STATISTICAL MECHANICS OF NEURAL NETWORKS

Studies of disordered systems have generated new insights into the cooperative behavior and emergent computational properties of large, highly connected networks of simple, neuron-like processors.

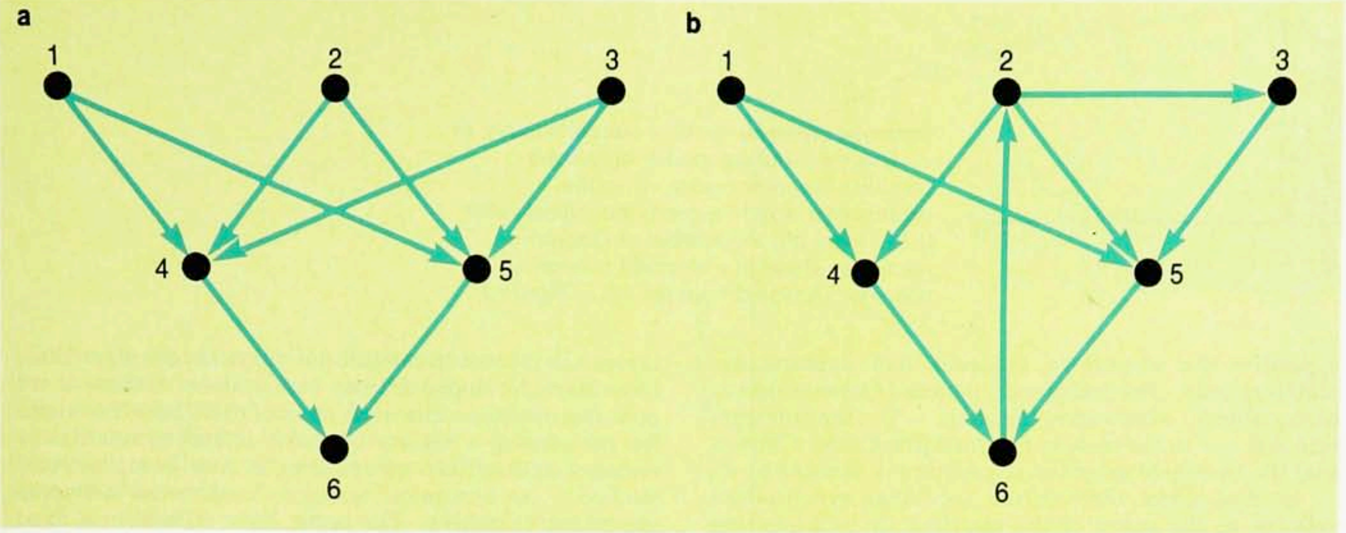
Haim Sompolinsky

A neural network is a large, highly interconnected assembly of simple elements. The elements, called neurons, are usually two-state devices that switch from one state to the other when their input exceeds a specific threshold value. In this respect the elements resemble biological neurons, which fire—that is, send a voltage pulse down their axons—when the sum of the inputs from their synapses exceeds a "firing" threshold. Neural networks therefore serve as models for studies of cooperative behavior and computational properties of the sort exhibited by the nervous system.

Neural network models are admittedly gross oversimplifications of biology. But these simple models are accessible to systematic investigations and may therefore shed light on the principles underlying "computation" in biological systems and on how those principles differ from the ones that we have so successfully mastered in building digital computers. In addition, psychologists use neural networks as conceptual models for understanding cognitive processes in the human mind. For theoretical physicists, understanding the dynamical properties of large, strongly coupled nonequilibrium systems such as neural networks is a challenging problem in its own right.

Attempts to model the working of the brain with networks of simple, formal neurons date back to 1943, when Warren McCulloch and Walter Pitts proposed networks of two-state threshold elements that are capable of performing arbitrary logic operations.[1] In 1949, Donald Hebb, the psychologist, proposed that neural systems can learn and form associations through selective modifications of the synaptic connections.[2] Several adaptive networks, so called because they could learn to perform simple recognition tasks, were studied in the 1960s. These included Frank Rosenblatt's feedforward network, called the perceptron,[3] and Bernard Widrow's adaptive linear machine, the Adaline.[4] A variety of network models for associative memory and pattern recognition have been investigated over the past two decades by several research groups, including those of Shun-ichi Amari,[5] Stephen Grossberg[6] and Teuvo Kohonen.[7] Physicists' interest in neural networks stems largely from the analogy between such networks and simple magnetic systems. The analogy was first pointed out in 1974 by William Little.[8] Recently activity in this direction was stimulated by the work of

Haim Sompolinsky is a professor of physics at the Racah Institute of Physics of the Hebrew University of Jerusalem.



Network architectures. **a:** A feedforward system with three layers. **b:** A neural circuit. A circuit contains feedback loops, such as the directed graph 2→3→5→6→2, that close on themselves. This closure gives rise to recurrent activity in the network. **Figure 1**

John Hopfield, who pointed out the equivalence between the long-time behavior of networks with symmetric connections and equilibrium properties of magnetic systems such as spin glasses.[9] In particular, Hopfield showed how one could exploit this equivalence to "design" neural circuits for associative memory and other computational tasks.

Spin glasses are magnetic systems with randomly distributed ferromagnetic and antiferromagnetic interactions. The low-temperature phase of these systems—the spin glass phase—is in many ways a prototype for condensation in disordered systems with conflicting constraints. Theoretical studies have revealed that in spin glasses with long-range interactions between the spins, the energy surface (the energy as a function of the system's state, or spin configuration) has a rich topology, with many local minima very close in energy to the actual ground state.[10] (See the article by Daniel S. Fisher, Geoffrey M. Grinstein and Anil Khurana on page 56.)

Neural systems share several features with long-range spin glasses. (I will use the term "neural systems" for assemblies of real neurons.) The spatial configuration of the two systems bears no resemblance to the crystalline order of pure magnets or solids. The couplings between spins in spin glasses can be both positive and negative, which is also true of the couplings between neurons. And just as each spin in a long-range spin glass is connected to many others, so is each neuron in most neural systems. For example, each neuron in the cortex is typically connected to about 10^4 neurons.[11]

Of course, the analogy between long-range spin glasses and neural systems is far from perfect. First, the connections in neural systems, unlike those in spin glasses, are not distributed at random, but possess correlations that are formed both genetically and in the course of learning and adaptation. These correlations alter the dynamical behavior of the system and endow it with useful computational properties. Another major difference is the asymmetry of the connections: The pairwise interactions between neurons are, in general, not reciprocally symmetric; hence their dynamic properties may be very different from those of equilibrium magnetic systems, in which the pairwise interactions are symmetric.

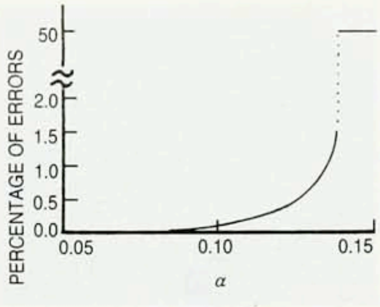
In this article I will describe how the concepts and tools of theoretical physics are being applied to the study of neural networks. As I have already indicated, the methods of equilibrium statistical mechanics have been particularly useful in the study of symmetric neural network models of associative memory. I will describe some of these models and discuss the interplay between randomness and correlations that determines a model's performance. The dynamics of asymmetric networks is much richer than that of symmetric ones and must be studied within the general framework of nonlinear dynamics. I will discuss some dynamical aspects of asymmetric networks and the computational potential of such networks. Learning—the process by which the network connections evolve under the influence of external inputs to meet new computational requirements—is a central problem of neural network theory. I will briefly discuss learning as a statistical mechanical problem. I will comment on the applications of neural networks to solving hard optimization problems. I will conclude with a few remarks about the relevance of neural network theory to the neurosciences.

Basic dynamics and architecture

I consider in this article neural network models in which the neurons are represented by simple, point-like elements that interact via pairwise couplings called synapses. The state of a neuron represents its level of activity. Neurons fire an "action potential" when their "membrane potential" exceeds a threshold, with a firing rate that depends on the magnitude of the membrane potential. If the membrane potential is below threshold, the neurons are in a quiescent state. The membrane potential of a neuron is assumed to be a linear sum of potentials induced by the activity of its neighbors. Thus the potential in excess of the threshold, which determines the activity, can be denoted by a local field

$$h_i(t) = \sum_{j=1}^{N} J_{ij} \frac{S_j(t)+1}{2} - \theta_i \qquad (1)$$

The synaptic efficacy J_{ij} measures the contribution of the activity of the jth, presynaptic neuron to the potential acting on the ith, postsynaptic neuron. The contribution



Errors per neuron increase discontinuously as $T \to 0$ in the Hopfield model, signaling a complete loss of memory, when the parameter $\alpha = p/N$ exceeds the critical value 0.14. Here p is the number of random memories stored in a Hopfield network of N neurons. (Adapted from ref. 16.) **Figure 2**

is positive for excitatory synapses and negative for inhibitory ones. The activity of a neuron is represented by the variable S_j, which takes the value -1 in the quiescent state and $+1$ in the state of maximal firing rate. The value of the threshold potential of neuron i is denoted by θ_i.

In this model there is no clock that synchronizes updating of the states of the neurons. In addition, the dynamics should not be deterministic, because neural networks are expected to function also in the presence of stochastic noise. These features are incorporated by defining the network dynamics in analogy with the single-spin-flip Monte Carlo dynamics of Ising systems at a finite temperature.[12] The probability that the updating neuron, which is chosen at random, is in the state, say, $+1$ at time $t + \mathrm{d}t$ is

$$P(h_i(t)) = \frac{1}{1 + \exp(-4\beta h_i)} \tag{2}$$

where h_i is the local field at time t, and $T \equiv 1/\beta$ is the temperature of the network. In the Monte Carlo process, the probability of a neuron's being in, say, state $+1$ at time $t + \mathrm{d}t$ is compared with a random number and the neuron's state is switched to $+1$ if the probability is greater than that number. The temperature is a measure of the level of stochastic noise in the dynamics. In the limit of zero temperature, the dynamics consists of single-spin flips that align the neurons with their local fields, that is, $S_i(t + \mathrm{d}t) = \mathrm{sign}(h_i(t))$. The details of the dynamics and, in particular, the specific form of $P(h)$ are largely a matter of convenience. Other models for the dynamics, including ones involving deterministic, continuous-time dynamics of analog networks, have also been studied.

The computational process in a neural network emerges from its dynamical evolution, that is, from flows in the system configuration space. The end products of this evolution, called attractors, are states or sets of states to which the system eventually converges. Attractors may consist of stable states, periodic orbits or the so-called strange attractors characteristic of chaotic behavior. Understanding the dynamics of the system involves knowing the nature of its attractors, as well as their basins of attraction and the time the system takes to converge to the attractors. In stochastic systems such as ours, one has to take into account the smearing of the trajectories by the stochastic noise.

The behavior of the network depends on the form of the connectivity matrix J_{ij}. Before specifying this matrix in detail, however, I will discuss the basic features of network architecture. Studies of neural networks have focused mainly on two architectures. One is the layered network (see figure 1a), in which the information flows forward, so that the computation is a mapping from the state of the input layer onto that of the output layer. The perceptron, consisting of only an input and an output layer, is a simple example of such a feedforward network. Although interest in the perceptron declined in the 1960s, interest in feedforward networks that contain hidden layers has revived in the last few years as new algorithms have been developed for the exploitation of these more powerful systems. The usefulness of multilayer networks for performing a variety of tasks, including associative memory and pattern recognition, is now being actively studied.[13] As dynamical systems, feedforward networks are rather primitive: The input layer is held in a fixed configuration and all the neurons in each subsequent layer compute their states in parallel according to the states of the preceding layer at the previous time step. I will focus on a different class of network models, namely, networks that contain feedback loops. These networks I term neural circuits (see figure 1b). Many structures in the cortex show extensive feedback pathways, suggesting that feedback plays an important role in the dynamics as well as in the computational performance. Feedback loops are essential also for the function of nervous systems that control stereotypical behavioral patterns in animals.[14] Besides their biological relevance, however, neural circuits are interesting because the long iterative dynamical processes generated via the feedback loops endow them with properties not obtained in layered networks of comparable sizes.

Symmetric circuits and Ising magnets

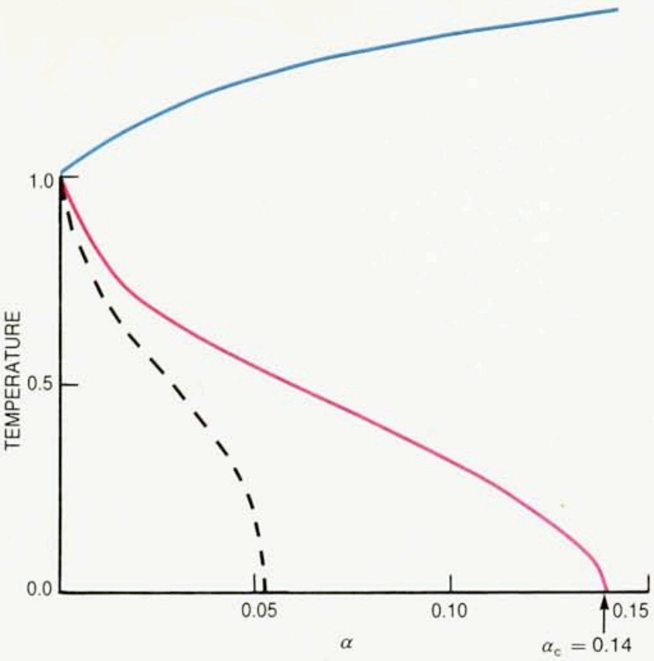
The dynamics may be rather complex for a general circuit of the type described above, but it is considerably simpler in symmetric circuits, in which the synaptic coefficients J_{ij} and J_{ji} are equal for each pair of neurons. In that case the dynamics is purely relaxational: There exists a function of the state of the system, the energy function, such that at $T = 0$ the value of this function always decreases as the system evolves in time. For a circuit of two-state neurons the energy function has the same form as the Hamiltonian of an Ising spin system:

$$E = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j - \sum_{i=1}^{N} h_i^0 S_i \tag{3}$$

The first term represents the exchange energy mediated by pairwise interactions, which are equal in strength to the respective synaptic coefficients. The last term is the energy due to the interaction with external magnetic fields, which, in our case, are given by

$$h_i^0 = \sum_j J_{ij} - 2\theta_i .$$

The existence of an energy function implies that at $T = 0$ the system flows always terminate at the local minima of E. These local minima are spin configurations in which every spin is aligned with its local field. At nonzero T, the notion of minima in configuration space is more subtle. Strictly speaking, thermal fluctuations will eventually carry any finite system out of the energy "valleys," leading to ergodic wandering of the trajectories. If the energy barriers surrounding a valley grow with the size of the system, however, the probability of escaping the



Phase diagram of the Hopfield model. The solid blue line marks the transition from the high-temperature ergodic phase to a spin glass phase, in which the states have negligible overlap with the memories. Memory phases, that is, valleys in configuration space that are close to the embedded memory states, appear below the solid red line. A first-order transition occurs along the dashed line; below this line the memory phases become the globally stable phases of the model. (Adapted from ref. 16.) **Figure 3**

valley may vanish in the thermodynamic limit, $N \to \infty$, at low temperatures. In that case energy valleys become disjoint, or disconnected, on finite time scales, and one says that ergodicity is broken. Each of these finite-temperature valleys represents a distinct thermodynamic state, or phase.

The analogy with magnetic systems provides a number of lessons that are useful in understanding the structure of the energy terrain and its implications for the dynamics of neural circuits. Ising ferromagnets with a constant positive value of J_{ij} are easily equilibrated even at relatively low temperatures. Their energy landscape has only two minima: one for each direction of the total magnetization. By contrast, a disordered magnet can rarely equilibrate to its low-temperature equilibrium state on a reasonable time scale. This is particularly true of spin glasses; their energy landscapes possess an enormous number of local minima, of higher energy than the ground state, that are surrounded by high-energy barriers. A spin glass usually gets stuck in one of these local minima and does not reach its equilibrium state when it is cooled to low temperatures.

I have already mentioned the presence of disorder and internal competition in many real neural assemblies. Indeed, most of the interesting neural circuits that have been studied are disordered and frustrated. (Frustration means that a system configuration cannot be found in which competing interactions are all satisfied.) Nevertheless there are applications of neural circuits in which the computational process is not as slow and painful as equilibrating a spin glass. Among important examples of such applications are models of associative memory.

Associative memory

Associative memory is the ability to retrieve stored information using as clues partial sets or corrupted sets of that information. In a neural circuit model for associative memory the information is stored in a special set of states of the network. Thus, in a network of N neurons a set of p memories are represented as p, N-component vectors S^μ, $\mu = 1, \ldots, p$. Each component S_i^μ takes the value $+1$ or -1 and represents a single bit of information. The models are based on two fundamental hypotheses. First, the information is stored in the values of the J_{ij}'s. Second, recalling a memory is represented by the settling of the neurons into a persistent state that corresponds to that memory, implying that the states S^μ must be attractors of the network.

Associative memory is implemented in the present models in two dynamic modes. Information is stored in the learning mode. In this mode the p memories are presented to the system and the synaptic coefficients evolve according to the learning rules. These rules ensure that at the completion of the learning mode, the memory states will be attractors of the network dynamics. In symmetric networks the J_{ij} are designed so that S^μ will be local minima of E. The stored memories are recalled associatively in the second, retrieval mode. In the language of magnetism, the J_{ij}'s are quenched and the dynamic variables are the neurons.

In the retrieval mode, the system is presented with partial information about the desired memory. This puts the system in an initial state that is close in configuration space to that memory. Having "enough" information about the desired memory means that the initial state is in the basin of attraction of the valley corresponding to the memory, a condition that guarantees the network will evolve to the stable state that corresponds to that memory. (An illustration of the recall process is given in the figure on page 23 of this issue.)

Some of the simplest and most important learning paradigms are based on the mechanism suggested by Hebb.[2] Hebb's hypothesis was that when neurons are active in a specific pattern, their activity induces changes in their synaptic coefficients in a manner that reinforces the stability of that pattern of activity. One variant of these Hebb rules is that the simultaneous firing of a pair of neurons i and j increases the value of J_{ij}, whereas if only one of them is active, the coupling between them weakens. Applying these rules to learning sessions in which the neural activity patterns are the states S^μ results in the following form for synaptic strengths:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} S_i^\mu S_j^\mu \qquad (4)$$

This simple quadratic dependence of J_{ij} on S^μ is only one of many versions of the Hebb rules. Other versions have also been studied. They all share several attractive features. First, the changes in the J_{ij}'s are local: They depend only on the activities in the memorized patterns of the presynaptic and postynaptic neurons i and j. Second, a new pattern is learned in a single session, without the need for refreshing the old memories. Third, learning is unsupervised: It is performed without invoking the existence of a hypothetical teacher. Finally, regarding the plausibility of these learning rules occurring in biological systems, it is encouraging to note that Hebb-like plastic changes in synaptic strengths have been observed in recent years in some cortical preparations.[15]

I turn now to the performance of a network for associative memory in the retrieval mode, assuming that the learning cycle has been completed and the J_{ij}'s have reached a given static limit. The performance is characterized by several parameters. One is the capacity, that is, the maximum number of memories that can be simultaneously stabilized in the network. The associative nature of the recall is characterized by the sizes of the basins of attraction of the valleys corresponding to the memory states. These basins are limited in size by the presence of other, spurious attractors, which in the case of symmetric networks are local minima of E other than the memory states. The convergence times within the basins determine the speed of recall. Another important issue is the robustness of the network to the presence of noise or to failures of neurons and synapses. A theoretical analysis of most of these issues is virtually impossible unless one considers the thermodynamic limit, $N \to \infty$, where answers can be derived from a statistical mechanical investigation. I discuss below two simple associative memory networks that are based on the Hebb rules.

The Hopfield model

The Hopfield model consists of a network of two-state neurons evolving according to the asynchronous dynamics described above. It has an energy function of the form given in equation 3, with symmetric connections given by the Hebb rule (equation 4) and $h_i^0 = 0$. The memories are assumed to be completely uncorrelated. They are therefore represented by quenched random vectors S^μ, each of whose components can take the values ± 1 with equal probability.[9]

The statistical mechanical theory of the Hopfield model, derived by Daniel Amit, Hanoch Gutfreund and myself at the Hebrew University of Jerusalem, has provided revealing insight into the workings of that system and set the stage for quantitative studies of other models of associative memory.[16] The theory characterizes the different phases of the system by the overlaps of the states within each phase with the memories, given by

$$M^\mu = \frac{1}{N} \sum_{i=1}^{N} S_i^\mu S_i \qquad (5)$$

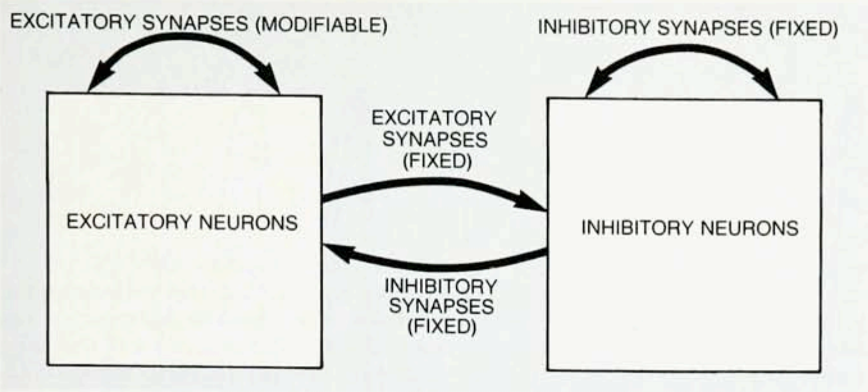
where S_i is the state of the neuron i in that phase. All the M^μ's are of order $1/N^{1/2}$ for a state that is uncorrelated with the memories, whereas a M^μ for, say, $\mu = 2$ is of order unity if a state is strongly correlated with memory 2.

To understand why this model functions as associative memory, let us consider for a moment the case $p = 1$, when there is only a single memory. Obviously, the states $S_i = S_i^1$ and $S_i = -S_i^1$ are the ground states of E, since in these states every bond energy $-J_{ij} S_i S_j$ has its minimum possible value, $-1/N$. Thus, even though the J_{ij}'s are evenly distributed around zero, they are also spatially correlated so that all the bonds can be satisfied, exactly as happens in a pure ferromagnet. In a large system, adding a few more uncorrelated patterns will not change the global stability of the memories, since the energy contribution of the random interference between the patterns is small. This expectation is corroborated by our theory. As long as p remains finite as $N \to \infty$, the network is unsaturated. There are $2p$ degenerate ground states, corresponding to the p memories and their spin-reversed configurations. Even for small values of p, however, the memories are not the only local minima. Other minima exist, associated with states that strongly "mix" several memories. These mixture states have a macroscopic value (that is, of order unity) for more than one component M^μ. As p increases, the number of the mixture states increases very rapidly with N. This decreases the memories' basins of attraction and eventually leads to their destabilization.

A surprising outcome of the theory is that despite the fact that the memory states become unstable if $p > N/(2 \ln N)$, the system provides useful associative memory even when p increases in proportion to N. The random noise generated by the overlaps among the patterns destabilizes them, but new stable states appear, close to the memories in the configuration space, as long as the ratio $\alpha = p/N$ is below a critical value $\alpha_c = 0.14$. Memories can still be recalled for $\alpha < \alpha_c$, but a small fraction of the bits in the recall will be incorrect. The average fraction of incorrect bits ε, which is related to the overlap with the nearest memory by $\varepsilon = (1 - M)/2$, is plotted in figure 2. Note that ϵ rises discontinuously to 0.5 at α_c, which signals the "catastrophic" loss of memory that occurs when the number of stored memories exceeds the critical capacity. The strong nonlinearity and the abundance of feedback in the Hopfield model are the reasons for this behavior.

Near saturation, when the ratio α is finite, the nature of the spurious states is different from that in the unsaturated case. Most of the states now are seemingly random configurations that are only slightly biased in the direction of the memories. The overlaps of these spurious states with the memories are all of order $1/N^{1/2}$. Their statistical properties are similar to those of states in infinite-range spin glasses.

A very important feature of the Hopfield model is the appearance of distinct valleys near each of the memories, even at nonzero temperatures. This implies that the energy barriers surrounding these minima diverge with N, and it also indicates that the memories have substantial basins of attraction. The existence of memory phases characterized by large overlaps with the memory states


EXCITATORY SYNAPSES (MODIFIABLE)
INHIBITORY SYNAPSES (FIXED)
EXCITATORY SYNAPSES (FIXED)
EXCITATORY NEURONS
INHIBITORY NEURONS
INHIBITORY SYNAPSES (FIXED)

Associative memory circuit with a biologically plausible architecture. The circuit consists of two neural populations: excitatory neurons, which excite other neurons into the active state, and inhibitory ones, which inhibit the activity of other neurons. Information is encoded only in the connections between the excitatory neurons. The synaptic matrix for the circuit is asymmetric, and for appropriate values of the synaptic strengths and thresholds, the circuit's dynamics might converge to oscillatory orbits rather than to stable states. **Figure 4**

implies that small amounts of noise do not disrupt the performance of the system entirely but do increase the inaccuracy in the retrieval. The full phase diagram of the model at finite α and T is shown in figure 3. The diagram shows that accurate memory phases exist even when they are not the global minima of E. This feature distinguishes the model from those encountered in equilibrium statistical mechanics: For a system to be able to recall associatively, its memory states must be robust local minima having substantial basins of attraction, but they do not necessarily have to be the true equilibrium states of the system.

The Willshaw model

From a biological point of view, the Hopfield model has several important drawbacks. A basic feature of the model is the symmetry of the connections, whereas the synaptic connections in biological systems are usually asymmetric. (I will elaborate on this issue later.) Another characteristic built into the Hopfield network is the up–down, or $S \to -S$, symmetry, which occurs naturally in magnetic systems but not in biology. This symmetry appears in the model in several aspects. For one, the external magnetic fields, h_i^0 of equation 3, are set to zero, and this may require fine tuning of the values of the neuronal thresholds. More important, the memories have to be completely random for the model to work, implying that about half of the neurons are active in each of the memory states. By contrast, the observed levels of activity in the cortex[17] are usually far below 50%. From the point of view of memory storage as well, there are advantages to sparse coding, in which only a small fraction of the bits are $+1$.

Another feature of the Hopfield model is that each neuron sends out about an equal number of inhibitory and excitatory connections, both having the same role in the storage of information. This should be contrasted with the fact that cortical neurons are in general either excitatory or inhibitory. (An excitatory neuron when active excites other neurons that receive synaptic input from it; an inhibitory neuron inhibits the activity of its neighbors.) Furthermore, the available experimental evidence for Hebb-type synaptic modifications in biological systems so far has demonstrated Hebb-type activity-dependent changes only of excitatory synapses.[15]

An example of a model that has interesting biological features is based on a proposal made by David Willshaw some 20 years ago.[18] Willshaw's proposal can be implemented in a symmetric circuit of two-state neurons whose dynamics are governed by the energy

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij}(S_i+1)(S_j+1) \qquad (6)$$

where

$$J_{ij} = \frac{1}{N}\Theta\Big(\sum_{\mu=1}^{p}(S_i^\mu+1)(S_j^\mu+1)\Big) - \frac{1}{N} \qquad (7)$$

where $\Theta(x)$ is 0 for $x=0$ and 1 for $x>0$. Thus the synapses in this model have two possible values. A J_{ij} is 0 if neurons i and j are simultaneously active in at least one of the memory states, and it is $-1/N$ otherwise. The memories are random except that the average fraction of active neurons in each of the states S^μ is given by a parameter f that is assumed to be small. This model is suitable for storing patterns in which the fraction of active neurons is small, particularly when $f \to 0$ in the thermodynamic limit. These sparsely coded memories are perfectly recalled as long as $p < \ln(Nf/\ln N)/f^2$. The capacity of the Willshaw model with sparse coding is far better than that of the Hopfield model. There are circuits for sparse coding that have a much greater capacity. For example, the capacity in some is given by $p < N/[f(-\ln f)]$.[19]

The learning algorithm implicit in equation 7 is interesting in that it involves only enhancements of the excitatory components of the synaptic interactions. The inhibitory component is uniform, $-1/N$ in equation 7, and its role is to suppress the activity of all neurons except those with the largest excitatory fields. This ensures that only the neutrons that are "on" in the memory state are activated. Furthermore, the same effect can be achieved by a model in which the inhibitory synaptic components represent not direct synaptic interactions between the N neurons of the circuit but indirect interactions mediated by other inhibitory neurons. This architecture, which is illustrated in figure 4, is compatible with the known architecture of cortical structures. Finally, information is stored in this model in synapses that assume only two values. This is desired because the analog depth of synaptic strengths in the cortex is believed to be rather small.

The two models discussed above serve as prototypes for a variety of neural networks that use some form of the Hebb rules in the design of the synaptic matrix. Most of these models share the thermodynamic features described above. In particular, at finite temperatures they already have distinct memory phases; near saturation, the memory states are corrupted by small fractions of erroneous bits, and spin glass states coexist with the memory states. Toward the end of this article I will discuss other learning strategies for associative memory.

Asymmetric synapses

The applicability of equilibrium statistical mechanics to the dynamics of neural networks depends on the condition that the synaptic connections are symmetric, that is, $J_{ij} = J_{ji}$. As I have already mentioned, real synaptic connections are seldom symmetric. Actually, quite often only one of the two bonds J_{ij} and J_{ji} is nonzero. I should also mention that the fully connected circuits, which have abundant feedback, and the purely feedforward, layered networks are two extreme idealizations of biological systems, most of which have both a definite direction of information flow and substantial feedback. Models of asymmetric circuits offer the possibility of studying such mixed architectures.

Asymmetric circuits have a rich repertoire of possible dynamic behaviors in addition to convergence onto a stable state. The "mismatch" in the response of a sequence of bonds when they are traversed in opposite directions gives rise, in general, to time-dependent attractors. This time dependence might propagate coherently along feedback loops, creating periodic or quasiperiodic orbits, or it might lead to chaotic trajectories characterized by continuous bands of spectral frequencies.

In asynchronous circuits, asymmetry plays a role in the dynamics in several respects. At $T=0$, either the trajectories converge to stable fixed points, as they do in the symmetric case, or they wander chaotically in configuration space. Whether the trajectories end in stable fixed points or are chaotic is particularly important in models of associative memory, where stable states represent the results of the computational process. Suppose one dilutes a symmetric network of Hebbian synapses, such as that in equation 4, by cutting the directed bonds at random, leaving only a fraction c of the bonds. Asymmetry is generated at random in the cutting process because a bond J_{ij} may be set to 0 while the reverse bond J_{ji} is not. The result is an example of a network with spatially unstructured asymmetry. Often one can model unstructured asymmetry by adding spatially random asymmetric synaptic components to an otherwise symmetric circuit. In the above example, the resulting synaptic matrix may be regarded as the sum of a symmetric part, $J_{ij}c$, which is the same as that in equation 4, and a random asymmetric part with a variance $(pc(1-c))^{1/2}/N$.

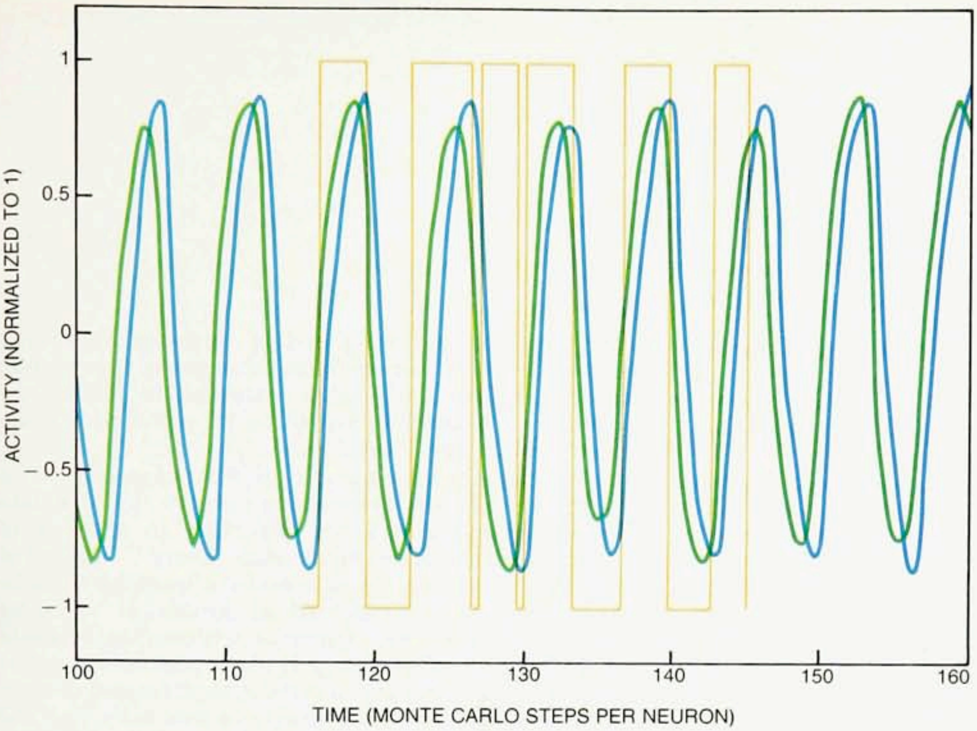
Recent studies have shown that the trajectories of large two-state networks with random asymmetry are always chaotic, even at $T=0$. This is because of the noise generated dynamically by the asymmetric synaptic inputs. Although finite randomly asymmetric systems may have some stable states, the time it takes to converge to those states grows exponentially with the size of the system, so that the states are[20] inaccessible in finite time as $N \to \infty$. This nonconvergence of the flows in large systems occurs as soon as the asymmetric contribution to the local fields, even though small in magnitude, is a finite fraction of the symmetric contribution as $N \to \infty$. In the above example of cutting the directed bonds at random, the dilution affects the dynamics in the thermodynamic limit only if c is not greater, in order of magnitude, than p/N.

The above discussion implies that the notion of encoding information in stable states to obtain associative memory is not valid in the presence of random asymmetry. When the strength of the random asymmetric component is reduced below a critical value, however, the dynamic flows break into separate chaotic trajectories that are confined to small regions around each of the memory states. The amount of information that can be retrieved from such a system depends on the amplitude of the chaotic fluctuations, as well as on the amount of time averaging that the external, "recalling," agent performs.

In contrast to the strictly random asymmetric synapses I discussed above, asymmetric circuits with appropriate correlations can have robust stable states at $T=0$. For instance, the learning algorithms of equation 10 (see below), in general, produce asymmetric synaptic matrices. Although the memory states are stable states of the dynamics of these circuits, the asymmetry does affect the circuits' performance in an interesting way. In regions of configuration space far from the memories the asymmetry generates an "effective" temperature that leads to nonstationary flows. When the system is in a state close to one of the memories, by contrast, the correlations induced by the learning procedure ensure that no such noise will appear. That the behavior near the attractors representing the memories is dynamically distinct from the behavior when no memories are being recalled yields several computational advantages. For instance, a failure to recall a memory is readily distinguished from a successful attempt by the persistence of fluctuations in the network activity.

Coherent temporal patterns

When studying asymmetric circuits at $T>0$ it is useful to consider the phases of the system instead of individual trajectories. Phases of asymmetric circuits at finite temperatures are defined by the averages of dynamic quantities over the stochastic noise as $t \to \infty$, where t is the time elapsed since the initial condition. This definition extends the notion of thermodynamic phases of symmetric circuits. The phases of symmetric circuits are always stationary, and the averages of dynamic quantities have a well-defined static limit. By contrast, asymmetric systems may exhibit phases that are time dependent even at nonzero temperatures. The persistence of time depend-



Periodic behavior of an asymmetric stochastic neural circuit of N inhibitory and N excitatory neurons having the architecture shown in figure 4. In the simulation, all connections between excitatory neurons had equal magnitude, $1/N$. The (excitatory) connections the excitatory neurons make with the inhibitory neurons had strengths $0.75/N$. The inhibitory neurons had synapses only with the excitatory neurons, of strength $0.75/N$. The "external fields" were set to 0 and the dynamics was stochastic, with the probability law given in equation 2. The neural circuit has stationary phases when $T > 0.5$, and nonstationary, periodic phases at lower temperatures. Results are shown for $N = 200$ at $T = 0.3$. The green curve shows the average activity of the excitatory population (that is, the activity summed over all excitatory neurons), the blue curve shows the corresponding result for the inhibitory population. The slight departure from perfect oscillations is a consequence of the finite size of the system. The instantaneous activity of individual neurons is not periodic but fluctuates with time, as shown here (orange) for one of the excitatory neurons in the circuit. **Figure 5**

ence even after averaging over the stochastic noise is a cooperative effect. Often it can be described in terms of a few nonlinearly coupled collective modes, such as the overlaps of equation 5. Such time-dependent phases are either periodic or chaotic. The attractor in the chaotic case has a low dimensionality, like the attractors of dynamical systems with a few degrees of freedom. The time-dependent phases represent an organized, coherent temporal behavior of the system that can be harnessed to process temporal information. A phase characterized by periodic motion is an example of the coherent temporal behavior that asymmetric circuits may exhibit.

Figure 4 shows an example of an asymmetric circuit that exhibits coherent temporal behavior. For appropriate sets of parameters, such a circuit exhibits a bifurcation at a critical value of T, such that its behavior is stationary above T_c and periodic below it, as shown in figure 5. Although the activities of single cells are fairly random in such a circuit, the global activity—that is, the activity of a macroscopic part of the circuit—consists of coherent oscillations (see figure 5). The mechanism of oscillations in the system is quite simple: The activity of the excitatory neurons excites the inhibitory population, which then triggers a negative feedback that turns off the excitatory neurons causing their activity. Such a mechanism for generation of periodic activity has been invoked to account for the existence of oscillations in many real nervous systems, including the cortex.

In general, the dynamical behavior should be more complex than the simple oscillations described above for it to be useful for interesting "computations." As in the case of static patterns, appropriate learning rules must be used to make sure that the complex dynamical patterns represent the desired computational properties. An interesting and important example of such learning rules are those used for temporal association, in which the system has to reconstruct associatively a temporally ordered sequence of memories. Asymmetric circuits can represent such a computation if their flows can be organized as a temporally ordered sequence of rapid transitions between quasistable states that represent the individual memories. One can generate and organize such dynamical patterns by introducing time delays into the synaptic responses.

In a simple model of temporal association the synaptic matrix is assumed to consist of two parts. One is the symmetric Hebb matrix of equation 4, with synapses with a short response time. The quick response ensures that the patterns S^μ are stable for short periods of time. The other component encodes the temporal order of the memories according to the equation

$$W_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} S_i^{\mu+1} S_j^{\mu} \tag{8}$$

where the index μ denotes the temporal order. The synaptic elements of this second component have a delayed response so that they do not disrupt the recall of the memories completely but induce transitions from the quasistable state S^{μ} to $S^{\mu+1}$. The composite local fields at time t are

$$h_i(t) = \frac{1}{2}\sum_{j=1}^{N} J_{ij}\,(S_j(t)+1) + \frac{\lambda}{2}\sum_{j=1}^{N} W_{ij}\,(S_j(t-\tau)+1) \tag{9}$$

where τ is the delay time and λ denotes the relative strength of the delayed synaptic input. If λ is smaller than a critical value λ_c, of order 1, then all the memories remain stable. However, when $\lambda > \lambda_c$, the system will stay in each memory only for a time of order τ and then be driven by the delayed inputs to the next memory. The flow will terminate at the last memory in the sequence. If the memories are organized cyclically, that is, if $S^p = S^1$, then, starting from a state close to one of the memories, the system will exhibit a periodic motion, passing through all the memories in each period. The same principle can be used to embed several sequential or periodic flows in a single network. It should be noted that the sharp delay used in equation 9 is not unique. A similar effect can be achieved by integrating the presynaptic activity over a finite integration time τ.

Circuits similar to those described above have been proposed as models for neural control of rhythmic motor outputs.[14] Synapses with different response times can also be used to form networks that recognize and classify temporally ordered inputs, such as speech signals. Whether biological systems use synapses with different time courses to process temporal information is questionable. Perhaps a more realistic possibility is that effective delays in the propagation of neural activity are achieved not by direct synaptic delays but by the interposition of additional neurons in the circuit.

Learning, or exploring the space of synapses

So far I have focused on the dynamics of the neurons and assumed that the synaptic connections and their strengths are fixed in time. I now discuss some aspects of the learning process, which determines the synaptic matrix. Learning is relatively simple in associative memory: The task is to organize the space of the states of the circuit in compact basins around the "exemplar" states that are known *a priori*. But in most perception and recognition tasks the relationship between the input (or initial) and the output (or final) states is more complex. Simple learning rules, such as the Hebb rules, are not known for these tasks. In some cases iterative error-correcting learning algorithms have been devised. Many of these algorithms can be formulated in terms of an energy function defined on the configuration space of the synaptic matrices. Synaptic strengths converge to the values needed for the desired computational capabilities when the energy function is minimized.

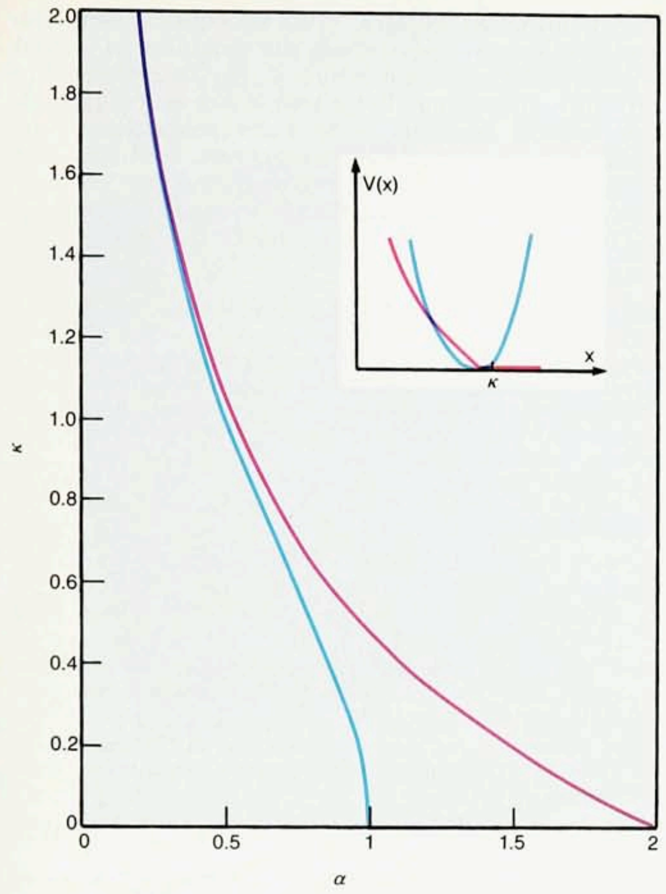
An illuminating example of this approach is its implementation for associative memory by the late Elizabeth Gardner and her coworkers in a series of important studies of neural network theory.[19] Instead of using the Hebb rules, let us consider a learning mode in which the J_{ij}'s are regarded as dynamical variables obeying a relaxational dynamics with an appropriate energy function. This energy is like a cost function that embodies the set of constraints the synaptic matrices must satisfy. Configurations of connections that satisfy all the constraints have zero energy. Otherwise, the energy is positive and its value is a measure of the violation of the constraints. An example of such a function is

$$E\{J_{ij}\} = \sum_{i=1}^{N}\sum_{\mu=1}^{p} V(h_i^{\mu} S_i^{\mu}) \tag{10}$$

where the h_i^{μ}'s, defined as in equation 1, are the local fields of the memory state S^{μ} and are thus linear functions of the synaptic strengths.

Two interesting forms of $V(x)$ are shown in figure 6. In one case the synaptic matrix has zero energy if the generated local fields obey the constraint $h_i^{\mu} S_i^{\mu} > \kappa$ for all i and μ, where κ is a positive constant. For the particular case of $\kappa \to 0$ the condition reduces to the requirement that all the memories be stable states of the neural dynamics. The other case represents the more stringent rquirement $h_i^{\mu} S_i^{\mu} = \kappa$, which means that not only are the memories stable but they also generate local fields with a given magnitude κ. In both cases κ is defined using the normalization that the diagonal elements of the square of the synaptic matrix be unity.

One can use energy functions such as equation 10 in conjunction with an appropriate relaxational dynamics to construct interesting learning algorithms provided that the energy surface in the space of connections is not too rough. In the case of equation 10 there are no local minima of E besides the ground states for reasonable choices of V, such as the ones described above. Hence simple gradient-descent dynamics, in which each step decreases the energy function, is sufficient to guarantee convergence to the desired synaptic matrix, that is, one with zero E, if such a matrix exists. Indeed, such a gradient-descent dynamics with the $V(x)$ as in the first of the two cases discussed above is similar to the perceptron learning algorithm;[3] the dynamics with the second choice of $V(x)$ is related to the Adaline learning algorithm.[4] However, energy functions that are currently used for learning in more complex computations are expected to have complicated surfaces with many local minima, at least in large networks. Hence the usefulness of applying them together with gradient-descent dynamics in large-scale problems is an important open problem.[4,13]



Capacity of a network of N neurons when random memories are stored by minimizing the energy function of equation 10. The lines mark the boundaries of the regions in the (κ,α) plane where synaptic matrices having zero energy exist for the two choices (shown in the inset) of the energy function in equation 10. In one case (red) the energy function constrains the local fields of the memory states to be bigger than some constant κ; in the other case (blue), the local fields are constrained to be equal to κ. (α is the ratio between the number p of stored memories and N. The capacity in the limit $N\rightarrow\infty$ is shown.) The red line terminates at $\alpha = 2$, implying that the maximum number of randomly chosen memories that can be embedded as stable states in a large network is $p = 2N$. **Figure 6**

Formulating the problem of learning in terms of energy functions provides a useful framework for its theoretical investigation. One can then use the powerful methods of equilibrium statistical mechanics to determine the number and the statistical properties of connection matrices satisfying the set of imposed constraints. For instance, one can calculate the maximum number of memories that can be embedded using function 10 for different values of κ. The results for random memories are shown in figure 6. Valuable information concerning the entropy and other properties of the solutions has also been derived.[19]

In spite of the great interest in learning strategies of the type described above, their usefulness as models for learning in biology is questionable. To implement function 10 using relaxational dynamics, for example, either all the patterns to be memorized must be presented simultaneously or, if they are presented successively, several, and often many, sessions of recycling through all of them are needed before they are learned. Furthermore, the separation in time between the learning phase and the computational, or recall, phase, is artificial from a biological perspective. Obviously, understanding the principles of learning in biological systems remains one of the major challenges of the field.

Optimization using neural circuits

Several tasks in pattern recognition can be formulated as optimization problems, in which one searches for a state that is the global minimum of a cost function. In some interesting cases, the cost functions can be expressed as energy functions of the form of equation 3, with appropriate choices of the couplings J_{ij} and the fields h_i^0. In this formulation, the optimization task is equivalent to the problem of finding the ground state of a highly frustrated Ising system. The mapping of optimization problems onto statistical mechanical problems has stirred up considerable research activity in both computer science and statistical mechanics. Stochastic algorithms, known as simulated annealing, have been devised that mimic the annealing of physical systems by slow cooling.[21] In addition, analytical methods from spin-glass theory have generated new results concerning the optimal values of cost functions and how these values depend on the size of the problem.[10]

Hopfield and David Tank have proposed the use of deterministic analog neural circuits for solving optimization problems.[9] In analog circuits the state of a neuron is characterized by a continuous variable S_i, which can be thought of as analgous to the instantaneous firing rate of a real neuron. The dynamics of the circuits is given by $\mathrm{d}h_i/\mathrm{d}t = -\partial E/\partial S_i$, where h_i is the local input to the ith neuron. The energy E contains, in addition to the terms in equation 3, local terms that ensure that the outputs S_i are appropriate sigmoid functions of their inputs h_i. As in models of associative memory, computation is achieved—that is, the optimal solution is obtained—by a convergence of the dynamics to an energy minimum. However, in retrieving a memory one has partial information about the desired state, and this implies that the initial state is in the proximity of that state. In optimization problems one does not have a clue about the optimum configuration; one has to find the deepest valley starting from unbiased configurations. It is thus not surprising that using two-state circuits and the conventional zero-temperature single-spin-flip dynamics to solve these problems is as futile as attempting to equilibrate a spin glass after quenching it rapidly to a low temperature. On the other hand, simulations of the analog circuit equations on several optimization problems, including small sizes of the famous "traveling salesman" problem, yielded "good" solutions, typically in timescales on the order of a few time constants of the circuit. These solutions usually are not the optimal solutions but are much better than those obtained by simple discrete algorithms.

What is the reason for the improved performance of the analog circuits? Obviously, there is nothing in the circuit's dynamics, which is the same as gradient descent, that prevents convergence to a local minimum. Apparently, the introduction of continuous degrees of freedom smooths the energy surface, thereby eliminating many of the shallow local minima. However, the use of "continuous" neurons is by itself unlikely to modify significantly the high energy barriers, which it takes changes in the states of many neurons to overcome. In light of this, one may question the advantage of using analog circuits to solve large-scale, hard optimization problems. From the point of view of biological computation, however, a relevant question is whether the less-than-optimal solutions that these networks find are nonetheless acceptable. Other crucial unresolved questions are how the performance of these networks scales with the size of the problem, and to what extent the performance depends on fine-tuning the circuit's parameters.

Neural network theory and biology

Interest in neural networks stems from practical as well as theoretical sources. Neural networks suggest novel architectures for computing devices and new methods for learning. However, the most important goal of neural network research is the advancement of the understanding of the nervous system. Whether neural networks of the types that are studied at present can compute anything better than conventional digital computers has yet to be shown. But they are certainly indispensable as theoretical frameworks for understanding the operation of real, large neural systems. The impact of neural network research on neuroscience has been marginal so far. This reflects, in part, the enormous gap between the present-day idealized models and the biological reality. It is also not clear to which level of organization in the nervous system these models apply. Should one consider the whole cortex, with its 10^{11} or so neurons, as a giant neural network? Or is a single neuron perhaps a large network of many processing subunits?

Some physiological and anatomical considerations suggest that cortical subunits of sizes on the order of 1 mm^3 and containing about 10^5 neurons might be considered as relatively homogeneous, highly interconnected functional networks. Such a subunit, however, cannot be regarded as an isolated dynamical system. It functions as part of a larger system and is strongly influenced by inputs both from sensory stimuli and from other parts of the cortex. Dynamical aspects pose additional problems. For instance, persistent changes in firing activities during performance of short-term-memory tasks have been measured. This is consistent with the idea of computation by convergence to an attractor. However, the large fluctuations in the observed activities and their relatively low level are difficult to reconcile with simple-minded "convergence to a stable state." More generally, we lack criteria for distinguishing between functionally important biological constraints and those that can be neglected. This is particularly true for the dynamics. After all, the characteristic time of perception is in some cases about one-tenth of a second. This is only one hundred times the "microscopic" neural time constant, which is about 1 or 2 msec.

To make constructive bridges with experimental neurobiology, neural network theorists will have to focus more attention on architectural and dynamical features of specific biological systems. This undoubtedly will also give rise to new ideas about the dynamics of neural systems and the ways in which it may be cultivated to perform computations. In the near future neural network theories will hopefully make more predictions about biological systems that will be concrete, nontrivial and susceptible to experimental verification. Then the theorists will indeed be making a contribution to the unraveling of one of nature's biggest mysteries: the brain.

* * *

While preparing the article I enjoyed the kind hospitality of AT&T Bell Labs. I am indebted to M. Abeles, P. Hohenberg, D. Kleinfeld and N. Rubin for their valuable comments on the manuscript. My research on neural networks has been partially supported by the Fund for Basic Research, administered by the Israeli Academy of Science and Humanities, and by the USA–Israel Binational Science Foundation.

References

1. W. S. McCulloch, W. A. Pitts, Bull. Math. Biophys. **5**, 115 (1943).
2. D. O. Hebb, *The Organization of Behavior*, Wiley, New York (1949).
3. F. Rosenblatt, *Principles of Neurodynamics*, Spartan, Washington, D. C. (1961). M. Minsky, S. Papert, *Perceptrons*, MIT P., Cambridge, Mass. (1988).
4. B. Widrow, in *Self-Organizing Systems*, M. C. Yovits, G. T. Jacobi, G. D. Goldstein, eds., Spartan, Washington, D. C. (1962).
5. S. Amari, K. Maginu, Neural Networks **1**, 63 (1988), and references therein.
6. S. Grossberg, Neural Networks **1**, 17 (1988), and references therein.
7. T. Kohonen, *Self Organization and Associative Memory*, Springer-Verlag, Berlin (1984). T. Kohonen, Neural Networks **1**, 3 (1988).
8. W. A. Little, Math. Biosci. **19**, 101 (1974). W. A. Little, G. L. Shaw, Math. Biosci. **39**, 281 (1978).
9. J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982). J. J. Hopfield, D. W. Tank, Science **233**, 625 (1986), and references therein.
10. M. Mezard, G. Parisi, M. A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, Singapore (1987). K. Binder, A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
11. V. Braitenberg, in *Brain Theory*, G. Palm, A. Aertsen, eds., Springer-Verlag, Berlin (1986), p. 81.
12. K. Binder, ed., *Applications of the Monte Carlo Methods in Statistical Physics*, Springer-Verlag, Berlin (1984).
13. D. E. Rumelhart, J. L. McClell and the PDP Group, *Parallel Distributed Processing*, MIT P., Cambridge, Mass. (1986).
14. D. Kleinfeld, H. Sompolinsky, Biophys. Jour., in press, and references therein.
15. S. R. Kelso, A. H. Ganong, T. H. Brown, Proc. Natl. Acad. Sci. USA **83**, 5326 (1986). G. V. diPrisco, Prog. Neurobiol. **22**, 89 (1984).
16. D. J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985). D. J. Amit, H. Gutfreund, H. Sompolinsky, Ann. Phys. N. Y. **173**, 30 (1987), and references therein.
17. M. Abeles, *Local Cortical Circuits*, Springer-Verlag, Berlin (1982).
18. D. J. Willshaw, O. P. Buneman, H. C. Longuet-Higgins, Nature **222**, 960 (1969). A. Moopen, J. Lambe, P. Thakoor, IEEE SMC **17**, 325 (1987).
19. E. Gardner, J. Phys. A **21**, 257 (1988). A. D. Bruce, A. Canning, B. Forrest, E. Gardner, D. J. Wallace, in *Neural Networks for Computing*, J. S. Denker, ed., AIP, New York (1986), p. 65.
20. A. Crisanti, H. Sompolinsky, Phys. Rev. A **37**, 4865 (1988).
21. S. Kirkpatrick, C. D. Gellat Jr, M. P. Vecchi, Science **220**, 671 (1983). ■