Relative entropy and learning rules

Minping Qian
Department of Probability and Statistics, Peking University, Beijing, People's Republic of China

Guanglu Gong
Department of Applied Mathematics, Tsinghua University, Beijing, People's Republic of China

John W. Clark
McDonnell Center for the Space Sciences and Department of Physics, Washington University, St. Louis, Missouri 63130
(Received 4 June 1990; revised manuscript received 28 September 1990)

The dynamics of a probabilistic neural network is characterized by the distribution $\nu(x'|x)$ of successor states x' of an arbitrary state x of the network. A prescribed memory or behavior pattern is represented in terms of an ordered sequence of network states $x^{(1)},x^{(2)},\ldots,x^{(l)}$. A successful procedure for learning this pattern must modify the neuronal interactions in such a way that the dynamical successor of $x^{(s)}$ is likely to be $x^{(s+1)}$, with $x^{(l+1)}=x^{(1)}$. The relative entropy G of the probability distribution $\delta_{x^{(s+1)},x'}$ concentrated at the desired successor state, evaluated with respect to the dynamical distribution $\nu(x'|x^{(s)})$, is used to quantify this criterion, by providing a measure of the distance between actual and ideal probability distributions. Minimization of G subject to appropriate resource constraints leads to "optimal" learning rules for pairwise and higher-order neuronal interactions. The degree to which optimality is approached by simple learning rules in current use is considered, and it is found, in particular, that the algorithm adopted in the Hopfield model is more effective in minimizing G than the original Hebb law.

I. INTRODUCTION

For the physicist, the novelty of neural networks as dynamical systems lies in the adaptive modification of the basic interactions between the neuronal units in response to the activity of these units. Thus the temporal development of the interactions depends on the dynamical states recently visited by the system. In more anthropomorphic terms, the neural network learns by example, or gains knowledge from experience. This aspect of neural networks is of course much more than a mere novelty: it endows them with remarkable potential for practical application in a wide range of information processing tasks including content-addressable memory storage and recall, constrained optimization, and pattern classification.

The choice of a suitable learning rule is essential to the design of a system which is to perform some useful pseudocognitive function. It is likewise of great interest to explore the properties of various hypothetical learning rules in the context of models of biological nerve networks. The literature on learning rules is large and diverse. The seminal notion is due to Hebb:[1] if neuron i is postsynaptic to neuron j, and neuron j repeatedly or persistently promotes the firing of i, then the efficiency of the synaptic coupling from j to i increases. Notable contributions to the theory of learning algorithms have been made by Rosenblatt,[2] Widrow and Hoff,[3] Caianiello,[4] Anderson,[5] Cooper,[6] Grossberg,[7] Kohonen,[8] Sutton and Barto,[9] Palm,[10] Hopfield,[11] and Peretto,[12] among many other workers. Some systematic discussions of learning rules may be found in Refs. 10 and 12–14.

The purpose of this article is to point out that the relative entropy[15] $G(P_1,P_2)$, an information-theoretic measure of the distance between two probability distributions P_1 and P_2, may be used to assess the effectiveness of proposed learning rules in approximate realization of a given set of attractors by a noisy neural network of binary threshold neurons. In general, the desired attractors may be terminal cycles as well as fixed points, permitting a richer memory map to be achieved than in the standard Hopfield model with symmetrical couplings.[11]

The analysis is based on the probabilistic neural network specified in Sec. II. To make the presentation more concrete, we adopt the stochastic, parallel updating scheme of the Little model.[16] The problem of adjusting the probabilistic dynamical map to conform, as closely as possible, with the desired attractor mapping is stated as a problem of minimization of the relative entropy of the two associated probability measures. In turn, it is seen that the latter problem reduces to the maximization of a certain set of functions involving the altered neuronal interactions. In Sec. III, an optimal solution is given for the case of pairwise interactions between the neuronal units, assuming that increments in these interactions due to learning remain bounded. This solution is local in the sense that the optimal change in the coupling from the presynaptic neuron j to the postsynaptic neuron i involves only the state of i at the given time t and the state of j one update earlier. We go on to consider the extent to which optimality is achieved by four simple local learning rules that have appeared in earlier work. These cases include (i) the original Hebb rule, which is symme-

trical in i and j and is only effective if both i and j are active at the relevant times; (ii) the symmetrical learning rule implemented in the Hopfield model of content-addressable memory,[11] which produces positive or negative increments in the $j \rightarrow i$ coupling if the firing states of i and j are respectively correlated or anticorrelated at the relevant times; and (iii) two asymmetrical rules, one of which comes into play if and only if the postsynaptic neuron is active, the other if and only if the presynaptic neuron is active. Among these, the symmetrical rule of the Hopfield model and the presynaptic asymmetric rule (studied, for example, in Ref. 17) are the most effective in minimizing G.

The analysis is generalized to higher-order interactions in Sec. IV. The important question of the effect that learning new information has an older knowledge is addressed in Sec. V, where we establish a condition under which an incremental learning rule of the class examined does not disturb previously learned attractors. The implications and limitations of our results are discussed in Sec. VI. In the Appendix we compare our conclusions regarding the relative merits of proposed learning rules with the findings of earlier analytic and computational studies carried out by Peretto.[12]

II. ATTRACTOR NETWORKS AND RELATIVE ENTROPY

We consider a system of N two-state neurons, the state of neuron i being denoted by a variable x_i (or y_i) which takes the value 1 when i is firing and 0 when it is silent. The states $x=(x_i)\equiv(x_1,x_2,\ldots,x_N)$ available to the system form the set of vertices of an N-dimensional hypercube $X=\{0,1\}^N$. [An alternative description may be framed in terms of Ising-spin state variables $2x_i-1$ (or $2y_i-1$).] In the deterministic, noiseless case the time development of the system state x is governed by a parallel, discrete-time threshold dynamics ϕ on X, where ϕ is a mapping from X into itself. The dynamical mapping ϕ is generated by the neuronal interactions through a *firing function,* or activation,[13] $F(x)=[F_i(x), i=1,\ldots,N]$, in accordance with the threshold conditions

$$(\phi x)_i=\Theta(F_i(x)), \quad i=1,2,\ldots,N \tag{1}$$

where $\Theta(u)$ is the usual step function (taking the value 1 for $u\geq 0$ and 0 otherwise). The firing function F involves the couplings and thresholds of the neurons in a manner that we need not specify at this point.

In general, the dynamics ϕ possesses several attractors $a_m \subset X$, $m=1,2,\ldots,n$, each with its own area (or basin) of attraction A_m. The attractor sets a_m may be stable fixed points or invariant sets (limit cycles). For all $x\in a_m$,

$$\phi x \in a_m, \quad m=1,2,\ldots,n \tag{2}$$

while for all $x\in A_m$, there exists an integer $q_m(x)$ such that

$$\phi^{q_m(x)} x \in a_m, \quad m=1,2,\ldots,n \ . \tag{3}$$

The information or knowledge stored in the network is embodied in its system of attractors. Thus each attractor is considered to represent some memory or behavior pattern which has been learned or otherwise acquired by the net. To recall a particular memory or elicit a particular response, the system is given a stimulus that places its state in the area of attraction of the corresponding attractor.

To add a new item of information to the store requires the creation of a new attractor a_{n+1} by a suitable alteration of the dynamics ϕ to a new dynamics $\tilde{\phi}$. This will involve manipulation of the neuronal couplings and thresholds to produce a suitably revised firing function, denoted $\tilde{F}(x)$. For example, if a_{n+1} consists of a single configuration $x^{(0)}$, the new dynamics should satisfy $\tilde{\phi}x^{(0)}=x^{(0)}$. More generally, if a_{n+1} is a periodic attractor specified by the ordered set $\{x^{(1)},x^{(2)},\ldots,x^{(l)}\}$, one would require $\tilde{\phi}x^{(s)}=x^{(s+1)}$, for $s=1,2,\ldots,l$, with $x^{(l+1)}=x^{(1)}$. However, it is not always possible to achieve learning goals of this kind, which involve the realization of associations of the form $\tilde{\phi}x=y$, corresponding to specified transitions $x\rightarrow y$. In general, the most we can expect is that $\tilde{\phi}x^{(0)}$ is *close* to $x^{(0)}$, or the $\tilde{\phi}x^{(s)}$ are *close* to their target states $x^{(s+1)}$, according to some useful criterion for "closeness."

We shall now introduce and examine such a criterion, within a probabilistic generalization of the network model. The successor ϕx to state x becomes random with a probability distribution given by[16,18]

$$\nu_{\phi x}(x')\equiv P(\phi x=x') \equiv \prod_{i=1}^{N}\{1+\exp[-\beta F_i(x)(2x_i'-1)]\}^{-1} \ . \tag{4}$$

The deterministic case

$$P[(\phi x)_i=\Theta(F_i(x)), i=1,2,\ldots,N]=1 \tag{5}$$

is regained in the noiseless, zero-temperature limit $\beta\rightarrow\infty$, $T=\beta^{-1}\rightarrow 0$, provided none of the F_i vanishes exactly.

Consider again the problem of creating an additional attractor $\{x^{(1)},x^{(2)},\ldots,x^{(l)}\}$. In the presence of noise and the original attractors, the ideal behavior

$$P(\tilde{\phi}x^{(s)}=x^{(s+1)})=1, \quad s=1,2,\ldots,l \tag{6}$$

cannot be expected. However, we can insist that the probability measure $\nu_{\tilde{\phi}x^{(s)}}(x')$, the distribution of $\tilde{\phi}x^{(s)}$, is as close as possible to the measure $\delta_{x^{(s+1)},x'}$ concentrated at $x^{(s+1)}$, for $s=1,..,l$.

To characterize the distance between two probability measures ν and μ we adopt the relative entropy[15] (also called the asymmetric divergence or information gain). Assuming that the measure μ is absolutely continuous with respect to ν, the relative entropy of μ with respect to ν is defined by

$$G[\mu,\nu]=\int \ln\frac{d\mu}{d\nu}\mu(dx') \ . \tag{7}$$

[Absolute continuity of μ with respect to ν implies that there is no event e for which $\nu(e)=0$ and $\mu(e)\neq 0$; if this condition fails, then $G=+\infty$.] The quantity G is posi-

tive semidefinite and vanishes if and only if the two measures coincide. As framed, the definition is quite general and includes the case that the space of states x is continuous. For a countable number of mutually exclusive states α, and probability distributions $P_1(\alpha)$ and $P_2(\alpha)$ over these states, the relative entropy of P_1 with respect to P_2 assumes the more familiar form

$$G[P_1,P_2]=\sum_\alpha P_1(\alpha)\ln\frac{P_1(\alpha)}{P_2(\alpha)} . \tag{8}$$

This form is seen, for example, in the Boltzmann machine learning algorithm.[19] A similar version of the relative entropy has been employed by Hopfield in developing learning procedures for analog perceptrons as well as reciprocally connected statistical networks.[20] It may be noted that in the α summation of expression (8), the logarithm of the ratio of the two probabilities is weighted with the "primary" probability $P_1(\alpha)$. In the present application, we use G to compare the probability distribution of successor states $\tilde{\phi}x^{(s)}$ of a given network state $x^{(s)}$, with the desired probability distribution $\delta_{x^{(s+1)},x'}$.

The effectiveness of a proposed learning rule may be judged by the extent to which the corresponding modified firing function $\tilde{F}(x)=F(x)+\Delta(x)$ *minimizes* the relative entropy $G[\delta_{x^{(s+1)},x'},\nu_{\tilde{\phi}x^{(s)}}(x')]$. Here $\Delta(x)\equiv[\Delta_i(x),i=1,\ldots,N]$ is the change $\tilde{F}(x)-F(x)$ in the firing function due to the learning rule. To state the optimization prescription in a less cumbersome and more generic form, we rewrite $x^{(s)}$ as x and the target successor state $x^{(s+1)}$ as y. Then, for the probabilistic updating law assumed in Eq. (4), it is seen that $\delta_{y,x'}$ is absolutely continuous with respect to $\nu_{\tilde{\phi}x}(x')$ and that the integral in (7) or the sum in (8) reduces to the single contribution corresponding to $x'=y$. [The other contributions, behaving like $0\ln(0)$, vanish in an appropriate limiting process.] Thus we have

$$\begin{aligned}G&=G[\delta_{y,x'},\nu_{\tilde{\phi}x}(x')]\\&=-\ln\prod_{i=1}^{N}(1+\exp\{-\beta[F_i(x)+\Delta_i(x)](2y_i-1)\})^{-1}\\&=\sum_{i=1}^{N}\ln(1+\exp\{-\beta[F_i(x)+\Delta_i(x)](2y_i-1)\}) .\end{aligned} \tag{9}$$

To minimize G, we need only maximize $\Delta_i(x)(2y_i-1)$, for $i=1,2,\ldots,N$.

III. MINIMIZING G FOR PAIRWISE INTERACTIONS

With pairwise interactions among neurons, the firing function of neuronal unit i is traditionally written in the form[2,14,16,18]

$$F_i(x)=\sum_j V_{ij}x_j-V_{0i} , \tag{10}$$

where V_{ij} represents the "two-body" synaptic coupling through which neuron j influences neuron i, and V_{0i} is the threshold assigned to i. As is customary, we consider only the couplings as modifiable, the thresholds being regarded as fixed. Thus

$$\Delta_i(x)=\sum_j \Delta V_{ij}x_j . \tag{11}$$

Under the restriction that any changes in the couplings are bounded in the sense that

$$-a\le\Delta V_{ij}x_j\le b \quad (a,b>0) , \tag{12}$$

the optimal solution of the above G-minimization problem is

$$\Delta V_{ij}=\begin{cases} b & \text{if } x_j(2y_i-1)=1\\ -a & \text{if } x_j(2y_i-1)=-1\\ \mathcal{A} & \text{if } x_j=0 ,\end{cases} \tag{13}$$

where $\mathcal{A}$ represents an arbitrary value. Condition (12) may be viewed as a constraint on the available resources (biological or technological). Note that the presence of x_j as a factor in this condition implies that when neuron j is inactive there is no restriction on the change in V_{ij}. (If we choose to remove this factor, the arbitrary increment $\mathcal{A}$ in (13) is restricted to the range $[-a,b]$.) In general, the parameters b and a may be synapse dependent and may even depend on the individual transitions in the behavior pattern that is to be learned. However, we do not make these dependences explicit [omitting corresponding indices such as ij, $n+1$, and (s)], and we ignore this complication entirely in Sec. V.

An important feature of the optimal learning rule (13) is its *locality*. It is local in "space," since the change in the pairwise coupling V_{ij} depends only on the states of the synapsing neurons i and j and not on the states of any other neurons. It is also local in time, since the indicated dependence is on the current state of the presynaptic neuron j and the target successor state of the postsynaptic neuron i.

Four simple learning algorithms in common use are specified in the Table I (Rules 1–4). They are not only local, but also of *separable* form in the state variables of neurons i and j, being proportional to a presynaptic factor x_j or $2x_j-1$ and a postsynaptic factor y_i or $2y_i-1$. The first rule listed is just the original Hebb law, and the last is the one adopted by Hopfield[11] and used in the majority of works on content-addressable memories that ex-

TABLE I. Modes of action of local learning rules. The separable rules are numbered 1–4 as in the text. Opt.(10) and Opt.(20) denote the optimal rules derived from the minimization of relative entropy G based on the respective expressions (10) and (20) for the firing function. The symbol ● indicates that the corresponding neuron is firing and ○ that it is not firing. The parameters $b>0$, $a>0$, and $\eta>0$ specify the learning rates associated with the four possible configurations $j\to i$.

	ΔV_{ij}			
$j\to i$	●→●	●→○	○→●	○→○
Rule 1	η	0	0	0
Rule 2	η	0	$-\eta$	0
Rule 3	η	$-\eta$	0	0
Rule 4	η	$-\eta$	$-\eta$	η
Opt.(10)	b	$-a$		
Opt.(20)	b	$-a$	$-a$	b

ploit statistical physics and the spin-glass analogy.[21,22] A mean-field analysis of the memory-storage properties of all four rules (and indeed of the *general* local learning rule for pairwise couplings) has been given by Peretto[12] (see the Appendix). He also calls attention to neurophysiological evidence that the second rule is implemented at certain excitatory synapses in vertebrate brains. The third learning algorithm has been employed in neural-network simulations of classical conditioning.[17] We note that the first and fourth rules are symmetrical in i and j, while the other two are not. We may further note that the Hebb rule is only effective in producing a synaptic change if *both* presynaptic and postsynaptic neurons assume active states. Changes via rule 2 are contingent on activity of the *postsynaptic* neuron, while rule 3 requires *presynaptic* activity.

The G-minimization criterion developed above provides the basis for a useful assessment of the efficacy of the four "separable" learning algorithms. How close do these rules come to the optimal rule (13)?

1. Hebb rule. The symmetrical rule

$$\Delta V_{ij}=\eta x_j y_i, \quad \eta>0 \tag{14}$$

conforms to (13) when $b=\eta$ and $a=0$. Hence this rule minimizes G under the restriction of pairwise interactions and the resource constraint

$$0\le \Delta V_{ij}x_j \le \eta \ . \tag{15}$$

It is incapable of producing negative increments in the couplings V_{ij} and is correspondingly limited in its ability to decrease G.

2. Postsynaptic asymmetrical rule. The asymmetrical rule

$$\Delta V_{ij}=\eta(2x_j-1)y_i, \quad \eta>0 \tag{16}$$

implies

$$\Delta V_{ij}x_j(2y_i-1)=\begin{cases}\eta & \text{if } x_j(2y_i-1)>0\\ 0 & \text{otherwise ,}\end{cases} \tag{17}$$

maximizing $\Delta V_{ij}x_j(2y_i-1)$ subject to $0\le\Delta V_{ij}x_j\le\eta$. Thus it is seen that the rule (16), like (14), minimizes G under the resource constraint (15) [or (15) with the factor x_j removed]. The negative increment of V_{ij} produced by (16) in the case $x_j=0$, $y_i=1$ has no effect because of the presence of x_j as factor in the generic form of Δ_i.

3. Presynaptic asymmetrical rule. The other asymmetrical rule

$$\Delta V_{ij}=\eta x_j(2y_i-1), \quad \eta>0 \tag{18}$$

corresponds to the special case $b=a=\eta$ in the optimal solution (13). This algorithm is "nearly" optimal. It does permit the reduction of G through negative increments of V_{ij}, but the upper and lower bounds on the change ΔV_{ij} are constrained to have the same magnitude. For arbitrary a and b, the best choice that can be made for η is $\min(a,b)$.

4. Hopfield rule. The symmetrical rule

$$\Delta V_{ij}=\eta(2x_j-1), \quad \eta>0(2y_i-1) \tag{19}$$

is just as effective in minimizing G as the presynaptic asymmetric rule, changing V_{ij} by $+\eta$ in the case $x_j=y_i=1$, and by a numerically equal but negative increment $-\eta$ in the other pertinent case $x_j=1$, $y_i=0$.

Thus, with respect to the G-minimization criterion proposed in Sec. II, the separable local rules 1 and 2 are equivalent, as are 3 and 4. These features are clear from Table I, since the behavior of ΔV_{ij} in the fourth and fifth columns is irrelevant. By virtue of their ability to implement appropriate negative shifts of V_{ij} (see entries in the third column of Table I), rules 3 and 4 are preferred over rules 1 and 2.

An exemplification of these general conclusions may be found in recent computer simulations of Witt and Clark.[17] In particular, it is demonstrated that the simple presynaptic asymmetrical rule (18) has a strong stabilizing effect in consolidation of classically conditioned responses of networks of threshold neurons containing multiple feedback loops, and is markedly superior to the original Hebb law (14).

The form (10) of the firing function F_i, which corresponds to a simplified description of the response of the dendritic tree and cell body of a neuron to impinging synaptic stimuli, goes back to the very inception of neural-network theory in the 1943 paper of McCulloch and Pitts.[23] As documented (for example) in Refs. 14, 24, and 25, it has been used in most of the studies of digital neural-network models predating the recent surge of activity[21,22] based on modern statistical methods developed for the treatment of analogous spin systems involving mixed ferromagnetic and antiferromagnetic interactions. In this later work, the firing function F_i is regarded as a local field acting on the spin representing neuron i and is commonly expressed (in our notation) in the form

$$F_i(x)=\sum_j V'_{ij}(2x_j-1)-V'_{0i} \ . \tag{20}$$

The new threshold parameter V'_{0i}, which corresponds to an external magnetic field in the spin analogy, is often taken to be zero. The two formulations (10) and (20) are equivalent if

$$V'_{ij}=\tfrac{1}{2}V_{ij}, \quad V'_{0i}=V_{0i}-\tfrac{1}{2}\sum_j V_{ij} \ . \tag{21}$$

On the other hand, in developing learning rules we may—as above—choose to leave the thresholds V_{0i} or V'_{0i} intact while allowing the synaptic interactions V_{ij} or V'_{ij} to undergo modification. The two formulations are then no longer equivalent, since the second relation of (21) cannot be maintained.

It is therefore of interest to see what solution to the G-minimization problem emerges when (20) is adopted in place of (10), while keeping V'_{0i} fixed. For this case we assume that changes in the couplings obey the constraint

$$-a\le\Delta V_{ij}\le b \quad (a,b>0) \tag{22}$$

in place of (12), with $\Delta V'_{ij}=\Delta V_{ij}/2$ in accordance with the convenient definition $V'_{ij}\equiv V_{ij}/2$. The optimal rule is now

$$\Delta V_{ij}=\begin{cases} b & \text{if } (2x_j-1)(2y_i-1)=1 \\ -a & \text{if } (2x_j-1)(2y_i-1)=-1\ . \end{cases} \tag{23}$$

Inspection of Table I shows immediately that among the familiar separable local rules, the Hopfield rule (rule 4) comes the nearest to (23). It falls short of optimality only in that it corresponds to the specialization $a=b=\eta$ and thus is incapable of taking advantage of any difference in magnitude that might exist between upper and lower bounds on the change ΔV_{ij}. Given $a\neq b$, the best match of the Hopfield rule with the optimal solution is achieved with $\eta=\min(a,b)$. The Hebb algorithm (rule 1) is unambiguously the worst of the four separable prescriptions. Whereas (23) implies synaptic changes in all possible combinations of the presynaptic and postsynaptic firing states of the two neurons i and j, rule 1 produces a change only for the case that both neurons are active. The other two simple rules (2 and 3) are intermediate in quality, yielding appropriately signed synaptic changes in two of the combinations of firing states and no changes in the other two. In those firing configurations where they are effective, these algorithms coincide with the Hopfield rule and thus do not distinguish between the magnitudes of positive and negative corrections.

IV. OPTIMAL LEARNING WITH HIGHER-ORDER INTERACTIONS

The foregoing analysis may be extended to a more general form of the firing function $F(x)$ that allows for higher-order or multiple-neuron interactions. For interactions of order K, where $2\leq K\leq N$, the form (10) generalizes to

$$F_i(x)=\sum_j V_{ij}x_j+\sum_{j_1,j_2} V_{i\{j_1j_2\}}x_{j_1}x_{j_2}+\cdots + \sum_{j_1j_2\cdots j_{K-1}} V_{i\{j_1j_2\cdots j_{K-1}\}}x_{j_1}x_{j_2}\cdots x_{j_{K-1}}-V_{0i}\ , \tag{24}$$

the quantity $\Delta_i(x)$ taking a similar form with $V_{\ldots}$ replaced by $\Delta V_{\ldots}$ and V_{0i} omitted. It is readily seen that an optimal solution of the G-minimization problem under the resource constraints

$$-a_k\leq \Delta V_{i\{j_1j_2\cdots j_k\}}x_{j_1}x_{j_2}\cdots x_{j_k}\leq b_k \qquad (k=1,\ldots,K-1) \tag{25}$$

is given by

$$\Delta V_{i\{j_1j_2\cdots j_k\}}=\begin{cases} b_k & \text{if } y_i=x_{j_1}=\cdots=x_{j_k}=1 \\ -a_k & \\ & \text{if } y_i=0,\ x_{j_1}=x_{j_2}=\cdots=x_{j_k}=1 \\ \mathcal{A} & \text{otherwise}\ , \end{cases} \tag{26}$$

where i and the j indices range over $1,\ldots,N$ and k runs from 1 to $K-1$. In the special case $a_k=0$, $b_k=\eta$, for all k, this solution can be achieved by the generalization

$$\Delta V_{i\{j_1j_2\cdots j_k\}}=\eta x_{j_1}x_{j_2}\cdots x_{j_k}y_i \quad (1\leq k\leq K-1) \tag{27}$$

of the Hebb rule, or by a corresponding generalization of the postsynaptic asymmetric rule of Table I. Likewise, in the special case $a_k=b_k=\eta$, the solution is matched by the rule

$$\Delta V_{i\{j_1j_2\cdots j_k\}}=\eta x_{j_1}x_{j_2}\cdots x_{j_k}(2y_i-1) \qquad (1\leq k\leq K-1)\ , \tag{28}$$

which extends the presynaptic asymmetric rule of Table I, or by a corresponding generalization of the Hopfield rule.

A parallel generalization of the analysis to multiple-neuron interactions may readily be performed based on the alternative form (20) for the firing function F_i.

V. SUCCESSIVE LEARNING

In models of learning, it is usually assumed that alterations of the synaptic interactions are simply cumulative. More specifically, in learning r transitions $x^{(1)}\rightarrow y^{(1)}, x^{(2)}\rightarrow y^{(2)},\ldots,x^{(r)}\rightarrow y^{(r)}$, the presynaptic asymmetric rule is implemented additively or incrementally, resulting in a net change of the form

$$\Delta V_{ij}=\lambda\sum_{q=1}^{r} x_j^{(q)}x_j^{(q)}\cdots x_{j_{K-1}}^{(q)}(2y_i^{(q)}-1) \tag{29}$$

in the case of Kth-order interactions, with λ some positive constant. A similar expression applies when any of the other simple learning rules of the preceding sections is chosen, or when one or another optimal solution [e.g., (13), (23), or (26)] is adopted.

In judging the performance of a memory system, an important consideration is the extent to which the acquisition of new items of information interferes with items engrammed previously. Upon learning a new attractor a_{n+1}, some of the fixed states or cycles a_m stored earlier may be "forgotten," i.e., no longer recalled by the network under approximate stimulus. It is natural to ask under what conditions the a_m with $m\leq n$ do remain attractors of the modified network dynamics $\tilde{\phi}$. A partial answer to this question is offered below.

Suppose the system has previously learned the associations defined by the transitions $x^{(1)}\rightarrow y^{(1)}, x^{(2)}\rightarrow y^{(2)},\ldots,x^{(r)}\rightarrow y^{(r)}$, in the sense that a dynamics ϕ has been realized such that in the noiseless limit $\phi x^{(q)}=y^{(q)}$ for all $q=1,\ldots,r$. Suppose further that an additional association $x^{(r+1)}\rightarrow y^{(r+1)}$ has been learned by means an incremental presynaptic asymmetric rule (29), so that (again in the noiseless limit) the new dynamics yields $\tilde{\phi}x^{(r+1)}=y^{(r+1)}$. Then the older memories (or associa-

tions) $x^{(q)} \rightarrow y^{(q)}$ can still be recalled, i.e., $\tilde{\phi} x^{(q)} = y^{(q)}$, $q = 1, \ldots, r$, if and only if

$$\lambda < \min_{\{q,i: y_i^{(q)} \neq y_i^{(r+1)}\}} \lambda_{q,i} , \tag{30}$$

where

$$\lambda_{q,i} = \frac{|F_i(x^{(q)})|}{((x^{(q)}, x^{(r+1)}))_K} . \tag{31}$$

Here we have introduced a scalar product

$$((x,y))_K \equiv \sum_j x_j y_j + \sum_{j_1,j_2} x_{j_1} x_{j_2} y_{j_1} y_{j_2} + \cdots + \sum_{j_1 j_2 \cdots j_{K-1}} x_{j_1} x_{j_2} \cdots x_{j_{K-1}} y_{j_1} y_{j_2} \cdots y_{j_{K-1}} . \tag{32}$$

Noting that the indicated storage of the old memories in ϕ implies $2y_i^{(q)} - 1 = \mathrm{sgn}[F_i(x^{(q)})]$, for all i and for $q = 1, \ldots, r$, the condition (30) is established by performing the following sequence of manipulations (for i and q such that $y_i^{(q)} \neq y_i^{(r+1)}$):

$$\begin{aligned} \tilde{F}_i(x^{(q)})(2y_i^{(q)} - 1) &= |F_i(x^{(q)})| + \Delta_i(x^{(q)})(2y_i^{(q)} - 1) \\ &> \lambda((x^{(q)}, x^{(r+1)})) + \lambda \left[\sum_j x_j^{(q)} x_j^{(r+1)} + \sum_{j_1,j_2} x_{j_1}^{(q)} x_{j_2}^{(q)} x_{j_1}^{(r+1)} x_{j_2}^{(r+1)} + \cdots \right. \\ &\quad \left. + \sum_{j_1 \cdots j_{K-1}} x_{j_1}^{(q)} \cdots x_{j_{K-1}}^{(q)} x_{j_1}^{(r+1)} \cdots x_{j_{K-1}}^{(r+1)} \right] (2y_i^{(q)} - 1)(2y_i^{(r+1)} - 1) \geq 0 . \end{aligned} \tag{33}$$

In reducing $\Delta_i(x^{(q)})$, we have made use of the definition $\Delta_i(x) = \tilde{F}_i(x) - F_i(x)$, the higher-order expression (24) for the firing function, and the assumed incremental learning rule. This analysis suggests that the parameters $\lambda_{q,i}$ may be used to measure the strength of a previously stored association q relative to a new association $r+1$.

VI. DISCUSSION

Considering networks of two-state neurons governed by a noisy dynamics, we have studied some variational aspects of the problem of learning a given behavior pattern, a process which, in the absence of noise, would correspond to the acquisition by the system of an associated attractor. In general, an attractor consists of a definite sequence of state transitions. The learning process will be enhanced if the probability distribution of successors of an arbitrary state $x^{(s)}$ on the attractor is brought as close as possible to the desired distribution localized at $x^{(s+1)}$. An appropriate scalar measure of the distance between these distributions is the relative entropy G. The learning problem is then reduced to a problem of adjusting the interactions among the neurons in such a way as to minimize G. We have presented explicit solutions of this optimization problem for the cases of pairwise (two-neuron) and arbitrary higher-order (multiple-neuron) interactions. The analysis is performed without any restriction on the symmetry of the interactions, but resource constraints have been imposed in the form of upper and lower bounds on changes in the synaptic couplings. The optimal solutions provide a basis for the assessment of four well-known separable local learning rules.

The formal development has been couched in terms of the Little model, which involves synchronous updating of all units in the assembly of neurons. However, it is easily seen that the results for the optimal learning rule and for the optimization properties of the common separable examples of Table I are considerably more general. In particular, the same analysis applies to the asynchronous dynamics of the Hopfield model,[11] with the sole modification that the state transitions involve single-neuron updates, so that the product over i in Eq. (4) reduces to a single factor. It may further be shown that the results generalize to continuous-time Markov models.

The generality of our approach also extends in other directions.

(a) As already indicated, the patterns to be learned may be ordered sequences of states $\{x^{(1)}, x^{(2)}, \ldots, x^{(l)}\}$ as well as single configurations $x^{(0)}$, corresponding respectively to the acquisition of limit-cycle and fixed-point attractors in an idealized noiseless case.

(b) Our main arguments based on G minimization entail no restrictions on the architecture of the system prior to a given learning attempt, i.e., the original couplings V_{ij} may be arbitrary.

(c) The system may operate at an arbitrary noise level β^{-1}.

The generality of applicability of the principle of minimum relative entropy is tempered by its broad character: This principle is based on an overall evaluation, through the single quantity G, of the likehood that the modified system will make correct transitions when placed in states belonging to an arbitrary candidate attractor. Accordingly, our analysis serves to illuminate only limited aspects of learning theory. Within the context of attractor networks,[22] such important practical matters as the sizes of areas of attraction and the rates of recall of acquired patterns have been left untouched, and nothing has been said about the actual learning time involved in the stepwise application of one or another algorithm. We have, in Sec. V, stated a simple result on the destablization of old patterns by newly acquired ones. However, the argument presented there is quite independent of the primary development based on the G-

minimization criterion. We have not sought to derive systematic, explicit results for the accuracy of recall of specific memories or for the storage capacities that might be allowed by the various learning rules.

The latter aspects of the learning problem—accuracy of recall and capacity under different learning algorithms—have been addressed in some detail by Peretto,[12] who has used mean-field theory to evaluate the memory storage capabilities of networks of two-state threshold neurons. Some very important differences from our approach are apparent. On the one hand, as stressed above, the principle of minimum relative entropy provides a very broad measure of the quality of proposed learning rules, applicable in a very general setting. On the other hand, Peretto's mean-field analysis is quite specific, involving (a) stochastically independent, single-configuration memory patterns, (b) a fully connected network architecture, containing an asymptotically large number of neurons, and (c) low noise. Consequently, his treatment leads to more incisive results for a much more restricted problem. Considering these important differences, there is essential agreement of the two approaches in indicating superior performance for the Hopfield rule (or something close to it) and the problematic nature of the original Hebb law. The Appendix contains further discussion of the Peretto study, together with an explicit comparison of its conclusions with those derived in Sec. III.

In connectionist language, the present work bears on the "credit assignment problem"[25] of multilayer nets or nets with feedback loops, although it does not attempt a full solution. Within this context, Hopfield[20] has examined the role of relative entropy in the construction of learning rules that enable such nets to capture probabilistically specified input-output relations. (Two restrictive cases were considered, both of which involve an architecture having three layers of neuronal units, with connections only between neurons in adjacent layers. One case is a feedback perceptron with analog units; the other is a symmetrically wired Boltzmann machine with noisy binary threshold units.) It is also important to recall that relative entropy plays an essential role in the original Boltzmann machine learning algorithm,[18] which provided one of the first solutions of the credit-assignment problem.

ACKNOWLEDGMENTS

This research was supported in part by the Chinese National Science Foundation and by the U.S. National Science Foundation, under Grant No. DMR-9002863. One of us (M.Q.) thanks the Department of Mathematics, Washington University, for kind hospitality during a leave from Peking University. We thank J.M.C. Chen for useful discussions.

APPENDIX: COMPARISON WITH PERETTO'S ANALYSIS

In this appendix, we outline Peretto's investigations[12] of local learning rules and examine his findings as they relate to corresponding conclusions derived from the principle of minimum relative entropy.

In Ref. 12, the primary criterion adopted for evaluation of the relative efficacies of various learning rules is maximum storage capacity consistent with accurate recall of particular memories. In contrast to our treatment, special assumptions are made regarding the nature of the patterns to be learned and the architecture of the network, and explicit analysis is mostly confined to the low-noise limit. The number N of neurons is taken to be asymptotically large, so that the analytical results obtained refer strictly to the thermodynamic limit. The expression (20) is adopted for the firing function (local field) of neuron i, with $V'_{0i}=0$.

A set of M single-configuration patterns is to be learned by the network, pattern μ being specified by $\sigma_i=\xi_i^{(\mu)}$, $i=1,\ldots,N$. Here we follow Peretto's use of the conventional notation in which the state variable of neuron i is denoted $\sigma_i^{(\mu)}$ and takes the value $+1$ if i is active and -1 if it is not. Ideally, under a proposed learning rule all the nominated collective states $\{\xi_i^{(\mu)}\}$ should become fixed points of the dynamics. In the presence of noise at temperature β^{-1}, this condition is translated into the statistical statement

$$\langle \sigma_i \rangle = \langle \tanh(\beta F_i) \rangle \ , \tag{A1}$$

where the angle brackets indicate a large-time average over a statistical ensemble, and a simplified description of synaptic noise (corresponding to that of Little[16]) has been implemented. An assumption crucial to the subsequent development is that the patterns μ are stochastically independent, the choices $+1$ and -1 for each component $\xi_i^{(\mu)}$ having equal probability (*random patterns*).

Peretto considers the most general class of learning rules for pairwise synaptic interactions:

$$\Delta V'_{ij}=N^{-1}(A\xi_i^{(\mu)}\xi_j^{(\mu)}+B\xi_i^{(\mu)}+C\xi_j^{(\mu)}+D) \ . \tag{A2}$$

Saturation of synaptic strengths is not taken into account, and the full network couplings V'_{ij} are obtained simply by summing (A2) over all patterns μ, and adding an extra pattern-independent term $N^{-1}D^0$ as a nonmodifiable ("nonplastic") component. This construction implies that, in general, every neuron interacts, asymmetrically, with every other neuron, and also experiences a self-interaction. The learning parameters A, B, C, and D, as well as the nonplastic parameter D^0, may in principle be synapse dependent and should then carry labels ij. Indeed, natural learning may well involve different rules of modification at excitatory and inhibitory synapses.[12] Such complications are not explored formally in Peretto's work.

Within the indicated framework, the limits of memory storage are investigated (i) for the case of a finite number M of patterns, $M/N=O(1/N)$; and (ii) for an infinite number of patterns, with the desired scaling $\alpha\equiv M/N=O(1)$, where α is called the load. In both cases, the strong connectivity assumed for the network is used to justify the application of the mean-field approximation, which (for example) allows the fixed-point condition (A1) to be replaced by $\langle \sigma_i \rangle = \tanh(\beta\langle F_i \rangle)$. The analysis in both cases focuses on the order parameter cor-

responding to a particular pattern, say pattern 1:

$$m^{(1)}=N^{-1}\sum_j \xi_j^{(1)}\langle\sigma_j\rangle\ . \tag{A3}$$

For good performance of the system as a content-addressable memory, this quantity should be as near to unity as possible, or the retrieval error fraction $(1-m^{(1)})/2$ should be close to zero—indicating that the system relaxes to a collective state near pattern 1 if it is initiated in that pattern. In pursuing mean-field theory, the specialization to random patterns allows one to neglect destabilizing terms arising from the other patterns $\mu\neq 1$ in case (i) and to invoke the self-averaging hypothesis (see below) in case (ii). Thereupon explicit coupled equations may be derived for the order parameter $m^{(1)}$ of Eq. (A3) and the order parameter $m^{(0)}=N^{-1}\sum_j\langle\sigma_j\rangle$, corresponding to the uniform field arising from $MD+D^0\neq 0$ [and for two additional order parameters in case (ii)]. The assumption that a given $\xi_i^{(\mu)}$ is equally likely to be $+1$ as -1 is exploited in the derivation of these coupled equations, and has the consequence that the learning parameters A, B, C, and D (together with the nonplastic parameter D^0 of V'_{ij}) enter only in the combinations

$$\begin{aligned} &s=B+MD+D^0,\quad t=C+A\ ,\\ &u=-B+MD+D^0,\quad v=C-A \end{aligned} \tag{A4}$$

in case (i) and in these combinations and A^2+C^2, MB^2, A, and C^2 in case (ii).

In Table II we collect the particular choices of the constants A, B, C, and D of the general form (A2) that correspond to the separable learning rules 1–4 defined in Sec. III, and to the optimal rule (23) given by G minimization for the relevant form (20) of the firing function. The labeling of rules in this table should not be confused with that used in Table II of Ref. 12.

In the case of a finite number of nominated memory patterns, Peretto finds that $|m^{(1)}|=1$ and $m^{(0)}=0$ are solutions of the fixed-pointed equations of the zero-noise problem if and only if the conditions $t=C+A>0$ and $v=C-A<0$ are met, which requires $A>0$ and $|C|<A$. The other solutions of the zero-noise equations do not yield acceptable memory properties. From Table II it is seen that rules 3 and 4 as well as the optimal rule satisfy the stated conditions (with $t=A$ and $v=-A$ in all three cases). Rules 1 and 2 yield $v=0$ and therefore must be excluded, although any small decrease of C relative to A would eliminate this defect. The next important question is that of stability around acceptable fixed-point solutions. In the zero-noise limit stability is guaranteed, but more generally it is desirable that the eigenvalues of the relaxation matrix be real and negative over the largest possible domains of any remaining adjustable parameters, for given fixed points $m^{(0)*}$ and $m^{(1)*}$ and noise level β^{-1}. As far as dependence on the learning parameters is concerned, the sizes of these domains are governed entirely by the parameter combinations $X=s+t+u-v$ and $Y=ut-sv$. The actual expression for the eigenvalues takes the form

$$\Lambda_\pm=\kappa[X-\kappa^{-1}\pm(X^2-8Y)^{1/2}]/\tau\ , \tag{A5}$$

where κ is a positive constant determined by β and the fixed-point solutions $m^{(0)*}$ and $m^{(1)*}$, and τ is the elementary time step. Thus reality requires $X^2-8Y\geq 0$, and negativity is favored by a more negative value of X. It is interesting that the two viable separable algorithms, namely the presynaptic rule (rule 3) and the Hopfield rule (rule 4), produce identical values for the key parameters X and Y, namely $X=2(A+D^0)$ and $Y=2AD^0$. Hence they must be judged equally effective in the present context. It may be recalled that according to the minimum-G criterion, the Hopfield rule is more nearly optimal than the presynaptic rule, when form (20) is used for the firing function. The optimal rule itself, which has a D admixture in the amount $D=N(b-a)/4$, gives $X=2(A+D^0)+2MD$ and $Y=2AD^0+2MAD$. If the bounds a and b appearing in the resource constraint (22) happen to coincide, one obtains the same behavior as for rules 3 and 4. However, if $b>a$, the optimal rule must be regarded as inferior to these in Peretto's problem, since the corresponding values of X and Y are larger algebraically by positive amounts $2MD$ and $2MAD$, respectively; while if $b<a$, the converse holds and the optimal rule is superior. For the sake of completeness, we quote the values of the key stability parameters of the other two separable rules: $X=2(A+D^0)+2MA$, $Y=2AD^0+2(M-1)A^2$ (Hebb case, rule 1) and $X=2(A+D^0)$, $Y=2AD^0$ (postsynaptic case, rule 2). Even if we could ignore the issue of acceptable fixed-point solutions for $m^{(1)}$ and $m^{(0)}$, the Hebb rule would still suffer from unfavorable stability properties when compared to rules 3 and 4 (and also rule 2).

Considering only the signs of s, t, u and v, Peretto asserts that the most favorable situation for stability is $-,+,-,-$. The second and fourth sign conditions are automatically met for presynaptic, Hopfield, and optimal rules. To satisfy the other two sign conditions, parameter choices $D^0<-A$, $D^0<0$, and $D^0<-MD$ must be made in the respective cases. For $|MD|>A$, the last is the strongest constraint of the three if $b>a$ and the weakest if $b<a$.

We now turn to Peretto's assessment of the efficacy of learning rules of the class (A2) for the problem in which an infinite number of patterns M is to be stored. This problem has been thoroughly studied by Amit and coworkers[21,22] for the case of symmetrical Hopfield couplings, by adapting methods developed for the equilibrium statistical mechanics of spin glasses, notably the replica trick. However, the general learning rule (A2) induces

TABLE II. Characterization of learning rules displayed in Table I in terms of the general representation (A2) used by Peretto. A "normalizing" factor $N/2$ has been removed from each entry.

	A	B	C	D
Rule 1	η	η	η	η
Rule 2	η	0	η	0
Rule 3	η	η	0	0
Rule 4	η	0	0	0
Opt.(20)	$(b+a)/2$	0	0	$(b-a)/2$

asymmetrical interactions, precluding the direct transference of these techniques. Accordingly, Peretto adopted a mean-field approach in conjunction with the self-averaging hypothesis. The latter ansatz amounts to the strong assumption that all statistical observables are sample independent in the thermodynamic limit, or, more to the point, that the memory storage capacity of a large neural net of the type under consideration can be found by averaging over a large number of realizations of the nominated patterns. The formal treatment based on the mean-field description and self-averaging was supplemented by rough scaling arguments and by computer simulations involving 200 and 400 neurons.

The scaling arguments are based on the elementary stability conditions needed to ensure that an arbitrarily chosen memory pattern (pattern 1) is a fixed point of the noiseless dynamics:

$$\xi_i^{(1)} F_i(\{\xi_j^{(1)}\}) > 0, \quad i = 1, \ldots, N \ . \tag{A6}$$

For random candidate patterns, destabilization of pattern 1 due to the local-field contributions from the other patterns $\mu \neq 1$, or due to the uniform field arising from $MD + D^0$, is predicted to occur when the combination of parameters

$$\frac{M}{N}\left[1+\left[\frac{B}{A}\right]^2+\left[\frac{C}{A}\right]^2+\frac{1}{M}\left[\frac{MD+D^0}{A}\right]^2\right]+\left[\frac{C}{A}\right]^2 \tag{A7}$$

becomes of order unity or greater. The best choice of parameters is clearly $A \neq 0$, $B = C = D = 0$, yielding the Hopfield rule, with its characteristic scaling behavior[21,22] $M_c \sim N$ for the memory capacity M_c(absent D^0). A nonzero value of any of the parameters B, C, or D entails a degradation of memory performance of some kind and in some degree, and hence the Hebb rule is obviously the worst of those shown in Table II. Studying the effects of the learning parameters B, C, and D individually relative to the Hopfield case, the structure of (A7) implies that C is the most destabilizing, and that a necessary requirement for stability is $|C| < A$, where, as usual, A is taken to be positive. This conclusion speaks against rules 1 and 2 of Table II, which were also found wanting in the finite-M analysis. The parameter B is the least damaging in that it does not destroy the scaling $M_c \sim N$ characteristic of the Hopfield rule. However, its presence does reduce the memory capacity from that of the Hopfield model. In particular, rule 3, which corresponds to $A = B > 0$, $C = D = 0$, yields the value $2M/N$ for (A7) (again, absent D^0) and is thus estimated to provide *half* the capacity allowed by the Hopfield rule, although it was seen to give a performance equivalent to that of rule 4 within the finite-M analysis. The parameter D is potentially quite dangerous, since an uncompensated D term in (A7) restricts the scaling behavior of the capacity to $M_c \sim \sqrt{N}$. Along with rule 1, the optimal rule of Table II may suffer from this restriction, if $a \neq b$; otherwise it coincides with the Hopfield rule and is immune. However, it is important to note that the effect of D is coupled with that of the nonplastic synaptic parameter D^0. The latter parameter has no impact on stability, unless it scales like M^n, with $n \geq 1$. But if $n = 1$ it may in fact be used to compensate the harmful influence of a D component (as for the optimal rule), by choosing $D^0 = -MD$.

Using the mean-field approximation with self-averaging, Peretto has constructed a set of coupled equations from which the order parameters $m^{(1)}$ and $m^{(0)}$ may be determined at arbitrary β, for any combination of learning parameters. For the parameters of the Hopfield rule, these equations reduce to those obtained by Amit and co-workers[21,22] based on the replica technique. A systematic study of the coupled equations in their full generality was not carried out. Rather, solutions were only discussed in the zero-noise limit, for the special case $A > 0$, $B = MD + D^0 = 0$. This case was chosen with the intent of isolating the destabilizing effects of the parameter C, putatively the most injurious to memory performance. The principal finding is that the efficacy of recall (as measured by $m^{(1)}$) and the critical load $\alpha_c \equiv M_c/N$ (beyond which retrieval fails catastrophically) experience steady degradation as $|C|/A$ rises from 0 to 1, at which point all memory storage abilities are lost. The critical value of $m^{(1)}$, i.e., the value of the order parameter at $\alpha = \alpha_c$, diminishes progressively from 0.97 to 0.5 as $|C|/A$ increases through the stated range. The results of this analytical approach are in "fair" agreement with the corresponding computer simulations, but significant deviations are evident, particularly at the larger values of $|C|/A$. Available computer simulations also provide some information on the effects of the other parameters, particularly B and D. As predicted by the scaling considerations, the harmful effect of B is weaker than that of C, implying (once more) that rule 3, though worse than rule 4, is better than rule 2 or rule 1. The deleterious influence of D is not found to be as serious as expected, in that it produces a selective destabilization of patterns in contrast to the even degradation caused by B and C. Accordingly, the optimal rule (23), with $a \neq b$, may not be much inferior to the Hopfield rule, especially if $b < a$ and D is thus negative. Moreover, as pointed out above, the auxiliary nonplastic parameter D^0 may be adjusted to reduce or remove the destabilizing effect of the D component.

In summary, the arguments presented in Ref. 12 point to the Hopfield memory storage algorithm (our rule 4) as the best rule of the general class (A2), at least within the somewhat restricted framework of the memory model considered. Of the four separable rules of Tables I and II, rule 3 is judged to be better than rule 2, and rule 1 (the simple Hebb prescription) is the worst of all. The relative performance of the optimal rule (23) emerging from G minimization for the corresponding firing function (20) depends critically on the difference in the resource constants b and a appearing in Eq. (22), since $2D/N = (b - a)/2$. In favorable circumstances, this algorithm is equivalent to the Hopfield rule and in some cases it can even be somewhat better (cf. Peretto's analysis for finite M).

Our results based on G minimization also favor the Hopfield rule among the elementary separable examples. The Hebb rule is again identified as the least effective, the

presynaptic and postsynaptic rules being of equivalent, intermediate effectiveness in reducing G. In comparison with the optimal rule (23), the Hopfield rule is faulted only in its inability to make full use of the available synaptic resources as specified by (22). On balance, it can be said that the two approaches to the assessment of learning algorithms are in qualified agreement. Differences in detailed conclusions may be attributed, on the one side, to the highly specific assumptions necessary for Peretto's treatment and, on the other, to the compromises that must be struck by a criterion so general as the principle of minimum relative entropy.

[1]D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949), p. 62.
[2]F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington, D.C., 1962).
[3]B. Widrow and S. D. Stearns, *Adaptive Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1985).
[4]E. R. Caianiello, J. Theor. Biol. **2**, 204 (1961).
[5]J. A. Anderson, Math. Biosci. **8**, 137 (1970).
[6]L. N. Cooper, in *Proceedings of the Nobel Symposium on Collective Properties of Physical Systems,* edited by B. Lundquist and S. Lundquist (Academic, New York, 1973), p. 252.
[7]S. Grossberg, Biol. Cybern. **23**, 121 (1976).
[8]T. Kohonen, *Associative Memory—A System-Theoretic Approach* (Springer-Verlag, Berlin, 1977); *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).
[9]R. S. Sutton and A. G. Barto, Psychol. Rev. **88**, 135 (1981).
[10]G. Palm, *Neural Assemblies: An Alternative Approach to Artificial Intelligence* (Springer-Verlag, Berlin, 1982).
[11]J. J. Hopfield, Proc. U.S. Natl. Acad. Sci. **79**, 2554 (1982).
[12]P. Peretto, J. Phys. (Paris) **49**, 711 (1988).
[13]D. E. Rumelhart *et al.*, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA, 1986), Vol. 1.
[14]J. W. Clark, in *Nonlinear Phenomena in Complex Systems,* edited by A. N. Proto (Elsevier, Amsterdam, 1989), p. 1.
[15]S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959); M. S. Pinsker, *Information and Information Stability of Random Processes* (Holden-Day, San Francisco, 1964); S. R. S. Varadhan, *Large Deviations and Applications, CBMS Regional Conference Series in Applied Mathematics* (Society of Industrial and Applied Mathematics, Philadelphia, 1984), Vol. 46.
[16]W. A. Little, Math. Biosci. **19**, 101 (1974).
[17]J. C. Witt and J. W. Clark, Math. Biosci. **99**, 77 (1990).
[18]J. W. Clark, Phys. Rep. **158**, 91 (1988).
[19]D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Sci. **9**, 147 (1985).
[20]J. J. Hopfield, Proc. U.S. Natl. Acad. Sci. **84**, 8429 (1987).
[21]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987).
[22]D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, England, 1989).
[23]W. S. McCulloch and W. Pitts, Bull. Math. Biophys. **5**, 115 (1943).
[24]J. W. Clark, J. Rafelski, and J. V. Winston, Phys. Rep. **123**, 215 (1985).
[25]J. D. Cowan and D. H. Sharp, Q. Rev. Biophys. **21**, 365 (1988).