Journal of Statistical Physics, Vol. 64, Nos. 3/4, 1991

Hierarchical Neural Networks for the Storage of Correlated Memories

Varsha Deshpande[1] and Chandan Dasgupta[1]

Received December 3, 1990; final March 20, 1991

A class of hierarchical neural network models introduced by Dotsenko for the storage and associative recall of strongly correlated memories is studied analytically and numerically. In these models, patterns stored in higher levels of the hierarchy represent generalized categories and those stored in lower levels describe finer details. We first show that the models originally proposed by Dotsenko have a serious flaw: they are not able to detect or correct errors in categorization which may be present in the input. We then describe three different models which attempt to overcome this shortcoming of the original models. In the first model, the interaction between different levels of the hierarchy has the form of an external field conjugate to memories stored in the lower level. In the second model, a three-spin interaction term is included in addition to the usual binary interactions of the Hopfield type. The third model makes use of a time delay mechanism to induce, if necessary, transitions between memory states and their complements. Detailed analytical and numerical studies of the performance of these models are presented. Our analysis shows that all three models are able to detect and also to correct in varying degrees any error in categorization that may be present in the input pattern.

KEY WORDS: Neural networks; associative memory; hierarchically correlated patterns; statistical mechanics of Ising systems; replica formalism; numerical simulation.

1. INTRODUCTION

Recently, much interest has been focused on neural network models of content-addressable memory.(1) A neural network is a large, highly connected assembly of simple computing elements (neurons). In the simplest models, each neuron is assumed to be a two-state threshold device having outputs

[1] Department of Physics, Indian Institute of Science, Bangalore 560 012, India.

$+1$ or -1, representing the active or the quiescent state, respectively. Such a neuron may be represented by an Ising spin. The state of the network at any instant of time is represented by the configuration of these Ising variables at that instant. The information (memories) stored in the network is embedded in the interconnections (synaptic efficacies) among the neurons. The time evolution of the network is governed by an assumed dynamics of the individual neurons. The system behaves like a content-addressable memory if the configurations representing the stored memories are locally stable attractors of the assumed dynamics. There are reasons to believe that such networks provide highly simplified models of some of the collective computational properties exhibited by the nervous system.

The neural network models which have received the greatest attention from physicists belong to a class of which the original Hopfield model[2] is the simplest example. In these models, the synaptic interaction matrix is assumed to be symmetric with zero as diagonal elements. One may then define an "energy function" (Hamiltonian) for the network which has the property that the most commonly assumed dynamics of the neurons corresponds to the rule that the state of a neuron is changed (the corresponding Ising spin is flipped) only if the energy is decreased in the process. Such a network functions as a content-addressable memory if the interaction matrix is chosen so as to make the memorized configurations local minima of the associated energy function. The performance of networks of this kind depends on many factors, such as the prescription (learning rule) used to define the synaptic matrix, the assumed dynamics of the neurons, and the number and statistical properties of the stored memories. In the original Hopfield model, the memories were assumed to be random binary strings (strings of elements each of which takes on the values $+1$ and -1 with equal probability) and the so-called "generalized Hebb rule" was used to construct the synaptic matrix. Extensive analytical and numerical studies[3,4] have shown that this model functions as an associative memory if the number of stored memories does not exceed about 15% of the total number of neurons. This restriction on the storage capacity is a major shortcoming of the Hopfield model. Another problem with the Hopfield model arises from the fact[1] that it performs very poorly if the patterns to be stored have strong correlations among themselves.

During the last few years, several attempts have been made to overcome these shortcomings of the Hopfield model. In particular, several neural network models with a hierarchical organization of correlated memories have been proposed.[5] The hierarchical (treelike) structure may be embedded either in the construction of the synaptic matrix or in the architecture of the network itself. The idea of a hierarchical organization of memorized data in which the memories are first grouped into categories, which are in turn grouped into supercategories and so on, is an intuitively appealing one. Also, there exists some neurobiological evidence[6] suggesting that the processing of information in the nervous system is performed in a hierarchical fashion. The known ultrametric organization of the low-lying energy states of certain long-ranged spin-glass models (see, e.g., ref. 7) has been an added incentive for the construction of hierarchical neural network models. In this paper, we consider a class of models proposed by Dotsenko[8] for the storage of an exponential number of highly correlated memories in a hierarchical structure. The basic idea in this approach is to divide the network into a number of "clusters" of neurons (Ising spins). These clusters form the lowest level of the hierarchy. A certain number of random binary patterns are stored in each of these clusters by using the Hebb rule. The signs of the sums of the spins (the magnetizations) in each of these clusters form the effective "spins" in the next level of the hierarchy. Again, these effective spins are grouped into clusters and several patterns are stored in each of these clusters. This construction may be continued to an arbitrary number of levels. For simplicity, we shall consider networks with only two levels. In that case, there is just one cluster in the second level. The patterns stored in this cluster represent different categories. The interaction matrix is constructed in such a way that the spins in the lower level clusters tend to converge to stored patterns which are consistent with one of the categories in the sense that the sign of the magnetization of the state reached by each lower-level spin-cluster matches that of the corresponding bit of one of the patterns stored in the upper level. An interesting visual representation of networks of this type may be given in terms of an analogy with image reconstruction. Let us consider an image composed of pixels each of which can be in one of the two states, black or white. These two states are identified with the $+1$ and -1 states of an Ising spin representing a two-state neuron in the lower level of the network. The pixels are divided into groups which are identified with the spin clusters in the lower level of the network. The sign of the magnetization of a cluster then represents the overall blackness or whiteness of the corresponding group of pixels, and the patterns stored in the upper level represent "coarse-grained" views of the image. Convergence to appropriate patterns in the lower-level clusters corresponds to the retrieval of finer details consistent with one of the stored coarse-grained patterns. This process is similar to a recently proposed renormalization-group approach[9] to image processing problems. It is also interesting to note that networks of this type have certain similarities with the hierarchical organization of neurons in the visual cortex of the brain.[10]

In the papers of Dotsenko,[8] two different prescriptions for the construction of networks with the properties described above were given. In this paper, we point out the fact that networks constructed according to these prescriptions have a serious shortcoming if, as considered in ref. 8, the patterns stored in each of the lowest-level clusters are chosen to be random binary strings. We show that such networks have locally stable states in which the lower-level clusters converge to memory states whose magnetizations do not have the correct signs as specified by the patterns stored at the upper level. If the interaction parameters are chosen to ensure convergence to stored patterns in the lower-level clusters, then all configurations in which one has convergence to memory states in the lower-level clusters are locally stable, irrespective of whether the signs of the magnetizations of these clusters match one of the stored patterns in the upper level or not. These networks thus do not have the capability of correcting errors in categorization. For an input configuration which is close to memory states in the lower level, but contains errors in the next level, these networks evolve to a stable state in which the errors in the upper level are not corrected. One of the attractive features of hierarchical memory models is the possibility of having a "hierarchy of errors." It is desirable for the network to have the property that in the process of associative recall, errors in determining the category should be less likely to occur than errors in the retrieval of finer details. The models proposed in ref. 8 do not achieve this goal. The arguments leading to this conclusion are described in Section 2. We also present numerical results supporting the analytic arguments. In subsequent sections, we describe three different ways of overcoming this shortcoming. In Section 3, we consider a network in which the interactions between the two layers correspond to the application of a field conjugate to patterns stored in the lower-level clusters. We show that this prescription leads to the desired behavior if the number of patterns stored in each of these clusters is small. We also present analytic and numerical results on the behavior of this model in the presence of fast synaptic noise ("finite temperature"). In Section 4, we consider a model in which a three-spin interaction is introduced in the lower-level clusters in order to break the symmetry between a stored pattern and its complement. This allows the network to choose the proper states in the lower-level clusters. An analytic calculation of the storage capacity of this network is described. We also present results of numerical simulations which demonstrate that this model functions as desired. Section 5 contains the description of a third model in which a set of connections with time delays are used to produce transitions between stored patterns and their complements, so that patterns consistent with one of the chosen categories may be selected at the lower level. The proper functioning of this model is also confirmed by simulations. Finally, in Section 6, we summarize the main results and discuss a few possible extensions of this work.

After the completion of this work, we came across a couple of papers which point out the shortcoming of the original Dotsenko model discussed here. In the paper by Dotsenko and Tirozzi,[11] the problem of convergence to patterns with wrong signs of the magnetizations of the lower-level cluster is avoided by storing patterns with large magnetizations in each of these clusters. The network studied by them is fundamentally different from the one considered here because in their model, the patterns stored in each lower-level cluster are strongly correlated (they in fact consider the limit of maximal correlations), whereas two patterns stored in a lower-level cluster of the model studied here are uncorrelated on the average. Thus, the analysis of Dotsenko and Tirozzi does not have much bearing on the problem we are considering. Willcox[12] proposes to overcome the problem in the original model by introducing in the dynamics a "tunneling" process in which the signs of all the spins in a lower-level cluster are reversed simultaneously. In contrast, we have concentrated here on the construction of models in which the simple "single-spin-flip" dynamics assumed in most neural-network models is retained and the interactions are chosen in order to achieve the desired behavior.

2. THE DOTSENKO MODEL AND ITS LIMITATIONS

We consider a two-level hierarchy in which the lower level consists of n_2 clusters, each containing n_1 neurons (Ising spins). These neurons are represented by the Ising variables $\{\sigma_i^\alpha\}$, $i = 1, 2,..., n_1$ and $\alpha = 1, 2,..., n_2$. In each of these clusters, p_1 random binary patterns (memories) are stored by using the Hebb rule. Thus, the intracluster interaction matrix has the form

$$J_{ij}^\alpha = \frac{1}{n_1} \sum_{\mu=1}^{p_1} \xi_i^{\mu\alpha} \xi_j^{\mu\alpha}, \qquad i \neq j \tag{1}$$

where μ labels the p_1 different memory states in each cluster and $\xi_i^{\mu\alpha} = +1$ or -1 with equal probability. We consider large values of p_1 and n_1, with the ratio $\alpha_1 = p_1/n_1$ having a value much smaller than α_c, the critical value calculated[4] for the Hopfield model. The construction of the second level of the hierarchy may be done in two different ways. In the first version of the model (Model I), only one set of neurons $\{\sigma_i^\alpha\}$ is present. The signs of the total spins (magnetizations) of the individual clusters form the effective spins $\{\rho_\alpha\}$, $\alpha = 1, 2,..., n_2$, in the second level:

$$\rho_\alpha = \text{sign}\left(\sum_{i=1}^{n_1} \sigma_i^\alpha\right) \tag{2}$$

The next step is to store a certain number p_2 (n_2, $p_2 \gg 1$, $p_2/n_2 \ll \alpha_c$) of random binary patterns $\{\phi_\alpha^\nu\}$, $\nu = 1, 2,..., p_2$ and $\alpha = 1, 2,..., n_2$, which represent the categories in the upper level. This should be done in a way that ensures that only states with $\rho_\alpha = P\phi_\alpha^\nu$, $\alpha = 1, 2,..., n_2$, $P = \pm 1$, are stable states of the dynamics of the network. In the original model, this goal is supposed to be achieved by using a Hebb-like rule to define an intercluster interaction matrix

$$J'_{\alpha\beta} = \frac{a}{n_2} \sum_{\nu=1}^{p_2} \phi_\alpha^\nu \phi_\beta^\nu, \qquad \alpha \neq \beta \tag{3}$$

where a is an adjustable parameter. The total Hamiltonian for the network then has the form

$$H = -\sum_{\alpha=1}^{n_2} \sum_{i>j} J_{ij}^\alpha \sigma_i^\alpha \sigma_j^\alpha - \sum_{\alpha>\beta} J'_{\alpha\beta} \left(\frac{1}{\sqrt{n_1}} \sum_i \sigma_i^\alpha\right)\left(\frac{1}{\sqrt{n_1}} \sum_j \sigma_j^\beta\right) \tag{4}$$

In the second version of the model (Model II) the upper level of the network consists of a different set of neurons ρ_α, $\alpha = 1, 2,..., n_2$. These neurons interact among themselves via the Hebb rule matrix $J'_{\alpha\beta}$ defined in Eq. (3). The connection between the two levels, which should ensure that each lower-level cluster converges only to memory states with the correct sign of the total magnetization, is supposed to be provided by a ferromagnetic interaction of each spin ρ_α in the upper level with all the spins $\{\sigma_i^\alpha\}$, $i = 1,..., n_1$, of one of the lower-level clusters. The Hamiltonian is then given by

$$H = -\sum_{\alpha=1}^{n_2} \sum_{i>j} J_{ij}^\alpha \sigma_i^\alpha \sigma_j^\alpha - \sum_{\alpha>\beta} J'_{\alpha\beta} \rho_\alpha \rho_\beta - h_h \sum_\alpha \rho_\alpha \sum_i \sigma_i^\alpha \tag{5}$$

where h_h is a positive constant of order unity. For both models, we assume the usual deterministic (zero-temperature) asynchronous dynamics defined by the rule

$$\sigma_i^\alpha(t+1) = \text{sign}[h_i^\alpha(t)] \tag{6}$$

where the local fields $\{h_i^\alpha\}$ are given by

$$\begin{aligned} h_i^\alpha &= \sum_{j \neq i} J_{ij}^\alpha \sigma_j^\alpha + \sum_{\beta \neq \alpha} J'_{\alpha\beta} \frac{1}{n_1} \sum_j \sigma_j^\beta \qquad &\text{(Model I)} \\ h_i^\alpha &= \sum_{j \neq i} J_{ij}^\alpha \sigma_j^\alpha + h\rho_\alpha \qquad &\text{(Model II)} \end{aligned} \tag{7}$$

The effects of fast synaptic noise may be taken into account by modifying the dynamics in the following way:

$$\begin{aligned} \sigma_i^\alpha(t+1) &= -\sigma_i^\alpha(t) \text{ with probability } 1 \qquad \text{if} \quad h_i^\alpha(t)\,\sigma_i^\alpha(t) < 0 \\ \sigma_i^\alpha(t+1) &= -\sigma_i^\alpha(t) \text{ with probability } \exp[-2h_i^\alpha(t)\,\sigma_i^\alpha(t)/T] \\ &\qquad \text{if} \quad h_i^\alpha(t)\,\sigma_i^\alpha(t) \geqslant 0 \end{aligned} \tag{8}$$

Here, the "temperature" T is a measure of the strength of the noise. The dynamics of the ρ spins in Model-II may be defined in an analogous manner.

In order to simplify the analysis of the behavior of these models, we assume, without any loss of generality, that $\sum_i \xi_i^{\mu\alpha} > 0$ for all μ, α. Then, for Model 1, the desired stable states are the ones in which one has convergence to stored patterns in the upper level as well as in each of the lower-level clusters:

$$\begin{aligned} &\sigma_i^\alpha = \xi_i^{\mu_\alpha\alpha} P_\alpha \qquad \text{for all} \quad i, \alpha \qquad [P_\alpha = \pm 1] \\ &\rho_\alpha = \operatorname{sign}\left[\sum_i \sigma_i^\alpha\right] = P_\alpha = \phi_\alpha^\nu \qquad \text{for all} \quad \alpha \end{aligned} \tag{9}$$

It is easy to show that if the parameter a [see Eq. (3)] is sufficiently small, then the states which satisfy the conditions given in Eq. (9) are stable under the dynamics specified in Eq. (6). For such a state, the local field h_i^α at the ith spin in the αth cluster is given by

$$h_i^\alpha = \phi_\alpha^\nu \xi_i^{\mu_\alpha\alpha} \pm O\left(\frac{\sqrt{p_1}}{\sqrt{n_1}}\right) + \phi_\alpha^\nu \frac{ab}{\sqrt{n_1}} \pm O\left(\frac{\sqrt{p_2}}{\sqrt{n_2}} \frac{a}{\sqrt{n_1}}\right) \tag{10}$$

Here, b is a numerical factor of order unity, defined by

$$\left\langle \left| \sum_{i=1}^{n_1} \xi_i^{\mu\alpha} \right| \right\rangle = b\sqrt{n_1} \tag{11}$$

where $\langle \cdots \rangle$ represents an average over the random distribution of the $\xi_i^{\mu\alpha}$. If the ratios p_1/n_1 and p_2/n_2 are small compared to α_c, then the correction terms in Eq. (10) arising from the interference of other memorized patterns may be neglected. Then, it is clear that the state $\{\sigma_i^\alpha = \xi_i^{\mu_\alpha\alpha}\phi_\alpha^\nu\}$ is locally stable (h_i^α has the same sign as that of $\xi_i^{\mu_\alpha\alpha}\phi_\alpha^\nu$) if $ab/\sqrt{n_1} < 1$. If $ab/\sqrt{n_1} > 1$, then the local field would have the same sign as that of ϕ_α^ν and this would result in a ferromagnetic alignment of the spins in each cluster. Thus, the value of a must satisfy the inequality $ab/\sqrt{n_1} < 1$ in order to ensure that the configurations which correspond to convergence to memories in the lower level are at least locally stable. We now show that if this inequality is satisfied, then the configuration

$$\begin{aligned} \sigma_i^\alpha &= -\phi_\alpha^\nu \xi_i^{\mu_\alpha \alpha}, \qquad && i = 1, 2,\ldots, n_1 \\ \sigma_i^\beta &= \phi_\beta^\nu \xi_i^{\mu_\beta \beta}, && \beta \neq \alpha \end{aligned} \tag{12}$$

is also locally stable. For this configuration, the local field at the ith spin in the αth is given by

$$h_i^\alpha \cong -\phi_\alpha^\nu \xi_i^{\mu_\alpha \alpha} + \phi_\alpha^\nu \frac{ab}{\sqrt{n_1}} \tag{13}$$

which has the same sign as $-\phi_\alpha^\nu \xi_i^{\mu_\alpha \alpha}$ if $ab/\sqrt{n_1} < 1$. Note that the configuration specified in Eq. (12) has convergence to memory states in each of the lower-level clusters, but an error in the αth bit at the upper level. Since this state is also locally stable, the network is not able to correct this error. Similarly, it can be shown that if $ab/\sqrt{n_1} < 1$, then all states with convergence to stored patterns in the lower-level clusters are locally stable, irrespective of whether the signs of the magnetizations of these clusters match one of the patterns $\{\phi_\alpha^\nu\}$ or not. This means that the system will converge to a pattern belonging to a correct category only if the input pattern has the correct signs for the magnetizations of all the lower-level clusters. This is, of course, not the behavior one desires for the model to exhibit. In order to have some capacity for error correction in the determination of the category, it is necessary to have a model with the property that states in which the signs of the magnetizations of the lower-level clusters are different from those specified by the patterns $\{\phi_\alpha^\nu\}$ are *not* locally stable. In other words, the model should be able to discriminate between a stored pattern $\{\xi_i^{\mu\alpha}\}$ and its complement $\{-\xi_i^{\mu\alpha}\}$ depending on whether the sign of $\sum_i \xi_i^{\mu\alpha}$ matches that of ϕ_α^ν or not. As discussed above, Model I does not achieve this goal.

It is easy to show that Model II also suffers from the same deficiency. By choosing the parameter a to be sufficiently large, one can always make the states $\{\rho_\alpha = \phi_\alpha^\nu\}$ stable in the upper layer. From Eqs. (6) and (7), it is readily seen that the state $\{\sigma_i^\alpha = \phi_\alpha^\nu \xi_i^{\mu_\alpha \alpha}\}$ which represents a convergence to appropriate memory states in all the lower-level clusters is locally stable if $h < 1$. However, for $h < 1$, all states represented by $\sigma_i^\alpha = P_\alpha \xi_i^{\mu_\alpha \alpha}$, $i = 1, 2,\ldots, n_1$, $\alpha = 1, 2,\ldots, n_2$, and $P_\alpha = \pm 1$ chosen arbitrarily, are also locally stable. Thus, this model also cannot correct errors in categorization. All states with convergence to memories in the lower-level clusters are locally stable irrespective of whether the signs of the magnetizations of the clusters correspond to one of the allowed categories or not.

We note here that Model II is somewhat easier to analyze than Model I, especially if we make the following assumptions about the time evolution of this model. We assume that inputs are given separately to the two levels, with ρ_α set equal to sign $(\sum_i \sigma_i^\alpha)$ at the beginning. The spins in the upper level are first allowed to relax until they reach a stable state. If the parameter a and the ratio p_2/n_2 are chosen properly, then the memory states $\{\phi_\alpha^\nu\}$ will be locally stable, and we assume that such a state is reached in the upper level. The spins in the lower level are then allowed to relax. The dynamics of the spins in the αth cluster is then governed by the effective Hamiltonian

$$H^\alpha = -\sum_{i>j} J_{ij}^\alpha \sigma_i^\alpha \sigma_j^\alpha - h\phi_\alpha^\nu \sum_i \sigma_i^\alpha \tag{14}$$

Since there is no intercluster interaction, each cluster may be considered separately and then we may drop the cluster index α. The problem then reduces to that of describing a single n_1-spin cluster interacting with a single spin ρ in the upper layer. In the remaining part of this paper, we shall use this simplified version of the problem. Cast in this form, an analysis of Model II reduces to that of the behavior of a model defined by the Hamiltonian

$$H = -\sum_{i>j} J_{ij} \sigma_i \sigma_j - h\phi \sum_i \sigma_i \tag{15}$$

where $\phi = +1$ or -1. This Hamiltonian describes a Hopfield model in a uniform magnetic field. The properties of this model can be calculated by a straightforward application of the replica formalism developed by Amit *et al.*[4] We consider the limit n_1, $p_1 \to \infty$ with $\alpha = p_1/n_1$ finite and define the usual order parameters

$$m = \left\langle \frac{1}{n_1} \sum_i \xi_i^{\mu_0} \bar{\sigma}_i \right\rangle \tag{16a}$$

$$q = \left\langle \frac{1}{n_1} \sum_i (\bar{\sigma}_i)^2 \right\rangle \tag{16b}$$

$$r = \left\langle \frac{1}{\alpha} \sum_{\mu \neq \mu_0} \left(\frac{1}{n_1} \sum_i \xi_i^\mu \bar{\sigma}_i \right)^2 \right\rangle \tag{16c}$$

where the bar represents a thermal average at temperature T [we consider the general stochastic dynamics of Eq. (8)]. Assuming replica symmetry, we obtain the following self-consistent equations for the order parameters in the $T \to 0$ limit:

$$m = \frac{1}{2}\left[\operatorname{Erf}\left(\frac{m+h}{(2\alpha r)^{1/2}}\right) + \operatorname{Erf}\left(\frac{m-h}{(2\alpha r)^{1/2}}\right)\right] \tag{17a}$$

$$C = \frac{1}{T}(1-q) = \frac{1}{(2\pi\alpha r)^{1/2}}\left[e^{-(m+h)^2/2\alpha r} + e^{-(m-h)^2/2\alpha r}\right] \tag{17b}$$

$$r = \frac{1}{(1-C)^2} \tag{17c}$$

The same equations are obtained for both $\phi = 1$ and $\phi = -1$. The first thing we note is that these equations are invariant under a change of sign of h. Thus, the sign of upper-level spin does not have any effect on the thermodynamic behavior of the spins in the lower-level cluster. Also, for a given h, if m_0 is a self-consistent solution of Eqs. (17a)–(17c), then $-m_0$ is an equally good solution. Thus, the uniform field term does not discriminate between the memory state $\{\sigma_i = \xi_i^{\mu_0}\}$ and its complement. This result is in agreement with the conclusion reached earlier in this section. We also note that in the $\alpha \to 0$ limit, Eq. (17a) becomes

$$m = \tfrac{1}{2}[\operatorname{Sign}(m+h) + \operatorname{Sign}(m-h)] \tag{18}$$

which has a nonzero solution, $m = \pm 1$ if $|h| < 1$. Thus, as noted earlier, convergence to a memory state in the $\alpha \to 0$ limit is possible only if $|h| < 1$. The main effect of the presence of the uniform field term is to reduce the critical storage capacity α_c beyond which no solution with $m \neq 0$ exists. Numerical solutions of Eq. (17) show that α_c decreases smoothly from 0.14 at $h = 0$ to zero at $h = 1$.

The conclusions reached in the arguments presented above have been verified by numerical simulations. For Model I, we carried out a simulation in which we had nine clusters at the lower level, each containing 51 spins ($n_1 = 51$, $n_2 = 9$). Four randomly chosen patterns were stored in each of the lower-level clusters and the upper cluster had two patterns stored in it ($p_1 = 4$, $p_2 = 2$). The value of the parameter a was set at 1.0. Initial configurations were chosen to lie close to stored patterns or their complements in the lower-level clusters, without any bias for the signs of the cluster magnetizations. The spins were updated in a random sequence according to the zero-temperature dynamics of Eq. (6) until convergence to a stable state was reached. In Table I, we present the results for the distribution (obtained from 500 runs) of the overlap of the final ρ-spin configuration with the closest stored pattern in the upper level. For a comparison, we also show in Table I the exact distribution that would have been obtained

Table I. Distribution of the Largest Overlap with Stored Patterns in the Upper Level of a Dotsenko Model (Model I)[a]

	Probability for	
Overlap	Dotsenko model	Random configurations
1/9	0.297	0.2422
3/9	0.553	0.4307
5/9	0.149	0.2505
7/9	0	0.0688
1	0	0.0078

[a] With $n_1 = 51$, $n_2 = 9$, $p_1 = 4$, $p_1 = 2$, $a = 1.0$. The distribution of the largest overlap for randomly chosen configurations is also shown for comparison.

if the overlaps were calculated for randomly chosen ρ-spin configurations. These two distributions are quite similar, indicating that the presence of the intercluster interaction has almost no effect toward ensuring convergence to stored patterns at the upper level. In contrast, a Hopfield model with nine spins and two memories shows convergence to a memory state in nearly 100% of runs starting from random inputs. In order to test the performance of Model II, we simulated the behavior of the system defined by the Hamiltonian of Eq. (15), with $n_1 = 100$, $p_1 = 4$, $\phi = 1$, $h = 0.2$, and $\sum_i \xi_i^\mu > 0$ for all μ. Starting from a randomly chosen initial configuration, the spins were updated until convergence to a stable state was reached. Overlaps of this state with the stored memories were calculated and the memory state for which the magnitude of the overlap is largest was identified. Figure 1 shows the distribution (obtained from 1000 runs) of the overlap with this memory state. This distribution is found to be more or less symmetric about zero, indicating that the uniform external field h does not accomplish the task of inducing convergence to only those memory states which have the correct sign of the magnetization. Similar results were obtained for $h = 0.4$. Higher values of h were found to induce ferromagnetic ordering in the lower-level cluster.

3. MODEL WITH A FIELD CONJUGATE TO MEMORY STATES

As discussed in Section 2, the behavior of Dotsenko-type models of the second kind may be understood by considering a single cluster at the lower level interacting with one spin in the upper level. We showed earlier that a ferromagnetic interaction between the upper-level spin and the

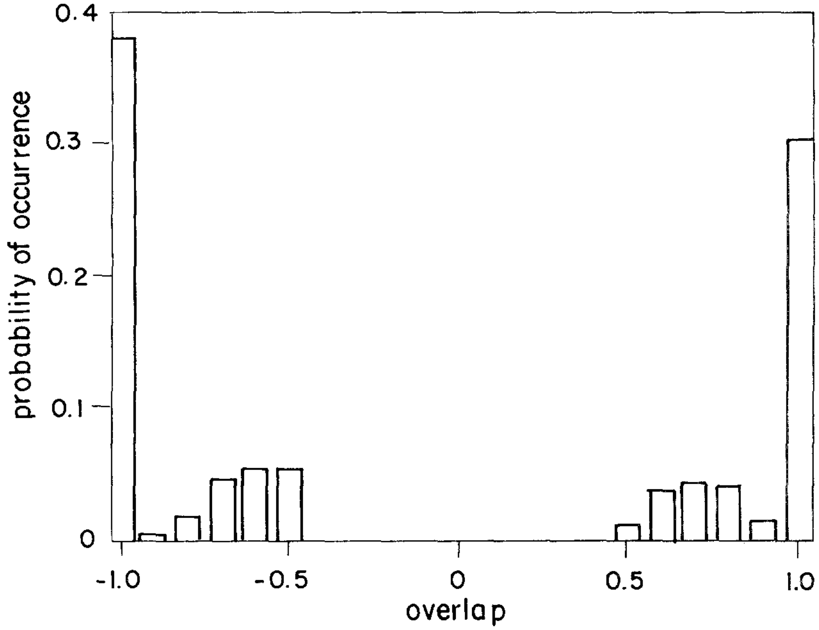

Fig. 1. Distribution of the largest (in magnitude) overlap with stored patterns for a lower-level cluster interacting ferromagnetically with a spin in the upper level [model defined in Eq. (15)]. The values of the parameters are $n_1 = 100$, $p_1 = 4$, $\phi = 1$, and $h = 0.2$. The histogram was obtained from 1000 runs with random inputs.

lower-level cluster does not discriminate effectively between a memory state and its complement if the memories are chosen to be random binary sequences. The main reason for this failure is the fact that each randomly selected memory state has a magnetization of order $\sqrt{n_1}$, so that the energy associated with a uniform field term is also of order $\sqrt{n_1}$ if we consider cluster states close to memories or their complements. The energy associated with the Hopfield term is, on the other hand, of order n_1, so that this term dominates over the effects of the uniform field. This observation suggests that if the interaction between the two levels is chosen to have the form of an external field conjugate to the patterns stored in the lower-level cluster, then it may discriminate effectively between a memory and its complement in the lower cluster. We therefore consider a model defined by the following Hamiltonian:

$$H = -\sum_{i>j} J_{ij}\sigma_i\sigma_j - h\phi \sum_i \sum_{\mu=1}^{p_1} \xi_i^\mu \sigma_i \tag{19}$$

where the matrix J_{ij} has the Hebb-rule form of Eq. (1) and $\phi = \pm 1$ represents the state of the spin at the upper level. For the sake of definiteness, we assume that $\sum_i \xi_i^\mu > 0$ for all μ. Then, the memory states $\{\sigma_i = \xi_i^\mu\}$ should be selected at the lower level if $\phi = 1$, and the complement states should be selected if $\phi = -1$. As shown below, the added field conjugate to the memory states achieves this goal if the number of patterns (p_1) stored in each cluster is small and the value of h is chosen properly. If p_1 is large, then the fields conjugate to all the other memories interfere with the retrieval of a particular memory and the network does not function as desired.

We first consider the situation where p_1 is of order unity and assume that $\phi = 1$. In a desired retrieval state $\{\sigma_i = \xi_i^\mu\}$, the local field at site i is given by

$$h_i = (1+h)\,\xi_i^\mu + h \sum_{\nu \neq \mu} \xi_i^\nu \pm O\left(\frac{1}{\sqrt{n_1}}\right) \tag{20}$$

Since the largest magnitude the second term can have is $h(p_1 - 1)$, this state will be stable if $1 + h > h(p_1 - 1)$, i.e., if $h < 1/(p_1 - 2)$.

Similarly, it is easy to show that the complement state $\{\sigma_i = -\xi_i^\mu\}$ will be unstable if $h > 1/p_1$. Thus, the field term would be able to make the desired choice if the value of h lies in the interval between $1/p_1$ and $1/(p_1 - 2)$. Since the width of this interval goes to zero for large values of p_1, the model being considered here would not be useful in the large-p_1 limit. The restriction to small p_1 values is, however, not a very serious one because a large number of correlated patterns may be stored in the network even if p_1 is small. For example, in a 1010-spin network with $n_1 = 100$, $n_2 = 10$, and $p_1 = 4$, the number of patterns one would be able to store and retrieve is 4^{10}, which is quite large.

Although the recall states $\{\sigma_i = \xi_i^\mu\}$ are locally stable if $h < 1/(p_1 - 2)$, they are not the globally stable states of the system. The thermodynamic behavior of this model for $p_1 \sim O(1)$ and $n_1 \to \infty$ may be worked out easily by using the analytic methods developed by Amit *et al.*[3] Defining the overlaps m^μ as

$$m^\mu = \left\langle \frac{1}{n_1} \sum_i \xi_i^\mu \bar{\sigma}_i \right\rangle \tag{21}$$

we obtain the following self-consistent equations for them:

$$m^\nu = \left\langle \xi^\nu \tanh\left\{\beta \sum_\mu (m^\mu + h)\,\xi^\mu\right\}\right\rangle \tag{22}$$

where $\beta = 1/T$ and ϕ has been set equal to unity. The free energy per spin has the form

$$f(\beta, h) = \frac{1}{2} \sum_{\mu} m_{\mu}^{2} - \frac{1}{\beta} \left\langle \ln 2 \cosh \left\{ \beta \sum_{\mu} (m^{\mu} + h)\, \xi^{\mu} \right\} \right\rangle \tag{23}$$

Using these equations, the thermodynamic behavior of the system may be analyzed for any value of $p_1 \sim O(1)$. Here, we explicitly work out the details for $p_1 = 4$. We confine our attention to solutions of Eq. (22) for which three of the four m^{μ} have the same value (m') and the fourth one may have a different value (m). Other types of solutions are possible, but those solutions have high values of the free energy and will not be considered here. The self-consistent equations for m and m' have the form

$$\begin{aligned} m = {} & \tfrac{1}{8}[\tanh \beta(m + 3m' + 4h) + 3 \tanh \beta(m + m' + 2h) \\ & + 3 \tanh \beta(m - m') + \tanh \beta(m - 3m' - 2h)] \\ m' = {} & \tfrac{1}{8}[\tanh \beta(m + 3m' + 4h) + \tanh \beta(m + m' + 2h)] \\ & + \tanh \beta(m' - m) + \tanh \beta(3m' - m + 2h)] \end{aligned} \tag{24}$$

The free energy of a solution of these equations may be calculated from Eq. (23), and the local stability of such a solution may be determined by calculating the eigenvalues of the matrix that describes small fluctuations about the solution. In the $T \to 0$ limit, the tanh functions appearing in Eq. (24) reduce to sign functions. Solutions of these equations may then be obtained by inspection. We find the following solutions at $T = 0$ for positive values of h:

(a) $m = 1$, $m' = 0$: This solution is possible if $h < 0.5$, as expected from the arguments described earlier in this section. At $T = 0$, it represents a state with perfect recall of a stored memory. It is locally stable near $T = 0$, but becomes unstable at higher temperatures (at $T \simeq 0.3$ for $h = 0.3$).

(b) $m = m' = 0.375$: This solution, which is present for all values of h, is a continuation of the symmetric disordered phase that occurs at high temperatures. It is unstable for $T < 0.5$. This state has the lowest free energy at high temperatures.

(c) $m = 0.75$, $m' = 0.25$: This solution is also present for all positive values of h. It is stable at low temperatures, and it becomes unstable at a temperature slightly higher than 0.5. It has the lowest free energy at low temperatures and therefore it is the globally stable low-temperature phase of the system.

(d) $m = 0$, $m' = 0.5$: This solution, which is also present for all values of h, is locally stable at low temperatures. Its free energy is slightly higher than that of solution (c).

(e) $m = -1$, $m' = 0$: This solution represents a convergence to the complement of a memory state. It is present only if $h < 0.25$, as expected. It has a high value of the free energy and it becomes unstable at a low temperature ($T \simeq 0.1$ for $h = 0.2$).

The system undergoes a weakly first-order phase transition from the high-temperature disordered phase [solution (b)] to a low-temperature ordered phase represented by solution (c) at a temperature slightly above 0.5. This transition is similar to that exhibited by the four-state Potts model in mean field theory. The external field h breaks the symmetry between a pattern and its complement, but preserves the symmetry under permutations of the labels representing the different memory states. This permutation symmetry is broken at the phase transition to the ordered state.

The calculations described above have to be modified if the number of memories p_1 is proportional to n_1. In this limit, the magnitude of the field should be chosen to be proportional to $1/\sqrt{n_1}$, so that the local field produced by this term is of order unity at each site. A replica-symmetric calculation of the thermodynamic behavior of this system yields the following self-consistent equations at $T \to 0$ for the order parameters m, q, and r defined earlier:

$$m = \mathrm{Erf}(m/(2\alpha r)^{1/2}) \tag{25a}$$

$$q \simeq 1 \tag{25b}$$

$$r = \frac{1 + h'^2}{(1 - C)^2}, \qquad C = \left(\frac{2}{\pi \alpha r}\right)^{1/2} e^{-m^2/2\alpha r} \tag{25c}$$

where $h = h'/\sqrt{n_1}$. As expected from the general arguments described in the first part of this section, the presence of the h term does not discriminate between a pattern and its complement in this limit. The main effect of the h field is to increase the "noise" arising from the interference of noncondensed patterns. This interference causes a decrease of the storage capacity of the network as h' is increased from zero. From a numerical solution of Eq. (25), we find that the critical value of $\alpha = p_1/n_1$ goes to zero at $h' \simeq 0.35$.

We have carried out a few numerical simulations to check some of the predictions of the analytic calculations. The histograms shown in Fig. 2 were obtained from a simulation with $n_1 = 101$, $p_1 = 4$, $\phi = 1$, $h = 0.3$, and $\sum_i \xi_i^\mu > 0$ for all μ. The chosen value of h lies in the region where the memory states $\{\sigma_i = \xi_i^\mu\}$ are locally stable, whereas their complements are

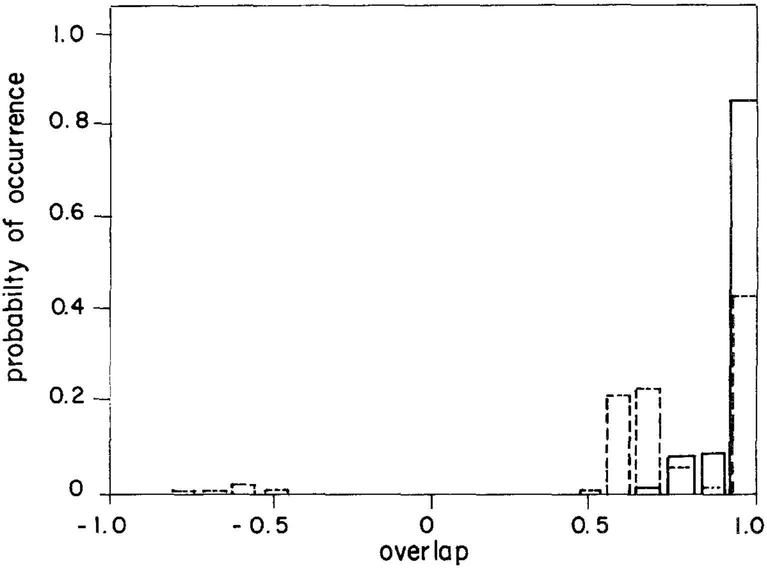

Fig. 2. Distribution of the largest (in magnitude) overlap with stored memories for the model with an external field conjugate to all the memorized patterns [see Eq. (19)]. The values of the parameters used are $n_1 = 101$, $p_1 = 4$, $h = 0.3$, $\phi = 1$. The first histogram (solid line) shows the distribution obtained from 400 runs, each starting from a 10% corrupted version of one of the memories. The second histogram (dashed line) shows the distribution obtained when the initial configurations are chosen to be 10% corrupted versions of the complements of the memories.

not. The histograms show the distributions of the overlap of the final stable configuration with the memory state for which the overlap has the largest magnitude. Large positive overlaps represent convergence to memory states, whereas large negative values correspond to unwanted convergence to the complement states. The first histogram (solid line) shows the results for 400 starting configurations obtained by reversing the signs of a randomly chosen small fraction (10%) of the spins representing one of the stored memories. We find that the system converges to the appropriate memory state in nearly all the runs. The second histogram (dashed line) shows the results obtained from 400 runs in which the initial states were 10% corrupted versions of the complements of the memory states. This histogram has very little weight at large negative overlaps, indicating that the field term effectively prevents the system from converging to the complement of a memory state. In a large fraction (~45%) of runs, the final state has nearly 100% overlap with one of the stored memories. This model, thus, is able to correct errors in categorization in nearly half of the runs. However, in a substantial number of runs, the system ends up in spurious states which do not have large overlaps with any of the stored memories. The possibility of getting stuck in spurious states is, of course, present in all Hopfield-like models. This problem is made worse by the fact that there is no way for the system to "realize" that the stable state it has reached is a spurious one. The same difficulty is present in the model we are considering here. We note, however, that in this model there may be a way for the system to find out whether the input pattern belongs to one of the stored categories or not. We find that the average number of updates needed to reach a stable state from an initial configuration close to the complement of a memory state is substantially larger (by a factor of ~ 2) than that needed when the initial state is close to a memory. Thus, by monitoring the "time" required to reach convergence, the system may be able to determine whether the input pattern belongs to one of the allowed categories or not.

4. MODEL WITH THREE-SPIN INTERACTIONS

In this section, we consider a model with three-spin interactions in the lower-level clusters. In the original Hopfield model, the Hamiltonian is quadratic in the spin variables and is therefore invariant under a simultaneous reversal of the signs of all the spins. For this reason, the Hopfield model does not discriminate between a stored pattern and its complement. A Hamiltonian with a three-spin interaction term would not be invariant under a simultaneous spin reversal. A model with such a Hamiltonian, therefore, would not have this degeneracy between a memory and its complement. Since we are interested here in networks which are able to discriminate between a memory and its complement, we consider a model with the Hamiltonian

$$H = -\sum_{i>j} J_{ij}\sigma_i\sigma_j - \frac{\lambda\phi}{3} \sum_{i \neq j \neq k} K_{ijk}\sigma_i\sigma_j\sigma_k \tag{26}$$

where $\phi = \pm 1$ represents, as before, the state of the spin in the upper level and the coefficients K_{ijk} of the three-spin term have the Hebb-rule-like form

$$K_{ijk} = \frac{1}{n_1^2} \sum_{\mu=1}^{p_1} \xi_i^\mu \xi_j^\mu \xi_k^\mu \tag{27}$$

The adjustable parameter λ determines the strength (relative to the usual Hopfield term J_{ij}) of the three-spin interaction. As before, we consider memory states with $\sum_i \xi_i^\mu > 0$. Then, for $\phi = \pm 1$, the states $\{\sigma_i = \pm \xi_i^\mu\}$ should be stable and their complements should be unstable. Since for $\phi = 1$, the energy associated with the three-spin interaction term is negative for the states $\{\sigma_i = \xi_i^\mu\}$ and positive for the spin-reversed states, the model defined in Eq. (26) is expected to provide the required discrimination between a memory and its complement if the coefficient λ is sufficiently large. The analysis presented later in this section shows that this is indeed the case.

A three-spin interaction term $K_{ijk}\sigma_i\sigma_j\sigma_k$ may be looked upon as describing a situation in which the sign of the interaction between a pair of spins (say i and j) depends on the state of a third spin (k). A similar situation is known to exist in biological networks,(13) where one finds that the effect of one neuron on another is sometimes determined by the state of a third neuron (a so-called interneuron). Also, connections similar in nature to three-spin interactions are often used in models of parallel distributed processing,(14) where they are called "Sigma-Pi" units.

Hopfield-like models with general p-spin interactions were studied by Gardner,(15) who showed that the storage capacity of a network with *only* p-spin interactions is proportional to the $(p-1)$th power of the number of spins. In the model considered here, both two-spin and three-spin interac-

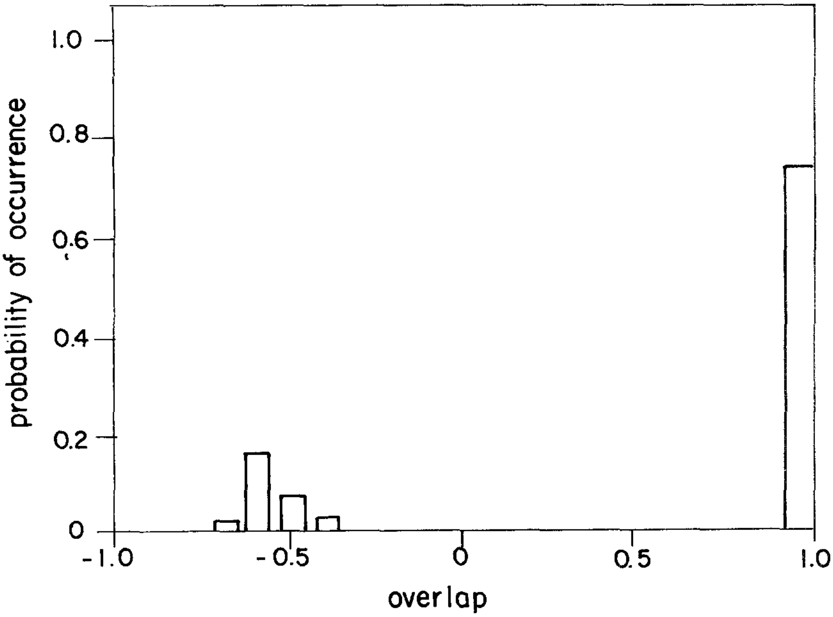

Fig. 3. Distribution of the largest (in magnitude) overlap with stored memories for the model with three-spin interactions [Eq. (26)]. The histogram was obtained from 200 runs, each starting from a 10% corrupted version of the complement of one of the memorized patterns. The parameters used are $n_1 = 51$, $p_1 = 4$, $\lambda = 1.2$, and $\phi = 1$.

tions are present. It is easy to show that the storage capacity of this model is proportional to n_1, the number of spins. The usual "signal-to-noise" analysis[1] tells us that the noise coming from the two-spin term due to the interference of noncondensed patterns is of order $(p_1/n_1)^{1/2}$, whereas the three-spin term generates a noise of order $(p_1/n_1^2)^{1/2}$. Since the signals generated by both two-spin and three-spin terms are of order unity, the noise coming from the two-spin term would overwhelm the signal if p_1 increases faster than linearly with n_1. If p_1 is of order n_1, then the three-spin term increases the signal without increasing the noise. The net effect is an increase of the storage capacity, so that the critical value of $\alpha = p_1/n_1$ is expected to be higher than that calculated for the Hopfield model with only two-spin interactions.

The thermodynamic behavior of the model defined in Eq. (26) may be analyzed by using a straightforward generalization of the replica method developed by Gardner. A replica-symmetric calculation (with $\phi = 1$) along these lines leads to the following self-consistent equations in the $T \to 0$ limit:

$$m = \mathrm{Erf}(k/(2\alpha r)^{1/2}) \tag{28a}$$

$$q \simeq 1 \tag{28b}$$

$$r = 1/(1-C)^2 \tag{28c}$$

$$C = \left(\frac{2}{\pi \alpha r}\right)^{1/2} \exp\left(\frac{-k^2}{2\alpha r}\right) \tag{28d}$$

$$k = m + \lambda m^2 \tag{28e}$$

For $\lambda \neq 0$, the self-consistent equation for m is no longer symmetric under $m \to -m$. Also, solutions with m close to -1 are clearly not possible if $\lambda > 1$. Thus, if λ is chosen to be greater than unity, then this model effectively prevents convergence to the complements of the memory states. The critical storage capacity of the network may be determined from numerical solutions of Eqs. (28a)–(28e). We find that α_c increases rapidly with λ, typical values being $\alpha_c \simeq 6.0$ for $\lambda = 0.5$ and $\alpha_c \simeq 13$ for $\lambda = 1$.

We have also carried out numerical simulations to test the performance of this model. These simulations are very similar to those described in Section 3, except that we took $n_1 = 51$ and $p_1 = 4$. The value of λ was set at 1.2. The histogram for the distribution of overlaps for runs in which the initial state was one of the memories corrupted by 10% is not shown because we obtained convergences to the target patterns in all the runs.

The histogram for the distribution of overlaps obtained from runs starting from 10% corrupted versions of complements of memory states shows that the system converges to a memory with the correct sign of the magnetization in about 70% of the cases. The problem of convergence to spurious states is, however, present in this model also. The "time" taken to reach convergence is again substantially longer when the initial state is close to the complement of a memory state. The general behavior of this model is thus similar to the one described in Section 3, except that it has the added advantage of being able to store a macroscopic number of memories in each of the lower-level clusters. This advantage, however, is obtained at the cost of introducing three-spin interactions.

5. MODEL WITH SYNAPSES WITH A TIME DELAY

The models we have studied so far all involve instantaneous synaptic interactions, so that we could write down appropriate Hamiltonians describing their properties. In this section, we consider a model that uses synaptic connections with a time delay to induce a transition in a lower-level cluster from a memory to its complement if the sign of the magnetization of the memory state does not match that of the corresponding spin in the upper level. This model makes use of the method of temporal sequence generation proposed by Kleinfeld[16] and by Sompolinsky and Kanter.[17] These authors consider models in which there are two different sets of synaptic interactions. The first set is of the Hopfield type, tending to stabilize the network in a memory state. The second set of connections tends to induce specified transitions from one memory state to another. A time delay is associated with this set of interactions, so that the system remains in a memory state for some time before making a transition to another one. We make use of the same principles to construct a model in which the system undergoes a transition from a memory state to its complement if the magnetization of the memory state does not have the "correct" sign. The local field acting on the ith spin at "time" t in this model is given by

$$h_i(t) = \sum_{j \neq i} J_{ij}\sigma_j(t) + \lambda\eta(t) \sum_{j \neq i} D_{ij}\sigma_j(t-\tau) \tag{29}$$

where the first term is the usual Hopfield one, λ is an adjustable parameter, and $\eta(t)$ is a variable that takes on the values 0, 1 according to whether the sign of the magnetization of the cluster at time t matches that of the corresponding spin ϕ in the upper level or not. A simple network to compute $\eta(t)$ will be described later. The matrix D_{ij} is constructed in a way that makes the second term in Eq. (29) tend to induce transitions from memory states to their complements:

$$D_{ij} = \frac{1}{n_1} \sum_{\mu=1}^{p_1} \bar{\xi}_i^\mu \xi_j^\mu \tag{30}$$

where we have used the notation $\bar{\xi}_i^\mu = -\xi_i^\mu$. The connections represented by D_{ij} are delayed in the sense that the local fields arising from them at time t depend on the spin configuration at an earlier time $t-\tau$. The time delay τ is an adjustable parameter, usually taken to be ~ 10 attempted updates per spin.

In order to understand the working of this network, let us consider a situation where at time $t = t_0$, the lower-level cluster has settled down into a memory state $\{\sigma_i = \xi_i^\mu\}$ under the action of the J_{ij} term. The configurations at earlier times are assumed to be uncorrelated with the memories, so that the contribution to the local field coming from the D_{ij} term is $\sim O(1/\sqrt{n_1})$, which may be neglected. If the magnetization of this state has the correct sign, then the second term in Eq. (29) is inoperative ($\eta = 0$) and the system remains in this state. If, on the other hand, the sign of the magnetization of the memory state does not match that of the upper-level spin, then $\eta = 1$ and the second term comes into play. The local field at the ith spin at time $t_0 + \tau$ is then given by

$$h_i(t_0 + \tau) = \xi_i^\mu + \lambda \bar{\xi}_i^\mu = (1 - \lambda)\, \xi_i^\mu \tag{31}$$

In writing Eq. (31), we have neglected correction terms arising from the interference of noncondensed patterns. This is correct if the number of stored patterns p_1 is of order unity. It is then obvious that if the parameter λ is chosen to have a value greater than unity, then the system undergoes a transition to the complement state $\{\sigma_i = -\xi_i^\mu\}$, which has the correct sign of the magnetization, at time $t_0 + \tau$. Thus, for $p_1 \sim O(1)$ and $\lambda > 1$, the network functions in the following way. If the initial configuration is close to a memory state with the correct sign of the magnetization, then the system converges to this memory state and stays there. If, on the other hand, the initial state is close to a memory with an incorrect sign of the magnetization, then the network still converges to this memory state, stays in it for a time $\sim \tau$, and then undergoes a spontaneous transition to the complement state which has the correct sign of the magnetization. This model, thus, is able to correct errors in categorization. The behavior of this model for $p_1 \sim O(n_1)$ is difficult to analyze exactly because the methods of equilibrium statistical mechanics are no longer applicable. The dynamics of an asymmetrically diluted version of this model may be analyzed exactly in the limit of extreme dilution by using methods developed by

Derrida *et al.*[18] and Gutfreund and Mezard.[19] This calculation[20] suggests that if p_1 is of order n_1, then transitions between a memory and its complement occur if $\lambda > \lambda_c$, with λ_c less than unity. Numerical simulations[20] of the fully connected model appear to confirm this prediction.

We now discuss how the variable $\eta(t)$ may be computed in a neural network. Consider the network shown in Fig. 4, where the lines with arrows represent one-way synaptic connections with strengths as indicated. The spins in a lower-level cluster are represented by $\sigma_1, \sigma_2,..., \sigma_{n_1}$ and ϕ represents the upper-level spin associated with this cluster. η_1, η_2, η_3, and η_4 are binary neurons (Ising spins) with thresholds given by the numbers in brackets. The dynamics of one of these neurons is governed by the rule

$$\eta_i(t+1) = \mathrm{Sign}(h_i(t) - h_{i0}) \tag{32}$$

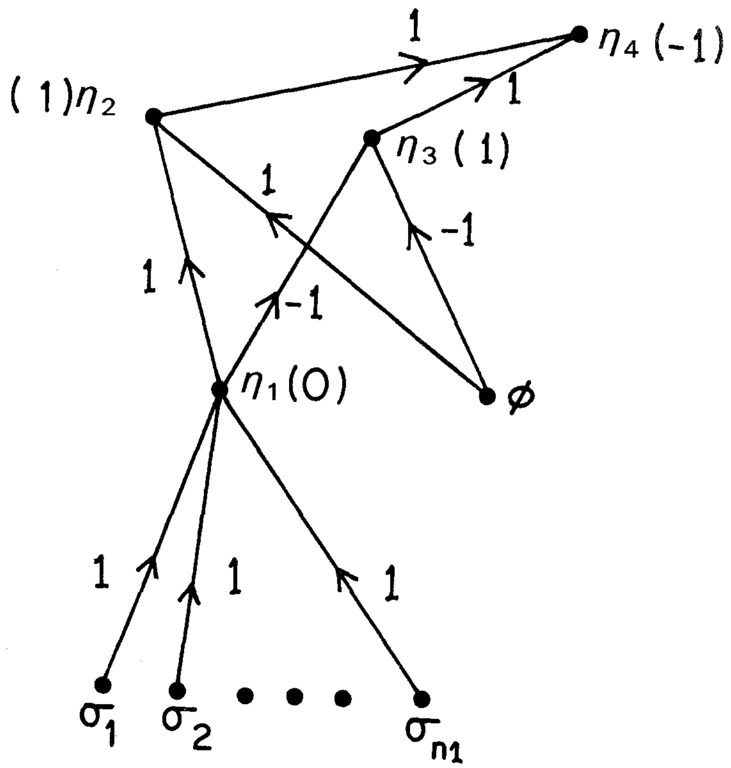

Fig. 4. Neural circuit for computing the quantity η defined in text [see the paragraph following Eq. (29)]. The lines with arrows represent one-way synaptic connections with strengths as indicated. η_1, η_2, η_3, and η_4 are binary neurons with thresholds given by the numbers in parentheses.

where $h_i(t)$ is the local field arising from interactions with other neurons and h_{i0} is the threshold. It is readily seen that $\eta_1 = \text{sign}(\sum_i \sigma_i)$, and if η_1 and ϕ have the same sign, then one of η_2, η_3 is in the $+1$ state and the other one is in the -1 state. The spin η_4 is then in the $+1$ state. In the other case, $\eta_1 = -\phi$, both η_2 and η_3 are in the -1 state and $\eta_4 = -1$. Thus, if we set $\eta = (1 - \eta_4)/2$, then η will have the desired property. With this definition of η, the D_{ij} term of Eq. (29) has the form $(\lambda/2)(1 - \eta_4) \sum_j D_{ij} \sigma_j$, which consists of two terms. The first term corresponds to a direct interaction between σ_i and σ_j, whereas the second one represents an indirect interaction between σ_i and σ_j proceeding via a third neuron, η_4. As mentioned in the preceding section, there is some biological evidence[13] for the existence of such indirect interactions between neurons in the cerebellum.

We have carried out simulations to test the performance of the model defined in Eq. (29). Figure 5 shows the distribution of final overlaps obtained from 500 runs with random inputs to a network with $n_1 = 101$, $p_1 = 5$, $\lambda = 1.3$, $\tau = 9$ attempted updates per spin, $\phi = 1$, and $\sum_i \xi_i^\mu > 0$ for all μ. As before, we have taken the overlap which has the largest magnitude among the overlaps of the final configuration with all the memory states

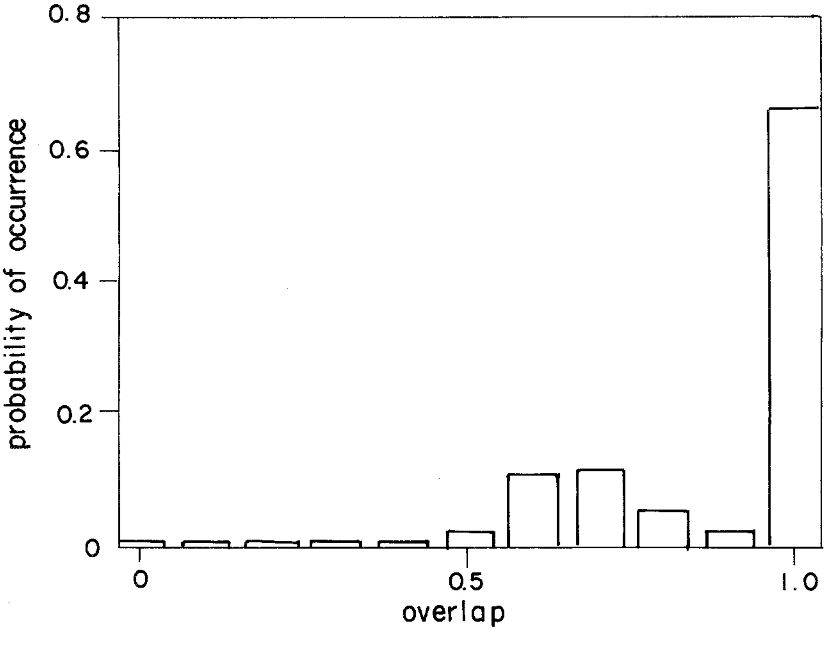

Fig. 5. Distribution of the overlap with the largest magnitude for the model with a time delay [see Eq. (29)]. The histogram was obtained from 500 runs with random inputs. The values of the parameters used are $n_1 = 101$, $p_1 = 5$, $\phi = 1$, $\lambda = 1.3$, and $\tau = 9$ units.

$\{\xi_i^\mu\}$. The large peak near 100% overlap indicates that this scheme functions quite well. The fact that there is no weight at negative values of the overlap tells us that this model effectively prevents the system from condensing into states with incorrect signs of the magnetization. In a few runs, the system gets stuck in spurious states which, perhaps, cannot be avoided in any model of this class if the initial states are chosen at random.

6. SUMMARY AND DISCUSSION

In this paper, we have used analytic and numerical methods to study the behavior of a class of hierarchical neural network models for the storage and retrieval of strongly correlated memories. This class of models was introduced by Dotsenko. In Section 2, we showed that the models considered by Dotsenko have a serious shortcoming if the patterns stored in each cluster at the lowest level of the hierarchy are, as originally proposed, random binary sequences. The problem is that these models are not able to detect or correct errors in categorization which may be present in the input pattern. In Sections 3–5 we described three different models which are constructed with a view toward overcoming this limitation of the original models. Analytic and numerical calculations discussed in detail in these sections show that all three models are able to recognize, and also to correct in a majority of cases, any error in categorization present in the input. The first model, which uses fields conjugate to the patterns, is the simplest of the three. It, however, has the disadvantage of being able to store only a small number of patterns in each lower-level cluster. The second and third models do not have this limitation, but they involve more complicated multispin interactions. The problems associated with spurious stable states are present, in varying degrees, in all three models. This problem is the least severe in the third model, which, in our numerical simulations, appears to always converge to a memory state with the correct sign of the magnetization if the initial state is close to a memory or its complement. The third model also has the nice feature of exhibiting two distinct time scale, one describing a convergence of the system to the nearest memory state and the other one, determined by the parameter τ, describing the process of correcting any error in categorization that may be present in a given input.

We finally note that the models studied here belong to a more general class of neural network models which may be constructed to address the following problem. Consider a network consisting of m clusters each containing n spins. Each cluster stores p patterns which may or may not be correlated among themselves. By combining the patterns stored in all the clusters, we get $(2p)^m$ distinct but correlated mn-bit patterns if a pattern and its complement are considered to be distinct. The problem is to determine a scheme for connecting the clusters (either directly or via a second set of spins) in such a way that only a selected subset of the $(2p)^m$ patterns correspond to stable memory states and the remaining ones are unstable. A network which has this property would be useful in many contexts. For example, the memories stored in each cluster may represent the letters in the alphabet, and the overall patterns composed of appropriate memories selected from the clusters may represent meaningful words. In another example, each cluster may store the digits 0–9 and the overall pattern may represent m-digit numbers to be memorized. It is hoped that the models studied in this paper would be useful in this more general context.

ACKNOWLEDGMENTS

This work was supported in part by the Indian National Science Academy through a research fellowship awarded to C.D.

REFERENCES

1. D. J. Amit, *Modelling Brain Function* (Cambridge University Press, 1988).
2. J. J. Hopfield, *Proc. Natl. Acad. Sci. USA* **79**:2554 (1982); **81**:3088 (1984).
3. D. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. A* **32**:1007 (1985).
4. D. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. Lett.* **55**:1530 (1985); *Ann. Phys.* **173**:30 (1987).
5. L. B. Ioffe and M. V. Feigelman, *JETP Lett.* **44**:189 (1986); N. Parga and M. A. Virasoro, *J. Phys.* (Paris) **47**:1857 (1986); G. Toulouse, S. Dehaene, and T. Changeux, *Proc. Natl. Acad. Sci. USA* **83**:1965 (1986); C. Cortes, A. Krogh, and J. A. Hertz, *J. Phys. A* **20**:4449 (1987); H. Gutfreund, *Phys. Rev. A* **37**:570 (1987).
6. M. M. Merzenich and J. H. Kaas, *Prog. Psychobiol. Physiol. Psychol.* **9**:1 (1980).
7. M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
8. V. S. Dotsenko, *J. Phys. C* **18**:L1017 (1987); *JETP Lett.* **44**:193 (1986); *Physica* **140A**:410 (1986).
9. B. Gidas, *IEEE Trans. Pattern Analysis Machine Intelligence* **2**:164 (1989).
10. D. H. Hubel and T. N. Wiesel, *Proc. R. Soc. Lond. B* **198**:1 (1977).
11. V. Dotsenko and B. Tirozzi, *Carr. Rep. Math. Phys.* (1989).
12. C. R. Willcox, *J. Phys. A* **22**:4707 (1989).
13. S. W. Kuffler and J. G. Nicholls, *From the Neuron to the Brain* (Sinauer, Sunderland, Massachusetts, 1976).
14. D. E. Rumelhert and J. L. MacClelland, *Parallel Distributed Processing*, Vol. 1 (MIT Press, Cambridge, Massachusetts, 1986).
15. E. Gardner, *J. Phys. A* **20**:3453 (1987).
16. D. Kleinfeld, *Proc. Natl. Acad. Sci. USA* **83**:9469 (1986).
17. H. Sompolinsky and I. Kanter, *Phys. Rev. Lett.* **57**:2861 (1986).
18. B. Derrida, E. Gardner, and A. Zippelius, *Europhys. Lett.* **4**:167 (1987).
19. H. Gutfreund and M. Mezard, *Phys. Rev. Lett.* **61**:235 (1988).
20. V. Deshpande and C. Dasgupta, to be published.