Associative recall of memory without errors

I. Kanter and H. Sompolinsky*
Department of Physics, Bar-Ilan University, 52 100 Ramat-Gan, Israel
(Received 9 June 1986)

A neural network which is capable of recalling without errors any set of linearly independent patterns is studied. The network is based on a Hamiltonian version of the model of Personnaz *et al.* The energy of a state of N (± 1) neurons is the square of the Euclidean distance—in phase space—between the state and the linear subspace spanned by the patterns. This energy corresponds to *nonlocal* updatings of the synapses in the learning mode. Results of the mean-field theory (MFT) of the system as well as computer simulations are presented. The stable and metastable states of the network are studied as a function of "temperature" T and $\alpha=p/N$, where p is the number of embedded patterns. The maximum capacity of the network is $\alpha=1$. For all α ($0\leq\alpha<1$) the embedded patterns are not only locally stable but are global minima of the energy. The patterns appear, as *meta*stable states, below a temperature $T=T_M(\alpha)$. The temperature $T_M(\alpha)$ decreases to zero as $\alpha\rightarrow 1$. The spurious states of the network are studied in detail in the case of random uncorrelated patterns. At finite p, they are identical to the mixture states of Hopfield's model. At finite α, a spin-glass phase exists as a metastable state. According to the replica symmetric MFT the spin-glass state becomes degenerate with the patterns at $\alpha=\alpha_g=1-2/\pi$ and *disappears above it.* Possible interpretations of this unusual result are discussed. The average radius of attraction R of the patterns has been determined by computer simulations, for sizes up to $N=400$. The value of R for $0<\alpha<1$ depends on the details of the dynamics. Results for both parallel and serial dynamics are presented. In both cases R is unity (the largest distance in phase space by definition) at $\alpha\rightarrow 0$ and decreases monotonically to zero as $\alpha\rightarrow 1$. Contrary to the MFT, simulations have not revealed, so far, any singularity in the properties of the spurious states at an intermediate value of α.

I. INTRODUCTION

A. Neural networks with local learning rules

Recently there has been an upsurge of interest in models of neural networks which exhibit associative memory.[1-7] In many of the models, the network consists of a highly connected system of spins (neurons) whose internal connections (synapses) are updated to facilitate the storage and the retrieval of information. Usually the synapses are designed so that a given set of states of the system become fixed attractors of its dynamic evolution. These states are the patterns which are memorized by the network. This memory is associative: starting from an initial state which partially resembles one of the patterns the system evolves fast into that pattern. We will assume for specifity that the neurons are two-state elements $S=\pm 1$. A network of N neurons has 2^N possible states. Out of these, a set of p states $\{\xi_i^\mu\}$ ($i=1,2,\ldots,N$; $\mu=1,2,\ldots,p$) constitutes the patterns or memories. Note that each ξ_i^μ is either $+1$ or -1. It represents the value of S_i in the μth pattern. The synapse between S_i and S_j is denoted by J_{ij}.

Most studies of neural networks focused on *local* learning rules: Each J_{ij} depends only on the values of ξ_i^μ and ξ_j^μ. Locality of the learning rules is a reasonable condition for biological networks. There, the synapses are modified presumably by the past activities of the neurons they connect. These past activities are represented by the "quenched" values ξ_i^μ and ξ_j^μ. In building artificial devices the restriction of locality can be lifted. Furthermore, it is not unreasonable that even in biological systems, some long-term synaptic modifications which depend on the history of activity of a large group of neurons might occur, in addition to the main local modifications. It is thus interesting to study the potential of networks with nonlocal learning rules.

Local learning rules have the common limitation that correlated patterns are difficult to learn. The overlaps between the patterns are characterized by the matrix,

$$C_{\mu\nu}\equiv N^{-1}\sum_{i=1}^{N}\xi_i^\mu\xi_i^\nu,\quad \mu,\nu=1,2,\ldots,p\ . \tag{1.1}$$

They act as an internal static noise which tends to misalign spins relative to the original patterns. When the patterns are uncorrelated random variables the combined overlap of a pattern with all the other patterns is of $O(\sqrt{p/N})$. Hence, when p is finite the patterns are effectively orthogonal. If, on the other hand, $\alpha\equiv p/N$ is *finite,* the stable states of the system contain[3] a finite fraction of errors, which increases with α. When α reaches a critical value α_c, there is a dramatic increase in the level of errors and the system ceases to provide associative memory. The value of α_c depends on the details of the model but it is always less than 0.14.[3,8]

B. Nonlocal model of Personnaz *et al.* (Ref. 5)

One potential advantage of nonlocal learning rules is that they can suppress the adverse effects of the overlaps

among the patterns. In fact, Personnaz *et al.*[5] have constructed a nonlocal model in which any set of patterns, correlated or not, can be memorized without errors as long as they are linearly independent. Their model consists of a set of synapses

$$J_{ij}=N^{-1}\sum_{\mu,\nu=1}^{p}\xi_i^\mu\xi_j^\nu(\underline{C}^{-1})_{\mu\nu} \tag{1.2}$$

where $\underline{C}^{-1}$ is the inverse of the overlap matrix $\underline{C}$, defined in Eq. (1.1). The dynamics of the system is such that a configuration is stable if all spins in it are parallel to their local fields,

$$S_i=\mathrm{sgn}\tilde{h}_i\ , \tag{1.3}$$

$$\tilde{h}_i=\sum_{j=1}^{N}J_{ij}S_j\ . \tag{1.4}$$

It is straightforward to see that

$$\sum_j J_{ij}\xi_j^\mu=\xi_i^\mu \tag{1.5}$$

whereas vectors $\{\phi_i\}$ which are orthogonal to all the patterns are eigenvectors of J with zero eigenvalue. Equation (1.5) implies that

$$\mathrm{sgn}\left[\sum_{j=1}^{N}J_{ij}\xi_j^\mu\right]=\xi_i^\mu\ , \tag{1.6}$$

hence all embedded patterns are stable states. This construction holds for any set of patterns, provided that they are linearly independent. Thus the patterns remain stable for all $\alpha<1$.

Several important questions regarding this model arise. Are the stored patterns stable also to simultaneous flips of many spins? What is the effect of fast stochastic noise (e.g., thermal fluctuations) on the stability of the patterns? What are the sizes of the basins of attraction of the patterns? What is the nature of the other, "spurious" states of the network?

C. Effect of self-coupling

The dynamics of the network of Personnaz *et al.*[5] is not governed by the energy function $-\frac{1}{2}\sum_i\tilde{h}_iS_i$. This is due to the presence of the *self-coupling* term $J_{ii}S_i$ in Eq. (1.4). This term is needed for the validity of the eigenvalue equation, (1.5). The self-coupling restricts severely the size of the basins of attraction of the patterns especially for large α. To see this consider a case where all spins except, say, S_1, are parallel to a pattern $\{\xi_i^\mu\}$. Then Eq. (1.3) for S_1 reduces to

$$S_1=\mathrm{sgn}\left[\sum_{j\,(\neq 1)}J_{ij}\xi_j^\mu+J_{11}S_1\right]$$

$$\simeq\mathrm{sgn}[(1-\alpha)\xi_1^\mu+\alpha S_1]\ . \tag{1.7}$$

We have used here the fact that

$$\sum_i J_{ii}=\sum_{\mu,\nu}(\underline{C}^{-1})_{\mu\nu}\left[\sum_i\xi_i^\mu\xi_i^\nu\right]=N\alpha\ , \tag{1.8}$$

which implies that the average of J_{ii} is

$$\langle\langle J_{ii}\rangle\rangle=\alpha \tag{1.9}$$

and the fluctuations of J_{ii} around this value are small $[O(N^{-1/2})]$ as $N\to\infty$. It is thus clear from Eq. (1.7) that if $\alpha>\frac{1}{2}$, each of the two configurations $S_1=\xi_1^\mu$ and $S_1=-\xi_1^\mu$ is stable. This means that above $\alpha=\frac{1}{2}$ if one starts from a configuration which differs from a memory by only one spin one does not flow to the full memory. Thus, although the memories are stable up to $\alpha=1$ the maximum capacity of the system for providing *associative* memory is

$$\alpha_c=\tfrac{1}{2}\ . \tag{1.10}$$

Similar considerations show that the presence of self-coupling terms reduces significantly the basin of attraction below $\alpha=\frac{1}{2}$. This is confirmed by numerical simulations which will be discussed in Sec. VI.

D. Outline of the paper

In the following section we introduce a modification of the model of Personnaz *et al.* by eliminating the self-coupling terms from the dynamics. This leads to a simple description of the network by an extensive energy function. A qualitative discussion of the model follows. In Sec. III we present the mean-field solution of the model at finite and zero temperature. The details of the mean-field theory which is based on the replica method are delegated to Appendixes B and C. Section IV is devoted to a discussion and presentation of simulation results regarding the basins of attraction of the memories. The role of different dynamic scenarios are also elaborated. Section V presents a summary of the main results.

II. A NONLOCAL MODEL WITH AN ENERGY FUNCTION

A. The model

In this paper we study a slightly modified version of the model of Personnaz *et al.* We consider a network which is governed by an energy function,

$$H=-\tfrac{1}{2}\sum_{i,j}J_{ij}S_iS_j\ , \tag{2.1}$$

where J_{ij} are defined by Eq. (1.2). Both the zero- and finite-temperature properties are considered. The attractors of the network are the local minima of H. They are configurations which satisfy

$$S_i=\mathrm{sgn}h_i\ , \tag{2.2}$$

$$h_i=\sum_{j\,(\neq i)}J_{ij}S_j\ . \tag{2.3}$$

At finite T, the long-time behavior of the system is determined by the Boltzmann distribution of states which gives rise to a free energy,

$$F=-\beta^{-1}\ln\mathrm{Tr}_{\{S_i\}}\exp(-\beta H),\quad \beta=1/T\ . \tag{2.4}$$

Note that the diagonal terms J_{ii} in Eq. (2.1) contribute

only a *constant* term, $-\frac{1}{2}\sum_i J_{ii} = -\frac{1}{2}N\alpha$, to H and do not affect the thermal or dynamic properties of the system. Thus, the relevant synaptic matrix, in the present model, is $J_{ij}(1-\delta_{ij})$. In the thermodynamic limit, the matrix can be replaced by $J_{ij}-\alpha\delta_{ij}$. In this limit the embedded patterns are eigenvectors of the synaptic matrix with an eigenvalue $1-\alpha$ and configurations which are orthogonal to them, are eigenvectors with an eigenvalue $-\alpha$.

In the investigation of the model (2.1) it is useful to define the following decomposition:

$$S_i = \sum_{\mu=1}^{p} a_\mu \xi_i^\mu + \delta S_i \ , \tag{2.5}$$

where $\{\delta S_i\}$ is orthogonal to all the patterns, i.e., $\sum_i \delta S_i \xi_i^\mu = 0$. In general, the coefficients a_μ are different from the more conventional order parameters, i.e., the overlaps m_μ,

$$m_\mu = N^{-1}\sum_i S_i \xi_i^\mu \ . \tag{2.6}$$

They are related by

$$a_\mu = \sum_\nu (\underline{C}^{-1})_{\mu\nu} m_\nu \ , \tag{2.7}$$

where $\underline{C}$ is the overlap matrix (1.1). The role of the order parameters a_μ is manifest through the local fields (2.3) which can be written as

$$\begin{aligned} h_i &= \sum_\mu \xi_i^\mu a_\mu - J_{ii} S_i \simeq \sum_\mu \xi_i^\mu a_\mu - \alpha S_i \\ &= (1-\alpha)\sum_\mu \xi_i^\mu a_\mu - \alpha\,\delta S_i \ . \end{aligned} \tag{2.8}$$

The Hamiltonian (2.1) can be written as

$$H = -\frac{N}{2}\sum_\mu a_\mu m_\mu \ . \tag{2.9}$$

These equations should be compared with the corresponding equations of Hopfield's model[1,3]

$$h_i = \sum_\mu \xi_i^\mu m_\mu - \alpha S_i, \quad H = -\frac{N}{2}\sum_\mu (m_\mu)^2 \ . \tag{2.10}$$

The advantage of Eqs. (2.8) and (2.9) in suppressing the effect of overlaps can be seen by considering the configuration $S_i = \xi_i^1$. Here $m_\mu - \delta_{1\mu} = (\underline{C}^{-1})_{1\mu}$ which is nonzero [of $O(1/\sqrt{N})$] *even* if the patterns are random, whereas

$$a_\mu = \delta_{1\mu} \tag{2.11}$$

regardless of the correlations of the patterns.

B. Global minima of H and the theory of random spaces

The local fields of the configuration $S_i = \xi_i^\mu$ are

$$h_i = \xi_i^\mu (1-J_{ii}) \sim \xi_i^\mu (1-\alpha) \ . \tag{2.12}$$

Despite the factor $1-J_{ii}$ all the patterns are stable, for all $p<N$. This is because

$$0 \le J_{ii} \le 1, \quad i = 1,2,\ldots,N \ , \tag{2.13}$$

as is proved in Appendix A. In fact, the patterns are not only stable to single spin flips but are *global minima* of H. Their energy per spin, H/N, is

$$E = -\tfrac{1}{2} \tag{2.14}$$

which is the lowest value of E. In addition, all spin configurations which are linear combinations of the patterns are also global minima of H, degenerate with the original memories.

The above properties become apparent in the following geometric formulation of H. Let us denote by Δ,

$$\Delta = \left[\sum_i (\delta S_i)^2\right]^{1/2} , \tag{2.15}$$

the Euclidean distance between the (N-dimensional) vector $\{S_i\}$ and the p-dimensional subspace spanned by the patterns. Squaring Eq. (2.5) and summing over i one obtains

$$\Delta^2/N = 1 - \sum_\mu a_\mu m_\mu \ . \tag{2.16}$$

In other words, Δ^2/N is the Hamming distance between a spin configuration and its projection on the space spanned by the patterns. Now Eq. (2.9) can be written as

$$H = -\frac{N}{2} + \tfrac{1}{2}\Delta^2 \ . \tag{2.17}$$

Thus H represents a global cost function which is the Euclidean distance (squared) of a corner of the unit N-dimensional hypercube from the subspace spanned by p randomly chosen corners. The dynamics[9] of the system consists of hopping from a given corner to its neighboring ones along paths which reduce Δ. A corner is a local minimum of H if it is closer to the subspace than all its neighboring corners. It is now trivial that the set of global minima of H consists of the patterns, their inversions, and all other spin configurations which lie in the linear space spanned by the patterns.

From the point of view of memory the linear combinations of the embedded patterns are *spurious* states. At first sight, the fact that these linear combinations are stable and degenerate with the memories might seem to be a serious weakness of the model. This however, is not the case, because the occurrence of states $\{S_i = \pm 1\}_{i=1}^{i=N}$ which are linear combinations of the memories is very rare in large N. In fact, given p, N-dimensional (± 1) vectors $\{\xi_i^\mu\}$ chosen at random, the probability P that the linear subspace spanned by them contains a (± 1) vector, different from the $\pm\{\xi_i^\mu\}$, vanishes as[10]

$$P \sim 4\binom{p}{3}\left(\frac{3}{4}\right)^N, \quad N \to \infty \tag{2.18}$$

for all $p/n < 1 - 7\ln 2/\ln N$. The origin of the result (2.18) is the fact that the dominant contribution to P comes from the probability of linear combinations of *three* of the patterns. In order that $\{\xi_i^1\},\{\xi_i^2\},\{\xi_i^3\}$ span another (± 1) vector the N triplets $(\xi_i^1,\xi_i^2,\xi_i^3)$ must contain only six out of the eight different triplets, yielding the result (2.18).

Indeed, none of the many spurious states that have been

found in the simulations of the model (with $N > 100$) was a linear combination of the memories. This will be discussed in Sec. IV.

III. MEAN-FIELD THEORY (MFT)

A. Finite-p limit

The behavior of the network in the limit of finite number of patterns and $N \to \infty$ is relatively simple. In this case the interactions J_{ij} are of order $1/N$, and therefore the local magnetizations $\langle S_i \rangle$ [$\langle\ \rangle$ denotes thermal average] obey the simple mean-field equations

$$\langle S_i \rangle = \tanh(\beta \langle h_i \rangle) = \tanh\left[\beta \sum_{\mu=1}^{p} \xi_i^\mu \langle a_\mu \rangle\right] . \tag{3.1}$$

See Eq. (2.8). The order parameters $\langle a_\mu \rangle$ are given by the self-consistent equations

$$\langle a_\mu \rangle = \sum_\nu (\underline{C}^{-1})_{\mu\nu} \langle m_\nu \rangle = \sum_\nu (\underline{C}^{-1})_{\mu\nu} \frac{1}{N} \sum_i \xi_i^\nu \langle S_i \rangle . \tag{3.2}$$

The free energy per spin is

$$f = \tfrac{1}{2} \sum_\mu \langle a_\mu \rangle \langle m_\mu \rangle - \beta^{-1} \left\langle\!\!\left\langle \ln 2 \cosh\left[\beta \sum_\mu \xi_i^\mu \langle a_\mu \rangle\right] \right\rangle\!\!\right\rangle$$

where $\langle\!\langle\ \rangle\!\rangle$ denotes average over ξ_i^μ.

The ground state of the free energy is always given by the Mattis states, which are of the type

$$\langle S_i \rangle = \xi_i^1 m, \quad \langle a_\mu \rangle = m \delta_{\mu 1}, \quad m = \tanh(\beta m) , \tag{3.3}$$

regardless of the correlations among the patterns. The nature of other solutions depends on the overlap matrix C. In the specific case of random uncorrelated patterns, $(\underline{C}^{-1})_{\mu\nu} - \delta_{\mu\nu} \sim O(N^{-1/2})$, which can be neglected *if p is finite*. In this case, $m_\mu = a_\mu$ and the present model and Hopfield's[1,3] model are identical. As p increases, the *total* contribution of the off-diagonal elements of $\underline{C}^{-1}$ to Eqs. (3.2) increase and this leads to different behavior of the two models in the limit of large p. This limit is studied in the following.

B. MFT in the finite $\alpha = p/N$ limit

In this paragraph we present the general features of the MFT of the model (2.1) in the limit where α remains finite. We concentrate here on the case of random uncorrelated patterns. The derivation of the MFT is presented in Appendixes B and C. It is based on the replica method and replica symmetry is assumed. A general solution of the MFT may have s macroscopic nonzero linear coefficients $\{\langle a_\nu \rangle\}$ where for specifity ν is assumed to run from 1 to s. In the case of *random* patterns these $\langle a_\nu \rangle$ are identical to the macroscopic overlaps $\langle m_\nu \rangle$ with the corresponding patterns. The rest $\{\langle a_\mu \rangle\}$ with $\mu > s$ are of $O(N^{-1/2})$ and oscillate in sign. This leads to the additional order parameter

$$r \equiv \frac{1}{\alpha} \left\langle\!\!\left\langle \sum_{\mu(>s)}^{N\alpha} \langle a_\mu \rangle^2 \right\rangle\!\!\right\rangle , \tag{3.4}$$

see Appendix C. Finally, the Edwards-Anderson order parameter is defined as usual by

$$q = \langle\!\langle \langle S_i \rangle^2 \rangle\!\rangle . \tag{3.5}$$

It is also useful to define the local susceptibility

$$C = \beta(1-q) . \tag{3.6}$$

In the MFT, the local magnetizations have the following form:

$$\langle S_i \rangle = \tanh\left[\beta J \left[\sqrt{r\alpha} z_i + \sum_{\nu=1}^{s} \langle a_\nu \rangle \xi_i^\nu\right]\right] . \tag{3.7}$$

The macroscopic overlaps generate local fields which are similar to the finite-p case. The coefficients $\langle a_\nu \rangle$ are determined self-consistently by Eq. (3.2), which in the case of random patterns is just $\langle a_\nu \rangle = N^{-1} \sum_i \xi_i^\nu \langle S_i \rangle$. The prefactor J equals

$$J = \frac{1}{2C}\{1 + C - [(1-C)^2 + 4\alpha C]^{1/2}\} , \tag{3.8}$$

where C is defined in Eq. (3.7). Note that J is always smaller than 1, and equals 1 at $\alpha = 0$.

The random uncondensed coefficients $\{a_\mu\}$ generate a random Gaussian local field $J\sqrt{r\alpha} z_i$ where $\langle\!\langle z_i^2 \rangle\!\rangle = 1$ and r is determined by

$$r\alpha = \frac{q - \sum_{\nu=1}^{s} \langle m_\nu \rangle \langle a_\nu \rangle}{1 + C - 2CJ} . \tag{3.9}$$

The free energy per spin is

$$-\beta f = -\tfrac{1}{2}(1-\alpha)\ln(1-\alpha) + \frac{\beta}{2}(1-J) + \tfrac{1}{2}\ln J + \frac{\alpha}{2}\ln(J^{-1}-1) - \frac{J^2 r\alpha\beta C}{2} - \frac{\beta J}{2} \sum_{\nu=1}^{s} \langle a_\nu \rangle \langle m_\nu \rangle + \left\langle\!\!\left\langle \ln 2 \cosh\left[\beta J \left[\sqrt{r\alpha} z + \sum_\nu \langle a_\nu \rangle \xi^\nu\right]\right]\right\rangle\!\!\right\rangle . \tag{3.10}$$

The symbol $\langle\!\langle\ \rangle\!\rangle$ refers here to the average over the discrete distribution of ξ^ν as well as a Gaussian integral over z. In the following paragraphs the various solutions to the MF equations are studied.

C. Retrieval states

The preceding equations have a solution in which only one $\langle a_\nu \rangle$ is nonzero. This solution is the most relevant

one for retrieval of memory and is termed a retrieval state. Assuming $\langle a_\nu \rangle = m\delta_{\nu 1}$ we observe that $r=0$ is a consistent solution. Substituting $r=0$ in Eq. (3.7) yields $\langle S_i \rangle = \xi_i^1 m$, hence

$$q = m^2 \tag{3.11}$$

which guarantees by Eq. (3.9) that indeed $r=0$. In fact, another solution which has one nonzero overlap and $r \neq 0$, exists but is unstable to variation of m. The macroscopic overlap m of the state with the pattern $\{\xi_i^1\}$ is, by Eq. (3.7),

$$m = \langle\langle \xi_i^1 S_i \rangle\rangle = \tanh(\beta J m) , \tag{3.12}$$

where J is given by Eq. (3.8) and $C=\beta(1-m^2)$. This state is very similar to the finite-p Mattis state, Eq. (3.3). In both cases, at $T=0$, $S_i=\xi_i^1$ implying a perfect retrieval of memory without errors for all $\alpha<1$. The zero T limit of f, Eq. (3.9), in the retrieval state, is $-\frac{1}{2}$ as expected, see Eq. (2.14).

The main difference between the finite-p and finite-α cases is the temperature dependence. The phase transition, in the finite-α case is of first order. The temperature $T_M(\alpha)$ at which a solution of Eqs. (3.12) and (3.8) with nonzero m first appears is plotted in Fig. 1. At this temperature the retrieval state is only metastable. It becomes a global minimum of the free energy at a lower temperature, $T_c(\alpha)$, see Fig. 1. Note that the two temperatures in the figure are divided by the factor $1-\alpha$. This is carried out because all the local fields in this state are reduced by a factor of $1-\alpha$, see Eq. (2.12). The discontinuity in m at T_M is shown in Fig. 2 as a function of α. It is evident that the transition is a strong first-order transition already at $\alpha \sim 0.1$.

The general MF equations have been derived for random, uncorrelated patterns, but *the results for the retrieval states are quite general and hold even if the patterns are correlated.* Although an analysis of stability to replica symmetry breaking (RSB) has not yet been calculated, it is most likely that the solution for the retrieval states is stable to RSB, since they are states in which the local fields are uniform (relative to ξ_i^μ). Also, the zero T limit of the results is evidently exact as was discussed in Sec. II.

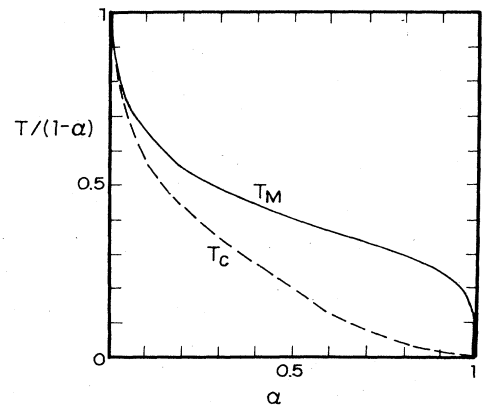

FIG. 1. Mean-field results for the critical temperatures of the retrieval states divided by $1-\alpha$: T_M is the temperature where the states appear, as free energy metastable states; T_c is the temperature where they become global minima.

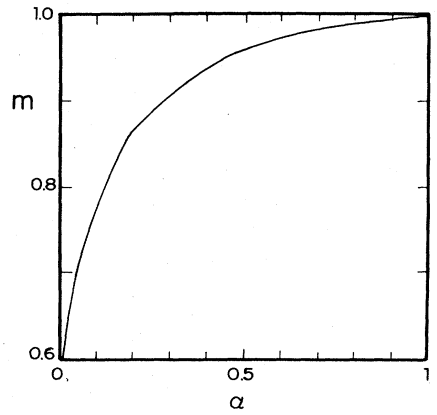

FIG. 2. Calculated value of the overlap m at T_M as a function of α. As $\alpha \to 0$ it approaches zero, indicating that the transition becomes of second order as in the finite-p case.

It should be noted that $T\tanh^{-1}\langle S_i \rangle$ is $Jm\xi_i^1$ whereas the exchange field $\langle h_i \rangle$ for this state is $(1-\alpha)m\xi_i^1$, see Eq. (2.8). Note that J is always less than $1-\alpha$. The physical origin for this difference is that $\langle S_i \rangle$ is induced by the part of h_i which does not include the contribution of S_i itself. In the finite-α case, this "reaction" term is exactly $(1-\alpha-J)m\xi_i^1$, yielding

$$\langle S_i \rangle = \tanh\{\beta[\langle h_i \rangle - (1-\alpha-J)m\xi_i^1]\} = \xi_i^1 \tanh(\beta J m) .$$

This can be shown explicitly[11] using an approach similar to that of Thouless, Anderson, and Palmer[12] for the SK (Ref. 13) model. As $T \to 0$, $C \to 0$ and $J \to 1-\alpha$ as expected.

The principal conclusion from the existence of a phase transition, at a finite T, to a retrieval state is that the embedded patterns are not only stable to single-spin flips but are stable to simultaneous flips of a large number (infinite in the $N \to \infty$ limit) of spins. It implies that for any $\alpha<1$, the patterns are surrounded by "infinite" energy barriers and have substantial basins of attraction. Therefore the predicted capacity of the present network is

$$\alpha_c = 1 . \tag{3.13}$$

D. Noisy mixture states

At sufficiently low values of α, locally stable MF solutions exist which have macroscopic overlaps $\langle a_\nu \rangle$ with several patterns. Since even at $T=0$, $\sum_{\nu=1}^{s} a_\nu m_\nu \leq 1$ and equals 1 only for $s=1$, it is implied by Eq. (3.9) that mixture states are noisy: their local fields include also a random Gaussian part. The origin of this can be observed by noting that

$$\operatorname{sgn}\left[\sum_{\nu=1}^{s} a_\nu \xi_i^\nu\right] = \sum_{\nu=1}^{s} a_\nu \xi_i^\nu + \sum_{\mu\,(>s)} a_\mu \xi_i^\mu + \delta S_i \ ,$$

where $a_\mu = O(N^{-1/2})$ and δS_i is orthogonal to all the patterns. Therefore

$$h_i - (1-\alpha)\sum_{\nu}^{s} a_\nu \xi_i^\nu = (1-\alpha) \sum_{\mu\,(>s)} a_\mu \xi_i^\mu - \alpha\, \delta S_i \neq 0 \ .$$

This noise vanishes as $\alpha \to 0$.

The mixtures states disappear rather quickly when α increases. For instance, at $T=0$, the mixture state with $a_1 = a_2 = a_3 \equiv a$ disappears at $\alpha_3 \simeq 0.1$. The value of a at α_3 is 0.497 compared with $a=0.5$ at $\alpha=0$.. Likewise, a solution with $a_1 = a_2 = \ldots = a_5 = a$ exists only below $\alpha_5 \simeq 0.05$. The value of a at α_5 is 0.374, compared with 0.375 at $\alpha = 0$.

E. Spin-glass (SG) state

The SG solution is characterized by not having any macroscopic overlaps. Equations (3.5)—(3.9) have a solution with $\langle a_\nu \rangle = 0$, $q \neq 0$, below a temperature $T_g(\alpha)$ which is shown in Fig. 3. Unlike the case in Hopfield's model, the SG solution appears discontinuously: both q, r and J exhibit a discontinuity at T_g. The zero-temperature limit of the SG solution is calculated in detail in Appendix B 3. The zero-temperature energy is shown in Fig. 4. As $\alpha \to 0$, $T_g \to 1$, the zero-temperature energy E_g approaches the value $-1/\pi$ and $r\alpha \to 2/\pi$. Thus, in the limit of $\alpha \to 0$ the SG phase coincides with the "high" mixture states of the finite-p case, just as in Hopfield's model.[3] As α increases from zero, T_g decreases very rapidly, $1 - T_g(\alpha) \propto \alpha^{1/4}$ (see Fig. 3). It vanishes at $\alpha = \alpha_g$,

$$\alpha_g = 1 - 2/\pi \sim 0.363 \ . \tag{3.14}$$

As $\alpha \to \alpha_g$ the SG energy becomes degenerate with the ground-state energy,

$$E_g(\alpha) \to -\tfrac{1}{2}, \quad \alpha \to \alpha_g \ . \tag{3.15}$$

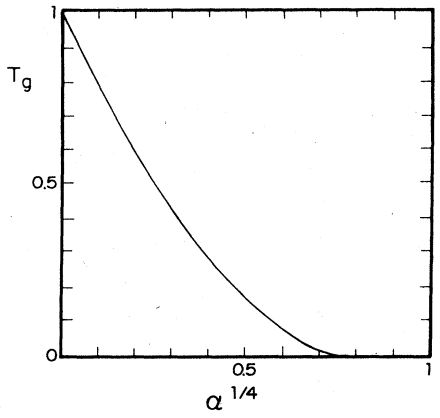

FIG. 3. Results of the MFT for the temperature below which an SG phase appears as a metastable state. Note that T_g vanishes at $\alpha = 1 - 2/\pi$.

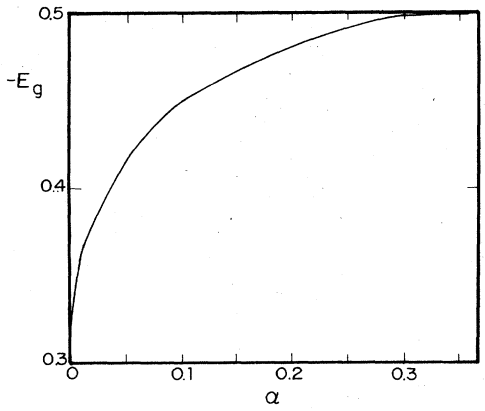

FIG. 4. Zero-temperature energy of the SG state as a function of α, according to the MFT.

F. Absence of an SG solution above α_g

The absence above α_g of solutions other than the retrieval states would suggest that for large value of α, ($\alpha_g < \alpha < 1$) spurious states do not exist or at least that their number is reduced considerably relative to smaller values of α. This would imply a remarkable increase in the basin of attraction of the memories as α increases. Another unusual feature of the above results is the degeneracy of the SG state and the ground state at α_g, see Eq. (3.15). Recalling the geometrical interpretation of the energy, Eq. (2.17), the result (3.15) means that as α increases, the distance between the spin configuration which corresponds to the SG phase and the linear space of the memories decreases until it vanishes completely at α_g. Note that, as $\alpha \to \alpha_g$,

$$\left\langle\!\!\left\langle \sum_{\mu=1}^{N\alpha} a_\mu^2 \right\rangle\!\!\right\rangle = r\alpha \to \frac{1}{1-\alpha_g} \ , \tag{3.16}$$

see Appendix B 3. Equation (3.16) implies that the SG solution *does not* merge, at α_g, with one of the embedded patterns [for which the left-hand side of Eq. (3.16) is necessarily one]. Instead it has small linear coefficients, a_μ of $O(1/\sqrt{N})$, with an infinite number of patterns. This would seem to suggest that at α_g the SG state becomes an exact linear combination of the embedded patterns. This interpretation, however, contradicts the fact [see Eq. (2.18)] that the probability of having even a single linear combination state is exponentially small in the thermodynamic limit, for all $\alpha < 1$. We have to bear in mind that the above replica symmetric SG solution is unstable, and has a negative entropy at $T=0$. It is thus quite possible that in the full mean-field theory including replica symmetry breaking, the value of α_g will be shifted to 1. Another possibility is that at α_g the energy of the SG phase is not exactly $-N/2$ but is higher by an amount which is not proportional to N. This would mean that at α_g the SG state is very close to the subspace spanned by the memories but does not lie in it.

IV. THE BASINS OF ATTRACTION OF THE MEMORIES

A. General discussion

Stability of the learned patterns in a neural network is not sufficient to guarantee the emergence of associative memory. An important requirement is that the stable patterns have a sizable basin of attraction. This ensures the recall of the full memory by an input which contains only partial information on it. The importance of investigating the basins of attraction of the patterns is exemplified by the discussion of Sec. I on the model of Personnaz *et al.*, Eqs. (1.2)—(1.4). This network ceases to provide associative memory above $\alpha=\frac{1}{2}$ because the radius of attraction of the patterns is zero although they are stable for all $\alpha<1$.

It is important to note that the basins of attraction are not isotropic in phase space. In other words, the flow of a state to a pattern is not determined solely by the Hamming distance, $d=1-m$, m being the overlap of the state and the pattern. Firstly, the flow depends also on the proximity of the state to the other stable states, i.e., the spurious states. Secondly, the flow depends in general also on the path that is taken. This means that the basins of attraction are sensitive to the details of the dynamics. For instance, serial and parallel dynamics may define different basins of attraction. The basins may be affected also by order of updating in the serial dynamics.

Given specific dynamic rules, the radius of attraction R of a pattern is defined, in the present work, as the largest Hamming distance within which *almost* all of the states (but not necessarily all of them) flow to the pattern. More specifically, *a state which is chosen at random subject to the constraint that it has an overlap m with the pattern,* will flow to it with a probability Prob(m), where

$$\lim_{N\to\infty}\text{Prob}(m,N)=\begin{cases}1, & m>1-R\\ <1, & m<1-R\ .\end{cases} \tag{4.1}$$

The present study of the basins of attraction is restricted to random uncorrelated patterns.

B. Finite-p limit

Certain aspects of the basins of attractions become considerably simpler in the limit of finite p and $N\to\infty$. In this limit, the present model is equivalent to the Hopfield model, and the only spurious state are the mixture states, as discussed in Sec. III. The closest spurious state to a pattern is the mixture of three patterns, e.g., $S_i=\text{sgn}(\xi_i^1+\xi_i^2+\xi_i^3)$. Its overlap with each of the three patterns is $\frac{1}{2}$. Thus, beyond a radius of $\frac{1}{2}$ some states will not flow to the patterns but to the mixture states. Nevertheless, the (relative) number of these states is negligible. Suppose a state is chosen so that it has an overlap m with a pattern and is random otherwise. Then its overlap with the rest of the patterns is only $O(1/\sqrt{N})$ and its overlap with the mixture state is $m/2$. If m retains a finite value as $N\to\infty$, the state will flow to the pattern with probability which approaches 1 as $N\to\infty$. Thus, for finite p, the radius of attraction of the patterns is, according to the definition (4.1),

$$R=1\ . \tag{4.2}$$

This holds also for the radii of attraction of metastable mixture states. The result (4.2) is relatively insensitive to the details of the dynamics of the system. Of course, most of the states in the phase space lie at distance 1 from the pattern and their flows depend in a complicated fashion on overlaps which are of $O(1/\sqrt{N})$ as well as on the details of the dynamics.

The interesting questions pertaining to the flows of states with initial overlaps which vanish as $N\to\infty$ are beyond the scope of the present work. Here we are mainly concerned with the reduction of R from unity as $\alpha=p/N$ increases.

C. Results of simulations

The radii of attraction of the embedded memories were measured by following the time evolution of states with varying *initial* overlaps. For a given set of p random patterns, an initial state is chosen so that $S_i=\xi_i^1$, $i=1,2,\ldots,mN$, and S_i with $i>mN$ are random (±1). The state evolves according to the dynamic rules (which will be specified below) until a stable state is reached. The process is repeated for different initial states. At a high value of m, states always flow to the patterns ξ^1. As m is reduced, a value m_0 is reached where substantial number of states (or all of them) flow to different fixed points. Averaging m_0 over different sets of patterns yields

$$R=1-\langle\langle m_0\rangle\rangle\ . \tag{4.3}$$

Simulations were performed with sizes $N=100$—400. The main limitations on size comes not from inverting the matrix $C_{\mu\nu}$ but rather from the evaluation of J_{ij}. In the present case, each one of the J_{ij}, Eq. (1.2), requires a time of order α^2N^2. In the Hopfield model,[1] this time is αN. At small values of α, where R is close to unity, finite-size fluctuations are big. In particular, the $O(1/\sqrt{N})$ overlaps of the initial state with the rest of the patterns are not negligible. To partially compensate this effect, we use as a definition of R,

$$R=\left\langle\!\!\left\langle\left[\frac{1-m_0}{1-m_1}\right]\right\rangle\!\!\right\rangle\ , \tag{4.4}$$

where m_1 is the largest overlap with the rest of the patterns, i.e., $m_1=\max\{m_\mu\}$, $\mu>1$. Thus, when the initial prescribed overlap with $\{\xi_i^1\}$ is the same as the random overlap with another pattern (i.e., $m_0=m_1$) the state is at the maximum distance ($R=1$) from $\{\xi_i^1\}$. As $N\to\infty$, $m_1\to0$ and Eq. (4.4) reduces to Eq. (4.3). The correction (4.4) suppresses the finite-size effects especially at small α.

1. *Parallel dynamics*

Parallel dynamics at $T=0$ consists of updating all the spins simultaneously according to

$$S_i(t+1)=\text{sgn}[h_i(t)],\quad i=1,2,\ldots,N\ , \tag{4.5}$$

where $h_i(t)$ is given by Eq. (2.3) with $\{S_i(t)\}$. The dashed line a of Fig. 5 shows the result for the radius of attraction. As $\alpha \to 0$, $R \to 1$ in agreement with the finite-p case. As α increases, R decreases monotonically and vanishes at $\alpha_c = 1$. Inside the basin of attraction the flows to the patterns are fast: starting from the edges of the basin, the typical time is 10—20 steps.

The dashed line b of Fig. 5 represents the numerical results for $R(\alpha)$ with the original model of Personnaz *et al.*,[5] again with parallel dynamics. The drastic effect of the self-coupling terms on R is clearly seen. In this case, R vanishes above $\alpha = \frac{1}{2}$ in agreement with the analysis of Sec. I.

In general, parallel dynamics[2] may lead to attractors which are cycles rather than fixed states. For symmetric connections $J_{ij} = J_{ji}$ such as in the present case, the cycles are at most of length 2 (i.e., the system may oscillate indefinitely between two states).[14,15] Indeed, we have found in our model that in almost all cases where the flows did not end up in the patterns they end up in cycles of length 2. The number of spins that flip during the cycle fluctuates from a few spins in some cases to most of the spins in others. The abundance of spurious cycles may be useful in that the system is able to distinguish through the dynamics between the desired attractors (i.e., the patterns) and the spurious ones.

Modifying the parallel dynamics by introducing memory terms affects strongly the flows of the system. This has been demonstrated by simulating (at $\alpha = 0.6$)

$$S_i(t+1) = \mathrm{sgn}[\tfrac{1}{2}h_i(t) + \tfrac{1}{2}h_i(t-1)] \,, \quad i = 1,2,\ldots,N \,, \tag{4.6}$$

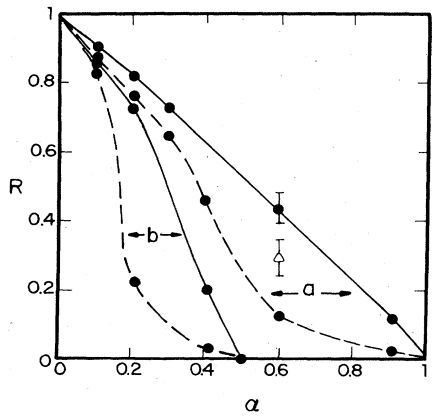

FIG. 5. Measurements of the radius of attraction R, by computer simulations. The lines are guides to the eye. Typical size of the statistical fluctuations is indicated. Lines denoted by a refer to the present model. Lines denoted by b refer to the model of Personnaz *et al.* (Ref. 5) which contains the self-coupling terms. Solid lines refer to serial single-spin-flip dynamics with a specific order of updating as described in the text (Sec. IV). Dashed lines refer to parallel dynamics. Open triangle results from parallel dynamics with memory, Eq. (4.6).

instead of Eq. (4.5). The dynamics of Eq. (4.6) led to a substantial increase in $R(\alpha)$, particularly at high values of α, as shown in Fig. 5. The origin of this improvement is twofold. Firstly, with Eq. (4.6) the only allowed cycles are of length 3 and those are very rare. More importantly, the memory has an effect somewhat similar to annealing. It enhances the influence of the strong ordering fields, produced by the initial overlap with the pattern, at the expense of the weak local noise. This enables the system to escape from shallow spurious valleys.

2. *Serial dynamics*

Simulations with $T = 0$ single-spin-flip dynamics yielded a radius of attraction slightly higher than that of the serial dynamics. This slight enhancement may be attributed to the absence of spurious cycles in the serial dynamics. The fact that the difference is rather small implies that the spurious states and cycles lie mostly in the same regions of phase space.

In the above mentioned single-spin-flip simulations the sequence of spin flips was random, i.e., uncorrelated neither with the particular form of the initial state nor with the nearby pattern. Substantial increase in R is obtained if the order of updating is such that the initial updatings are more likely to increase the overlap with the pattern than to decrease it. The "ideal" order would be to update first the spins which are antiparallel to the pattern. This, however, is impossible since the system does not know in advance where the errors are. However, one may envisage a situation where the system does identify quickly regions where errors are more likely to be. In the context of pattern recognition these regions might be, for instance, the boundaries of the figure.

Recall that in our simulations the *initial state* is $S_i = \xi_i^1$, $1 \le i \le mN$ and the rest are random. Thus the region $mN < i \le N$ represents an area where errors are more likely to occur. We have measured R for a serial dynamics in which the initially "random" spins $\{S_i, mN < i \le N\}$ are flipped before the "parallel" ones $\{S_i, 1 \le i \le mN\}$. The results are presented by the solid line a of Fig. 5. They follow approximately a straight line.

$$R(\alpha) \simeq 1 - \alpha, \quad 0 \le \alpha \le 1 \,. \tag{4.7}$$

A similar increase in R is seen in the model of Personnaz *et al.*,[5] Fig. 5(b). A mixture of serial and parallel dynamics would also substantially increase R (relative to that of the parallel dynamics) as long as the spins which are more likely to be in error (i.e., the random spins) are flipped before the rest of the system.

It should be noted that changes in the details of the dynamics affect the final destiny only if they are made in the first one or two sweeps across the system. Modifying the dynamics of the subsequent steps does not usually change the end of the flow. Finally, in terms of steps per spin the flows to the pattern (inside the basin of attraction) are very fast in the serial dynamics: the typical time of flow from the edges of the basin is 5—7 steps per spin.

D. Properties of the spurious states

The preceding results clearly imply an increase in the number of spurious states as a function of α. This is an apparent contradiction with the prediction that the SG state disappears above $\alpha_g \simeq 0.363$ [see Eq. (3.14)]. Most of the spurious states, obtained from initial states which were at a distance $d \gtrsim R$ from a memory, have a substantial overlap with the memory and hence are not pure SG states. To obtain a more direct comparison with the prediction regarding the SG phase, we have studied the spurious stable states which are obtained at the end flows of *random* initial states. At very small α, the states flow almost always to either one of the memories or to one of the mixture states. Above $\alpha \simeq 0.1$ mixture states do not appear, in agreement with the results of Sec. III C. Instead, most of the spurious states have very little [of $O(1/\sqrt{N})$] overlap with individual patterns. Are these states related to the SG phase? In Fig. 6 we plot the energy of these spurious states as a function of α. At small α, the results are close to the theoretical curve, Fig. 4. At large values of α the measured energy deviates strongly from $E_g(\alpha)$. In particular, the measured energy approaches $-\frac{1}{2}$ only as $\alpha \to 1$ whereas $E_g(\alpha) \to -\frac{1}{2}$ as $\alpha \to \alpha_g$.

The observed spurious states do not exhibit any singular behavior at any value of α between 0 and 1, and certainly do not disappear as α increases. One possible explanation is that the observed spurious states at $T=0$ are local minima which are stable only to a small number of spin flips whereas the mean-field theory refers to macroscopically stable states. Indeed, most of the observed spurious valleys are shallow. Random flips of few spins are sufficient to get out of them. Nevertheless, this does not exclude the possibility that there is an underlying truly metastable SG state. In fact, preliminary finite-T Monte Carlo simulations have indicated that there is a temperature T^* below which a SG state is stable for long times and $T^*(\alpha)$ decreases monotonically to zero as $\alpha \to 1$. This seems to support the suggestion made at the end of Sec. III, namely that the prediction of Eq. (3.14) may be an artifact of the approximation of replica symmetry, and that in the full mean-field theory, the SG state exists (as a metastable state) at all α, with $T_g(\alpha)$ which vanishes at $\alpha = 1$.

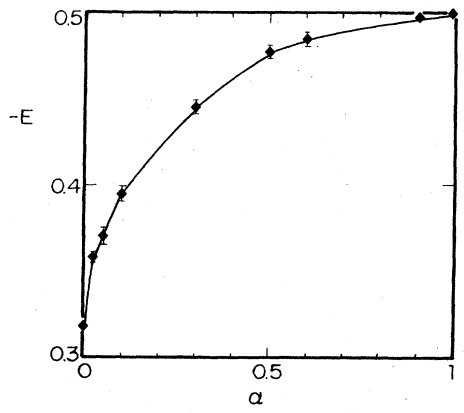

FIG. 6. Measurements of the energies per spin of spurious stable states as a function of α, by simulations at $T=0$. These spurious states which are obtained from initial random states have only small overlap with the patterns.

Finally, we emphasize that none of the numerous spurious states that have been observed has an energy $-\frac{1}{2}$, i.e., is a linear combination of the original patterns. This clearly implies that the occurrence of "linear combination" states is very rare, as expected, and in any case does not affect the basin of attraction of the memories.

V. SUMMARY

In this work we have studied the performance of a neural network which has the ability to retrieve perfectly any p linearly independent patterns for all $p < N$, N being the number of neurons. The model is based on a Hamiltonian version of the network of Personnaz *et al.*[5] The facts that the patterns are retrieved without errors and that correlated patterns can also be stored is a great advantage. The price is that the learning rules are nonlocal which makes the model unattractive from a biological point of view. Nevertheless, the model is interesting not only because of its potential practical applications but also because several general features of neural networks can be more easily studied. In particular, the study of the basin of attraction as well as the roles of the different dynamic rules is more readily investigated in a model where the attractors themselves are unambiguously identified.

The stability of the patterns has been built into the model. The main questions have been the robustness of this stability to thermal fluctuations and the attraction of the patterns. Here again, statistical mechanical methods proved to be very useful. We have shown that below a critical temperature $T_M(\alpha)$, states which are fully correlated with the patterns appear. The states are stable to flips of up to $O(N)$ spins and are separated from each other by infinitely high barriers (in the $N \to \infty$ limit). At $T=0$ they become the ground states of the energy function. The temperature $T_M(\alpha)$ decreases to zero as $\alpha \to 1$, hence the maximum capacity of the network is $\alpha_c = 1$.

In addition to the retrieval states, other locally stable states exist. At finite p, they are mixture states, identical to those in Hopfield's model. At finite α, spurious stable states appear which have zero overlaps with the patterns and have the usual properties of SG states. According to the replica symmetric theory the SG states disappear above $\alpha_g \sim 0.36$ but the numerical evidence is that they exist at all values of $\alpha < 1$. From the point of view of the SG theory the structure of valleys is rather unusual. The ground states are well defined "Mattis states" with no SG features. The energy surface in the immediate neighborhood of each of the $2p$ ground states is smooth even in an energy scale of $O(1)$ as is demonstrated by the $T=0$ simulations. Far from these regions, the energy surface is very rough and is dominated by an enormous SG-like metastability. This is analogous to the situation in most *real glasses* where the metastable glass states are well separated from the crystalline ground states.

The enormous increase in the number of spurious states

as α increases affects the radius of attraction of the memories. This radius has been defined in a statistical sense, namely, the maximum Hamming distance away from a pattern at which the probability of a flow to the pattern is unity in the $N\to\infty$ limit. The radius of attraction $R(\alpha)$ is unity as $\alpha\to 0$ and decreases monotonically to zero as $\alpha\to\alpha_c$. These limits are quite general. The value of $R(\alpha)$ for $0<\alpha<1$ depends on the details of the dynamics. Parallel dynamics is fast but has a relatively small R at large values of α. In addition, almost all of the spurious attractors in a parallel dynamics are cycles of length 2. In a serial dynamics cycles do not exist and the resultant $R(\alpha)$ is higher than the parallel one. A very substantial improvement in R is achieved if the spins which are *known* to *initially* align with the pattern are the *last* ones to flip.

The network considered here is governed by an extensive cost function which is the square of the Euclidean distance from the random linear space spanned by the patterns. All states which are linear combinations of the original patterns are therefore ground states as well. In a linear network, this would have implied the existence of an enormous flat region in the bottom of the energy surface. However, in the case of two-state neurons which are considered here, the linear combinations states are very rare (in a large system). This is another demonstration of the important role of the nonlinearity in neural networks.

Finally, the above results refer to a network whose synaptic matrix does not contain self-coupling terms. Inclusion of the self-coupling terms reduces the radius of attraction and leads to the vanishing of $R(\alpha)$ above $\alpha=\frac{1}{2}$. The reduction in the radius of attraction due to self-coupling terms probably goes beyond the present specific model.

Note Added: After the completion of this work we have received a report of some interesting work by Personnaz, Guyon, and Dreyfus (see Ref. 5) in which they define an energy function identical to our Eq. (2.1), and discuss its geometrical interpretation. However, they consider in detail only parallel dynamics *with* the self-coupling terms. We are grateful to them for sending us their results before publication.

ACKNOWLEDGMENTS

We wish to thank Professor D. J. Amit, Professor H. Gutfreund, and Professor G. Toulouse for interesting discussions. We are grateful to Dr. G. Kalai and Dr. N. Linial for most helpful discussions on random linear spaces. We are particularly indebted to Dr. Andrew Odlyzko for communicating to us the proof of the result (2.18) and for correcting an error which appeared in an earlier version of the paper. We thank AT&T Bell Laboratories for providing the oppurtunity of interacting with Dr. Odlyzko. This work has been supported in part by the Fund of Basic Research administered by the Israeli Academy of Science and Humanities.

APPENDIX A: PROOF OF EQ. (2.13)

From Eqs. (1.1) and (1.2) it follows that

$$\sum_{j=1}^{N} J_{ij}^2 = N^{-2}\sum_{j=1}^{N}\sum_{\mu,\nu=1}^{p}\sum_{\gamma,\delta=1}^{p}\xi_i^\mu\xi_j^\nu\xi_i^\gamma\xi_j^\delta(\underline{C}^{-1})_{\mu\nu}(\underline{C}^{-1})_{\gamma\delta}$$
$$= J_{ii}, \quad i=1,2,\ldots,N . \tag{A1}$$

Therefore

$$J_{ii}-J_{ii}^2=\sum_{j(\neq i)} J_{ij}^2\geq 0, \quad i=1,2,\ldots,N , \tag{A2}$$

which implies that

$$0\leq J_{ii}\leq 1, \quad i=1,2,\ldots,N . \tag{A3}$$

APPENDIX B: REPLICA MEAN-FIELD THEORY

1. Derivation of the mean-field theory (MFT)

The MFT is derived using the replica method. The averaged free energy per spin is given as

$$f=\lim_{n\to 0}\lim_{N\to\infty}\left[-\frac{1}{\beta N n}(\langle\langle Z^n\rangle\rangle-1)\right], \tag{B1}$$

where

$$Z^n=\mathrm{Tr}_{\{S_i^\rho\}}\exp\left[\frac{\beta J}{2N}\sum_{i,j=1}^{N}\sum_{\mu,\nu=1}^{p}(\underline{C}^{-1})_{\mu\nu}\xi_i^\mu\xi_j^\nu\sum_{\rho=1}^{n}S_i^\rho S_j^\rho\right], \tag{B2}$$

$\underline{C}$ is the overlap matrix (1.1), and $\langle\langle\ \rangle\rangle$ stands for averaging over the ξ's. Using a Gaussian transformation yields

$$\langle\langle Z^n\rangle\rangle=\mathrm{Tr}_{S^\rho}\left\langle\!\!\left\langle\int_{-\infty}^{\infty}\prod_{\rho=1}^{n}\prod_{\mu=1}^{p}dx_\mu^\rho\exp\left[-\tfrac{1}{2}\sum_{\rho=1}^{n}\sum_{\mu,\nu=1}^{p}C_{\mu\nu}x_\mu^\rho x_\nu^\rho+(\beta/N)^{1/2}\sum_{i=1}^{N}\sum_{\mu=1}^{p}\sum_{\rho=1}^{n}x_\mu^\rho S_i^\rho\xi_i^\mu\right]\right\rangle\!\!\right\rangle , \tag{B3}$$

where we have neglected the Jacobian $|\det\underline{C}|^{-1/2}$. Substituting the expression (1.1) of $\underline{C}$ in Eq. (B3) and linearizing the resultant quadratic form by a Gaussian transformation, Eq. (B3) reduces to

$$\langle\langle Z^n\rangle\rangle=\left\langle\!\!\left\langle\mathrm{Tr}_{S^\rho}\int\prod_{\rho=1}^{n}\prod_{\mu=1}^{p}dx_\mu^\rho\int_{-i\infty}^{i\infty}\prod_{i=1}^{N}dy_i^\rho\exp\left[+\tfrac{1}{2}\sum_{i=1}^{N}\sum_{\rho=1}^{n}(y_i^\rho)^2+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{p}\sum_{\rho=1}^{n}\xi_i^\mu x_\mu^\rho(y_i^\rho+\sqrt{\beta}S_i^\rho)\right]\right\rangle\!\!\right\rangle . \tag{B4}$$

Following Ref. 3 we divide the patterns into two sets: the first s patterns (with s finite) are macroscopically condensed and the rest $(p-s)$ which are uncondensed. For simplicity we will consider here explicitly only the case $s=1$. The variables ξ_i^1 need not be averaged explicitly since the free energy is self-averaging with respect to them. The rest $p-1$ patterns have to be averaged explicitly leading to a term

$$\sum_{i=1}^{N}\sum_{\mu=2}^{p}\ln\cosh\left[\sum_{\rho=1}^{n}x_\mu^\rho(y_i^\rho+\sqrt{\beta}S_i^\rho)/\sqrt{N}\right].$$

Since these modes are uncondensed, x_μ^ρ is of $O(1)$, and only the terms quadratic in $x_\mu^\rho/\sqrt{N}$ survive. This leads to

$$\langle\langle Z^n\rangle\rangle=\mathrm{Tr}_{S^\rho}\int\prod_{\rho=1}^{n}\prod_{\mu=1}^{p}dx_\mu^\rho\int_{-i\infty}^{i\infty}\prod_{i=1}^{N}\prod_{\rho=1}^{n}dy_i^\rho\exp\left[+\tfrac{1}{2}\sum_{i=1}^{N}\sum_{\rho,\sigma=1}^{n}y_i^\rho y_i^\sigma(\delta_{\rho\sigma}+X_{\rho\sigma})-\sqrt{\beta}\sum_{\rho=1}^{n}\sum_{i=1}^{N}y_i^\rho S_i^\rho\right.$$
$$\left.+\sum_{\rho=1}^{n}m^\rho\sum_{i=1}^{N}\xi_i^1(y_i^\rho+\sqrt{\beta}S_i^\rho)+\tfrac{1}{2}\beta n\right], \tag{B5}$$

where

$$X_{\rho\sigma}=\frac{1}{N}\sum_{\mu=2}^{p}x_\mu^\rho x_\mu^\sigma \tag{B6}$$

and $m^\rho\equiv x_1^\rho/\sqrt{N\beta}$ is the order parameter corresponding to the macroscopic overlap with the pattern $\{\xi_i^1\}$, see below. Integrating $\{y_i^\alpha\}$ yields

$$\langle\langle Z^n\rangle\rangle=\mathrm{Tr}_{S^\rho}\int\prod_{\rho=1}^{n}\prod_{\mu=1}^{p}dx_\mu^\rho\exp\left[-\frac{N}{2}\mathrm{Tr}\ln(\underline{I}+\underline{X})-\frac{\beta N}{2}\sum_{\rho,\sigma=1}^{n}Q_{\rho\sigma}(\underline{I}+\underline{X})^{-1}{}_{\rho\sigma}\right.$$
$$\left.-\frac{N\beta}{2}\sum_{\rho,\sigma=1}^{n}m^\rho m^\sigma(\underline{I}+\underline{X})^{-1}{}_{\rho\sigma}+\beta\sum_{i=1}^{N}\sum_{\rho,\sigma=1}^{n}S_i^\rho\xi_i^1 m^\sigma(\underline{I}+\underline{X})^{-1}{}_{\rho\sigma}+\tfrac{1}{2}\beta n\right], \tag{B7}$$

where

$$Q_{\rho\sigma}=\frac{1}{N}\sum_{i=1}^{N}S_i^\rho S_i^\sigma\,. \tag{B8}$$

Finally, introducing Lagrange multipliers $\hat{X}_{\rho\sigma}$ and $R_{\alpha\beta}$ for the constraints (B6) and (B8), respectively, and integrating $\{X_\mu^\alpha,\mu>1\}$ one obtains for f, Eq. (B1), the following expression:

$$-\beta f=-\tfrac{1}{2}\mathrm{Tr}\ln(\underline{I}+\underline{X})+\frac{\beta}{2}\mathrm{Tr}[\underline{Q}(\underline{I}+\underline{X})^{-1}]+\tfrac{1}{2}\mathrm{Tr}(\underline{\hat{X}}\underline{X})-\frac{\alpha}{2}\mathrm{Tr}\ln\underline{\hat{X}}-\tfrac{1}{2}\alpha\beta^2\mathrm{Tr}(\underline{R}\underline{Q})-\frac{\beta}{2}\sum_{\rho,\sigma=1}^{n}m^\rho m^\sigma(\underline{I}+\underline{X})^{-1}{}_{\rho\sigma}$$
$$+\ln\mathrm{Tr}_{S^\rho}\exp\left[\tfrac{1}{2}\alpha\beta^2\sum_{\rho,\sigma=1}^{n}R_{\rho\sigma}S^\rho S^\sigma+\sum_{\rho,\sigma=1}^{n}\beta m^\rho S^\sigma(\underline{I}+\underline{X})^{-1}{}_{\rho\sigma}\right]+\frac{\beta}{2}-\tfrac{1}{2}(1-\alpha)\ln(1-\alpha)-\frac{\alpha}{2}\,. \tag{B9}$$

The last two constants are $(1/2N)\mathrm{Tr}\ln C$ which comes from the neglected Jacobian of the transformation leading to Eq. (B5). The variable S^ρ in Eq. (B9) is a single-site variable $S_i^\rho\xi_i^1$. Note that the matrix $R_{\rho\sigma}$ is nonzero only for $\rho\neq\sigma$. The order parameters $\underline{Q}$, $\underline{R}$, $\underline{X}$, $\underline{\hat{X}}$, and m are determined by the saddle point equations for $f(\underline{Q},\underline{R},\underline{X},\underline{\hat{X}},m)$.

2. Replica symmetric solution

In the replica symmetric saddle point, the order parameters are

$$Q_{\rho\sigma}=\delta_{\rho\sigma}+q(1-\delta_{\rho\sigma})\,, \tag{B10}$$
$$R_{\rho\sigma}=R(1-\delta_{\rho\sigma})\,, \tag{B11}$$
$$X_{\rho\sigma}=x_0\delta_{\rho\sigma}+x\,, \tag{B12}$$
$$\hat{X}_{\rho\sigma}=\hat{x}_0\delta_{\rho\sigma}+\hat{x}\,, \tag{B13}$$
$$m^\rho=m\,. \tag{B14}$$

Substituting this Ansatz in Eq. (B9), yields at the $n\to0$ limit,

$$-\beta f=-\tfrac{1}{2}\ln(1-x_0)-\frac{x}{2(1+x_0)}-\frac{\alpha}{2}\ln\hat{x}_0-\frac{\alpha}{2}\frac{\hat{x}}{\hat{x}_0}+\tfrac{1}{2}(x_0\hat{x}_0+x\hat{x}_0+\hat{x}x_0)-\frac{\beta}{2}\left[\frac{1}{1+x_0}-\frac{(1-q)x}{(1+x_0)^2}\right]-\frac{\beta}{2}\frac{m^2}{1+x_0}$$

$$+\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\ln 2\cosh\left[\beta\sqrt{R\alpha}z+\frac{\beta m}{1+x_0}\right]+\frac{\beta}{2}-\tfrac{1}{2}(1-\alpha)\ln(1-\alpha)\ . \tag{B15}$$

Differentiating with respect to x_0, x, $\hat{x}_0$, $\hat{x}$, q, R, and m yields after some algebra the following saddle point equations:

$$1+x_0=\frac{1}{2(1-\alpha)}\{1+C+[(1-C)^2+4\alpha C]^{1/2}\}\ , \tag{B16}$$

where $C=\beta(1-q)$,

$$x=\beta(q-m^2)x_0/[1-2\alpha+2x_0(1-\alpha)-C]\ , \tag{B17}$$

$$\hat{x}_0=\alpha/x_0\ , \tag{B18}$$

$$\hat{x}=-\alpha x/x_0^2\ , \tag{B19}$$

$$R=x/\alpha\beta(1+x_0)^2 \tag{B20}$$

$$q=\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\tanh^2\{\beta[\sqrt{R\alpha}z+m/(1+x_0)]\}\ , \tag{B21}$$

$$m=\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\tanh\{\beta[\sqrt{R\alpha}z+m/(1+x_0)]\}\ . \tag{B22}$$

A straightforward generalization of the above equations to the case where s patterns have macroscopic condensations yield Eqs. (3.7)—(3.10), where we have denoted

$$r\equiv R(1+x_0)^2=x/\beta\alpha\ , \tag{B23}$$

$$J\equiv(1+x_0)^{-1}\ . \tag{B24}$$

3. Solutions at $T=0$

At $T=0$ the above MF equations reduce to

$$C=J^{-1}\sqrt{2/\pi r\alpha}\exp\left[-\frac{m^2}{2r\alpha}\right]\ , \tag{B25}$$

$$m=\mathrm{erf}(m/\sqrt{2r\alpha})\ , \tag{B26}$$

$$r=\frac{1-m^2}{2+C-2CJ}\ , \tag{B27}$$

and J is given by Eq. (3.8). Here C is $\lim_{T\to 0}[\beta(1-q)]$. The energy per spin at $T=0$ equals

$$E=-\tfrac{1}{2}(1-J)-\tfrac{1}{2}r\alpha J^2C-\tfrac{1}{2}Jm^2\ . \tag{B28}$$

The retrieval state is given by $m^2=1$, $r=0$, and $E=-\frac{1}{2}$. The SG state is defined by $m=0$. Solution of the above equations with $m=0$ yields

$$C=\frac{1}{(1-\gamma)}\{1-2\alpha+[(1-2\alpha)^2+\gamma-1]^{1/2}\}\ , \tag{B29}$$

$$\gamma\equiv\frac{\pi^2\alpha(1-\alpha)}{2(\pi-2)}\ . \tag{B30}$$

Equation (B29) has a solution with positive C only for

$$0\le\alpha<\alpha_g=1-2/\pi\ . \tag{B31}$$

As $\alpha\to\alpha_g$, γ approaches unity, C diverges as $(4/\pi-1)/(1-\gamma)$, J vanishes as $2/\pi C$, and $r\alpha=2/\pi J^2C^2$ approaches $\pi/2$.

Equations (B23)—(B25) also have a solution with $0<m<1$ and $r\neq 0$. This is true also for finite T. However, calculating explicitly $\chi=\partial m/\partial h$ we find that this solution has a negative χ, hence, is not a minimum (with respect to m). On the other hand, both the above retrieval state and SG solution have a positive χ.

APPENDIX C: INTERPRETATION OF ORDER PARAMETERS

In this Appendix the physical meaning of the order parameters $\chi_{\rho\sigma}$ [Eq. (B6)] is clarified. We introduce external fields conjugate to the variables a_μ, i.e., we add to the replica Hamiltonian a term

$$-\sum_{\rho=1}^{n}\sum_{\nu,\mu=1}^{p}h_\mu^\rho\underline{C}^{-1}{}_{\mu\nu}\sum_{i=1}^{N}\xi_i^\nu S_i^\rho\ . \tag{C1}$$

Differentiating the averaged free energy with respect to h_μ^ρ yields,

$$\frac{\partial f}{\partial h_\mu^\rho}=\langle\!\langle\ \langle a_\mu^\rho\rangle\ \rangle\!\rangle\ , \tag{C2}$$

$$\frac{1}{N}\frac{\partial^2 f}{\partial h_\mu^\rho\,\partial h_\nu^\sigma}=\langle\!\langle\ \langle a_\mu^\rho a_\nu^\sigma\rangle\ \rangle\!\rangle\ . \tag{C3}$$

On the other hand, we can absorb (C1) in $\langle\!\langle Z^n\rangle\!\rangle$ by shifting $X_\mu^\rho\to\bar{X}_\mu^\rho+\tilde{h}_\mu^\rho/\sqrt{N\beta}$ where

$$\tilde{h}_\mu^\rho=\sum_\nu(\underline{C}^{-1})_{\mu\nu}h_\nu^\rho\ .$$

Differentiating $\langle\!\langle Z^n\rangle\!\rangle$ with respect to h_μ^ρ and comparing with Eqs. (C2) and (C3) in the $n\to 0$ limit, one finds

$$\langle\!\langle\ \langle a_\mu^\rho\rangle\ \rangle\!\rangle=\langle\!\langle\ \langle X_\mu^\rho\rangle\ \rangle\!\rangle/\sqrt{\beta N}\ , \tag{C4}$$

$$\langle\!\langle\ \langle a_\mu^\rho a_\nu^\sigma\rangle_c\ \rangle\!\rangle=\langle\!\langle\ \langle x_\mu^\rho x_\nu^\sigma\rangle_c-(\underline{C}^{-1})_{\mu\nu}\delta^{\rho\nu}\rangle\!\rangle/N\beta\ . \tag{C5}$$

Thus, x_μ are the order parameters which measure the coefficients a_μ in the linear expansion of the state $\langle S_i\rangle$, Eq. (2.5). In particular, specializing to the replica symmetric theory [Eqs. (B10)—(B14) and (B21)] one obtains, for a state with no more than one macroscopic overlap,

$$m=a_1=x_1/\sqrt{\beta N}\ , \tag{C6}$$

$$\left\langle\!\left\langle \sum_{\mu=2}^{p} \langle a_\mu\rangle^2 \right\rangle\!\right\rangle = \frac{1}{\beta}\sum_{\mu=2}^{p}\langle x_\mu^\rho x_\mu^\sigma\rangle(1-\delta^{\rho\sigma})=\frac{x}{\beta}=r\alpha\ . \tag{C7}$$

Additional information is obtained if one does not absorb $\tilde{h}_\mu^\rho$ by shifting x_μ^ρ but instead, calculate $\langle\!\langle Z^n\rangle\!\rangle$ in the presence of $\tilde{h}_\mu^\rho$ using the same method as in Appendix B and taking derivatives with respect to $\tilde{h}_\mu^\rho$ at the end. Comparing the result with the equality $\partial^2 f/\partial\tilde{h}_\mu^\rho\partial\tilde{h}_\nu^\sigma = N\langle\!\langle\langle m_\mu^\rho m_\nu^\sigma\rangle\rangle\!\rangle$ one obtains

$$T\left\langle\!\left\langle \sum_{\mu=2}^{p}\langle m_\mu^\rho m_\mu^\sigma\rangle\right\rangle\!\right\rangle = [\underline{Q}(\underline{I}+\underline{X})^{-2}\underline{X}\underline{Q}+\alpha\underline{T}\underline{Q}]_{\rho\sigma}\ . \tag{C8}$$

In particular, in the replica symmetric SG solution, Eq. (C8) reduces to

$$\left\langle\!\left\langle \sum_{\mu=1}^{p}\langle m_\mu\rangle^2\right\rangle\!\right\rangle = 2qCJ^2x_0+TxC^2(2-x_0)J^3+\alpha q\ . \tag{C9}$$

Comparison with the expression for the $T=0$ SG energy [Eq. (B26)] one finds that the right-hand side (rhs) of Eq. (C9) equals at $T=0$ to $-2E_g$. Comparing with Eq. (2.9), we conclude that in the SG state at $T=0$,

$$\left\langle\!\left\langle \sum_{\mu=1}^{p} m_\mu^2\right\rangle\!\right\rangle = \left\langle\!\left\langle \sum_{\mu=1}^{p} a_\mu m_\mu\right\rangle\!\right\rangle = -2E_g\ . \tag{C10}$$

However, $\langle\!\langle\sum_{\mu=1}^{p} a_\mu^2\rangle\!\rangle = r\alpha$ is larger than Eq. (C11). Whether the equality (C10) holds for the SG state beyond the replica symmetric approximation is not known at present.

*Permanent address: Racah Institute of Physics, Hebrew Univeristy, Jerusalem 91904, Israel.

[1]J. J. Hopfield, Proc. Nat. Acad. Sci. USA **79**, 2554 (1982); **81**, 3088 (1984).

[2]P. Peretto, Biol. Cybern. **50**, 51 (1984); in *Disordered Systems and Biological Organization,* edited by E. Bienenstock (Springer-Verlag, Berlin, 1986).

[3]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Phys. Rev. Lett. **55**, 1530 (1985); Ann. Phys. (to be published).

[4]W. Kinzel, Z. Phys. B **60**, 205 (1985); E. Domany, R. Meir, and W. Kinzel (unpublished).

[5]L. Personnaz, I. Guyon, and G. Dreyfus, J. Phys. (Paris) Lett. **46**, L-359 (1985).

[6]J. P. Nadal, G. Toulouse, S. Dehaene, and J. P. Changeux, Europhys. Lett. **1**, 535 (1986). M. Mezard, J. P. Nadal, and G. Toulouse (unpublished); G. Toulouse, S. Dehaene, and J. P. Changeux, Proc. Nat. Acad. Sci. USA (to be published).

[7]G. Parisi (unpublished).

[8]C. Andrea, D. J. Amit, and H. Gufreund (unpublished); H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).

[9]Here we refer to a single-spin-flip dynamics.

[10]A. Odlyzko (unpublished).

[11]H. Sompolinsky (unpublished).

[12]D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).

[13]D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[14]G. Parisi, Phys. Rev. Lett. **50**, 1946 (1983).

[15]A. E. Goles, in *Disordered Systems and Biological Organization,* edited by E. Bienenstock (Springer-Verlag, Berlin, 1986).

[16]A. Frumkin and E. Moses, Phys. Rev. A **34**, 714 (1986).