverable (Sejnowski, 1987). The third is the principle of differences. Given the limited accuracy of signals in neurons, mechanisms that depend on accurate, absolute values are not feasible. Information storage can be made more compact when differences between signals—effectively error signals—are used to make changes at synapses.

Hebb's learning rule has led to a fruitful line of experimental research and a rich set of network models. The Hebb synapse is a building block for many different neural network algorithms. As experiments refine the parameters for Hebbian plasticity in particular brain areas, it should become possible to begin refining network models for those areas. There is still a formidable gap between the complexity of real brain circuits and the simplicity of the current generation of network models. As models and experiments evolve the common bonds linking them are likely to be postulates like the Hebb synapse that serve as algorithmic building blocks.

Acknowledgment

This chapter was prepared with the support of the Mathers Foundation.

REFERENCES

- Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, 10, 25–61.
- Alkon, D. L. (1987). *Memory traces in the brain*. Oxford: Oxford University Press.
- Anderson, J. A. (1970). Two models for memory organization using interacting traces. *Mathematical Biosciences*, 8, 137–160.
- Anderson, J. A., & Mozer, M. C. (1981). Categorization and selective neurons. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory*. Hillsdale, N.J.: Erlbaum.
- Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature*, 306, 21–26.
- Barrow, H. G. (1987). Learning receptive fields. In M. Caudill & C. Butler (Eds.), *Proceedings of the First International Conference on Neural Networks* (Vol. 4, pp. 115–121). San Diego: SOS Press.
- Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, 4, 229–256.
- Bear, M., Cooper, L. N., & Ebner, F. F. (1987). A physiological basis for a theory of synapse modification. *Science*, 237, 42–48.

- Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2, 32-48.
- Brindley, G. S. (1964). The use made by the cerebellum of the information that it receives from sense organs. *International Brain Research Organization Bulletin*, 3, 80.
- Brown, T. H., Ganong, A. H., Kariss, E. W., Keenan, C. L., & Kelso, S. R. (1989). Long-term potentiation in two synaptic systems of the hippocampal brain slice. In J. H. Byrne & W. O. Berry (Eds.), *Neural models of plasticity*. Orlando, Fla.: Academic Press.
- Carew, T. J., Hawkins, R. D., & Kandel, E. R. (1983). Differential classical conditioning of a defensive withdrawal reflex in *Aplysia californica*. *Science*, 219, 397-400.
- Chattarji, S., Stanton, P., & Sejnowski, T. J. (1988). Commissural, but not mossy fiber, synapses exhibit both associative long-term potentiation (LTP) and depression (LTD) in the CA3 region of the hippocampus. *Society for Neuroscience Abstracts*, 14, 567.
- Chauvet, G. (1986). Habituation rules for a theory of the cerebellar cortex. *Biological Cybernetics*, 55, 201-209.
- Churchland, P. S., & Sejnowski, T. J. (1988). Neural representations and neural computations. In L. Nadel (Ed.), *Neural connections and mental computation* (pp. 15-48). Cambridge, Mass.: MIT Press.
- Collingridge, G. L., Kehl, S. L., & McLennan, H. (1983). *Journal of Physiology (London)*, 334, 33.
- DiPrisco, G. V. (1984). Hebb synaptic plasticity. *Progress in Neurobiology*, 89, 98-102.
- Divko, R., & Schulten, K. (1986). Stochastic spin models for pattern recognition. In J. S. Denker (Ed.), *AIP Conference Proceedings 151: Neural Networks for Computing* (pp. 129-134). New York: AIP.
- Dolphin, A. C., Errington, M. L., & Bliss, T. V. P. (1982). *Nature*, 297, 496.
- Farley, J., & Alkon, D. L. (1985). Cellular mechanisms of learning, memory and information storage. *Annual Review of Psychology*, 36, 419-494.
- Finkel, L. H., & Edelman, G. M. (1987). Population rules for synapses in networks. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Synaptic function*. (pp. 711-757). New York: Wiley.
- Fregnac, Y., & Imbert, M. (1984). Development of neuronal selectivity in the primary visual cortex of the cat. *Physiology Review*, 64, 325.
- Fregnac, Y., Shulz, D., Thorpe, S., & Bienenstock, E. (1988). A cellular analogue of visual cortical plasticity. *Nature*, 333, 367-370.
- Gelperin, A. (1986). Complex associative learning in small neural networks. *Trends in Neurosciences*, 9, 323-328.
- Gelperin, A., Hopfield, J. J., & Tank, D. W. (1985). The logic of Limax learning. In A. Selverston (Ed.), *Model neural networks and behavior* (pp. 237-261). New York: Plenum.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3, 79-92.
- Gluck, M. A., & Thompson, R. F. (1987). Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review*, 94, 176-191.
- Hawkins, R. D., & Kandel, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, 91, 375-391.
- Hebb, D. O. (1949). *Organization of behavior*. New York: Wiley.
- Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 448-453). Silver Spring, Md.: IEEE Computer Society Press.
- Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (pp. 282-317). Cambridge, Mass.: MIT Press.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective compu-

- tational abilities. *Proceedings of the National Academy of Sciences USA*, 79, 2554–2558.
- Hopfield, J. J. (1984). Neurons with graded response have collective computation abilities. *Proceedings of the National Academy of Sciences USA*, 81, 3088–3092.
- Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science*, 233, 625–633.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interactions, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154.
- Hutchinson, J., Koch, C., Luo, J., & Mead, C. (1988). *IEEE Computer*, 21(3), 52–64.
- Ito, M. (1982). Cerebellar control of the vestibulo-ocular reflex—around the flocculus hypothesis. *Annual Review of Neuroscience*, 5, 275–296.
- Ito, M. (1984). *The cerebellum and neural control*. New York: Raven.
- Kandel, E. R., Klein, M., Hochner, B., Shuster, M., Siegelbaum, S. A., Hawkins, R. D., Glanzman, D. L., & Castellucci, V. F. (1987). Synaptic modulation and learning: New insights into synaptic transmission from the study of behavior. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Synaptic function* (pp. 471–518). New York: Wiley.
- Kelso, S. R., & Brown, T. H. (1986). Differential conditioning of associative synaptic enhancement in hippocampal brain slices. *Science*, 232, 85–87.
- Kelso, S. R., Ganong, A. H., & Brown, T. H. (1986). Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences USA*, 83, 5326–5330.
- Kienker, P. K., Sejnowski, T. J., Hinton, G. E., & Schumacher, L. E. (1986). Separating figure from ground with a parallel network. *Perception*, 15, 197–216.
- Kleinschmidt, A., Bear, M. F., & Singer, W. (1986). *Neuroscience Letters*, 26 (Suppl.), S58.
- Klopf, A. H. (1987). *A neuronal model of classical conditioning*. (Technical Report AFWAL-TR-87-1139). Dayton, Ohio: Wright-Patterson Air Force Base Aeronautical Laboratories.
- Kohonen, T. (1970). Correlation matrix memories. *IEEE Transactions on Computers*, C-21, 353–359.
- Kohonen, T. (1984). *Self-organization and associative memory*. New York: Springer-Verlag.
- Komatsu, Y., Fujii, K., Maeda, J., Sakaguchi, H., & Toyama, K. (1988). Long-term potentiation of synaptic transmission in kitten visual cortex. *Journal of Neurophysiology*, 59, 124–141.
- Levy, W. B., Anderson, J. A., & Lehmkuhle, W. (1984). *Synaptic change in the nervous system*. Hillsdale, N.J.: Erlbaum.
- Levy, W. B., Brassel, S. E., & Moore, S. D. (1983). Partial quantification of the associative synaptic learning rule of the dentate gyrus. *Neuroscience*, 8, 799–808.
- Linsker, R. (1986). From basic network principles to neural architecture: Emergence of orientation columns. *Proceedings of the National Academy of Sciences USA*, 83, 8779–8783.
- Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature*, 217, 104–107.
- Lynch, G., & Baudry, M. (1988). Structure–function relationships in the organization of memory. In M. S. Gazzaniga (Ed.), *Perspectives in memory research and training* (pp. 23–92). Cambridge, Mass.: MIT Press.
- Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, 202, 437–470.
- Marr, D. (1982). *Vision*. San Francisco: Freeman.
- McNaughton, B. L., Douglas, R. M., & Goddard, G. V. (1978). Synaptic enhancement in fascia dentata: Cooperativity among coactive afferents. *Brain Research*, 157, 277.
- McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10, 408–415.
- Miles, F. A., & Lisberger, S. G. (1981). Plasticity in the vestibulo-ocular reflex: A new hypothesis. *Annual Review of Neuroscience*, 4, 273–299.
- Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, D. E. J., Sutton, R. S., & Barto, A. G.

- (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing, and interstimulus intervals. *Behavioral Brain Research*, 21, 143-154.
- Mpitsos, G. J., & Cohna, C. S. (1986). Discriminative behavior and Pavlovian conditioning in the mollusc. *Journal of Neurobiology*, 17, 469-486.
- Reiter, H. O., & Stryker, M. P. (1987). A novel expression of plasticity in kitten visual cortex in the absence of postsynaptic activity. *Society for Neuroscience Abstracts*, 13, 1241.
- Rolls, E. T. (1987). Information representation, processing and storage in the brain: Analysis at the single neuron level. In J.-P. Changeux & M. Konishi (Eds.), *Neural and molecular mechanisms of learning* (pp. 503-540). Berlin: Springer-Verlag.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. Cambridge, Mass.: MIT Press.
- Sahley, C., Rudy, J. W., & Gelperin, A. (1981). An analysis of associative learning in a terrestrial mollusc: Higher-order conditioning, blocking, and a transient US pre-exposure effect. *Journal of Comparative Physiology*, 144, 1-8.
- Sahley, C. L., Rudy, J. W., & Gelperin, A. (1984). Associative learning in a mollusc: A comparative analysis. In D. Alkon & J. Farley (Eds.), *Primary neural substrates of learning and behavioral change* (pp. 243-258). New York: Cambridge University Press.
- Sejnowski, T. J. (1977a). Statistical constraints on synaptic plasticity. *Journal of Mathematical Biology*, 69, 385-389.
- Sejnowski, T. J. (1977b). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, 4, 303-321.
- Sejnowski, T. J. (1981). Skeleton filters in the brain. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 189-212). Hillsdale, N.J.: Erlbaum.
- Sejnowski, T. J. (1987). Computational models and the development of topographic projections. *Trends in Neurosciences*, 10, 304-305.
- Sejnowski, T. J., & Hinton, G. E. (1987). Separating figure from ground with a Boltzmann machine. In M. A. Arbib & A. R. Hanson (Eds.), *Vision, brain and cooperative computation* (pp. 703-724). Cambridge, Mass.: MIT Press.
- Sejnowski, T. J., & Tesauro, G. (1989). The Hebb rule for synaptic plasticity: Algorithms and implementations. In J. N. Byrne & W. O. Berry (Eds.), *Neural model of plasticity* (pp. 94-103). Orlando, Fla.: Academic Press.
- Shepherd, G. M., Brayton, R. K., Miller, J. P., Segev, I., Rinzel, J., & Rall, W. (1985). Signal enhancement in distal cortical dendrites by means of interactions between active dendritic spines. *Proceedings of the National Academy of Sciences USA*, 82, 2192-2195.
- Sherman, S. M., & Spear, P. D. (1982). Organization of visual pathways in normal and deprived cats. *Physiological Reviews*, 62, 738.
- Singer, W. (1987). Activity-dependent self-organization of synaptic connections as a substrate of learning. In J. P. Changeux & M. Konishi (Eds.), *The neural and molecular bases of learning* (pp. 301-336). New York: Wiley.
- Stanton, P., Jester, J., Chattarji, S., & Sejnowski, T. J. (1988). Associative long-term depression (LTD) or potentiation (LTP) is produced in the hippocampus dependent upon the phase of rhythmically active inputs. *Society for Neuroscience Abstracts*, 14, 19.
- Steinbuch, K. (1961). Die Lernmatrix. *Kybernetik*, 1, 36-45.
- Stent, G. W. (1973). A physiological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences USA*, 70, 997-1001.
- Sutton, R. S. (1987). *A temporal-difference model of classical conditioning*. (Technical Report TR87-509.2). Waltham, Mass.: GTE Labs.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88, 135-170.
- Tesauro, G. (1986). Simple neural models of classical conditioning. *Biological Cybernetics*, 55, 187-200.

- Tesauro, G. (1988). A plausible neural circuit for classical conditioning without synaptic plasticity. *Proceedings of the National Academy of Sciences USA*, 85, 2830–2833.
- Thompson, R. F. (1986). The neurobiology of learning and memory. *Science*, 233, 941–947.
- Toulouse, G., Dehaene, S., & Changeux, J. P. (1986). Spin glass model of learning by selection. *Proceedings of the National Academy of Sciences USA*, 83, 1695–1698.
- von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in striate cortex. *Kybernetik*, 14, 85.
- Wigstrom, H., & Gustafsson, B. (1985). On long-lasting potentiation in the hippocampus: A proposed mechanism for its dependence on coincident pre- and postsynaptic activity. *Acta Physiologica Scandinavica*, 123, 519.
- Wilshaw, D. (1981). Holography, associative memory, and inductive generalization. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 83–104). Hillsdale, N.J.: Erlbaum.
- Woody, C. D. (1982). *Memory, learning and higher function*. Berlin: Springer-Verlag.