ANNALS OF PHYSICS 173, 30–67 (1987)

Statistical Mechanics of Neural Networks near Saturation

DANIEL J. AMIT AND HANOCH GUTFREUND

Racah Institute of Physics, Hebrew University, Jerusalem, Israel

AND

H. SOMPOLINSKY*

Bar Ilan University, Ramat Gan, Israel

Received February 3, 1986; revised August 5, 1986

The Hopfield model of a neural network is studied near its saturation, i.e., when the number p of stored patterns increases with the size of the network N, as $p = \alpha N$. The mean-field theory for this system is described in detail. The system possesses, at low α, both a spin-glass phase and $2p$ dynamically stable degenerate ferromagnetic phases. The latter have essentially full macroscopic overlaps with the memorized patterns, and provide effective associative memory, despite the spin-glass features. The network can retrieve patterns, at $T = 0$, with an error of less than 1.5% for $\alpha < \alpha_c = 0.14$. At α_c the ferromagnetic (FM) retrieval states disappear discontinuously. Numerical simulations show that even above α_c the overlaps with the sored patterns are not zero, but the level of error precludes meaningful retrieval. The difference between the statistical mechanics and the simulations is discussed. As α decreases below 0.05 the FM retrieval states become ground states of the system, and for $\alpha < 0.03$ mixture states appear. The level of storage creates noise, akin to temperature at finite p. Replica symmetry breaking is found to be salient in the spin-glass state, but in the retrieval states it appears at extremely low temperatures, and is argued to have a very weak effect. This is corroborated by simulations. The study is extended to survey the phase diagram of the system in the presence of stochastic synaptic noise (temperature), and the effect of external fields (neuronal thresholds) coupled to groups of patterns. It is found that a field coupled to many patterns has a very limited utility in enhancing their learning. Finally, we discuss the robustness of the network to the relaxation of various underlying assumptions, as well as some new trends in the study of neural networks.

1. INTRODUCTION

A. *The Context*

The Hopfield–Little model [1, 2] is a thermodynamic extension of the McCulloch–Pitts [3] program, for the realization of the basic computational functions of a network of neurons. It is based on the schematization of the network as a system of interconnected two state units. The question is whether such a system can perform computations and whether it can serve as associative memory.

* Present address: Racah Institute of Physics, Hebrew University, Jerusalem.

0003-4916/87 $7.50

In a previous study (Ref. [4], to be referred to as I) we have shown that the Hopfield–Little model, extended to include temperature, provides an effective framework for associative memory. In this model one considers a fully interconnected network of N McCulloch–Pitts neurons (Ising spins). The two states of each neuron are represented by $+1$ (active neuron) and -1 (passive). The "world" can present 2^N different patterns to this network: each in the form of an N-bit word, which is nothing but the full enumeration of the possible states of the N neurons, at any particular moment.

The dynamics of the network prescribes a trajectory in the space of the 2^N possible states of the system. At every moment in time the network "goes" through one of the the N-bit "words" it can represent. It is, therefore, natural to identify retrieval, as opposed to accidental transience, as the *persistence of a pattern under the dynamics*. This means that a retrieved (and hence memorized pattern is an attractor.

Guided by Hebb's [5, 1, 2] hypothesis about learning, one locates memory in the synapses, i.e., in the distribution of values of the synaptic efficacies. The synaptic efficacies are mapped onto exchange couplings in the spin system. Since retrieval is related to the persistence under the dynamics, learning can take place *locally*. The changes in each synaptic efficacy depend only on the recurrent states of the pre- and post-synaptic neurons connected by it. The basic goal of the synaptic modifications is to create attractors for subsequent dynamic processes, which are to retrieve the learnt information.

The appearance of attractors connects rather naturally to thermodynamic nonergodicity, such as would appear below a phase transition. Once a ferromagnetic system, for example, is cooled below its Curie temperature it would usually drift rapidly to one of the phases of broken symmetry. The statistical mechanical analogy promises that the nonergodicity will be of a collective, nonlocal, character. Robustness is, therefore, a likely feature.

Finally, undesired persistent patterns (spurious attractors) may also appear, as a result of the strong nonlinearity of the system. They can be suppressed by a modest amount of stochastic inernal noise. This noise is represented by a finite temperature T, where T is a measure of the level of synaptic noise.

In this paper we present the statistical mechanics of Hopfield's model near saturation. It is an extension of I, in which the properties of the network far from saturation have been investigated. Preliminary presentation of some of the results of the present work has been given in Ref. [6].

B. *The Hopfield Model*

Hopfield's model of associative memory consists of a system of N Ising spins, whose dynamics is a heat-bath Monte Carlo process, governed by the Hamiltonian

$$H = -\tfrac{1}{2} \sum_{i \neq j} J_{ij} S_i S_j \tag{1.1}$$

(see, e.g., I). The spin $S_i(=+1, -1)$ represents the two states (active or passive, respectively) of the neuron. The bonds J_{ij} represent the synaptic efficacies between pairs of neurons. The effect of learning is to modify the bonds, so that the learnt patterns become dynamically stable configurations of the network.

Hebb's learning rule suggests an ansatz for the J_{ij}'s [2],

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu. \tag{1.2}$$

Each ξ_i^μ is an independent, quenched, random variable, which takes the values $+1$ and -1 with equal probability. For each μ the pattern $\{\xi_i^\mu\}$ represents a configuration of the network, which has been learnt. The system provides associative memory if these p patterns are indeed dynamically stable configurations, i.e., if the $2p$ configurations $\{S_i=\xi_i^\mu\}$, or $\{S_i=-\xi_i^\mu\}$, are stable states of the system.

The performance of the system depends on several factors:

(1) The sizes of the basins of attraction of the embedded patterns;

(2) The number and the properties of additional, dynamically stable, spurious states;

(3) The storage capacity, i.e., the maximum number of patterns p that can be embedded in the network, without destroying their own stability.

Neural networks conain internal noise, which gives rise to occasional spontaneous activity (or inactivity) of neurons. Well–learnt patterns are those that are stable even in the presence of the noise. We have incorporated this feature into the model (1.1) by considering the statistical mechanics of the system at finite temperature $T\ (=\beta^{-1})$.

In I we have studied this model in the "unsaturated" limit: p remains finite as the size of the system, N, grows to infinity. The main properties of the network, in this limit, are:

(1) At $T=T_c=1$ the system undergoes a second order phase transition, from a disordered phase, in which the dynamics is ergodic, to a phase of broken ergodicity. Below T_c, $2p$ thermodynamically stable, degenerate states appear. They have been referred to as Mattis-states. Each one of these states is correlated macroscopically with a single learnt pattern.

(2) The correlations of the states of the system with the embedded patterns are natural order parameters. They are measured by the overlaps

$$m^\mu=\frac{1}{N}\sum_i\langle S_i\rangle\,\xi_i^\mu \tag{1.3}$$

where $\langle\cdots\rangle$ denotes a thermal average. The μth Mattis-state has $m^\nu=m\delta^{\nu\mu}$, $\langle S_i\rangle=\xi_i^\mu m$, and $m=th(\beta m)$. At $T=0$, $m=1$ and $\langle S_i\rangle=\xi_i^\mu$.

(3) Between $T=1$ and $T=0.46$ the Mattis-states are the only stable states.

Below $T=0.46$ additional dynamically stable "spurious" states appear. These states (the "mixture states") correspond to local fields which are specific linear combinations of several patterns. As T decreases to zero, the numer of mixture states increases. At $T=0$ their number increases (at least) exponentially with p.

(4) At all T, the mixture states are, at best, metastable. Their energies are higher than that of the Mattis-states by an *extensive* amount. Both the Mattis-states and the metastable mixture states are, locally, highly stable: they are surrounded by free energy barriers of order N.

(5) The basins of attraction of the "learnt" patterns are enormous. An initial state, which has an overlap greater than $O(1/\sqrt{N})$ with a single stored pattern, and random overlaps, i.e., of $O(1/\sqrt{N})$, with all the others, will flow very rapidly to the pattern.

C. *The Storage Capacity of the Network*

The perfect stability of the stored patterns for small values of p/N (at $T=0$) is due to the fact that the local field acting on the spins, in the state $S_i=\xi_i^\nu$, is

$$h_i=\sum_{j\neq i} J_{ij}S_j=\frac{1}{N}\sum_{j\neq i}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu\xi_j^\nu\cong\xi_i^\nu(1+\delta_i). \tag{1.4}$$

The noise

$$\delta_i=\frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu}\xi_i^\mu\xi_j^\mu\xi_i^\nu\xi_j^\nu \tag{1.5}$$

is a random variable with variance

$$\overline{(\delta_i)^2}=(p-1)/N,$$

which is negligible in the limit $N\to\infty$ and finite p. When p increases, the noise generated by the random overlaps between the patterns, becomes increasingly important. Eventually, it destroys completely the stability of the stored patterns.

A naive guess, based on Eq. (1.5), would give that stability will be destroyed when p increases linearly with N,

$$p=\alpha N \tag{1.6}$$

since a finite value for α leads to δ_i of order unity.

However, more careful arguments, based on probability theory, lead to the conclusion that already for

$$p>\frac{N}{2\ln N} \tag{1.7}$$

the original patterns become unstable [7, 8] (see, e.g., Sect. 5). On the other hand, Hopfield [1] conclued, on the basis of simulations and Gaussian noise arguments

that the associative memory provided by the system is substantially degraded only when $p > \alpha_c N$, where $\alpha_c \simeq 0.1 - 0.2$.

D. *Outline of the Paper*

In this paper we study the model (1.1), (1.2) in the limit in which $\alpha = p/N$ is finite, when N becomes very large. The point of view adopted here is that the network can be useful as an associative memory, even if the stability of the embedded patterns is no longer guaranteed. While the original patterns may be destabilized, new stable states may develop, very near the original learnt patterns. The stability analysis of the stored patterns is replaced, therefore, by a direct investigation into the nature of the *true* stable states of the system, in the presence of finite internal synaptic noise.

This is a rather delicate task, and it is here that mean-field theory of spin-glasses [9] is very useful. Our version of this theory is formulated in Sect. 2, using the replica method. The general self-consistent mean-field equations are simplified in Section 3, by considering replica symmetric solutions. The properties of the metastable states at $T = 0$ are described in Section 4, and at finite temperature in Section 5. The addition of external sources, marking certain groups of patterns, is considered in Section 6. The features of replica symmetry breaking (RSB), as well as the analog of the Almeida–Thouless line [10] of the Sherrington–Kirkpatrick [9] (SK) model are discussed in Section 7. In Section 8, we present a brief description of computer simulations performed to test the theoretical predictions. Finally, Section 9 comprises a discussion of the robustness of the model and of a few of its emerging extensions.

E. *Summary of Results*

(1) The most important result is that there exists a range of α where the system provides very effective retrieval of memory. This is a consequence of the fact that, despite the existence of a finite amount of noise which destabilizes the stored patterns, the modified stable states remain very near the original patterns, provided

$$\alpha < \alpha_c \simeq 0.14. \tag{1.8}$$

These metastable states which are correlated with only a single memory are termed *retrieval states*. As long as $\alpha < \alpha_c$ the overlap between them and the original pattern is greater than $m \simeq 0.97$, which is the value of the overlap at α_c. As α decreases to zero, m increases (at $T = 0$) exponentially fast towards 1,

$$1 - m \simeq \exp(-1/2\alpha), \qquad \alpha \to 0. \tag{1.9}$$

(2) The maximum storage capacity of the network depends on the precision which one requires from retrieval. If retrieval of a pattern with a small percentage of errors, i.e., spins misaligned with the original patterns, is acceptable then the storage capacity is given by $p = N\alpha_c$.

If one requires retrieval with a vanishing *fraction* of errors as $N \to \infty$, then α must

vanish with increasing N. Finally, Eq. (1.9) implies that retrieval of patterns free of errors occurs when $\alpha < 1/(2 \ln N)$.

(3) Retrieval states exist as thermodynamic, metastable states also at finite T, for $T < T_M(\alpha)$. The maximum temperature $T_M(\alpha)$ decreases from unity at $\alpha = 0$, to zero at $\alpha = \alpha_c$ (see, e.g., Fig. 2). Thus, retrieval of memory is stable against a small amount of fast stochastic noise even when α is finite, but the maximum allowed level of noise decreases to zero as the system approaches its maximum capacity.

(4) At finite α there are two classes of *spurious states*, namely metastable states other than the single pattern retrieval states. At sufficiently small α there are *mixture* states, which have finite overlaps with several patterns. In addition, there is a *spin-glass* (SG) phase, at all finite α. This phase, which has a vanishingly small $(O(1/\sqrt{N\alpha}))$ overlap with the memories, appears at $T < T_g = 1 + \sqrt{\alpha}$ (see, e.g., Fig. 2).

(5) The spin-glass phase in the present model is of the same nature as that in the SK model [9]. Replica symmetry is broken [10] in this phase, signalling the existence of many hierarchically organized spin-glass pure states [11, 12]. Dynamic relaxation in such a phase is expected to be anomalously slow [13]. On the other hand, in the retrieval states, replica symmetry breaking sets in only at $T < T_R$, where $T_R < T_M$ in almost the entire range of $\alpha < \alpha_c$, and

$$T_R \simeq \exp(-1/2\alpha) \qquad \text{as} \quad \alpha \to 0.$$

Even at $T = 0$ the effect of replica symmetry breaking on the retrieval states is expected to be rather weak. Hence, also at finite α, retrieval of memory remains a fast process.

(6) The existence of a critical value $\alpha_c \simeq 0.14$, below which memory retrieval is still possible and rather efficient, has been demonstrated by numerical simulations. A new feature which emerges from the simulations is the *remanent* behavior, at $\alpha > \alpha_c$. Starting from one of the patterns, at $T = 0$, the system flows into a locally stable state, with a small but finite overlap with the pattern. For $\alpha = \alpha_c^+$, the remanent overlap is $m \simeq 0.35$. This value is much smaller than the value $m \simeq 0.97$ at $\alpha = \alpha_c^-$, but is still higher than the overlap with a random initial state, which is $m_R \simeq 0.08$. As $\alpha \to \infty$ both overlaps approach the value of the remanent magnetization (at $T = 0$) of the SK model [9], $m_R \simeq 0.15$ [14]. The sensitivity of the remanent overlap, at $\alpha > \alpha_c$, to finite temperature is expected to be significantly higher than that of the retrieval states at $\alpha < \alpha_c$.

(7) The application of a static magnetic field conjugate to one of the embedded patterns enhances its retrieval. It can be retrieved (with a low fraction of errors) even when $\alpha > 0.14$. However, a field which is conjugate to several patterns increases the internal noise on each of these patterns. Consequently, the enhancement of their retrieval is rather modest, and decreases with the increasing number of "marked" patterns.

2. Mean-Field Theory

We proceed to study the statistical mechanics of the hamiltonian equations (1.1), (1.2), in the limit of finite α. In this case the low temperature phase will have weak random overlaps with most of the patterns of typical magnitude $O(1/\sqrt{N})$. This can be realized from the ground state energy per spin, which is

$$E=\frac{1}{2}\alpha-\frac{1}{2}\sum_{\mu=1}^{p}(m^{\mu})^2, \tag{2.1}$$

where m^{μ} are the overlaps defined in Eq. (1.3). It is, however, possible that one, or a finite number of overlaps condense macroscopically, i.e., that their magnitude remain finite as $N\to\infty$.[1] To take into account this possibility, we introduce external fields, conjugate to a *finite* number of patterns ($\{\xi_i^{\nu}\}$, $\nu=1,\ldots,s$), adding a term

$$H_h=-\sum_{\nu=1}^{s}h^{\nu}\sum_i\xi_i^{\nu}S_i \tag{2.2}$$

to the Hamiltonian. This is, of course, nothing but Bogolyubov's method of quasi-averages [23].

To average over the ξ's we employ the "replica method" [9]. The averaged free energy per spin would be

$$f=\lim_{n\to 0}\lim_{N\to\infty}\frac{-1}{\beta nN}(\langle\langle Z^n\rangle\rangle-1), \tag{2.3}$$

where $\langle\langle\cdots\rangle\rangle$ is the quenched average over the ξ's, and

$$\begin{aligned}\langle\langle Z^n\rangle\rangle=&\left\langle\left\langle \mathrm{Tr}_{S^{\rho}}\exp\left[\frac{\beta}{2N}\sum_{ij\mu\rho}(\xi_i^{\mu}S_i^{\rho})(\xi_j^{\mu}S_j^{\rho})\right.\right.\right.\\ &\left.\left.\left.-\frac{1}{2}\beta pn+\beta\sum_{\nu}h^{\nu}\sum_{i\rho}\xi_i^{\nu}S_i^{\rho}\right]\right\rangle\right\rangle\\ =&\exp(-\beta pn/2)\left\langle\left\langle \mathrm{Tr}_{S^{\rho}}\int\prod_{\mu\rho}(dm_{\rho}^{\mu}/\sqrt{2\pi})\right.\right.\\ &\times\exp\beta N\left[-\frac{1}{2}\sum_{\mu\rho}(m_{\rho}^{\mu})^2+\sum_{\mu\rho}m_{\rho}^{\mu}\frac{1}{N}\sum_i\xi_i^{\mu}S_i^{\rho}\right]\\ &\left.\left.\times\exp\beta N\left[-\frac{1}{2}\sum_{\nu\rho}(m_{\rho}^{\nu})^2+\sum_{\nu\rho}(m_{\rho}^{\nu}+h^{\nu})\frac{1}{N}\sum_i\xi_i^{\nu}S_i^{\rho}\right]\right\rangle\right\rangle\end{aligned} \tag{2.4}$$

[1] If the number of *macroscopically* condensed overlaps is to diverge as $N\to\infty$ so would the energy *per spin*, which is excluded.

in which h^ν are the external fields coupled to projections on the first s patterns, and ρ is the replica index, $\rho = 1,..., n$. The sums $\sum_\nu$ and $\sum_\mu$ are over the first s patterns and over the remaining $p-s$ patterns, respectively.

Averaging over the first s ξ's retains the discrete nature of these random variables, while the averaging over the infinite number of other patterns becomes, eventually, gaussian. This we perform in steps:

(1) The first exponential in (2.4) is averaged over the $p-s$ "high" ξ's, to give

$$\% = \exp\left[-\frac{\beta N}{2}\sum_{\mu\rho}(m_\rho^\mu)^2 + \sum_{i\mu}\ln \operatorname{ch}\beta \sum_\rho m_\rho^\mu S_i^\rho\right]. \tag{2.5}$$

To obtain a well-defined thermodynamic limit we rescale the integration variables

$$m_\rho^\mu \to m_\rho^\mu/\sqrt{N}.$$

The leading terms in (2.5), as $N \to \infty$, are then

$$\% = \exp\beta\left[-\frac{1}{2}\sum_{\mu\rho}(m_\rho^\mu)^2 + \frac{\beta}{2N}\sum_{\rho\sigma\mu i} m_\rho^\mu m_\sigma^\mu S_i^\rho S_i^\sigma\right]. \tag{2.6}$$

(2) The m_ρ^μ's are integrated out leading to

$$\begin{aligned}
&\int \prod_{\mu\rho}(dm_\rho^\mu/\sqrt{2\pi})\% \\
&\quad = \int \prod_{(\rho,\sigma)} dq_{\rho\sigma} \exp\left[-\frac{p}{2}\operatorname{Tr}\ln\left[(1-\beta)\mathbf{I} - \beta\mathbf{q}\right]\right] \\
&\qquad \times \prod_{(\rho,\sigma)} \delta\left(q_{\rho\sigma} - \frac{1}{N}\sum_i S_i^\rho S_i^\sigma\right) \\
&\quad = \int \prod_{(\rho,\sigma)} dr_{\rho\sigma} \prod_{\rho\sigma} dq_{\rho\sigma} \exp\left[-\frac{1}{2}p\operatorname{Tr}\ln\left[1-\beta)\mathbf{I} - \beta\mathbf{q}\right]\right]. \\
&\qquad \times \exp N\left[-\frac{1}{2}\alpha\beta^2\sum_{\rho\sigma} r_{\rho\sigma}q_{\rho\sigma} + \frac{1}{2}\alpha\beta^2 N^{-1}\sum_{i\rho\sigma} r_{\rho\sigma}S_i^\rho S_i^\sigma\right],
\end{aligned} \tag{2.7}$$

where the subscript (ρ, σ) implies $\rho < \sigma$ and $\mathbf{q}$ is the matrix $q_{\rho\sigma}$. Recall that there are only $n(n-1)/2$ independent variables $q_{\rho\sigma}$ and $r_{\rho\sigma}$. The trace refers to the sum over the diagonal replica indices of the matrix.

The physical meaning of the parameters m_ρ^ν, $q_{\rho\sigma}$, and $r_{\rho\sigma}$ is determined from the saddle point equations. One has for the overlaps of a state with a stored pattern

$$m_\rho^\nu = \frac{1}{N}\left\langle\!\!\left\langle \sum_i \xi_i^\nu \langle S_i^\rho\rangle \right\rangle\!\!\right\rangle. \tag{2.8}$$

The Edwards–Anderson order parameter [15]

$$q_{\rho\sigma} = \left\langle\!\!\left\langle N^{-1} \sum_i \langle S_i^\rho \rangle \langle S_i^\sigma \rangle \right\rangle\!\!\right\rangle. \tag{2.9}$$

The Lagrange multiplier $r_{\rho\sigma}$ $(\rho \neq \sigma)$ is

$$\begin{aligned} r_{\rho\sigma} &= \alpha^{-1} \sum_{\mu > s} \left\langle\!\!\left\langle \left[N^{-1} \sum_i \xi_i^\mu \langle S_i^\rho \rangle \right] \left[N^{-1} \sum_i \xi_i^\mu \langle S_i^\sigma \rangle \right] \right\rangle\!\!\right\rangle \\ &= \alpha^{-1} \sum_{\mu = s+1}^{\alpha N} \langle\!\langle m_\rho^\mu m_\sigma^\mu \rangle\!\rangle \end{aligned} \tag{2.10}$$

which in the replica symmetric theory (see Sect. 3.), is identified as the mean square, random overlap of a configuration of spins with the "high" patterns.

(3) Inserting (2.6) and (2.7) in (2.4) and taking account of the fact that the quenched averaging over the finite number of ξ^ν's can be effected by self averaging (see, e.g., I, Sect. A), one finds

$$\begin{aligned} \langle\!\langle Z^n \rangle\!\rangle = \exp(-\beta p n/2) \int \prod_\nu dm_\rho^\nu \int \prod_{(\rho,\sigma)} dq_{\rho\sigma} dr_{\rho\sigma} \\ \times \exp N \Bigg[-\frac{1}{2} \beta \sum_{\nu\rho} (m_\rho^\nu)^2 \\ -\frac{1}{2} \alpha \operatorname{Tr} \ln \left[(1-\beta) \mathbf{I} - \beta \mathbf{q} \right] - \frac{1}{2} \alpha \beta^2 \sum_{\rho \neq \sigma} r_{\rho\sigma} q_{\rho\sigma} \\ + \left\langle\!\!\left\langle \ln \operatorname{Tr}_{S^\rho} \exp \left[\frac{1}{2} \alpha \beta^2 \sum_{\rho \neq \sigma} r_{\rho\sigma} S^\rho S^\sigma \right. \right.\right. \\ \left.\left.\left. + \beta \sum_{\nu\rho} (m_\rho^\nu + h^\nu) \xi^\nu S^\rho \right] \right\rangle\!\!\right\rangle_\xi \Bigg], \end{aligned} \tag{2.11}$$

where $\langle\!\langle \cdots \rangle\!\rangle_\xi$ denotes the average over the patterns $\{\xi_i^\nu\}$. As $N \to \infty$, the integrand in Eq. (2.11) is dominated by its saddle point, leading to the following averaged free energy per spin,

$$\begin{aligned} f = \tfrac{1}{2}\alpha + (\alpha/2\beta n) \operatorname{Tr} \ln[(1-\beta)\mathbf{I} - \beta \mathbf{q}] + (1/2n) \sum_{\nu\rho} (m_\rho^\nu)^2 \\ + (\alpha\beta/2n) \sum_{\rho \neq \alpha} r_{\rho\sigma} q_{\rho\sigma} - (1/n\beta) \langle\!\langle \ln \operatorname{Tr}_S \exp(\beta H_\xi) \rangle\!\rangle_\xi \end{aligned} \tag{2.12}$$

in the limit $n \to 0$. The operator H_ξ is given by

$$H_\xi = (\alpha\beta/2) \sum_{\rho \neq \sigma} r_{\rho\sigma} S^\rho S^\sigma + \sum_{\nu\rho} (m_\rho^\nu + h^\nu) \xi^\nu S^\rho. \tag{2.13}$$

Equations (2.12), (2.13) give a complete description of the stable states of the system, in the limit $N \to \infty$. This mean-field theory, is exact because the network is fully connected, as in the SK model [9]. The stationary states are obtained by varying f with respect to the order parameters m_ρ^ν, $q_{\rho\sigma}$, $r_{\rho\sigma}$.

It should be remarked, in concluding this section, that all the free energies described by this theory are of $O(N)$, including the barriers between the various stable saddle points.

3. Replica Symmetric Theory

Most of our discussion will take place within the replica symmetric theory. The breaking of replica symmetry, while important in principle, will turn out to have only a marginal effect on the bulk of the results. The restriction of the free energy, (2.12), to the symmetric phase is affected by the choices

$$m_\rho^\nu = m^\nu, \tag{3.1a}$$

$$q_{\rho\sigma} = q, \qquad \rho \neq \sigma, \tag{3.1b}$$

$$r_{\rho\sigma} = r, \qquad \rho \neq \sigma. \tag{3.1c}$$

These are substituted in (2.12) and (2.13) and the limit $n \to 0$ taken. The result is

$$\begin{aligned} f = {} & \frac{1}{2}\alpha + \frac{1}{2}\sum_\nu (m^\nu)^2 + \frac{\alpha}{2\beta}\left[\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right] \\ & + (\alpha\beta r/2)(1-q) - \beta^{-1}\int (dz/\sqrt{2\pi})\exp(-z^2/2) \\ & \times \left\langle\!\!\left\langle \ln 2ch\beta\left[\sqrt{\alpha r}z + \sum_\nu (m^\nu + h^\nu)\,\xi^\nu\right]\right\rangle\!\!\right\rangle_\xi . \end{aligned} \tag{3.2}$$

The details of the limiting process are left to Appendix A.

Variation of f, Eq. (3.2), with respect to m^ν, q, and r, leads to the equations for the stationary states. All the solutions, which are local minima of f, are stationary states of the dynamical process, with barriers of $O(N)$.

The equations are

$$m^\nu = \langle\!\langle \xi^\nu th\beta[\sqrt{\alpha r}z + (\vec{m}+\vec{h})\cdot\vec{\xi}]\rangle\!\rangle, \tag{3.3}$$

$$q = \langle\!\langle th^2\beta[\sqrt{\alpha r}z + (\vec{m}+\vec{h})\cdot\vec{\xi}]\rangle\!\rangle, \tag{3.4}$$

$$r = q/(1-\beta+\beta q)^2 \tag{3.5}$$

The vectors $\vec{m}$, $\vec{h}$, and $\vec{\xi}$ have s components each, corresponding to the condensed patterns. The average $\langle\!\langle \cdots \rangle\!\rangle$ refers to the combined average over the discrete ξ^ν's and over the gaussian noise z.

Note that the local field consists of two parts: a feromagnetic part $\vec{m}\cdot\vec{\xi}$, resulting from the s condensed overlaps, and a spin-glass part $\sqrt{r\alpha}z$, generated by the random overlaps with the rest of the patterns.

In concluding this section we remark that at the saddle points the values of the parameters m^ν, q, and r are just the quantities defined in Eqs. (2.8), (2.9), and (2.10), respectively.

4. Magnetic and Spin-Glass Ordering at $T=0$

To approach $T=0$ in the replica symmetric theory we start by noting that

$$\begin{aligned}\int (dz/\sqrt{2\pi})\exp(-z^2/2)\, th\beta(\sqrt{\alpha r}z+x)&\\ =\sqrt{2/\pi}\int_0^{x/\sqrt{\alpha r}} dz\exp(-z^2/2)+O(T)&\\ \equiv \operatorname{erf}(x/\sqrt{2\alpha r})+O(T).&\end{aligned}\tag{4.1}$$

Inserting the leading term in Eq. (3.3) leads to

$$m^\nu = \langle\!\langle \xi^\nu \operatorname{erf}[(\vec{m}+\vec{h})\cdot\vec{\xi}/\sqrt{2\alpha r}]\rangle\!\rangle_\xi,\tag{4.2}$$

where the average is over the discrete distribution of ξ^ν, $\nu=1,\ldots,s$.

Even at this early stage it is tempting to compare Eq. (4.2) with Eq. (2.7) of I, which reads

$$\vec{m} = \langle\!\langle \vec{\xi} th(\beta\vec{m}\cdot\vec{\xi})\rangle\!\rangle_\xi\tag{4.3}$$

for the condensed overlaps, in a system with a finite number of patterns at finite temperature. In fact, the error function is quite similar to the hyperbolic tangent. Hence, the zero temperature magnetization for finite α behaves as if the system were at a finite temperature $\sqrt{2\alpha r}$, provided r is finite for finite m. Namely, the random overlaps with the noncondensed patterns act as gaussian noise, which introduces errors even at $T=0$. As $\alpha\to 0$ (4.2) tends to the finite p result,

$$\vec{m} = \langle\!\langle \vec{\xi}\operatorname{sgn}[(\vec{m}+\vec{h})\cdot\vec{\xi}]\rangle\!\rangle,$$

provided $\alpha r\to 0$ as $\alpha\to 0$, when $m\neq 0$.

Equation (3.4) yields in the limit $T\to 0$,

$$C\equiv\beta(1-q)\to\sqrt{2/\pi\alpha r}\langle\!\langle\exp\{-[(\vec{m}+\vec{h})\cdot\vec{\xi}]^2/2\alpha r\}\rangle\!\rangle.\tag{4.4}$$

A. *The Retrieval States*

States with $\vec{m}\neq 0$ are termed ferromagnetic (FM) states. Those with a single nonvanishing overlap, e.g., $m^\nu = m\delta^{\nu,1}$, were referred to as Mattis-states in I. Here they

are named "retrieval" states. They are Mattis-like, but in the case of finite α, m is less than 1 even at $T=0$. For such a state, Eq. (4.2) reads

$$m = \mathrm{erf}(m/\sqrt{2\alpha r}) \tag{4.5}$$

and C of Eq. (4.4) is given by

$$C = \sqrt{2/\pi\alpha r}\, \exp(-m^2/2\alpha r). \tag{4.6}$$

Substituting in Eq. (3.5) one has

$$r = (1-C)^{-2}. \tag{4.7}$$

The three equations (4.5), (4.6), and (4.7) can be reduced to one equation, for the variable $y = m/\sqrt{2\alpha r}$

$$y = \mathrm{erf}(y)/[\sqrt{2\alpha} + (2/\sqrt{\pi})\, e^{-y^2}]. \tag{4.8}$$

Equation (4.8) always has a solution $y = m = 0$. This is a spin-glass (SG) solution, with no macroscopic overlaps with any of the patterns. It will be discussed in the next subsection. For $\alpha > \alpha_c = 0.138$ this is the only solution.

For $\alpha < \alpha_c$ FM solutions $-m \neq 0 -$ of Eq. (4.8) appear. There are $2p$ such solutions. They appear with an overlap which already at α_c takes the value

$$m = 0.967$$

The dependence of $(1-m)/2$ on α is depicted in Fig. 1.

These (FM) solutions have macroscopic overlaps with the embedded patterns which are very close to unity, allowing retrieval. It should be emphasized that, while such a state has a mocroscopic projection on a given pattern, it will have vanishingly small overlaps with each of the other patterns. Those random overlaps will be of $O(1/\sqrt{N})$. Thus, to the extent that these are dynamically stable states, the identification of the retrieved pattern is umambiguous.

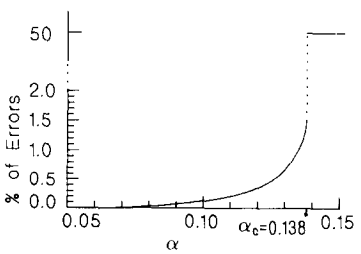

FIG. 1. Average percentage of errors, $N_e/N = (1-m)/2$, in the retrieval states, as a function of α at $T=0$.

The average energy per spin in a configuration $\{S_i\}$, which has an overlap m per spin with one of the ξ's is

$$\begin{aligned} E &= \langle\langle -(1/2N^2)\sum_{\mu}\sum_{ij}\xi_i^{\mu}\xi_j^{\mu}S_iS_j\rangle\rangle + \frac{\alpha}{2} \\ &= -\frac{m^2}{2} + \frac{\alpha}{2}(1-r), \end{aligned} \tag{4.9}$$

where use has been made of Eq. (2.9), as the definition of r in a replica symmetric solution. This expression can also be derived as the $T \to 0$ limit of Eq. (3.2).

At $\alpha = \alpha_c \cong 0.138$, $E \cong -0.5014$. As $\alpha \to 0$, for finite m, it follows from (4.6) and (4.7) that $r \to 1$, and hence $E \to -0.5$. In other words, for finite α the system discovers that it pays slightly in energy to relax a small fraction of the spins, from full alignment with a pattern, and to increase r somewhat, to accommodate the random correlations with the other patterns.

At $T=0$ $q=1 \neq m^2$. This difference—$(q-m^2)$—vanishes as $\alpha \to 0$, and the retrieval states become *bona fide* Mattis-states.

B. *The* SG *solution*

In the SG state ($y=0$), one finds from Eqs. (4.6), (4.7),

$$r = (1+\sqrt{2/\pi\alpha})^2. \tag{4.10}$$

Substitution in Eq. (4.9) gives for the energy of this state

$$E_{\mathrm{SG}} = -\frac{1}{\pi} - \sqrt{2\alpha/\pi}. \tag{4.11}$$

At $\alpha = \alpha_c$, the SG state is the global minimum of the energy, with $E_{\mathrm{SG}} = -0.615$. As α decreases, E_{SG} increases and at $\alpha_M = 0.051$ it exceeds the energy of the retrieval solution. Below α_M the retrieval state is the ground state of the system.

As $\alpha \to 0$, $E_{\mathrm{SG}} \to -1/\pi$. Also, from (4.10) and (4.6) it follows that $r \to 2\pi/\alpha$ and $C = \sqrt{(2/\pi\alpha r)} \to 1$. In this limit the SG solution resembles the symmetric mixed states *for finite* p, when p increases to infinity and the number of nonzero components is of the order of p (see I, Sect. II). In particular, in a symmetric solution with p components, each $m^{\mu} = \sqrt{2/\pi p}$, which vanishes as $p \to \infty$, yet

$$r = \alpha^{-1}\sum_{\mu=1}^{p}(m^{\mu})^2 = 2/\pi\alpha,$$

which is exactly the value of r in the SG state for small α. Moreover, as $p \to \infty$ before $T \to 0$, the value of C $(=\beta(1-q))$ in the symmetric mixed state approaches unity, as in the SG solution.

The similarity of the SG state for small α and the mixed states for finite, large p indicates that the former develops from the "melting" of the mixture states as the

number of mixed patterns increases to infinity. This "melting" of the metastable states is due to the fact that the energy differences per spin, as well as the barriers separating two finite-p states, vanish like $1/p$ as $p \to \infty$. Such barriers are given by the energy difference per spin between a state with p(odd) nonzero m's and one with $p-1$ (even) nonzero ones. In the notation of I (see I, Eqs. (2.33) and (2.34)), the energy *per spin* of a state with p nonzero m's is f_p and, for large p,

$$|f_p - f_{p-1}| \simeq 1/\pi p.$$

Hence, the energy barriers are of $O(1)$, or rather of $O(\alpha^{-1})$. When $\alpha = O(N^{-1})$ (p strictly finite) the energy barriers are of $O(N)$ again.

C. *Storage Capacity at* $T=0$

The fact that an FM, retrieval state appears for finite α (< 0.14), implies that the system can serve as an associative memory, even if p is $O(N)$. At first glance this appears to be in contradiction with probabilistic estimates, which put the storage capacity at $N/(2 \ln N)$ [7] or $N/(4 \ln N)$ [8]. As was already mentioned in the Introduction, this discrepancy is only apparent.

To observe this in detail note that if the overlap of a state with a given pattern is m, then the average number of errors is given by

$$N_e = \frac{N}{2}(1-m). \tag{4.12}$$

This is the number of spins which do not align with the embedded patterns. It is seen from Fig. 1 that N_e/N starts at about 1.5% and tends to zero rapidly (in fact, exponentially) as α decreases. To estimate this decrease we return to Eq. (4.5). In the region of α where an FM retrieval state exists, $m \simeq 1$. As $\alpha \to 0$, the argument on the r.h.s. of (4.3) becomes very large and, using the asymptotic behavior of the error function

$$\operatorname{erf}(x) \simeq 1 - (\sqrt{\pi}x)^{-1} e^{-x^2}$$

and the fact that for finite m and $\alpha \to 0$, $r \to 1$ one has

$$N_e/N \simeq \sqrt{\alpha/2\pi}\, e^{-(1/2\alpha)}. \tag{4.13}$$

For a finite value of α, N_e/N remains finite, and becomes vanishingly small as $\alpha \to 0$. Hence, if one allows a small finite *fraction* of missaligned spins, one can retrieve patterns even if $p = \alpha N$ ($\alpha < 0.14$). To recapitulate, despite this finite, small fraction of misaligned spins the system still permits effective retrieval, since the overlaps with *all* other patterns are vanishingly small, as $N \to \infty$.

On the other hand, if one requires, as in refs. [7, 8] that N_e itself be fixed and small (or zero), as $N \to \infty$, then α must vanish with increasing N. Taking,

$$\alpha = (2x \ln N)^{-1}$$

one finds, from Eq. (4.13), that

$$N_c \simeq N^{1-x}.$$

Hence, $x=1$, or $\alpha=(2\ln N)^{-1}$, is the borderline, in aggreement with ref. [7]. Finally, if we demand that entire sets of p patterns be reproduced without error, with finite probability, then we must have

$$pN_c \simeq O(1)$$

hence $x \geqslant 2$, which implies $p \leqslant N/(4\ln N)$, as in ref. [8].

While there are no mean-field solutions with finite m for $\alpha \geqslant 0.14$, there may be states which are stable against single spin flips and which have finite overlaps with the stored patterns. For finite p, any state which is stable against single spin-flips is also a metastable solution of the mean-field equations [4]. As $p \to \infty$ this is no longer the case. This issue will be discussed in the context of simulations, in Section 8.

D. *Mixture States*

Next, consider symmetric solutions with more than one nonvanishing component of $\vec{m}$. In other words, we look for extrema of the free energy for which overlaps with several patterns have condensed with equal amplitude. In the notation of I, $m_i = \pm m_n$ for $i=1,\ldots,n$ and $m_i=0$ for $i>n$. Equations (4.2) and (4.4), with $\vec{h}=0$, can be written as

$$nm_n = \langle\langle z_n \operatorname{erf}(z_n m_n/\sqrt{2\alpha r})\rangle\rangle, \tag{4.14}$$

$$C = \sqrt{2/\pi\alpha r}\,\langle\langle e^{-z_n^2/2\alpha r}\rangle\rangle, \tag{4.15}$$

where, as in I,

$$z_n = \sum_{\mu=1}^{n} \xi^{\mu}. \tag{4.16}$$

Defining $y_n \equiv m_n/\sqrt{2\alpha r}$, Eq. (4.14) together with Eq. (4.7) can again be reduced to a single equation, which reads

$$ny_n = \langle\langle z_n erf(z_n y_n)\rangle\rangle/[\sqrt{2\alpha} + (2/\sqrt{\pi})\langle\langle \exp(-z_n^2 y_n^2)\rangle\rangle]. \tag{4.17}$$

For any given value of n, this equation is only slightly more complicated than (4.8). As n increases one finds critical values α_n of α, above which $y_n=0$ is the only solution of (4.17). Just below α_n a symmetric solution with $m_n \neq 0$ appears *discontinuously*. Only the solutions with odd values of n are locally stable. Their degeneracy is the same as for finite p.

Thus, for example, $\alpha_3 \approx 0.03$, and $m_3(\alpha_3) \approx 0.496$, to be compared with $m_3=0.5$ for finite p (see, e.g., I). These (odd n) symmetric solutions evolve from the zero temperature, metastable, odd n states of finite p. They remain metastable as α

increases, until they disappear. The value of α_n decreases with n. For large n the binomial probability distribution of z_n approaches a Gaussian distribution. Equation (4.7) with z_n distributed according to a Gaussian has nonzero solutions only for $\alpha = 0$, and considering the deviations from the Gaussian one finds

$$\alpha_n \simeq \frac{1}{n}.$$

In other words, the noise created by the random overlaps with the uncondensed patterns acts, like temperature, to smooth the free energy surface, by successively eliminating metastable mixture states: states with high n are eliminated first.

As for finite p, all symmetric mixture states with an even number of condensed overlaps are unstable. As p increases, the first symmetric states to be destablized are the ones with odd n, nearest to p. These states evolve into the SG state. As α increases, more and more of the symmetric mixture states join the basin of attraction of the SG state, until at $\alpha = 0.138$ even the Mattis-like states merge into this basin.

Nonsymmetric mixture states appear as well. We have not studied them in any detail, but for values of α above which the $n = 3$ states are unstable, no mixture states are stable.

5. The Network at Finite Temperature

A. *The* SG *Phase*

The transition from the disordered, paramagnetic phase to the SG phase is of second order. To find the transition temperature T_g, we expand Eqs. (3.4) and (3.5), with $\vec{m} = \vec{h} = 0$, in powers of q and r. The transition temperature T_g is determined by the leading order

$$q \simeq \beta^2 \alpha r \simeq \frac{\beta^2 \alpha q}{(1-\beta)^2} + O(q^2) \tag{5.1}$$

which yields

$$T_g = 1 + \sqrt{\alpha}. \tag{5.2}$$

For $T < T_g$,

$$q \simeq \beta^2 \alpha r \simeq T_g - T. \tag{5.3}$$

The SG phase is stable to long range FM order ($\vec{m} \neq 0$) at all T. To see this, we compute the FM susceptibility from Eq. (3.3), and find,

$$\chi^{\mu\nu} = \frac{\partial m^\mu}{\partial h^\nu} = \frac{C}{1-C}\delta^{\mu\nu}, \tag{5.4}$$

where $C=\beta(1-q)$ (Eq. (4.4)). Explicit calculation gives, for $\alpha>0$, $C<1$, for all T, and hence χ remains positive and bounded. Nevertheless, FM retrieval states do appear at low T. They appear, like at $T=0$, via a first-order transition, as will be discussed below.

B. *The Retrieval States*

Retrieval states are described by

$$m^{\mu}=m\delta^{\mu\nu}.$$

In that case Eqs. (3.3), (3.4) reduce (for $\vec{h}=0$) to

$$m=\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} th[\beta(\sqrt{\alpha r}z+m)], \tag{5.5}$$

$$q=\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} th^2[\beta(\sqrt{\alpha r}z+m)], \tag{5.6}$$

and r is given by Eq. (3.5). These equations are solved numerically, and where multiple solutions appear, their free energies, Eq. (3.2), are compared. This yields the phase diagram depicted in Fig. 2.

Above $T=1$ there are no FM solutions, for any value of α. Below T_g, Eq. (5.3), one enters a SG phase. For $\alpha<\alpha_c\simeq 0.14$, one finds the line $T_M(\alpha)$, below which the retireval solutions appear, as locally stable states. When they appear they have a well-developed magnetization (macroscopic overlap) with the patterns, as do metastable states in a φ^6-theory [16]. It should be emphasized that the appearance of these metastable states implies a discontinuous change in the dynamical behavior of the system.

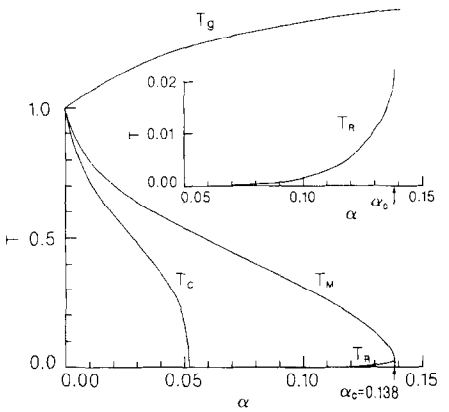

FIG. 2. Plots of critical temperatures of the *SG* and the retrieval *FM* states as a function of α. T_g is the transition temperature to the *SG* state, T_M is the temperature at which *FM* states first appear. T_c is the first order transition at which these states become global minima. Replica symmetry is broken below T_R, which is displayed on an expanded scale in the inset.

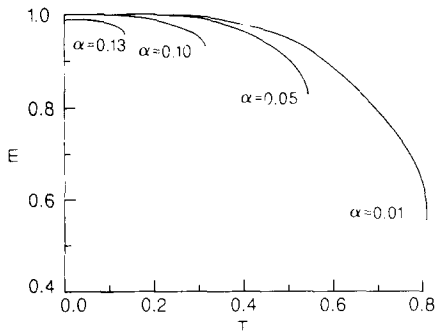

FIG. 3. Magnetization curves of the *FM* retrieval states as a function of T, for several values of α.

As α approaches α_c the line T_M approaches the $T=0$ axis continuously. The low temperature expansion of Eqs. (5.5) and (5.6) shows that T_M vanishes linearly, according to

$$T_M \simeq \frac{1}{C_0 \alpha_c}(\alpha_c - \alpha), \tag{5.7}$$

where $C_0 \simeq 0.18$ is the value of C, defined in Eq. (4.6), at α_c. The slope of T_M near α_c is very steep, about 40.[2]

For $\alpha < 0.051$, as T is lowered, a first-order phase transition takes place, from the SG phase to the FM retrieval phase, reminiscent again of the φ^6-theory. The transition temperature $T_c(\alpha)$ is determined by equating the free energies of the SG and the FM phases. Below this line the retrieval phases are globally stable, down to $T=0$.

Curves of the magnetization *vs* T for several values of α are depicted in Fig. 3. They give the amplitude of the overlap of a single retrieval state with the corresponding learnt pattern.

C. *The Phase Diagram near* $T=1, \alpha=0$

Since both $T_M(\alpha)$ and $T_c(\alpha)$ are lines associated with discontinuous transitions, there is no small parameter near a generic point on these lines. However, both lines run to $T=1$ as $\alpha \to 0$, and in the vicinity of this point the magnetization (overlap) is small. The reason is that for p finite ($\alpha=0$) one has a continuous transition at $T=1$, and the FM Mattis-states emerge globally stable (this has been shown in I).

The structure of the theory can be studied in this region by an expansion of Eqs. (3.3)–(3.5), with $\vec{h}=0$ and a single nonvanishing overlap, in powers of m, α

[2] Note that this very steep slope has led in [6] to the mistaken impression that the line T_M disappears discontinuously.

and t ($\equiv 1-T$). Searching for nonzero solutions for m, the three equations become, correspondingly,

$$t = \tfrac{1}{3}m^2 + \alpha r, \tag{5.8}$$

$$q = m^2 + \alpha r, \tag{5.9}$$

$$r = \frac{q}{(q-t)^2}. \tag{5.10}$$

Eliminating q and r in favor of m and writing

$$\tau \equiv t/\sqrt{\alpha}; \qquad y \equiv \tfrac{2}{3}m^2/\sqrt{\alpha} \tag{5.11}$$

these equations reduce to

$$g(y) \equiv \tfrac{1}{2}y^3 - \tau y^2 + y + \tau = 0 \tag{5.12}$$

for $y > 0$.

This equation has either two positive solutions, or none. The disappearance of these two solutions gives T_M. Hence, τ_M is the value of τ for which $g(y)$ vanishes at the same value of y as does its derivative. The result is

$$T_M \simeq 1 - 1.95\sqrt{\alpha}. \tag{5.13}$$

The line $T_c(\alpha)$, in Fig. 4, also starts at $T=1$, $\alpha=0$, and is also of the form $T_c \simeq 1 - A\sqrt{\alpha}$. It is determined, for a given α, by the value of τ for which the free energy, Eq. (3.2) with $m=0$, equals the free energy for $m \neq 0$, which solves Eqs. (5.8)–(5.10).

Using the notations of (5.11) one writes the above condition in the form

$$f(m) - f(0) = \frac{\alpha}{2}\left[\ln y - \tau y - \frac{\tau}{y} + 2\tau + \frac{1}{4}y^2 + \frac{1}{2}\right] = 0 \tag{5.14}$$

which has to be solved together with Eq. (5.9). The result is

$$T_c \simeq 1 - 2.6\sqrt{\alpha}.$$

In conclusion, it should be noted that the SG phase described above is unstable to replica symmetry breaking (RSB), at all $T < T_g$. The retrieval states become unstable to RSB below a temperature $T_R(\alpha)$, see Fig. 2. The line $T_R(\alpha)$, the analog of the Almeida–Thouless line [10] in the SK model [9], marks the appearance of many degenerate states around each of the retrieval states. All the states associated with one of the retrieval states are highly correlated and have the same macroscopic overlap with that state. This is the subject of Section 7.

6. The Effect of External Fields

External fields conjugate to memorized patterns may appear in two contexts. First, the fields may enhance the learning of a subset of memories. This could be

achieved by an appropriate tuning of thresholds as part of the learning process. Second, external fields may be introduced as a representation of external stimuli initiating retrieval. The implications of the introduction of the fields in the two scenarios may be quite different. Here we focus on the first of the two.

We introduce on spin i an external field of the form

$$h_i = \sum_\nu h^\nu \xi_i^\nu \tag{6.1}$$

by adding to the Hamiltonian a term

$$H_h = -\sum_i \sum_\nu h^\nu \xi_i^\nu S_i \tag{6.2}$$

as in Eq. (2.2). The associated patterns will be referred to as "marked" patterns and the question to be addressed is whether it is possible to affect the quality of their retrieval.

We first study the case of a field conjugate to a single pattern, i.e., $h_i = h\xi_i^1$, at $T=0$. Subsequently, we discuss the case of a field coupled to several patterns, at $T=0$. In the last part of this section the phase diagram of a network with a single marked pattern is described.

A. *Field on a Single Pattern at* $T=0$

The equation for the magnetization in the presence of external fields is Eq. (4.2), which in the present case reads:

$$m = \mathrm{erf}[(m+h)/\sqrt{2\alpha r}], \tag{6.3}$$

where, $r = (1-C)^{-2}$, and

$$C = 1 + \sqrt{2/\pi\alpha r}\, e^{-(m+h)^2/2\alpha r}. \tag{6.4}$$

At large values of α, only the SG state is stable. In the presence of the external field it has a nonzero overlap, m, with the marked pattern, which develops continuously from zero, linearly in the field h. As α decreases, a metastable state appears at $\alpha_c(h)$, Fig. 4, as in the case of $h=0$. This is an FM retrieval state, with a high value of m. As α is further decreased, this state becomes globally stable at $\alpha_T(h)$. Finally, at even lower α, the SG state, i.e., the state with low m, disappears and only the FM state persists. This takes place at $\alpha_{SG}(h)$.

At $h=0$, $\alpha_c = 0.138$, $\alpha_T = 0.051$, and $\alpha_{SG} = 0$. The curves of $\alpha_c(h)$, $\alpha_T(h)$, and $\alpha_{SG}(h)$ meet at one point: $h = h_c \simeq 0.37$ (Fig. 4). For $h > h_c$, there is only one stable state for any α.

The increase of α_c with increasing h, implies that the field allows for the retrieval of the marked pattern, even when the network is oversaturated ($\alpha > \alpha_c(0)$). Moreover, at a given α, the quality of the retrieval is improved by the application of the field. This is shown in Fig. 5. The overlaps of the FM state and the SG state with the marked pattern are plotted as a function of α for different values of h.

As an example consider the case $h=0.2$. The marked pattern can be retrieved in

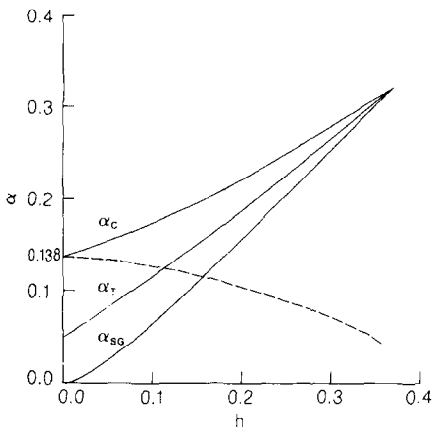

FIG. 4. Curves of: α_c (h), below which the marked retrieval state appears, α_T below which the retrieval state becomes a global minimum and α_{SG} below which the *SG* state disappears. The dashed curve represents $\alpha_c(h)$ for the unmarked patterns.

a network with α up to $\alpha_c = 0.22$. At this value of α the retreival has 2.5% errors $(m(\alpha_c) = 0.95)$, which is a higher level of errors than the 1.5% at $h = 0$, with $\alpha = \alpha_c(0)$. However, when $\alpha = 0.2$, the marked pattern is retrieved with less than 1% error.

It should be recalled that RSB is strong in the SG (low m) states. Hence, the lower curves of m in Fig. 5, which are based on a replica symmetric theory, give only a qualitative picture.

The application of a static external field to any configuration of the network, learnt or random (which has not been learnt), will induce a nonzero overlap with that state. For example, if

$$h_i = h\eta_i, \tag{6.5}$$

where $\{\eta_i\}$ is a configuration of the network, uncorrelated with any of the $\{\xi_i\}$ that are embedded in the couplings J_{ij}, then for finite values of p $(\alpha = 0)$ there is a metastable state at $T = 0$, with $\{S_i = \eta_i\}$, for arbitrary h. However, as soon as α

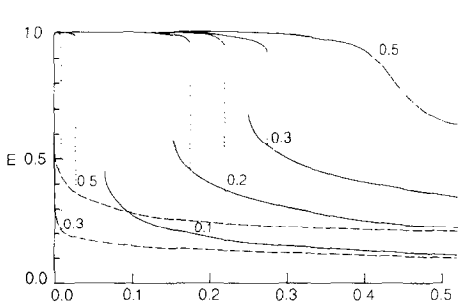

FIG. 5. Magnetization curves $m(\alpha)$ at fixed h. The solid lines correspond to the marked pattern. The dashed lines represent the magnetization of random patterns, which have not been learnt.

becomes finite the overlap of this metastable state with $\{\eta_i\}$ decreases very sharply. The potential usefulness of a field coupled to a memorized pattern, is in the fact that the overlap of the FM state, condensed into that pattern remains high even when α is not very small.

Figure 5 presents a comparison of the magnetization induced by a field conjugate to a random pattern, i.e., such as (6.5). The equations for the magnetization are derived, in this case, by the procedure outlined in Sections 2 and 3, except that now *all* the ξ's are averaged in the transition from Eq. (2.4) to (2.5). The result is,

$$m = \operatorname{erf}\left(\frac{h}{x}\right) \tag{6.6}$$

and $x \equiv \sqrt{2\alpha r}$ is determined by

$$x = \sqrt{2\alpha} + (2/\sqrt{\pi})\, e^{-h^2/x^2}. \tag{6.7}$$

For small α, Eq. (6.7) has two stable solutions, in one of which x is very small and hence $m \simeq 1$. However, this state disappears discontinuously at a value of $\alpha = \alpha^*(h)$, which is much smaller than $\alpha_c(h)$, (e.g., $\alpha^*(0.3) = 0.008$ and $\alpha^*(0.5) = 0.027$). Curves of m *vs* α for the case of a random configuration are shown in Fig. 5 for $h = 0.3$ and 0.5.

It should be realized that the increase in the value of α_c, described above, applies only to the marked pattern: the one coupled to h. In fact, the presence of h on one pattern produces random noise for the other patterns. This results in a reduction in the value of α_c for the other patterns, as is depicted by the dashed curve in Fig. 4. Consequently, the improvement in the retrieval of a marked pattern may come at the expense of the ability to retrieve the rest of the patterns. In other words, if the network is slightly below saturation, the marking of a pattern will reduce the error in its retrieval. However, it may at the same time make the retrieval of all the rest of the memorized patterns impossible, by reducing the critical α below the actual storage level. On the other hand, if the network is initially above saturation, a marked pattern can be made retrievable, while the damage to the rest of the memorized patterns is irrelevant, since they were in the dark to start with.

B. *Field Coupled to Several Patterns*

Applying the field with equal intensity to n patterns, via

$$h_i = h \sum_{\nu=1}^{n} \xi_i^{\nu} \tag{6.8}$$

one asks about the quality of FM retrieval states: those states with $m \simeq 1$ on one of the marked patterns. These states will also have a finite, hopefully small, overlap m' with the other $n-1$ marked patterns. The values of m and m' are determined from the equations

$$m = \langle\!\langle \operatorname{erf}\{[(m+h) + (m'+h)\, z_{n-1}]/\sqrt{2\alpha r}\} \rangle\!\rangle, \tag{6.9}$$

$$(n-1)\, m' = \langle\!\langle z_{n-1} \operatorname{erf}\{[(m+h)\, \xi + (m'+h)\, z_{n-1}]/\sqrt{2\alpha r}\} \rangle\!\rangle \tag{6.10}$$

as usual, $r=(1-C)^{-2}$ and

$$C=\sqrt{2/\pi\alpha r}\ll\exp\{-[(m+h)\,\xi+(m'+h)\,z_{n-1}]^2/2\alpha r\}\gg, \tag{6.11}$$

where ξ is a random variable ± 1 and z_n is a sum of n such independent random variables. Equation (6.9) implies that for $n>1$, the field generates an additional noise on m, which is due to the random overlaps of $\{\xi_i^1\}$ with the remaining $n-1$ marked patterns.

Expanding Eqs. (6.9) and (6.10) for small h one finds

$$m=\mathrm{erf}[(m+h)/\sqrt{2\alpha r}]-(n-1)\,Cmh^2/\alpha r+O(h^3), \tag{6.12}$$

$$m'=[C/(1-C)]\,h+O(h^2), \tag{6.13}$$

and C is given by Eq. (6.4).

The second term on the r.h.s. of Eq. (6.12) represents the noise generated by the magnetizations of the $n-1$ unretrieved marked patterns. This noise has a strong effect on the value of $\alpha_c(n, h)$, below which retrieval is possible. At sufficiently small h and fixed n both m and $\alpha_c(n, h)$ increase with h. However, as h increases, the effect of the noise takes over and leads to a reduction of m and α_c. To enhance retrieval the field must be much smaller than $1/\sqrt{n}$. In Fig. 6, $\alpha_c(n, h)$ is plotted for $n=1-5$. One observes that already for $n=3$ the the maximum increase in storage capacity is about 15% (at $h\simeq 0.1$). The conclusion is that the external field is not a very useful mechanism for enhancing learning, beyond a very few patterns.

C. *External Fields at Finite Temperature*

We discuss briefly the case of a field conjugate to a single pattern at finite temperature. The phase diagram in the (T, h) plane, for a fixed small α ($\alpha=0.1$) is shown in Fig. 7. At high T the phase is paramagnetic. The overlap m is nonzero and

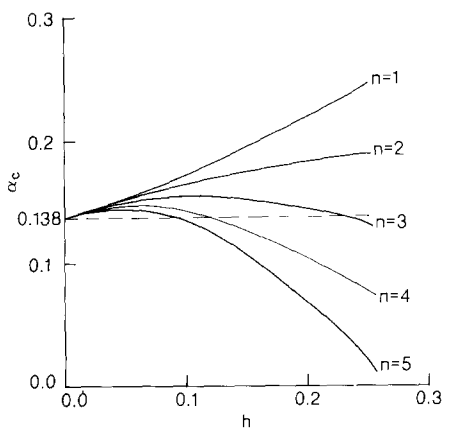

FIG. 6. Plots of $\alpha_c(h)$ when the field is applied to n patterns, for several values of n.

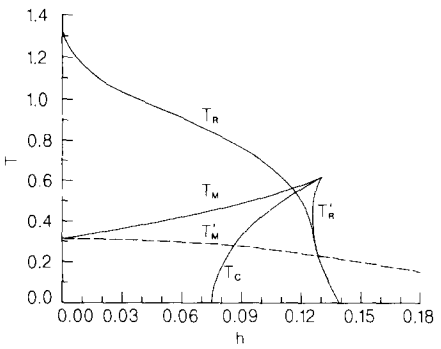

FIG. 7. The phase diagram in the T–h plane, with a single marked pattern. T_R is the transition to the *SG* phase, T_M the appearance of the *FM* retrieval state, T_c the *FM* retrieval state becomes globally stable, to the right of T'_R, the low magnetization state merges with retrieval state, the dashed line T'_M denotes the appearance of the unmarked *FM* retrieval states.

is proportional to h, for small h. Likewise, q and r are proportional to h^2. But the thermodynamic state is ergodic.

As T decreases below $T_R(h, \alpha)$, the system enters an SG phase of broken ergodicity, in which many different degenerate states with huge barriers exist. The temperature T_R is the analog of the Almeida–Thouless line in the SK model. It is determined by RSB, which is described in Section 7.

As T is further decreased, one crosses the line T_M, into a phase which possesses a metastable retrieval state having a large overlap with the "marked" pattern. For $T_M > T > T'_M$ it is only the marked pattern which is retrievable. The other $2(p-1)$ states, associated with the unmarked patterns, become metastable and retrievable on crossing the curve T'_M, at lower temperature. As $h \to 0$, both $T_M(h)$ and $T'_M(h)$ approach the same limit $T_M(0)$, as expected.

The marked pattern becomes a global minimum below $T_c(h)$. Note that for $\alpha = 0.1$ the line T_c approaches a finite value, $h = 0.075$, at $T = 0$. For $\alpha < 0.051$ this line crosses the T axis, as implied by the $h = 0$ results in Fig. 4.

The region of the global stability of the retrieval state is bounded on the high field side by the line T'_R. This boundary denotes the merging of the SG state with the retrieval state. To the right of this line the system is ergodic, with a unique state which has a large overlap with the marked pattern. This region is part of the paramagnetic phase, and connects smoothly with the low-h high-T regime.

In concluding we recall that the application of external fields is equivalent, in biological terms, to the modification of neural activation thresholds.

7. REPLICA SYMMETRY BREAKING (RSB)

Standard wisdom has it that below the line $T_g = 1 + \sqrt{\alpha}$, when SG ordering sets in, the system becomes unstable to RSB [10]. The SG state, with vanishing

overlaps with all stored patterns, represents a large number (infinite as $N \to \infty$) of almost degenerate stable states. They are organized ultrametrically, with all the accompanying features [12]. This collection of states is not the main focus of our inquiry, since it is not useful for retrieval, except insofar as to provide a wastebasket for unidentified external stimuli.

But the possibility of RSB raises questions about the stability of the FM retrieval phases, be they stable or metastable. Three issues come to mind:

Could the replica symmetric FM phase be stable at $T=0$?

If not, where in the $h-T$ plane (for a given α) does it become unstable? That is, what is the analog of the Almeida–Thouless line [10]?

What is the nature of the RSB in the FM phase of the present model, if indeed it takes place? What elements should be added to Parisi's treatment of of RSB in the SK model [11]?

A. *The Entropy at* $T=0$

An indication of RSB at $T=0$ may be obtained by computing the entropy of the replica symmetric phase. Here it is to be done for phases with a finite m. Inserting the free energy, Eq. (3.2), in

$$S = -\frac{\partial f}{\partial T}$$

and taking the limit $T \to 0$, keeping in mind that

$$q \simeq 1 - CT + O(T^3),$$

where C is the limit of $\beta(1-q)$ as $T \to 0$ (Eq. (4.4)), one arrives at

$$S = -\frac{\alpha}{2}[\ln(1-C) + C/(1-C)] \tag{7.1}$$

which is negative for all $C < 1$.

For $\vec{h}=0$

$$C = \sqrt{2/\pi\alpha r}\, \langle\!\langle \exp[-(\vec{m}\cdot\vec{\xi})^2/2r\alpha] \rangle\!\rangle \tag{7.2}$$

(see Eq. (4.4)). It is easily verified that (7.2) together with (4.7) implies $C < 1$. Thus the entropy at $T=0$ is negative for all FM states, leading to the conclusion that the replica symmetric phase is unstable.

This entropy should be compared with the $T=0$ value of the entropy in the *SG* phase, as well as in the SK model. In the SG phase, the value of C can be derived from Eq. (4.11). It is

$$C = [1 + \sqrt{\pi\alpha/2}]^{-1} \simeq 0.68 \tag{7.3}$$

at $\alpha = \alpha_c = 0.138$. This leads to to $S \simeq -0.07$. The value of S in the SK model is obtained from (7.1) and (7.3) in the limit $\alpha \to \infty$. One finds

$$S = -1/2\pi \simeq -0.16,$$

while in the FM retrieval states, at $\alpha = \alpha_c$,

$$C = \sqrt{2/\pi\alpha_c r}\exp\left[-\frac{1}{2\alpha_c r}\right] \simeq 0.06$$

giving $S \simeq -1.4 \times 10^{-3}$.

The $T = 0$ FM entropy vanishes exponentially with decreasing α. In fact, for small α, $r \simeq 1$, $m \simeq 1$, and $C \simeq \sqrt{2/\pi\alpha}\exp(-1/2\alpha)$, which gives

$$S \simeq -\frac{1}{4}\alpha C^2 \simeq -\frac{1}{2\pi}\exp(-1/\alpha); \tag{7.4}$$

all the above serves as an indication that RSB in the retrieval states is very weak.

B. *The Almeida–Thouless Line*

To investigate the stability of the retrieval FM states against RSB we return to the general expression for the free energy, Eq. (2.9), and compute the matrix of its second derivatives, with respect to $q_{\rho\sigma}$ and $r_{\rho\sigma}$ at the replica symmetric point, much like in ref. [10]. The details are left to Appendix B. We find that the stability against fluctuations in the direction of RSB is associated with the sign of the quantity x_λ,

$$x_\lambda = [1-\beta(1-q)]^2 - \alpha\beta^2 \ll (1-\langle S\rangle^2)^2 \gg \tag{7.5}$$

which in turn determines the sign of the "replicon" eigenvalue λ. In the retrieval phases, m and q are given by Eqs. (3.3)–(3.5). The condition for the onset of the RSB instability is $\lambda = 0$, i.e., $x_\lambda = 0$, which reads, in terms of m and h,

$$\alpha\beta^2 \ll ch^{-4}\beta(\sqrt{\alpha r}z + m + h) \gg = [1-\beta(1-q)]^2. \tag{7.6}$$

This is the generalization of the Almeida–Thouless equation [10] to the neural network. The four equations have to be solved simultaneously, to give the line.

We first discuss the zero field case. In the paramagnetic phase ($q = m = 0$) the quantity

$$x_\lambda = (1-\beta)^2 - \alpha\beta^2 \tag{7.7}$$

is positive for all $T > T_g = 1 = 1 + \sqrt{\alpha}$. In the low T FM phase equation (7.6) is satisfied along a line of transition temperatures $T_R(\alpha)$, which is represented in Fig. 2. Its maximum value is where it intersects $T_M(\alpha)$, $\alpha = 0.137$. There $T_R \simeq 0.07$. It decreases to zero exponentially with α

$$T_R \simeq \sqrt{8\alpha/9\pi}\exp(-1/2\alpha). \tag{7.8}$$

In the interval $0<\alpha<0.137$ $T_M(\alpha)>T_R(\alpha)$. Hence, for most of the interval $0<\alpha<\alpha_c$, the FM phase is stable against RSB in a finite range of temperatures below T_M. This supports our assertion that RSB effects in the retrieval states are relatively weak.

In a finite field h and a fixed value of α, Eq. (7.6) yields a transition line, below which RSB occurs. The equations have been solved numerically and the results for $\alpha=0.1$, 0.2, and 0.4 are shown in Fig. 8.

At $\alpha=0.1$ the line of RSB consists of two disconnected parts. The upper part corresponds to the SG phase, as described in Section 6C, and is presented as T_R in Fig. 7. The dotted curve represents the line below which the FM retrieval solution appears (the line T_M in Fig. 7). This solution becomes unstable against RSB at the lower part of the curve. Note the exaggerated temperature scale for this part of the line.

A similar behavior of the RSB line characterizes the entire range $\alpha<\alpha_c\simeq 0.14$. The curve for $\alpha=0.2$ represents the behavior in the interval $\alpha_c<\alpha<0.32$. In this interval the AT line still consists of two disconnected parts, but the SG and the FM coexist in a range of values of h, which does not reach $h=0$. The FM solution appears below the dotted line T_M. For $\alpha>0.32$ there is only one solution and the AT curve is continuous and monotonic. The case of $\alpha=0.4$ is shown in Fig. 8.

C. *Phases with Broken Replica Symmetry*

The nature of the RSB in the present model is similar to that of the SK model. Following Parisi [11], we assume that $q_{\rho\sigma}$ and $r_{\rho\sigma}$ consist of a hierarchy of blocks, which lead as $n\to 0$ to a parametrization of the phase by continuous functions $q(x)$ and $r(x)$, $(0\leqslant x\leqslant 1)$. The magnetic order parameters $\vec{m}_\sigma$ are assumed to be

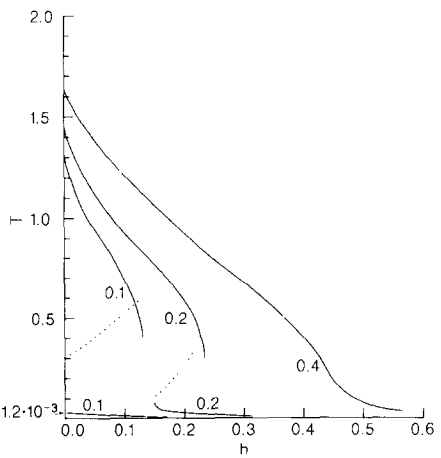

FIG. 8. The *RSB* transition lines for several values of α. The dotted lines are T_M-lines, below which the *FM* retrieval state appears.

independent of the replica index. Next, the free energy, Eq. (2.12), is written as a functional of $q(x)$ and $r(x)$. The quadratic term becomes

$$\frac{1}{n}\sum_{\rho\sigma} q_{\rho\sigma} r_{\rho\sigma} = -\int_0^1 dx\, q(x)\, r(x). \tag{7.9}$$

To evaluate the Tr ln term one notes that the spectrum of a general Parisi matrix $q(x)$ consists of bands [17] of degenerate eigenvalues ε, parametrized by a variable x

$$\varepsilon_x = -xq(x) - \int_x^1 q(y)\, dy. \tag{7.10}$$

The degeneracy of ε_x is $-n\, dx/x^2$, and x varies from $n(n \to 0)$ to 1. This accounts for

$$-n\int_n^1 dx/x^2 = n - 1$$

eigenvalues. In addition there is a nondegenerate eigenvalue

$$\varepsilon_0 = \int_n^1 dx\, q(x).$$

Hence,

$$\mathrm{Tr}\ln(1-\beta-\beta\mathbf{q}) = -\frac{n\alpha}{2}\int_n^1 \frac{dx}{x^2}\ln[1-\chi(x)] + \frac{\alpha}{2}\ln[1-\chi(n)+nq(n)] \tag{7.11}$$

where the local susceptibilities [13] at the scale x, $\chi(x)$ are defined as

$$\chi(x) = \beta\left[1 - xq(x) - \int_x^1 q(y)\, dy\right]. \tag{7.12}$$

Finally, the term $\ln \mathrm{Tr}\exp(\beta H_\xi)$ (as a functional of $r(x)$) is identical to the corresponding term in the SK model (as a functional of $q(x)$).

The equations for $\vec{m}$ and $q(x)$ are the usual saddle-point equations, Eqs. (2.8), (2.9), in which the smallest size common block of (ρ, σ) is of scale x. The equation for $r(x)$ reads

$$r(x) = r(0) + \int_{x_0}^{x} \frac{q'(y)\, dy}{[1-\chi(y)]^2}, \tag{7.13}$$

$$r(0) = \frac{q(0)}{[1-\chi(0)]^2}. \tag{7.14}$$

Note that both $q(x)$ and $r(x)$ are expected to be constant in the same interval of x,

$$q(x)=q(0), \qquad r(x)=r(0), \qquad 0\leqslant x\leqslant x_0,$$
$$q(x)=q(1), \qquad r(x)=r(1), \qquad x_1\leqslant x\leqslant 1.$$

An explicit solution for $q(x)$ and $r(x)$ at arbitrary temperature is very difficult to derive, as in the SK case. Nevertheless a few results can be deduced, mostly in the SG phase.

The SG *phase.* In this phase $q(0)=r(0)=x_0=0$. Expansion, near T_g, in powers of t ($t\equiv 1-T/T_g \ll \sqrt{\alpha}$) leads to

$$q(x)\simeq r(x)\propto \frac{x}{1+1/\sqrt{\alpha}}, \qquad 0\leqslant x\leqslant x_1\simeq \frac{t}{1+1/\sqrt{\alpha}},$$
$$q(x)=q(1), \qquad r(x)=r(1), \quad x_1\leqslant x\leqslant 1$$

with $q(1)\simeq r(1)\simeq t$. Note that $x_1\to 0$ as $\alpha\to 0$, (keeping $t/\sqrt{\alpha}$ small), which implies that replica symmetry is restored in this limit, as expected.

From the property $q(x)\to 0$ as $x\to 0$ it follows that both the equilibrium *local* susceptibility and the FM one are independent of T below T_g,

$$\chi(0)=\beta\left[1-\int_0^1 q(x)\,dx\right]=\frac{1}{T_g}=\frac{1}{1+\sqrt{\alpha}}, \tag{7.15}$$

$$\chi^{\mu\mu}=\frac{\chi(0)}{1-\chi(0)}=\frac{1}{\sqrt{\alpha}}; \tag{7.16}$$

see Eq. (5.4). Equations (7.15) and (7.16) imply that the phase is marginal. In addition, x_λ of Eq. (7.5) is expected to be zero, at all $T<T_g$. Hence, at low T,

$$1-q(1)\propto T^2/\sqrt{\alpha}$$

and this implies that the nonequilibrium (i.e., single valley) susceptibilities vanish as

$$\chi(1)=\beta[1-q(1)]\simeq T/\sqrt{\alpha}, \tag{7.17}$$

$$\chi^{\mu\mu}=\frac{\chi(1)}{1-\chi(1)}\simeq T/\sqrt{\alpha}, \tag{7.18}$$

when $T/\sqrt{\alpha}\to 0$.

To conclude, the SG phase has all the essential properties of the SG phase in the SK model, including the high degeneracy of pure states, marginality and anomalously slow relaxation.

The retrieval phases. In these phases $q(0)$, $r(0)$, and x_0 are positive. In fact, since RSB takes place when m is already close to unity, $q(1)\simeq r(1)\simeq 1$. Hence, the variation of $q(x)$ (and consequently of $r(x)$; see Eqs. (7.13) and (7.14)) contributes very little to the equations for quantities such as m or χ. This lends support to our

assertion that the replica symmetric theory is a very good approximation as far as the properties relevant for retrieval are concerned.

Nevertheless, the very occurrence of RSB implies that the energy landscape of the basin of each of the retrieval phases has features which are similar to the SG phase. In particular, each of the retrieval phases represents many degenerate retrieval states. All of them have the same macroscopic overlap m, but they differ in the location of the errors. These states are organized in an ultrametric structure [12]. The energy barriers between them are most probably not greater than $O(\sqrt{N})$. However, since all the overlaps between these states are close to unity, the effect of this structure on the retrieval performance of the system is negligible.

Finally, we note that the statistical mechanical interpretation of the order function $q(x)$ in our case is identical to that in the SK model [11]. One defines an overlap q between two copies of the system

$$q = \frac{1}{N} \sum_i S_i^\rho S_i^\sigma .$$

The ensemble averaged probability distribution of q is given by

$$P(q) = \frac{dx}{dq} .$$

In a similar fashion one defines an "error overlap" of two copies of the system, which are in the same *macroscopic* FM *phase*

$$\alpha r = \sum_{\mu > s} m_\rho^\mu m_\sigma^\mu .$$

The average probability distribution of r is given by

$$P(r) = \frac{dx}{dr} . \tag{7.19}$$

8. Numerical Simulations

The main result of Section 4 is the existence of a sharp value of α ($\alpha_c = 0.138$) above which there are no dynamically stable or metastable FM retrieval states. It was emphasized there that this result does not exclude the possibility of states with a finite overlap with one of the stored patterns, which are stable against all flippings of a single spin or even a finite cluster of spins. If such states exist, then an initial spin configuration, which corresponds to a stored pattern may evolve dynamically to a stationary state with $m \neq 0$, even at $\alpha > \alpha_c$.

To investigate this possibility and to explore the meaning of α_c we performed computer simulations on networks with different values of N and p.

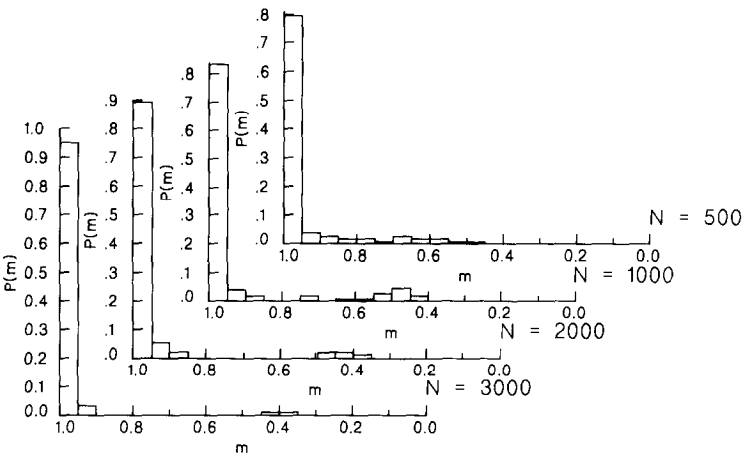

FIG. 9. Histograms of the overlaps of the retrieved state with the initial state, which is one of the learnt patterns, for $\alpha = 0.14$, for several values of N.

Starting from one of the stored patterns, all the spins of the network are checked consecutively for stability. If a spin-flip occurs, all the local fields are updated, before the next spin is tested. This corresponds to the asynchronous (Hopfield) dynamics. The process is repeated until a stable configuration is reached. The distribution of the overlaps of the final states with the initial memory states is then measured.

The results for $\alpha = 0.14$ and 0.16 and several values of N are shown in Figs. 9, 10. For $N = 500$ there is no qualitative difference between these two cases. There is a sharp peak near $m = 1$ and a broad, almost uniform, distribution extending to lower values of m. A sharp difference between the two figures develops as N

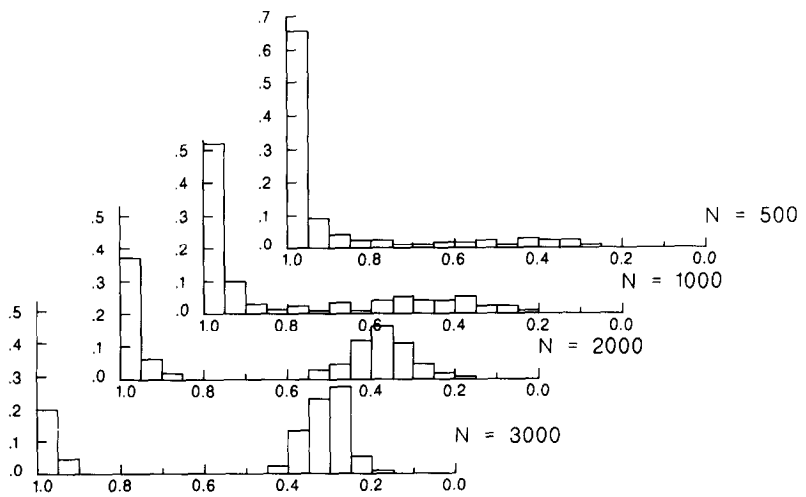

FIG. 10. Same as Fig. 9 for $\alpha = 0.16$.

increases. In the case of $\alpha = 0.14$, the peak near $m = 1$ grows and when N reaches $\simeq 3000$, it contains almost the entire weight of the distribution. The average value of m at $N = 1000$ (averaged over 200 patterns) is 0.972 ± 0.01, and, within the error bars, remains unchanged at higher N. From this we conclude that the actual value of α_c is somewhat higher than 0.138, predicted by the replica symmetric theory. The measured value of m is consistent with the theoretical predictions of Section 4, for $\alpha < \alpha_c$.

Note that the tail of the distribution does not shrink smoothly as N increases. Instead, a small peak centered about a low value of m ($m \simeq 0.4$) remains in intermediate size networks. This effect is more pronounced at $\alpha = 0.16$.

For $\alpha = 0.16$ as N increases there appears a clear separation between the peak near $m = 1$ and a lower peak around $m \simeq 0.35$, with a large gap in m where no stable states exist. The weight of the high-m peak is gradually transferred to the lower-m one with increasing N. To check whether the high-m part of $p(m)$ disappears as $N \to \infty$ we have performed a finite-size scaling analysis of $P(N, \alpha)$, for $\alpha = 0.15$ and 0.16, where P is the area under the high-m peak. We find that P decreases exponentially, at a rate which depends on α. The data fits a curve of the form

$$P = A \exp[B(\alpha_c - \alpha)\, N]. \tag{8.1}$$

A good fit was obtained with the parameters, where $A = 0.97 \pm 0.05$, $B = (2.8 \pm 0.3) \cdot 10^{-2}$, and $\alpha_c = 0.145 \pm 0.01$, see Fig. 11. This form of exponential finite size correction is reminiscent of first order transitions. The value of α_c obtained by this analysis is slightly higher than $\alpha_c = 0.138$, obtained from the replica symmetric theory. The small discrepancy may be attributed to RSB. Equation (8.1) indicates that the larger the value of α, the faster P decreases with N.

The low-m peak of $p(m)$, for $\alpha > \alpha_c$, becomes sharper as $N \to \infty$, but the position of its center does not change appreciably with N. This implies that the *average* overlap $\bar{m}$, at $T = 0$, does not vanish as $N \to \infty$, even above α_c. The phenomenon

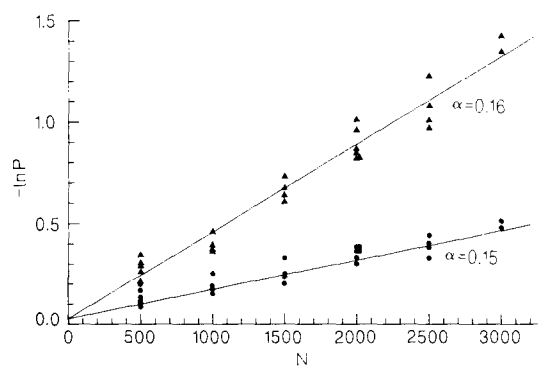

FIG. 11. Simulation data for the finite-size scaling analysis, used to determine α_c. P is the weight of the high-m peak. Each data point represents an average over 100 patterns.

persists at higher values of α as well. For instance, at $\alpha = 1$ we find for large N $\bar{m} \simeq 0.28$. As α increases $\bar{m}(\alpha)$ decreases, but our data is consistent with a nonzero value for $\bar{m}$ as $\alpha \to \infty$.

To understand the origin of the finite (nonzero) value of $\bar{m}$ above α_c, we note that although the thermodynamic state, which is the SG, has $m = 0$, there is an exponentially large number of states which are stable against all single spin-flips. Many of these states have a small, but finite, overlap with the stored patterns. In fact, for any finite α there is an exponentially large number of locally stable states which have a finite overlap with any configuration of the network, irrespective of whether it had been learnt. Consequently, single spin-flip dynamics, at $T = 0$ and finite α, will always lead to finite remanent magnetization, i.e., the final state will have a finite overlap with the initial state.

We have also measured the remanent magnetization, m_R, of a random (not learnt) initial state. We find that m_R is smaller than $\bar{m}$ even at $\alpha > \alpha_c$. For instance, at $\alpha = 0.16$, $m_R \simeq 0.08$, and for $\alpha = 1$, $m_R \simeq 0.12$. This suggests that even above saturation the dynamics of the system distinguishes between a random pattern and the embedded ones.

As α increases $\bar{m}$ decreases and m_R increases. Both tend to a common nonzero limit, which is the value of remanent magnetization, $m_R \simeq 0.15$, of the SK spin-glass [14]. More details and analysis of the numerical simulations will be published elsewhere.

9. Discussion

The increasing detail at which the properties of the Hopfield network can be analyzed, opens the road for deepening and widening the scope of the investigation.

Several aspects of the present results deserve further study. One is the extent to which breaking of replica symmetry modifies the quantitative analysis of the performance of the system. We have suggested that the difference between the simulation results, $\alpha_c = 0.145 \pm 0.01$ and the replica symmetric prediction $\alpha_c = 0.138$, is due to RSB. To clarify this question, we have undertaken a study of the Parisi mean-field equations for this model.

Another issue is the role of external fields. In the present work a field was added as an additional mode of learning of one or several patterns. We have shown that the field generates also random noise, which limits the scope of the resulting enhancement. An external field may also appear as part of the retrieval mechanism. This means that some neurons are kept aligned, by the external input, during the dynamic process of retrieval. The outcome of this scenario is currently under study.

One of the main results of our study is that for $\alpha < \alpha_c$, retrieval states coexist with the SG state. The effect of this "spurious" state on the basin of attraction of the retrieval states deserves further numerical study.

Widening the scope involves the relaxation of the restrictions underlying the model as analyzed here. This is an essential step, if the model is to stand a chance

with biological systems. Some aspects of the robustness have been already probed by Hopfield [1]. They are being looked at more systematically. Some of the main issues are:

(1) *Attrition of synapses.* The assumption that all neurons are interconnected is rather extreme. One way to probe it, is to dilute the synaptic connections. This was done, keeping the network symmetric, both analytically [18] and numerically. The behavior of the network is found to change very gradually. Up to 50% dilution α_c decreases from 0.14 to about 0.09. The retrieval quality, as measured by the overlap at α_c, decreases from 0.97 to 0.93.

(2) *Clipping.* What if synapses cannot contain all the detailed information (an integer between $-p$ and $+p$) for a network storing p patterns? An extreme case is that of clipped synapses,

$$J_{ij} = \frac{\sqrt{p}}{N} \operatorname{sgn}\left(\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \right).$$

Analysis of this case [18] has shown that α_c decreases from 0.14 to 0.1 and m_c from 0.97 to 0.95.

(3) *Asymmetry.* The most artificial restriction, perhaps, is that of the symmetry of the synaptic connections. It has provided a vital underpinning to the analytical statistical mechanical technique. To relax it one may resort to numerical simulations. Preliminary results indicate that diluting synapses asymmetrically, produces a network with retrieval performance similar to that of the network with symmetric dilution at the corresponding dilution levels. Again, an astounding roboustness.

(4) *Spin-glass noise at synapses.* It has been proposed [19] that prior to learning neurons may be connected randomly, creating an SK spin-glass. How does this type of static noise effect the network? At finite p, as long as $[\eta_{ij}^2]$ (the mean square gaussian noise) is less than $\simeq 0.2J/N$, where J is the amplitude of the Hebbian term, the retrieval state of a stored pattern has $m > 0.97$ [18].

Finally, a word about extensions. The next generation of models of neural networks will have to deal with the issue of memorizing correlated patterns. It has been shown that networks can be constructed, which can store an *arbitrary* set of patterns. Those may be useful as devices of artificial intelligence, but the nonlocal process required for the addition of a pattern, makes them rather unlikely candidates for biological systems [20].

One of the animating ideas in this context has been that of hierarchical ordering [19, 21, 22]. None has yet proved its feasibility. Very briefly one may describe the proposal of Toulouse, Dehaene, and Changeux [19] as a superposition of the "Hebb" learning rule on top of an innate spin-glass network. If the two are independent, then it has been shown [18] that it acts either as a noisy Hopfield model or a spin-glass. On the other hand, it has been proposed that the learning process modifies the spin-glass coupling by reinforcing the ground states of the spin-glass,

by the spin-glass dynamics itself, thus producing a store of hierarchically organized memories [12]. While this is an attractive idea, its realization has yet to be demonstrated. One of the main concerns is the proliferation of high metastable states, which are probably not ultrametrically organized. Reaching the true ground states of the SG will require prohibitively long times.

Virasoro and Parga [21] have proposed a complementary approach to the utilization of the hierarchical morphology of SG ground states. They propose a learning and preprocessing mechanism which directly constructs J_{ij}'s known to have a preassigned hierarchical organization of ground states. Dotsenko [22] has proposed a third approach to the generation of a network with hierarchically organized memories. He constructs a hierarchical model of groups of neurons, each storing uncorrelated patterns. At each consecutive level the total magnetizations of the groups of the previous level are coupled à la Hopfield. It would be interesting to study in more detail the performance of these hierarchical networks. A particularly important question is at which stage in the retrieval process the system takes advantage of the hierarchical structure. Ultimately, the attractivity of the hierarchical correlations is due to the possibility that they may precipitate the recognition of classes, at times much shorter than that required by detailed identification of patterns.

Acknowledgments

In the course of the preparation of this article we have benefitted extensively from discussions with Drs. G. Toulouse, J. Hopfield, M. Virasoro, M. Mezard, L. Peliti, G. Weisbuch, and G. Dreyfus. The work of D.J.A. and H. S. has been supported in part by the Fund for Basic Research of the Israel Academy of Science and Humanities.

Appendix A

Here we summarize some of the details involved in taking the $n \to 0$ limit of Eq. (2.9), leading to Eq. (3.2).

The matrix $(1-\beta)\mathbf{I} - \beta\mathbf{q}$, with $\mathbf{q}$ given by (3.1b), has one eigenvalue $1-\beta-(n-1)\beta q$ and $(n-1)$ eigenvalues $1-\beta+\beta q$. Hence,

$$\lim_{n\to 0} \frac{1}{n} \operatorname{Tr} \ln[(1-\beta)\mathbf{I} + \beta\mathbf{q}] = \lim_{n\to 0} \frac{1}{n} \{(n-1)\ln(1-\beta+\beta q) + \ln[1-\beta-(n-1)\beta q]\} \tag{A.1}$$
$$= \ln(1-\beta+\beta q) - \beta q/(1-\beta+\beta q).$$

With r and q given by (3.1),

$$\lim_{n\to 0} \frac{1}{n} \sum_{\rho \neq \sigma} r_{\rho\sigma} q_{\rho\sigma} = -qr. \tag{A.2}$$

Finally,

$$\frac{1}{n}\ll \ln \mathrm{Tr}_s \exp \beta H_\xi \gg_\xi$$

$$= \frac{1}{n}\left\langle\!\!\left\langle \ln \mathrm{Tr}_s \exp\left(\frac{1}{2}\alpha\beta^2 r \sum_{\rho\delta} s^\rho s^\delta - \frac{1}{2} n\alpha\beta^2 r\right)\right.\right.$$

$$\left.\left. \cdot \exp\left[\beta \sum_\rho (\vec{m}+\vec{h})\cdot \vec{\xi} s^\rho\right]\right\rangle\!\!\right\rangle_\xi$$

$$= -\frac{1}{2}\alpha\beta^2 r$$

$$+\frac{1}{n}\left\langle\!\!\left\langle \ln \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2 + n \ln 2ch\beta[\sqrt{\alpha r}z + (\vec{m}+\vec{h})\cdot\vec{\xi}]\right\}\right\rangle\!\!\right\rangle_\xi$$

$$\xrightarrow[n\to 0]{} -\frac{1}{2}\alpha\beta^2 r + \ll \ln 2ch\beta[\sqrt{\alpha r}\, z + (\vec{m}+\vec{h})\cdot\xi]\gg. \tag{A.3}$$

Appendix B

In this Appendix we present an evaluation of the eigenvalue corresponding to the replicon mode, in the replica symmetric phase. The stability matrix is of dimension $n(n-1)\times n(n-1)$ and has the general structure

$$C = \begin{bmatrix} A^{\alpha\beta,\gamma\delta} & \delta^{\alpha\beta,\gamma\delta} \\ \delta^{\alpha\beta,\gamma\delta} & B^{\alpha\beta,\gamma\delta} \end{bmatrix} \tag{B.1}$$

with

$$A^{\rho\sigma,\gamma\delta} = \frac{\partial^2(nf)}{\partial q_{\rho\sigma}\partial q_{\gamma\delta}} = -\alpha\beta^2[(1-\beta-\beta q)^{-1}_{\rho\gamma}(1-\beta-\beta q)^{-1}_{\sigma\delta}$$

$$+ (1-\beta-\beta q)^{-1}_{\rho\delta}(1-\beta-\beta q)^{-1}_{\sigma\gamma}], \tag{B.2a}$$

$$B^{\rho\sigma,\gamma\delta} = \frac{\partial^2(nf)}{\partial q_{\rho\sigma}\partial q_{\gamma\delta}} = -\alpha^2\beta^3 \ll [\langle S^\rho S^\sigma S^\gamma S^\delta\rangle - \langle S^\rho S^\sigma\rangle\langle S^\gamma S^\delta\rangle] \gg \tag{B.2b}$$

$$\delta^{\rho\sigma,\gamma\delta} = \frac{\partial^2(nf)}{\partial q_{\rho\sigma}\partial r_{\gamma\delta}} = \alpha\beta(\delta_{\rho\gamma}\delta_{\sigma\delta} + \delta_{\rho\delta}\delta_{\sigma\gamma}). \tag{B.2c}$$

The matrix A has three different types of matrix elements, when computed in the replica symmetric state, and those depend in turn on two parameters. They are

$$A^{\rho\sigma,\rho\sigma} = -\alpha\beta(C^2_{\rho\rho} + C^2_{\rho\sigma}), \tag{B.3a}$$

$$A^{\rho\sigma,\rho\gamma} = -\alpha\beta(C_{\rho\rho}C_{\rho\sigma} + C^2_{\rho\sigma}), \tag{B.3b}$$

$$A^{\rho\sigma,\gamma\delta} = -\alpha\beta(2C^2_{\rho\sigma}), \tag{B.3c}$$

and

$$C_{\rho\sigma} = \frac{\beta q}{[1-\beta(1-q)]^2}, \qquad \rho \neq \sigma, \tag{B.4a}$$

$$C_{\rho\rho} = C_{\rho\sigma} + [1-\beta(1-q)]^{-1}, \tag{B.4b}$$

where q has the usual meaning it had in the replica symmetric theory (see, e.g., Sect. 3).

The averages ($\langle \cdots \rangle$) in Eq. (B.2b), for the matrix B, are taken with the hamiltonian equation (2.13), in the replica symmetric phase.

Without elaborate justification we probe the stability of the replicon mode (10), namely,

$$\delta q_{\rho\sigma} = \eta_{\rho\sigma}, \qquad \delta r_{\rho\sigma} = x\eta_{\rho\sigma} \tag{B.5}$$

with the condition

$$\sum_{\sigma} \eta_{\rho\sigma} = 0 \qquad \text{for all} \quad \rho.$$

This condition is ensured, for all values of n, by

$$\eta_{\rho\sigma} = \eta, \qquad \rho, \sigma \neq 1, 2, \tag{B.6a}$$

$$\eta_{1\rho} = \eta_{2\rho} = \tfrac{1}{2}(3-n)\,\eta, \qquad \rho \neq 1, 2 \tag{B.6b}$$

$$\eta_{12} = \tfrac{1}{2}(2-n)(3-n)\,\eta, \tag{B.6c}$$

$$\eta_{\rho\rho} = 0. \tag{B.6d}$$

The eigenvalue equation for the replicon mode reads

$$\sum_{\gamma\delta} (A^{\rho\sigma,\gamma\delta} + x\delta^{\rho\sigma,\gamma\delta})\,\eta_{\gamma\delta} = \lambda\eta \qquad (\rho, \sigma \neq 1, 2), \tag{B.7}$$

$$\sum_{\gamma\delta} (xA^{\rho\sigma,\gamma\delta} + \delta^{\rho\sigma,\gamma\delta})\,\eta_{\gamma\delta} = \lambda x\eta \qquad (\rho, \sigma \neq 1, 2). \tag{B.8}$$

Substituting in these equation A, B, and δ, as given by (B.3)–(B.6), taking the limit $n \to 0$, and denoting

$$\tilde{\lambda} \equiv \lambda/2\alpha\beta,$$

one arrives at the two couplled equations

$$x = \tilde{\lambda} + \frac{1}{[1-\beta(1-q)]^2}, \tag{B.9}$$

$$x[\alpha\beta^2 \ll (1 - \langle S \rangle^2)^2 \gg + \tilde{\lambda}] = 1, \tag{B.10}$$

corresponding to (B.7) and (B.8).

The resulting quadratic equation for $\tilde{\lambda}$ has the solutions

$$\tilde{\lambda}_{\pm} = -\tfrac{1}{2}(u+v) \pm [\tfrac{1}{4}(u+v)^2 + 1 - uv]^{1/2} \tag{B.11}$$

with

$$u \equiv \alpha\beta^2 \langle\langle (1 - \langle S \rangle^2)^2 \rangle\rangle, \tag{B.12}$$

$$v \equiv [1 - \beta(1-q)]^{-2}. \tag{B.13}$$

For $T > T_g = 1 + \sqrt{\alpha}$, $q = 0$, and

$$uv = \frac{\alpha\beta^2}{(1-\beta)^2} < 1.$$

Hence, $\lambda_- < 0$ and $\lambda_+ > 0$. The eigenvalue λ_- does not change sign at low temperature and its sign *must* be corrected by a proper choice of the integration contour.

It is the change of sign of λ_+ which signals the instability to RSB. This change of sign occurs when $uv = 1$, leading to Eq. (7.6).

REFERENCES

1. J. J. HOPFIELD, *Proc. Natl. Acad. Sci. U.S.A.* **79** (1982), 2554; **81** (1984), 3088; J. J. HOPFIELD, D. I. FEINSTEIN, AND R. G. PALMER, *Nature* **304** (1983), 158.
2. W. A. LITTLE, *Math. Biosci.* **19** (1974), 101; W. A. LITTLE AND G. L. SHAW, *Math. Biosci.* **39** (1978), 281.
3. W. S. MCCULLOCH AND W. A. PITTS, *Bull. Math. Biophys.* **5** (1943), 115.
4. D. J. AMIT, H. GUTFREUND, AND H. SOMPOLINSKY, *Phys. Rev. A* **32** (1985), 1007.
5. D. O. HEBB, "The Organization of Behavior," Wiley, New York, 1949.
6. D. J. AMIT, H. GUTFREUND, AND H. SOMPOLINSKY, *Phys. Rev. Lett.* **55** (1985), 1530.
7. G. WEISBUCH AND F. FOGELMAN-SOULIÉ, *J. Phys. Lett.* **46** (1985), L623.
8. R. J. MCELIECE, E. C. POSNER, E. R. RODMICH, AND S. S. VENKATESH, Caltech preprint, (1986).
9. S. KIRKPATRICK AND D. SHERRINGTON, *Phys. Rev. B* **17** (1978), 4384.
10. J. R. L. DE ALMEIDA AND D. J. THOULESS, *J. Phys. A* **11** (1978), 983.
11. G. PARISI, *Phys. Rev. Lett.* **50** (1983), 1946.
12. M. MEZARD, G. PARISI, N. SOURLAS, G. TOULOUSE, AND M. VIRASORO, *J. Phys.* **45** (1984) 843.
13. H. SOMPOLINSKY, *Phys. Rev. Lett.* **47** (1981), 935.
14. W. KINZEL, "Remanent Magnetization of the Infinite Range Ising Spin Glass," IFK, Julich preprint, 1985.
15. S. F. EDWARDS AND P. W. ANDERSON, *J. Phys. F* **5** (1975), 965.
16. R. BAUSCH, *Z. Phys.* **254** (1972), 81.
17. H. SOMPOLINSKY, G. KOTLIAR, AND A. ZIPPELIUS, *Phys. Rev. Lett.* **52** (1984), 392.
18. H. SOMPOLINSKY, *Phys. Rev. A.* (RC), Sept., 1986.
19. G. TOULOUSE, S. DEHAENE, AND J. P. CHANGEUX, *Proc. Natl. Acad. Sci.* **83** (1986), 1695.
20. L. PERSONNAZ, I. GUYON, AND G. DREYFUS, *J. Phys. Lett.* **46** (1985), L359.
21. M. VIRASORO, Ultrametricity, Hopfield model and all that, *in* "Disordered Systems and Biological Organization" (E. Bienenstock, Ed.) les Houches, 1985; N. PARGA AND M. VIRASORO, The ultrametric organization of memories in a neural network, Trieste preprint, 1985.
22. V. S. DOTSENKO, *J. Phys. C* **18** (1985), L1017.
23. N. N. BOGOLYUBOV, **26** Physica (*Suppl.*) (1960), *S*1; and *Phys. Abh. S. V.* **6**, (1962), 229.