Journal of Statistical Physics, Vol. 71, Nos. 3/4, 1993

Theory of the Dynamics of the Hopfield Model of Associative Memory

Prabodh Shukla[1]

Received April 21, 1992; final November 24, 1992

We present an analysis of the parallel dynamics of the Hopfield model of the associative memory of a neural network without recourse to the replica formalism. A probabilistic method based on the signal-to-noise ratio is employed to obtain a simple recursion relation for the zero temperature as well as the finite temperature dynamics of the network. The fixed points of the recursion relation and their basins of attraction are found to be in fairly satisfactory agreement with the numerical simulations of the model. We also present some new numerical results which support our recursion relation and throw light on the nature of the ensemble of the network states which are optimized with respect to single spin flips.

KEY WORDS: Hopfield model; parallel dynamics; probabilistic method; signal-to-noise ratio.

1. INTRODUCTION

The Hopfield model of the associative memory of a neural network has been investigated extensively numerically as well as analytically.[1-4] Numerically, the results which have been most firmly established are the following. A network of N neurons can store αN uncorrelated patterns, where $\alpha \leqslant \alpha_c = 0.14$. The overlap of the retrieved memory with the corresponding stored memory is $M^* = 1$ in the limit $\alpha \to 0$, and $M^* = 0.97$ in the limit $\alpha \to 0.14$. The basins of attraction of the stored memories gradually shrink as more and more patterns are stored. The preceding results are based on the zero-temperature dynamics. Numerical work on the finite-temperature dynamics of the model is somewhat less extensive, but it indicates that an increase in temperature reduces the storage capacity of the network below $0.14N$; at the same time the quality of recall is

[1] Department of Physics, North Eastern Hill University, Shillong 793003, India.

improved. Subject to this proviso, the behavior of the network under finite-temperature dynamics is similar to the zero-temperature dynamics upto a critical temperature above which the network ceases to function as an associative memory device.

Theoretical understanding of the Hopfield model has been developed within the general framework of nonlinear dynamics and the methods of equilibrium statistical mechanics. The Hopfield model has symmetric connections (synapses) between pairs of formal neurons. Therefore the system is expected to evolve to a locally minimum energy state which can be studied by the methods of equilibrium statistical mechanics. Amit *et al.*[2] have studied in detail the statistical mechanics of the Hopfield model by utilizing the concepts and tools developed in the theory of spin glasses based on the replica method. There are two solutions in the replica-symmetric approximation. One is the memory retrieval solution with a large M^*. This has $\alpha_c = 0.137$ and $M^* = 0.967$ at $\alpha = 0.137$. The other solution is the spin-glass solution with $M^* = 0$. These solutions are quite close to the numerical results. However, at zero temperature, the large-M^* solution is the lowest energy solution only for $\alpha < 0.05$ (approximately), and for $\alpha > 0.05$ the spin-glass solution has lower energy. Thus, strictly speaking, the replica method's prediction for the storage capacity of the Hopfield model (with zero-temperature dynamics) is approximately $\alpha_c = 0.05$ against the numerical result $\alpha_c = 0.14$. This discrepancy remains even when replica symmetry breaking[5] is taken into account. A calculation[6] based on one level of replica symmetry breaking extends the region of the large-M^* solution from $\alpha_c = 0.137$ to $\alpha_c = 0.144$, but the solution does not correspond to minimum energy. Higher levels of replica symmetry breaking get increasingly more cumbersome to calculate, and have not been evaluated so far. There have been some attempts (ref. 5, Chapter XIII, p. 394; ref. 7) to avoid the replica method, but so far most of these attempts have at best produced results equivalent to those obtained by the replica method. In view of this it is desirable to explore alternate methods of analysis for understanding the Hopfield dynamics.

This paper presents analytical results on the zero-temperature as well as the finite-temperature dynamics of the Hopfield model (with parallel spin updating) without recourse to the replica method. Our method has the advantage of being simple and physically transparent, and it predicts results which are fairly close to the numerical results. We also present some new numerical results which support our analysis, and give important insight regarding the ensemble of equilibrium states which are approached by the Hopfield dynamics. Our numerical results show that the replica-symmetric prediction for the minimum energy of the system is slightly in error. The significance of these results goes beyond the small numerical discrepancies between the theory and numerical simulation. They indicate that the ensemble of equilibrium states approached by the Hopfield dynamics is not necessarily the canonical ensemble assumed in the replica method treatment of the equilibrium statistical mechanics of the Hopfield model. They also suggest that the most characteristic feature of the Hopfield dynamics may not be to decrease the system's energy, but rather to increase its entropy. We point out that an energy-conserving Hopfield dynamics may also possess the properties of an associative memory.

2. THE MODEL

Although the Hopfield model of associative memory is well known, we recall it briefly for the sake of completeness, and also to set up our notation. The model is characterized by the following Hamiltonian:

$$\mathbf{H} = -\frac{1}{2}\sum_{i,j} \mathbf{J}_{ij} S_i S_j (1-\delta_{ij}) \tag{2.1}$$

where

$$\mathbf{J}_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^\mu \xi_j^\mu$$

Here $\{\xi_i^\mu = \pm 1;\ i = 1, 2,..., N;\ \mu = 1, 2,..., p\}$ are p N-bit patterns stored in the memory of the network, and $\{S_i = \pm 1;\ i = 1, 2,..., N\}$ is a "sight" that jogs the memory of the network leading to the retrieval of the pattern, say $\{\xi_i^{\mu_0}\}$, which is presumably closest to $\{S_i\}$. The Kronecker delta symbol in the expression for $\mathbf{H}$ ensures that the self-interaction terms $i = j$ are excluded. It should be noted that the memories $\{\xi_i^\mu\}$ lose their individual identity when stored in the synaptic matrix $\mathbf{J}_{ij}$, i.e., all stored memories have the same address. Retrieval of a memory in the Hopfield model is based on the content of the memory (which may be partially corrupted). It is different from the usual computer process of fetching a memory by the hardware address of the memory. The prompting sight $\{S_i\}$ sets the network into motion; the motion eventually settles down to a fixed-point attractor of the dynamics; the fixed-point pattern is taken to be the retrieved memory. Thus, the memory is addressed by its content, which is incorporated in $\{S_i\}$. The equations of motion in their simplest form are the following discrete-time equations of motion:

$$S_i(t+1) = \operatorname{sign} \sum_j \mathbf{J}_{ij} S_j(t)(1-\delta_{ij}) \tag{2.2}$$

The dynamics of Eq. (2.2) is completely deterministic. It does not incorporate any stochastic noise. Drawing upon the notions of statistical mechanics, the dynamics of Eq. (2.2) can be viewed as the dynamics of the network at zero temperature. Stochastic noise can be incorporated by defining the network dynamics in analogy with the single-spin-flip Monte Carlo dynamics of Ising spins at finite temperature. Let $P(h_i)$ be the probability that $S_i(t+1)$ is $+1$, and $1-P(h_i)$ be the probability that $S_i(t+1)$ is -1. Then

$$P(h_i)=[1+\exp(-2\beta h_i)]^{-1} \tag{2.3}$$

Here h_i is the local field at site i at time t given by the argument of the sign function in Eq. (2.2), and $T=\beta^{-1}$ is the temperature of the network. It is easy to see that in the limit of zero temperature (2.2) and (2.3) yield the same dynamics.

3. DYNAMICS

Let us focus on the energy of the network states at each time step t. The energy of any state $\{S_i(t)\}$ in the Hopfield model is given by

$$E(\alpha, t)= -\frac{1}{2}\sum_{i,j} \mathbf{J}_{ij}S_i(t)\, S_j(t)(1-\delta_{ij}) \tag{3.1}$$

Let R_μ denote the normalized $(-1\leqslant R_\mu\leqslant 1)$ overlap of the pattern $\{S_i\}$ with $\{\xi_i^\mu\}$:

$$R_\mu(t)=\frac{1}{N}\sum_i \xi_i^\mu\cdot S_i(t) \tag{3.2}$$

The energy (3.1) can be rewritten as

$$E(\alpha, t)=Ne(\alpha, t) \tag{3.3}$$

where

$$e(\alpha, t)= -\frac{1}{2}\left[\sum_\mu R_\mu^2(t)-\alpha\right] \tag{3.4}$$

A remark regarding the quantity $e(\alpha, t)$ is in order. The notation suggests that $e(\alpha, t)$ depends only on α and t. This is strictly correct only in the limit $N\to\infty$ owing to the self-averaging property of the Hopfield model. In numerical simulations with finite-size systems $e(\alpha, t)$ is necessarily a sample-dependent quantity in the sense that it depends on the choice of the starting state as well as the specific realizations of the αN uncorrelated stored patterns. In the Hopfield model, as in all frustrated systems, the fluctuations in the energy of the system are sizable even in the equilibrium state. Thus special care has to be exercised in constructing ensemble averages of numerical results before they can be compared with the predictions of a theoretical formula.

3.1. Zero-Temperature Dynamics

For simplicity we shall first discuss the zero-temperature dynamics. In the context of associative memory, the interesting trajectories of Eq. (2.2) belong to the case when one of the overlaps R_μ, say R_{μ_0}, is much larger than the other overlaps. In this case,

$$\sum_\mu R_\mu^2(t)(1-\delta_{\mu\mu_0}) = \alpha - 2e(\alpha, t) - R_{\mu_0}^2(t) \tag{3.5}$$

We will focus on the development in time of the overlap of $\{\xi_i^{\mu_0}\}$ with $\{S_i(t)\}$. We have

$$\begin{aligned}
\xi_i^{\mu_0} S_i(t+1) &= \xi_i^{\mu_0} \operatorname{sign} \sum_j \mathbf{J}_{ij} S_j(t)(1-\delta_{ij}) \\
&= \operatorname{sign} \xi_i^{\mu_0} \sum_j \mathbf{J}_{ij} S_j(t)(1-\delta_{ij}) \\
&= \operatorname{sign} \xi_i^{\mu_0} \sum_\mu \sum_j \xi_i^\mu \xi_j^\mu S_j(t)(1-\delta_{ij}) \\
&= \operatorname{sign} \left\{ \sum_\mu (\xi_i^{\mu_0} \xi_i^\mu) \sum_j \xi_j^\mu S_j(t)(1-\delta_{ij}) \right\} \\
&= \operatorname{sign} \left\{ R_{\mu_0}(t) + \sum_\mu \xi_i^{\mu_0} \xi_i^\mu R_\mu(t)(1-\delta_{\mu\mu_0}) - \alpha \xi_i^{\mu_0} S_i \right\} \\
&= \operatorname{sign} \left[M_{\mu_0}(t) + \sum_\mu \xi_i^{\mu_0} \xi_i^\mu M_\mu(t)(1-\delta_{\mu\mu_0}) \right] \\
&\qquad (i = 1, 2, \ldots, N)
\end{aligned} \tag{3.6}$$

where

$$M_\mu(t) = \frac{1}{N} \sum_j \xi_j^\mu S_j(t)(1-\delta_{ij}) \tag{3.7}$$

The above equations contain the full dynamics of the Hopfield model. There is no approximation so far. The object of the following analysis is to obtain a macroscopic description of Eq. (3.6) which incorporates the essential physics of the microscopic dynamics. For this purpose, we will reduce the set of Eqs. (3.6) to a single equation for $R_{\mu_0}(t)$, i.e., for the overlap of $\{S_i(t)\}$ with the stored memory closest to it.

A comparison of Eqs. (3.2) and (3.7) shows that R_μ and M_μ differ by a very small quantity of the order of N^{-1} which goes to zero in the thermodynamic limit. The important difference is that, unlike R_μ, the quantity $M_\mu(t)$ in Eq. (3.6) does not involve the ith bit of the stored picture and is therefore independent of the multiplicative prefactor $\xi_i^{\mu_0}\xi_i^\mu$. This will be important later when we apply the central limit theorem to the right-hand side of Eq. (3.6).

We consider parallel updating of spins, and sum Eq. (3.6) over i. This yields

$$NR_{\mu_0}(t+1)=N_+ - N_-; \qquad N_+ + N_- = N$$

Here N_+ and N_- denote the number of times the quantity in the square brackets in Eq. (3.6) is positive or negative respectively. The quantity in the square brackets consists of a "signal" term $M_{\mu_0}(t)$ and a "noise" term which consists of a sum of $p-1$ terms which are independent because of the multiplicative prefactor $\xi_i^{\mu_0}\xi_i^\mu$. The multiplicative prefactors are independent of M_μ and different multiplicative factors are independent of each other because the stored patterns are uncorrelated with each other. Therefore the central limit theorem can be applied to the noise term, and we may take it to have a Gaussian distribution with average zero and standard deviation σ given by

$$\sigma^2 = \sum_\mu M_\mu^2(1-\delta_{\mu\mu_0}) \tag{3.8}$$

Now it is easy to see that

$$\frac{N_\pm}{N} = \frac{1}{2}\left[1 \pm \operatorname{erf}\left(\frac{M_{\mu_0}(t)}{2^{1/2}\sigma}\right)\right]$$

or

$$R_{\mu_0}(t+1) = \operatorname{erf}\left(\frac{M_{\mu_0}(t)}{2^{1/2}\sigma}\right) \tag{3.9}$$

In the above equation R_{μ_0} can be replaced by M_{μ_0} without introducing any error in the thermodynamic limit. Thus we get

$$M_{\mu_0}(t+1) = \operatorname{erf}\left(\frac{M_{\mu_0}(t)}{2^{1/2}\sigma}\right) \tag{3.10}$$

where

$$\sigma^2 = \alpha - 2e(\alpha, t) - M_{\mu 0}^2 \tag{3.11}$$

Equation (3.5) has been used in writing Eq. (3.11). Equation (3.10) is a closed-form equation for a single parameter $M_{\mu_0}(t)$. As there is only one parameter now, the subscript μ_0 may be dropped for convenience. We get

$$M(t+1) = \operatorname{erf}\left(\frac{M(t)}{\{2[\alpha - 2e(\alpha, t) - M^2(t)]^{1/2}\}}\right) \tag{3.12}$$

Equation (3.12) is our main result for the $T = 0$ dynamics of the Hopfield model. A special form of this equation valid at the fixed point, and with the fixed point value of $e(\alpha)$ set equal to -0.5, i.e., $e(\alpha) = e^*(\alpha) = -0.5$, was proposed by Kohring.[8] The advantage of the full equation is that it allows us to understand the main features of the Hopfield dynamics in terms of a single parameter $e(\alpha, t)$ which we can interpret as the energy of the system per spin in the thermodynamic limit. We have verified by direct numerical computation for large N and p (we performed numerical tests for $N = 4000$ and $p = 120$) that Eq. (3.6), which contains all the microscopic details of the dynamics without approximation, gives the same result as Eq. (3.12) if $e(\alpha, t)$ is calculated from (3.6) and substituted in (3.12).

The long-time behavior of Eq. (3.12) is determined by a fixed-point value $e^*(\alpha)$. Equation (3.12) has a fixed-point solution with a large overlap $M^* > 0.97$ for $\alpha - 2e^*(\alpha) < 1.14$. This fixed point may be associated with the memory retrieval phase. A novel feature of Eq. (3.12) is that it is the quantity $\alpha_c - 2e^*(\alpha_c)$, rather than α_c alone which determines the memory retrieval phase. The transition occurs at $\alpha_c - 2e^*(\alpha_c) = 1.14$. An analytic calculation of the fixed-point energy $e^*(\alpha)$ appears to be a difficult task. However, we may compare the prediction of Eq. (3.12) with the replica-symmetric result for the quantity $e^*(\alpha)$ at the threshold of the storage capacity of the network. If we ignore the stability problem of the replica solution, the replica-symmetric result for $e^*(\alpha_c)$ is $e^*(\alpha_c) = -0.5014$. If we use this value in the equation $\alpha_c - 2e^*(\alpha_c) = 1.14$, we get $\alpha_c = 0.137$, which agrees with the replica-symmetric result for α_c. Thus, we are able to recover the replica prediction for α_c in our formalism if we use the replica result for $e^*(\alpha_c)$. This is interesting because it could not have been seen beforehand.

Equation (3.12) also provides a qualitative idea of the basins of attraction. It should be kept in mind that we have derived (3.12) with the help of probabilistic arguments based on the signal-to-noise ratio. Our derivation is therefore applicable when the signal is larger than the noise, or $M^2(t) > [\alpha - 2e(\alpha, t) - M^2(t)]$. Let us assume that our starting pattern at $t = 0$ has an overlap $M(t = 0)$ with one of the stored patterns, and is uncorrelated with the other patterns. Then it is easy to see that $e(\alpha, t=0) = -0.5M^2(t=0)$. Thus, at the first time step, Eq. (3.12) takes the form

$$M(t=1) = \operatorname{erf}\left(\frac{M(t=0)}{(2\alpha)^{1/2}}\right)$$

The above equation tends to increase any initial overlap $M(t=0)$ for $\alpha < 2/\pi$. Numerical simulations also show a similar trend at the initial stages of the dynamics. However, only the trajectories with $M^2(t=0) > \alpha$ (signal larger than noise) may be expected to lead to the retrieval of the memory. In other words, only those starting patterns which have $M(t=0) > \alpha^{1/2}$ may be expected to lie inside the basin of attraction of the corresponding memory. This expectation is in fairly good agreement with the numerical simulation results.

It would be nice to have a simple dynamical equation for the quantity $e(\alpha, t)$. This requires equations for the smaller overlaps $M_\mu(t)$, $\mu \neq \mu_0$. We may attempt to construct these equations by writing an expression for $\xi_i^\mu S_i(t+1)$ similar to Eq. (3.6). However, in this case the noise term dominates over the signal term. The noise term is dominated by the largest overlap term. The term gives $\xi_i^\mu S_i(t+1) = \operatorname{sign}(\xi_i^{\mu_0} \xi_i^\mu)$. In this approximation we find $M_\mu(t) = \pm N^{-1/2}$ independent of t, and $e(\alpha, t) = -\frac{1}{2} M_{\mu_0}^2(t)$, which gives

$$M(t+1) = \operatorname{erf}\left(\frac{M(t)}{(2\alpha)^{1/2}}\right) \tag{3.13}$$

Unfortunately, the above approximation misses the subtle correlation effects of the Hopfield dynamics whose effect appears to be to increase the quantity $\sigma^2 = \sum_\mu M_\mu^2(1 - \delta_{\mu\mu_0})$ during the evolution of the pattern. Equation (3.13) gives a second-order transition at $\alpha_c = 2/\pi$ against the observed first-order transition at $\alpha_c = 0.14$.

3.2. Numerical Results and a Model Recursion Relation

We have studied the quantity $e(\alpha, t)$ numerically in the special case when the initial pattern is one of the stored patterns. We performed numerical simulations on finite-size systems with $N = 100$, 200, and 400. In each case αN uncorrelated patterns were generated and stored in the network memory. Then each stored pattern was tested under the zero-temperature Hopfield dynamics to see if it approached a fixed point. The procedure was repeated till we had obtained 1000 attractors for each α. The large number of attractors tested offset to some extent the small size of the networks ($N \leqslant 400$) which we could study conveniently on our PC 486. Table I shows the results of the average energy of the initial state and the final state (fixed-point pattern) for $N = 400$ and several values of α. The results for other values of N are similar.

Our numerical results indicate that there are perhaps two prominent aspects of the Hopfield dynamics. If the starting pattern is not already very close to a stored memory, then the first step of the parallel dynamics transforms the pattern very nearly into the stored memory. This step lowers the energy of the network significantly when the initial overlap is small. Subsequent steps of the dynamics appear to increase the entropy of the system without lowering the energy very much. This is done by taking a small fraction off the large overlap and distributing it among the overlaps with the other stored pictures. If the starting pattern has an overlap M_0 with one of the stored patterns and is uncorrelated with the other stored patterns, then we can make a very useful approximation for the quantity $e(\alpha, t)$. This approximation is as follows:

$$\begin{aligned} e(\alpha, t) &= -\tfrac{1}{2}M_0^2 && \text{for} \quad t = 0 \\ &= -\tfrac{1}{2} && \text{for} \quad t \geqslant 1 \quad \text{and} \quad \alpha \leqslant 0.14 \end{aligned} \tag{3.14}$$

Figure 1 shows the fixed points of Eq. (3.12) with $e(\alpha, t)$ given by Eq. (3.14). The upper curve shows the fixed-point value M^*, while the lower curve shows the minimum value of the initial overlap M_0 which leads to the memory retrieval. Figure 1 is in fair agreement with the available numerical results for the basins of attraction.[4] The results in ref. 4 are for sequential dynamics, but similar trends are observed in parallel dynamics.

It is also interesting to note that an energy-conserving parallel dynamics of the Hopfield model also possesses the main features of an associative memory. By an energy-conserving dynamics we mean a dynamics where at each step the p overlaps are constrained such that $e(\alpha, t) = -0.5$. The fixed points of Eq. (3.12) with $e(\alpha, t)$ set equal to -0.5

Table I

α	$\langle e(\alpha, t=0)\rangle$	$\langle e^*(\alpha)\rangle$
0.05	-0.500 ± 0.001	-0.500 ± 0.001
0.10	-0.500 ± 0.001	-0.500 ± 0.001
0.11	-0.500 ± 0.001	-0.500 ± 0.001
0.12	-0.500 ± 0.001	-0.500 ± 0.001
0.13	-0.500 ± 0.001	-0.500 ± 0.001
0.14	-0.500 ± 0.001	-0.500 ± 0.001

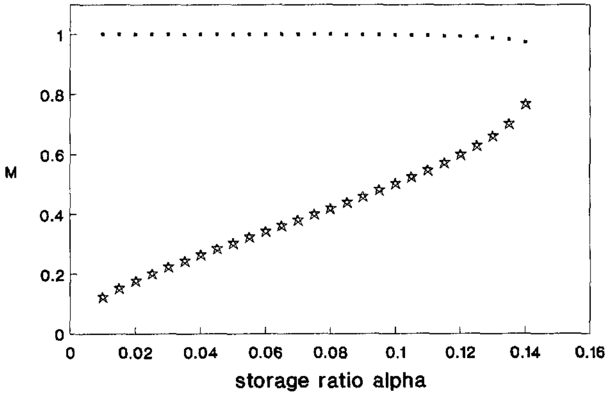

Fig. 1. Basins of attraction for memory retrieval with Eqs. (3.12) and (3.14): The lower curve shows the minimum initial overlap which leads to memory retrieval. The upper curve shows the overlap of the retrieved memory with the stored pattern.

are described in ref. 8. In the limit $\alpha = 0$, the nontrivial stable fixed point is reached for an initial overlap $M(t=0) \geqslant 0.765$. The basin of attraction of the nontrivial fixed point decreases with increasing storage ratio α. Thus, Eq. (3.12) with $e(\alpha, t) = -0.5$, i.e., the energy-conserving Hopfield dynamics, contains the main qualitative features of the non-energy-conserving dynamics. This is remarkable because it is generally thought that the most characteristic feature of the Hopfield dynamics is to move the system to states of lower energy. For sequential updating of spins, the Hopfield dynamics certainly moves the system to states of lower energy. However, this is not necessarily so for parallel updating. The more characteristic feature of (parallel updating) Hopfield dynamics seems to be an increase in entropy due to the increase in the smaller overlaps rather than a decrease in the energy of the system.

Before concluding this section, we mention that each fixed point of Eq. (3.12) in fact represents a very large number of stable solutions of Eq. (2.2), i.e., all distinct configurations which are compatible with the fixed-point value of the order parameter M^*. These configurations are optimized with respect to each single spin flip. The quantity $S_i(t+1)\,S_i(t)$ measures whether the spin $S_i(t)$ is flipped at the next time step or not. If $S_i(t+1)\,S_i(t) > 0$, then the spin is not flipped. Equivalently, if $\sum_j \mathrm{J}_{ij} S_j(t)$ $S_i(t)(1-\delta_{ij}) > 0$, then $S_i(t+1) = S_i(t)$. This condition again leads to the iterative map (3.12). Therefore the fixed points of (3.12) correspond to patterns which have been optimized with respect to each individual spin flip. There is a very large number of such patterns. For example, at $\alpha_c = 0.14$, the number of configurations corresponding to $M^* = 0.97$ scales as $\exp(0.08N)$.

3.3. Finite-Temperature Dynamics

The idea behind introducing temperature in the network dynamics is to make it more stochastic. At a finite temperature T, $S_i(t+1)$ need not be $+1$ even if the local field $h_i(t)$ is positive. The probability distribution for $S_i(t+1)$ is given by Eq. (2.3). The distribution depends on the temperature T and the local field h_i. The local field itself is distributed randomly from site to site. Thus the expectation value of $S_i(t+1)$ over the network involves two averages, an average over $P(h_i)$ which yields $\tanh(\beta h_i)$, and a further average over the distribution of local fields. As in the case of zero-temperature dynamics, we shall focus on the expectation value of $\xi_i^{\mu_0} S_i(t+1)$:

$$\begin{aligned} M_{\mu_0}(t+1) &= \langle\langle \xi_i^{\mu_0} S_i(t+1) \rangle\rangle \\ &= \langle\langle \xi_i^{\mu_0} \tanh[\beta h_i(t)] \rangle\rangle \\ &= \langle\langle \tanh[\beta \xi_i^{\mu_0} h_i(t)] \rangle\rangle \\ &= \left\langle\!\!\left\langle \tanh\left\{\beta\left[M_{\mu_0}(t) + \sum_{\mu} \xi_i^{\mu_0}\xi_i^{\mu} M_\mu(t)(1-\delta_{\mu\mu_0})\right]\right\}\right\rangle\!\!\right\rangle \end{aligned} \tag{3.15}$$

Following the same steps which took us from Eq. (3.6) to (3.12), we obtain

$$M(t+1) = \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{+\infty} dz \exp\left[-\frac{1}{2}\left(\frac{z-M(t)}{\sigma}\right)^2\right] \tanh\left(\frac{z}{T}\right) \tag{3.16}$$

where σ is given by Eq. (3.11),

$$\sigma = [\alpha - 2e(\alpha, t) - M^2(t)]^{1/2}$$

Equation (3.16) is our main result for finite-temperature dynamics, just as Eq. (3.12) was for $T=0$. It can be verified that Eq. (3.16) reduces to Eq. (3.12) when $T=0$. We can also check that in the limit $\sigma \to 0$ (as when only one pattern is stored and the Hopfield model reduces to an Ising model of a ferromagnet) we obtain the familiar mean-field equation $M(t+1) = \tanh[\beta M(t)]$, which has a second-order transition at $T_c = 1$. The role of finite α is to lower the transition temperature and more importantly to make the transition first order. Equation (3.16) can be studied numerically to obtain the phase boundary T_c vs. α below which the network can function as an associative memory. This is shown in Fig. 2 for $M_0 = 1$ and $e(\alpha, t)$ given by (3.14). The main qualitative difference between the phase diagram so obtained and the phase diagram obtained by Amit *et al.*[2] is that we do not have two kinds of ferromagnetic phases predicted by the replica theory. As far as we are aware, the numerical

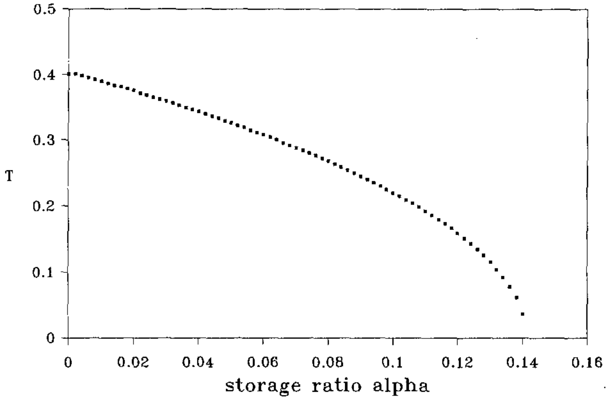

Fig. 2. Phase diagram obtained from Eqs. (3.16) and (3.14) for $M_0 = 1$. The phase boundary marks a first-order transition. The low-temperature phase corresponds to memory retrieval.

simulations do not indicate two types of memory retrieval phases. The phase diagram of Fig. 2 appears to be in fair agreement with the numerical simulations of the finite-temperature dynamics.

4. CONCLUSION

We have presented a simple and reasonably satisfactory theory of the Hopfield dynamics without recourse to the replica method. As we have already mentioned, there are several good reasons for avoiding the replica method. The replica formalism and particularly the notion of replica symmetry breaking is rather unphysical. The simplest solution in the replica method, i.e., the replica-symmetric solution, is unstable at $T = 0$ for $\alpha > 0.05$. Moreover, the theory of Amit *et al.*[2] based on the replica method focuses only on the equilibrium states of the system, which are assumed to form a canonical ensemble. On the other hand, the attractors of the Hopfield dynamics form a very large number of nearly degenerate states which are separated from each other by infinitely high barriers and it is doubtful if the standard techniques of equilibrium statistical mechanics and the canonical ensemble can be applied to the ensemble of the attactor states. In this background, it is rather satisfying that a simple analysis of the Hopfield dynamics based on the signal-to-noise ratio gives a reasonably satisfactory understanding of the basic phenomena.

Finally, we must also mention the points where our analysis lacks rigor, and those aspects of the numerical simulations which it fails to explain in its present form. We have argued that the noise has a Gaussian distribution. This is not rigorously correct. The reason is that the Hopfield dynamics builds up correlations between the signal and the noise. If the noise at a site i is larger than the signal and of opposite sign, then the spin at that site is flipped opposite to $\xi_i^{\mu_0}$. This contributes to the reduction in the signal at the next time step, and may make the noise distribution somewhat skewed. We will not go into this further, but suffice it to say that the approximation (3.14), which amounts to increasing the width of the noise distribution from α to $\alpha + (1 - M^2)$, captures the main trend of the parallel dynamics. The other point is that our theory predicts $M^* = 0$ for $\alpha > 0.14$. Numerical simulations[9] show that $M^* = 0.20$ in a small region beyond $\alpha = 0.14$. Although these results are based on sequential updating of spins, it is likely that similar results also hold for parallel dynamics. This aspect of numerical simulations is not explained by any theory at present. It requires further numerical as well as analytical study of the Hopfield dynamics which goes beyond the scope of the present paper.

REFERENCES

1. J. J. Hopfield, *Proc. Natl. Acad. Sci. USA* **79**:2254 (1982); **81**:3088 (1984).
2. D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. Lett.* **55**:1530 (1985); *Ann. Phys.* **173**:30 (1987).
3. B. M. Forrest, *J. Phys. A* **21**:245 (1988).
4. H. Horner, D. Bormann, M. Frick, H. Kinzelbach, and A. Schmidt, *Z. Phys. B* **76**:381 (1989).
5. M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
6. A. Crisanti, D. J. Amit, and H. Gutfreund, *Europhys. Lett.* **2**:337 (1986).
7. H. Reiger, M. Schreckenberg, and J. Zittartz, *Z. Phys. B* **72**:523 (1988).
8. G. A. Kohring, *Europhys. Lett.* **8**:697 (1989).
9. G. A. Kohring, *J. Stat. Phys.* **59**:1077 (1990).