Neural networks with hierarchically correlated patterns

H. Gutfreund*
Institute for Theoretical Physics, University of California, Santa Barbara, California 93106
*and The Racah Institute of Physics, the Hebrew University, Jerusalem, 91904 Israel**
(Received 12 January 1987; revised manuscript received 9 September 1987)

The Hopfield model of neural networks is extended to allow for the storage and retrieval of hierarchically correlated patterns. The overlaps between these patterns form an ultrametric tree. Intermediate states, which serve as ancestors for the following levels, are generated at each level of the tree. The states belonging to each level are stored, by a modified learning rule, in a series of identical networks, one for each level. The retrieval of a particular pattern is preceded and assisted by the successive retrieval of its ancestors. The performance of this scheme is studied analytically and numerically.

I. INTRODUCTION

The binary version of Hopfield's standard model for associative memory[1] consists of a system of N Ising spins whose dynamics is governed by the Hamiltonian

$$H=-\tfrac{1}{2}\sum_{\substack{i,j\\(j\neq i)}} J_{ij}S_iS_j \ , \tag{1.1}$$

where $S_i=\pm 1$ represents the two possible states of the ith neuron, and the interactions J_{ij}, which connect pairwise all the neurons in the network, represent the synaptic efficacies between them. A set of p patterns $\{\xi_i^\mu\}$ $(i=1,\ldots,N;\ \mu=1,\ldots,p)$, in which ξ_i^μ is either $+1$ or -1, is embedded in the J_{ij}'s, via the "learning" rule

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

These patterns are the "memories" stored in the network. A fair amount of understanding of the properties of this model as an associative memory has been achieved by analytical analysis supplemented by numerical simulations.[2,3]

One of the restrictions of the model represented by (1.1) and (1.2) is that each ξ_i^μ is $+1$ or -1 with equal probability. Thus, in a large network, the average bias or "magnetization" of the learned patterns vanishes,

$$\frac{1}{N}\sum_i \xi_i^\mu \equiv \langle\langle \xi_i^\mu \rangle\rangle = 0 \ . \tag{1.3}$$

Moreover, there are no correlations between the patterns, namely,

$$\langle\langle \xi_i^\mu \xi_i^\nu \rangle\rangle = 0 \quad \text{for } \mu\neq\nu \ . \tag{1.4}$$

This is a very unsatisfactory situation. More realistic networks should be able to store correlated patterns. A model proposed recently by Personnaz *et al.*[4] and studied analytically by Kanter and Sompolinsky[5] is free of such restrictions and has the capability to store any set of linearly independent patterns. However, the price one pays there is a more complicated, nonlocal, dependence of the J_{ij}'s on the stored patterns.

A modified version of Hopfield's model, which incorporates the storage and retrieval of patterns with a finite bias, and hence with finite correlations between the patterns, was proposed recently by Amit *et al.*[6] (to be referred to as AGS). The main virtue of this model is that it retains the simplicity of the learning rule. Every element ξ_i^μ of the learned patterns is chosen independently with probability

$$P(\xi_i^\mu)=\tfrac{1}{2}(1+a)\delta(\xi_i^\mu-1)+\tfrac{1}{2}(1-a)\delta(\xi_i^\mu+1) \ , \tag{1.5}$$

where $-1<a<1$. For patterns generated in this way one gets

$$\langle\langle \xi_i^\mu \rangle\rangle = a \ , \tag{1.6}$$

$$\langle\langle \xi_i^\mu \xi_i^\nu \rangle\rangle = a^2 \quad \text{for } \mu\neq\nu \ . \tag{1.7}$$

Such biased patterns can be stored in a neural network by modifying the synaptic efficacies [Eq. (1.2)] to

$$J_{ij}=\frac{1}{N}\sum_\mu (\xi_i^\mu-a)(\xi_j^\mu-a) \ . \tag{1.8}$$

With this modification the stored patterns are retrieved with a small fraction of errors, up to a storage level of $p=\alpha_c(a)N$. However, the storage capacity, characterized by $\alpha_c(a)$, decreases sharply with a, roughly as $(1-|a|)^2$. In addition, one finds that the energy landscape becomes dominated by spurious states which have finite overlaps with several of the learned patterns. As the bias a increases, the number of such states increases, they become the absolute minima of the energy, and, for rather low values of a, the critical storage level α_c of the spurious states becomes significantly higher than that of the stored patterns. Finally, and related to the previous problem, one finds that, although the learned patterns themselves are stable below $\alpha_c(a)$ (up to a small fraction of errors), their basins of attraction, and hence the fault tolerance of the network, decrease sharply with a. All these problems are overcome in AGS by also modifying the dynamics, so that it becomes con-

sistent with the bias of the stored patterns. This is done by constraining the dynamical process to states with an overall magnetization Na. Such a constraint may be imposed rigidly by requiring that at each state

$$\frac{1}{N}\sum S_i = a \ , \tag{1.9}$$

or by adding a penalty term for deviations from the rigid constraint to the Hamiltonian

$$H = -\tfrac{1}{2}\sum_{i,j} J_{ij} S_i S_j + \frac{g}{2N}\left[\sum_i S_i - Na\right]^2 . \tag{1.10}$$

For large values of g one recovers the rigid constraint.

The model studied by AGS is suited to treat the minimal correlations induced by the constant bias of the patterns. In the present paper, I generalize this approach to a set of hierarchically correlated patterns.

The organization of objects with well-defined relations of similarity into a hierarchical (ultrametric) tree arises naturally in many cases of data classification and analysis.[7] Several proposals to incorporate such a structure in a neural network have been made before. One of the main goals of the "selectionist" approach, promoted by Toulouse *et al.*,[8] is to make use of the ultrametric space spanned by the ground states of the SK (Sherrington-Kirkpatrick[20]) spin-glass for the storage of hierarchically correlated patterns. Two different models for storing a hierarchical tree of memories in a neural network were proposed by Parga and Virasoro[9] and by Dotsenko.[10] Both models are motivated by the recent understanding of the detailed microstructure of the ultrametric properties of the SK spin-glass.[11] Parga and Virasoro propose a learning rule which is closely related to the form of the couplings in the SK model found in Ref. 11. Dotsenko, also following in the footsteps of Ref. 11, describes the hierarchy of patterns in terms of spin clusters with given magnetizations, and proceeds to assign a hierarchy of connections—first within a cluster and then between clusters at higher and higher levels of the tree. The performance of these models with respect to the storage capacity, quality of retrieval, speed of convergence, and size of the basins of attraction has not been demonstrated. In fact, numerical simulations[12,13] of the last two models indicate a very poor storage capacity. This can be anticipated from a signal-to-noise analysis (similar to that performed in Sec. III A). An interesting extension and generalization of the Parga-Virasoro scheme has recently been suggested by Feigelman and Ioffe.[14]

The scheme proposed here is different from the above both in the learning rule and in the dynamics of retrieval. It retains the simplicity of the J_{ij}'s, which are only modified to indicate the cluster to which a particular pattern belongs. The paper begins with the description of a particular procedure for generating a hierarchical tree of patterns (Sec. II). The patterns produced at the highest level are the memories to be stored in the network. Analysis of the proposed learning rule (Sec. III) reveals the same problems encountered in AGS before the introduction of constrained dynamics. To avoid these problems in the present case, I propose a hierarchy of dynamical constraints which are implemented in an architecture based on several networks—one for each level of the tree (Sec. IV). The retrieval of a particular memory is achieved in the last network and is assisted by the retrieval of its ancestors at the various levels by the preceding networks. The analysis of Secs. III and IV is supplemented by numerical simulations in Sec. V.

Although I do not want to venture a neurobiological interpretation of the proposed scheme, I believe that a succession of networks, each of which plays a partial role in performing a certain task, has neurobiological appeal.

II. GENERATION OF A HIERARCHICAL TREE OF PATTERNS

One can derive many procedures to generate a set of hierarchically correlated patterns. Two such procedures are described in Ref. 9. They are similar to the procedure described below. There is, however, one difference. In Ref. 9 only the patterns themselves, namely, the points at the highest level of the hierarchical tree have any significance, while the branching points remain formal constructs and play no role in the model. In contrast, in the model studied here the branching points are represented by real states, which serve as ancestors for the subsequent generations and appear explicitly in the learning rule and in the dynamics of retrieval.

The hierarchical tree of patterns is constructed as follows. At the first level of the hierarchy one generates p_1 patterns $\{\xi_i^\mu\}$, $\mu = 1, \ldots, p_1$, where every component ξ_i^μ is chosen independently with the probability given by Eq. (1.5). These patterns serve as ancestors for the next level. At the second level a new correlation parameter b, $0 < b < 1$, is specified. One then generates from each pattern $\{\xi_i^\mu\}$, p_2 descendents $\{\xi_i^{\mu\nu}\}$, $\nu = 1, \ldots, p_2$, choosing their components with the probability

$$P(\xi_i^{\mu\nu}) = \tfrac{1}{2}(1+\xi_i^\mu b)\delta(\xi_i^{\mu\nu}-1) + \tfrac{1}{2}(1-\xi_i^\mu b)\delta(\xi_i^{\mu\nu}+1) \ . \tag{2.1}$$

This rule implies that a component ξ_i in a pattern belonging to the group descending from the μth ancestor has a higher probability to be equal to ξ_i^μ than to $-\xi_i^\mu$ (since $b > 0$). This process can be continued by specifying at the kth level of the hierarchy a correlation parameter a_k, and using each of the p_{k-1} patterns $\{\xi_i^{\alpha_1 \cdots \alpha_{k-1}}\}$ generated at the previous level to produce p_k descendents $\{\xi_i^{\alpha_1 \cdots \alpha_k}\}$, using the probability law (2.1).

The correlations between the patterns in the second generation are given by

$$\langle\langle \xi_i^{\mu\nu}\xi_i^{\mu'\nu'}\rangle\rangle = b^2 \quad (\mu=\mu', \ \nu\neq\nu')$$
$$= a^2 b^2 \quad (\mu\neq\mu') \ . \tag{2.2}$$

Thus, the patterns are grouped into clusters with high correlations between patterns within the same cluster and lower correlations between patterns in different clusters. The correlations of the patterns with the ancestors are

$$\langle\langle \xi_i^{\mu\nu}\xi_i^{\mu'}\rangle\rangle = b \quad (\mu=\mu')$$
$$= a^2 b \quad (\mu\neq\mu') \tag{2.3}$$

and the average bias of the patterns is

$$\langle\langle \xi_i^{\mu\nu}\rangle\rangle = ab \ . \tag{2.4}$$

In the case $a=0$, there is no correlation between patterns in different clusters ($\mu\neq\mu'$) and only patterns within the same cluster are correlated. Moreover, in this case, the patterns have on the average the same number of $+1$ and -1 bits.

The generalization of Eqs. (2.2) and (2.4) to a hierarchy of k levels is

$$\langle\langle \xi_i^{\alpha_1,\ldots,\alpha_k}\xi_i^{\alpha_1',\ldots,\alpha_k'}\rangle\rangle = a_k^2 a_{k-1}^2 \cdots a_{k-l}^2$$
$$(\alpha_1=\alpha_1',\ldots,\alpha_{k-l-1}=\alpha_{k-l-1}') \tag{2.5}$$

and

$$\langle\langle \xi_i^{\alpha_1\cdots\alpha_k}\rangle\rangle = \prod_{j=1}^{k} a_j \ , \tag{2.6}$$

where a_j is the correlation parameter at the jth level. Equation (2.5) represents a set of overlaps typical of the ultrametric clustering of the generated patterns.

III. THE LEARNING RULE

The treatment will henceforth be restricted to a hierarchy of two levels. The generalization to any number of levels is straightforward. The correlations between the patterns are determined by the parameters a and b. The number of patterns is $p=p_1p_2$ (p_1 clusters, p_2 patterns in each cluster) and the storage level is characterized by the parameter $\alpha=p/N$. The following learning rule is proposed to store the patterns in the network:

$$J_{ij}=\frac{1}{N}\sum_{\mu,\nu}(\xi_i^{\mu\nu}-\xi_i^{\mu}b)(\xi_j^{\mu\nu}-\xi_j^{\mu}b)$$
$$(\mu=1,\ldots,p_1;\ \nu=1,\ldots,p_2) \ . \tag{3.1}$$

Note that the model discussed in AGS [Eqs. (1.5) and (1.8)] is a special case of the present scheme, in which p patterns are generated from the single ancestor $\{\xi_i^1\}=(1,1,\ldots,1)$.

A. Signal-to-noise analysis

Let us first examine the conditions for the stability of the stored patterns as derived from a signal-to-noise ratio analysis. The local field on neutron i in configuration $\{\xi_i^{11}\}$ is

$$h_i \equiv \sum_{\substack{i,j\\(j\neq i)}} J_{ij}\xi_j^{11} = \xi_i^{11}(1-\xi_i^1\xi_i^{11}b)(1-b^2)$$
$$+ \sum{}'_{\mu,\nu}\left[(\xi_i^{\mu\nu}-\xi_i^{\mu}b)\times\frac{1}{N}\sum_{\substack{i,j\\(j\neq i)}}(\xi_j^{\mu\nu}-\xi_j^{\mu}b)\xi_j^{11}\right], \tag{3.2}$$

where the apostrophe indicates that the term $\mu=1$, $\nu=1$, which appears separately, is excluded from the sum. The first term represents the signal. It is of the same sign as ξ_i^{11} and hence stabilizes the pattern. The second term constitutes the destabilizing noise. The particular choice of the synaptic efficacies implies, through Eqs. (2.2) and (2.3), that the average of the noise term vanishes. The mean square of the noise term is

$$R^2=\frac{(N-1)(p_1p_2-1)}{N^2}(1-b^2)^2\simeq\alpha(1-b^2)^2 \ . \tag{3.3}$$

The lowest value of the signal term is $S=(1-b^2)(1-|b|)$, and thus the signal-to-noise ratio is

$$\frac{S}{R}=\frac{(1-|b|)}{\sqrt{\alpha}} \ . \tag{3.4}$$

Following the arguments of Ref. 15 one concludes from the probability distribution of the noise that the patterns are perfectly stable for

$$\alpha<\alpha_c=\frac{(1-|b|)^2}{2\ln N} \ . \tag{3.5}$$

This is the storage capacity for perfect retrieval. It depends on the total number of patterns and not on how they are grouped into the different clusters. It does not depend on a, and is determined only by the parameter b, which represents the strongest correlation among the patterns. In its form, Eq. (3.5) is identical to the corresponding result in AGS.

The most important feature of Eq. (3.1) is that it ensures that the noise term in Eq. (3.2) has a zero average. Note that this term does not average to zero if one naively follows the prescription of AGS and subtracts in Eq. (3.1) the bias of the patterns ab instead of $\xi_i^{\mu}b$. Analysis of this case shows that the patterns become destabilized when p_2 exceeds

$$p_2^c=(1-ab)+\frac{(1-ab)(1-a^2b^2)}{(b^2-a^2b^2)} \ , \tag{3.6}$$

which, for $a=0$, reduces to

$$p_2^c=1+\frac{1}{b^2} \ . \tag{3.7}$$

This restriction to a finite, usually small, number of patterns within each cluster is independent of the size of the network.

B. Mean-field analysis

The properties of the network for finite $\alpha=p/N$ are derived from the free energy associated with the Hamiltonian

$$H=-\frac{1}{2N}\sum_{\substack{i,j\\(i\neq j)}}\sum_{\mu,\nu}(\xi_i^{\mu\nu}-\xi_i^{\mu}b)(\xi_j^{\mu\nu}-\xi_j^{\mu}b)S_iS_j \ . \quad (3.8)$$

The replica method is used to perform the quenched averaging over the ξ's. Thus, the free energy is computed from

$$\langle\langle Z^n\rangle\rangle = e^{-\beta\alpha Nn(1-b^2)} \times\left\langle\left\langle \mathrm{Tr}_{S^\rho}\exp\frac{\beta}{2N}\sum_{\substack{\mu,\nu,\\ \rho}}\left[\sum_i(\xi_i^{\mu\nu}-\xi_i^{\mu}b)S_i^\rho\right]^2\right\rangle\right\rangle , \quad (3.9)$$

where ρ is the replica index. The calculation follows in the footsteps of AGS. The quadratic term in the exponential is linearized by a Gaussian transformation. One or several patterns are singled out as candidates for condensation and the ξ's of the remaining patterns are averaged out. This average must, in the present case, be performed with some care because the components ξ_i^μ of the ancestors of the condensed patterns couple to the uncondensed patterns belonging to the same cluster. Without repeating the details of the calculation, I write down the resulting saddle-point equations for the order parameters, for the case of a single condensed pattern $\{\xi_i^{\mu\nu}\}$,

$$m^{\mu\nu}=\langle\langle(\xi^{\mu\nu}-\xi^{\mu}b)\tanh\beta[\sqrt{\alpha r}\,z+m^{\mu\nu}(\xi^{\mu\nu}-\xi^{\mu}b)]\rangle\rangle \ , \quad (3.10)$$

$$q=\langle\langle\tanh^2\beta[\sqrt{\alpha r}\,z+m^{\mu\nu}(\xi^{\mu\nu}-\xi^{\mu}b)]\rangle\rangle \ , \quad (3.11)$$

$$r=\frac{q(1-b^2)^2}{[1-\beta(1-b^2)(1-q)]^2} \ . \quad (3.12)$$

Now $\langle\langle\cdots\rangle\rangle$ denotes averaging over $\xi^{\mu\nu}$, ξ^{μ}, and over the Gaussian variable z. The average over $\xi^{\mu\nu}$ and ξ^{μ} is performed with probabilities, Eqs. (2.1) and (1.5), respectively. The order parameters q and r are the same as in Refs. 2 and 6: q is the Edwards-Anderson spin-glass order parameter and r is related to the random overlaps with all the uncondensed patterns. The retrieval order parameter $m^{\mu\nu}$ is given by

$$m^{\mu\nu}=\frac{1}{N}\sum_i(\xi_i^{\mu\nu}-\xi_i^{\mu}b)\langle S_i\rangle \ . \quad (3.13)$$

The quality of retrieval is characterized by the overlap of the equilibrium state $\{\langle S_i\rangle\}$ with the stored pattern,

$$\frac{1}{N}\sum_i\xi_i^{\mu\nu}\langle S_i\rangle=m^{\mu\nu}+\frac{b}{N}\sum_i\xi_i^{\mu}\langle S_i\rangle \ . \quad (3.14)$$

To calculate this quantity, one has first to solve Eqs. (3.10)–(3.12) and then to compute

$$\langle\langle\xi^{\mu}\langle S\rangle\rangle\rangle\equiv\frac{1}{N}\sum_i\xi_i^{\mu}\langle S_i\rangle = \langle\langle\xi^{\mu}\tanh\beta[\sqrt{\alpha r}\,z+m^{\mu\nu}(\xi^{\mu\nu}-\xi^{\mu}b)]\rangle\rangle \ . \quad (3.15)$$

To compare the mean-field equations, obtained here, with the results of AGS, I shall now take the zero-temperature limit. Performing all the averages, one finds that as $\beta\to\infty$, Eq. (3.10) becomes

$$m=\tfrac{1}{2}(1-b^2)[\mathrm{erf}(A_+)+\mathrm{erf}(A_-)] \ , \quad (3.16)$$

where the superscript (μ,ν) has been dropped, and

$$A_\pm=\frac{m(1\pm b)}{\sqrt{2\alpha r}} \ . \quad (3.17)$$

Equations (3.11) and (3.12) become

$$r=\frac{(1-b^2)^2}{[1-(1-b^2)C]^2} \ , \quad (3.18)$$

where

$$C\equiv\lim_{\beta\to\infty}\beta(1-q) =\frac{1}{\sqrt{2\pi\alpha r}}[(1+b)\exp(-A_+^2)+(1-b)\exp(-A_-^2)] \ . \quad (3.19)$$

Finally,

$$\langle\langle\xi S\rangle\rangle=\tfrac{1}{2}(1+b)\mathrm{erf}(A_-)-\tfrac{1}{2}(1-b)\mathrm{erf}(A_+) \ . \quad (3.20)$$

The general conclusions derived previously from the signal-to-noise analysis follow also from the mean-field theory: (a) The properties of the network are determined by $p=p_1p_2$, and not by p_1 and p_2 individually; (b) they depend only on strongest correlations in the system, characterized by b; (c) with the redefinition of the order parameter $m^{\mu\nu}$ [Eq. (3.13)] the mean-field equations are identical in form with those obtained in AGS (this is true at all T).

The implication of point (c) is that one encounters here all the problems found there: (i) low storage capacity, decreasing with b roughly as $(1-|b|)^2$; (ii) dominance of the energy landscape by spurious states; (iii) very small basins of attraction.

IV. CONSTRAINED DYNAMICS

In the attempt to retrieve a particular pattern $\{\xi_i^{\mu\nu}\}$, it is natural to restrict the dynamics to configurations which, like the pattern itself, satisfy

$$\frac{1}{N}\sum_i\xi_i^{\mu}S_i=b \ . \quad (4.1)$$

The problem with such a constraint is that one has to first identify the ancestor pattern $\{\xi_i^{\mu}\}$. This can be achieved in the following scheme. Assume a succession of identical networks, one for each level of the hierarchy. In the case of a two-level hierarchy there are two

such networks. The ancestor patterns are stored in the first one, using the learning rule (1.8), and the memories are stored in the second network with the learning rule (3.1). An input pattern $\{\xi_i^{\mu\nu}\}$, possibly with a fraction of wrong bits, is "shown" first to the first network. For a wide range of parameters a and b the input pattern will be in the basin of attraction of $\{\xi_i^{\mu}\}$ and will flow, usually very fast—after 2–3 updatings per site—to a stationary state in the close neighborhood (a small Hamming distance) of $\{\xi_i^{\mu}\}$. Roughly, the requirements are that b is not too small, so that the overlap with $\{\xi_i^{\mu}\}$ is sufficiently large, and that a is not too large, to ensure a clear separation from the basins of attraction of the other ancestor patterns. The lower bound on b depends on p_1/N, which will usually be well below the saturation level. For example, for $p_1/N=0.05$, one finds[16] $b \gtrsim 0.2$. The simplest case is when $a=0$. Otherwise, especially for low values of b and high values of p_1/N, one has to impose, during the retrieval process at this level, the constraint $\langle\langle S_i \rangle\rangle = a$.

Once $\{\xi_i^{\mu}\}$ has been retrieved, that information is transferred to the second network as an external field (or threshold bias) h_i on each neuron,

$$h_i = h\xi_i^{\mu} . \tag{4.2}$$

Note that to this end only N connections between the two networks are required. The input pattern is now used to start the dynamics at the second level, which proceeds in the presence of the external field h_i. The effect of this field on the mean-field equations (3.10), (3.11), and (3.15) is to add the term $\beta h \xi^{\mu}$ to the argument of the hyperbolic tangent. The $T=0$ equations remain the same, provided that Eq. (3.17) is modified to read

$$A_{\pm} = \frac{m(1\pm b)\mp h}{\sqrt{2\alpha r}} . \tag{4.3}$$

There is one value, $h=h_0$, for which Eq. (3.20) is equal to b. Equations (3.16)–(3.19) with $h=h_0$ were obtained in AGS. There, the parameter b (a in the notation of AGS) represents the bias of patterns and the equations describe the effect of the rigid constraint $\langle\langle S_i \rangle\rangle = b$. In the present case, of a different type of correlations between the patterns and modified J_{ij}, the same equation describes the rigid constraint, Eq. (4.1).

Figure 1 shows the storage capacity as a function of h for several values of b. The maximum is reached at $h=h_0$. The value of h_0 depends on b; however, if one operates below maximum storage there is a fairly broad range of values of h which are appropriate for a range of values of b. This is demonstrated in Fig. 2. At a storage level of $\alpha=0.1$, all the learned patterns can be retrieved for all values of b and h within the hatched area. For example, when $b=0.5$ the stored patterns are stable for $0.24 \le h \le 0.62$; however, their basins of attraction depend on h.

V. NUMERICAL SIMULATIONS

The dependence of the size of the basins of attraction on the various parameters was studied by numerical

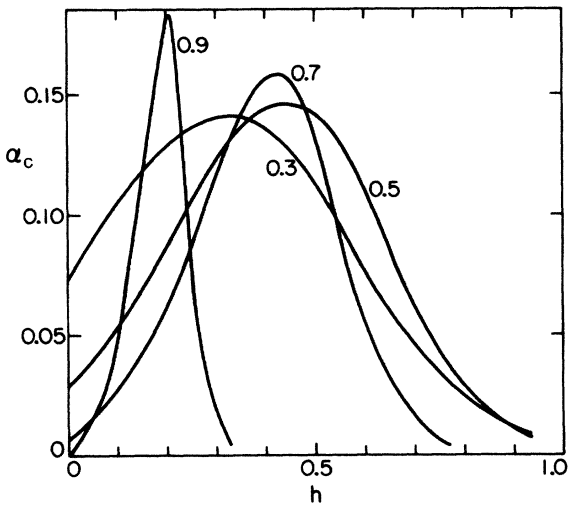

FIG. 1. Storage capacity α_c as a function of the external field conjugate to the ancestor state, for several values of the correlation parameter b.

simulations. In the simulations it was assumed that the ancestor pattern $\{\xi_i^{\mu}\}$ had already been retrieved. This defines the external field acting on the spins in the second network, in which $p=p_1p_2$ patterns generated with the procedure described in Sec. II have been stored. The dynamics is defined by

$$S_i = \mathrm{sgn}\left[\sum_{i,j} J_{ij}S_j + h\xi_i^{\mu}\right] , \tag{5.1}$$

where J_{ij} is given by Eq. (3.1). The initial configuration $\{S_i^{\mathrm{in}}\}$ has an overlap η with one of the patterns belonging to the cluster μ,

$$\frac{1}{N}\sum_i S_i^{\mathrm{in}}\xi_i^{\mu\nu} = \eta . \tag{5.2}$$

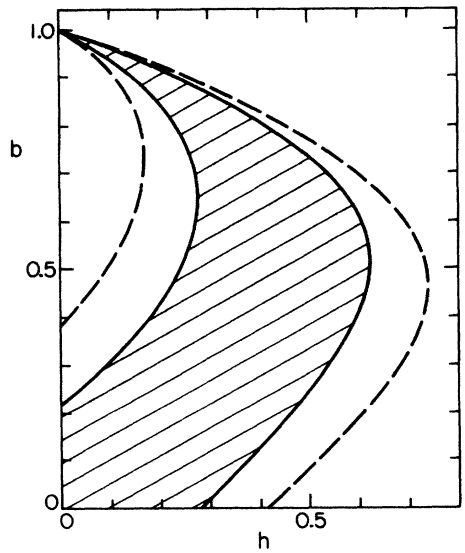

FIG. 2. Range of values of the external field h and the correlation parameter b, which ensure the stability of the learned patterns for a level of storage. (a) $\alpha=0.1$, the hatched area; (b) $\alpha=0.05$, the area between the dashed curves.

The spins are updated sequentially using Eq. (5.1) until a stationary state is reached. A successful retrieval is defined as the case when the final state is identical to the pattern $\{\xi_i^\mu\}$ up to 2% of error.

Figure 3 shows the results of such simulations for several values of h in the case of $N=500$, $p_1=5$, $p_2=10$, $a=0$, and $b=0.5$. Each entry in the figure represents an average over 500 trials. The optimal (for storage capacity) value h_0 is close to 0.45. When h increases above this value the basin of attraction decreases. It is clear why large values of h imply smaller basins of attraction. There are three contributions to the local field on spin i: a random disorienting field due to the overlap with the uncondensed patterns, a field which tends to align the spin configuration with the stored pattern, and the external field h_i, the effect of which is to align the spin configuration with the ancestor pattern. The latter is necessary to overcome, together with the other ordering field, the disorienting effect of the random field. In the present case this happens for $h \geq 0.24$. The two aligning fields now compete, and for large values of h the overlap between the initial configuration and the stored pattern has to be sufficiently large. Otherwise, the external field is dominant and the spin configuration flows away from the target pattern. At $h \geq 0.62$ this happens even when the initial configuration is the stored pattern itself. Figure 4 shows similar results for $N=1000$, $p_1=10$, $p_2=10$, $a=0$, $b=0.7$.

Extensive simulations have been performed for various combinations of the parameters a and b, and various pairs of p_1, p_2 for a given p. It was found that the behavior of the system does not depend significantly on a or on the grouping of the p patterns into separate clusters. One also finds that for $h \simeq 0.3$, one gets roughly the optimal basins of attraction for $0.2 \lesssim b \lesssim 8$. Thus, the value of h can be a property of the system, and does not have to be tuned in each case, provided that the data

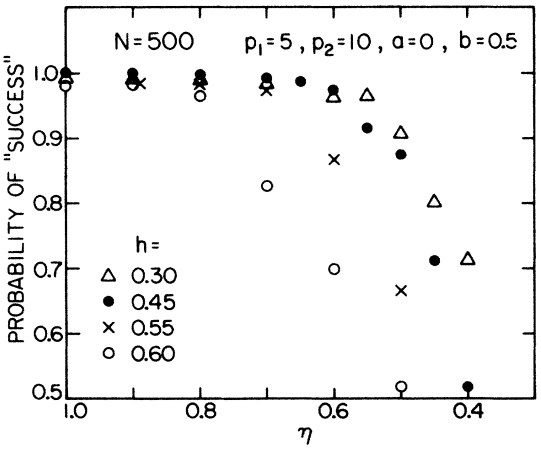

FIG. 3. Results of numerical simulations of the effect of h on the size of the basins of attraction. Successful retrieval is defined as convergence to a pattern with less than 2% error compared to the target pattern; η is the overlap of the initial configuration with the target pattern. Each entry is a result of 500 trials.

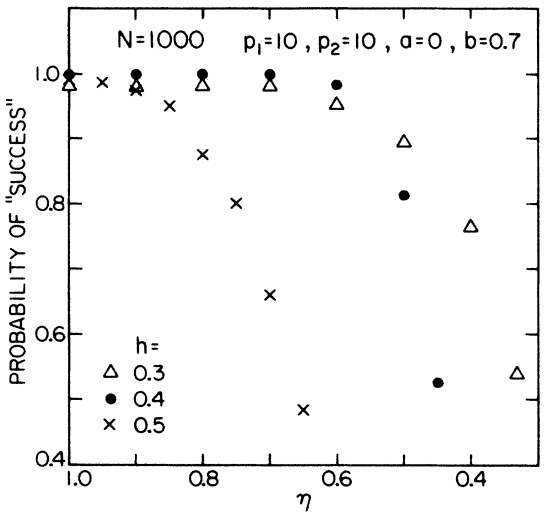

FIG. 4. Same as Fig. 3, for a different set of parameters.

are characterized by a correlation parameter b in this range. The typical size of the basin attraction for $p=0.1N$ is characterized by an initial overlap $\eta \simeq 0.6$. This is comparable to, though slightly worse than, what is found in the standard model with uncorrelated patterns,[16] where for $\alpha=0.1$ one gets $\eta \simeq 0.55$ for the same requirement on the quality of retrieval adopted here (less than 2% of error).

VI. DISCUSSION

In this paper I have outlined a particular scheme for storing and retrieving a set of hierarchically correlated patterns. The learning rule is a simple generalization of the one described in AGS. The new element is the retrieval process based on a series of networks, each one imposing dynamically a threshold bias on the subsequent network. Although the model was analyzed in detail only in the case of a two-level hierarchy, the generalization to any number of levels is straightforward. The model can also be extended to treat a more general hierarchical tree with different numbers of patterns in each cluster at each level, and also with a distribution of correlation parameters in each generation.

The model presented here depends on a hierarchical organization based on global relations between the patterns (a different type of a hierarchical organization is described below). Section II specifies a particular procedure for generating such a tree of patterns. In general, one faces the inverse problem. Suppose that a given set of patterns has an underlying ultrametric structure of Hamming distances. To store this set of patterns in the scheme proposed in this paper, one has to know the properties of the hierarchical tree, namely, the grouping of the patterns into the clusters at different levels, the correlation parameters, as well as the intermediate states at the branching points, which are not part of the given set of patterns. There exist classification algorithms for grouping data into a tree structure. They are referred to in Ref. 7. Once such a grouping is accomplished, one can easily deduce the correlation parameters from Eq.

(2.2). The ancestor pattern of a given group at the kth level of the tree is then labeled by $\alpha_1 \cdots \alpha_{k-1}$, and given by

$$\xi_i^{\alpha_1 \cdots \alpha_{k-1}} = \mathrm{sgn} \sum_{\alpha_k} \xi_i^{\alpha_1 \cdots \alpha_k} . \tag{6.1}$$

Coming back to the proposed structure, in which the states belonging to each level are stored in a series of identical networks, one should point out that this scheme is very uneconomical. The same size of network is required to store the p_1 ancestor patterns at the first level as is used at the highest level to store the $p_1 p_2 \cdots p_k$ memories (in a k-level tree). It is possible to modify the present model so as to make more efficient use of the available storage capacity at each level. One way to achieve that suggests itself in the case of several different sets of hierarchically correlated patterns. Each of them will be stored in a separate network, but they can share the same network for storing their ancestor patterns. This model will, in the general case, result in a hierarchical tree of networks. Without carrying the analogy too far, such a structure is reminiscent of the hierarchical organization of cortical areas in the visual system of the owl monkey,[17] where neurons in the lowest areas ($V1, V2$) respond to several domains of information such as shape and motion, while the response in the areas at higher levels is more specific.

The implementation of such a scheme of a hierarchy of networks requires a distinction between meaningful and meaningless activity of a neural network. Imagine that the ancestor states of two hierarchical trees, A and B, are stored at the first level, and suppose that responding to a particular stimulus, the network at this level converged to a pattern ξ_A^α. The descendents of this pattern are stored at the next level in network A, and one of them will be retrieved by the mechanism described in the present paper. But network B, which stores the descendents of the second tree, is exposed at the same time to the same external stimulus as A, and it will also converge to some fixed point, which is in general one of the exponentially many spurious states[2] and does not correspond to anything that has been learned before. How does the system "know" that the fixed point in network A is meaningful and the one reached in network B should be ignored as meaningless? Several schemes for such a distinction have been suggested. They are discussed in detail in Ref. 18. The simplest such scheme is based on the difference in the time required to reach a learned pattern starting within its basin of attraction, and the time needed to converge to one of the spurious fixed points, starting with a random input. An alternative scheme was proposed by Parisi.[19] Adding asymmetric connections affects very slightly the stability of the learned patterns, but has a drastic effect on the spurious states. In asymmetric networks, random inputs, which are not within the basin of a stored pattern, in general, lead to chaotic behavior and do not converge to fixed points.

Another model which makes better use of the available storage capacity at all levels of the tree is based on a different type of organization of the patterns into the respective clusters, when the correlations between patterns belonging to the same cluster are localized in a specific part of the site space, and not distributed over all the spins, as before. The extreme case is when all the patterns which belong to the same cluster have, say, the first N_1 bits in common. The remaining strings of $N - N_1$ bits characterizing the different patterns are uncorrelated. Likewise, there are no correlations between the vectors of N_1 bits defining the clusters. If there are more than two levels, then all the clusters, for which the first N_2 of these N_1 bits are identical, belong to the same supercluster. This classification can be extended to any number of levels. Such a hierarchy of patterns can be stored (with a different and even simpler learning rule) in a series of networks along the lines described before, except that now the ancestor states are shorter than the patterns at subsequent levels, and they are assigned to smaller and smaller networks. The retrieval of a particular pattern proceeds as before. First the ancestor pattern is detected in the smaller network and then an external field conjugate to this pattern is applied to the first N_1 spins of the full network. The retrieval of the target pattern is now assisted by the presence of an external field acting on a finite fraction of the spins. The details of this model will be described elsewhere.

ACKNOWLEDGMENTS

I am indebted to Haim Sompolinsky for helpful discussions and suggestions. I am also grateful for illuminating discussions on various aspects of the subject matter of the paper to Dana Ballard, John Hopfield, Christoff von der Malsburg, Sara Solla, and Miguel Virasoro. This research was supported in part by the National Science Foundation under Grant No. PHY82-17853, supplemented by funds from the National Aeronautics and Space Administration, and performed at the University of California at Santa Barbara.

*Permanent address.

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

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Phys. Rev. Lett. **55**, 1530 (1985); Ann. Phys. (NY) **173**, 30 (1987).

[3] H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).

[4] L. Personnaz, I. Guyon, and G. Dreyfus, J. Phys. (Paris) Lett. **46**, L359 (1985).

[5] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).

[6] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **35**, 2293 (1987).

[7] R. Rammal, G. Toulouse, and M. A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).

[8] G. Toulouse, S. Dehaene, and J. P. Changeux, Proc. Nat. Acad. Sci. U.S.A. **83**, 1695 (1986).

[9] N. Parga and M. A. Virasoro (unpublished); M. A. Virasoro, in *Disordered Systems and Biological Organization,* 1984 Les

Houches Lectures, edited by E. Bienenstock (North-Holland, Amsterdam, 1985).

[10]V. S. Dotsenko, J. Phys. C **18**, L1017 (1985).

[11]M. Mezard and M. A. Virasoro, J. Phys. (Paris) **46**, 1293 (1985).

[12]M. Moore (private communication).

[13]K. McLaughlin (private communication).

[14]M. V. Feigelman and L. B. Ioffe, Landau Institute Report No. 16, 1986 (unpublished).

[15]G. Weisbuch and F. Fogelman-Soulie, J. Phys. (Paris) Lett. **46**, L623 (1985).

[16]Detailed discussion of the basins of attraction of the Hopfield model and some of its modified versions will be presented elsewhere.

[17]D. Ballard, Behav. Brain Sci. **9**, 67 (1986).

[18]D. J. Amit, in *Proceedings of the Conference on Physics of Structure Formation, Tübingen, 1986,* edited by W. Guttinger (Springer-Verlag, Berlin, 1987).

[19]G. Parisi, J. Phys. A **19**, L675 (1986).

[20]S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).