Hopfield model with self-coupling

Manoranjan P. Singh
Laser Physics Division, Centre for Advanced Technology, Indore 452013, India
(Received 4 June 2001; published 25 October 2001)

We have studied analytically the retrieval performance of a Hopfield model in the presence of self-coupling in the synaptic matrix. We find, contrary to expectations from some earlier studies based on the counting of fixed points, that negative self-coupling causes deterioration in the retrieval performance of the network. On the other hand, it is possible to enhance the retrieval performance by having a positive self-coupling of appropriate magnitude.

DOI: 10.1103/PhysRevE.64.051912 PACS number(s): 87.18.Sn, 64.60.Cn, 75.50.Lk, 05.50.+q

I. INTRODUCTION

Neural network models of associative memory have received a great deal of attention in recent years [1,2]. Consequently, various learning rules have been designed for storing and recalling patterns. These learning rules are used to obtain the off-diagonal elements of the synaptic matrix corresponding to the interaction strength between two different model neurons. The diagonal elements of the synaptic matrix which represent the self-interaction strength of the model neurons are generally considered to be a nuisance [2,3] and are set equal to zero. However, there are reasons to study the neural network models with nonzero self-interaction terms. First, nonzero self-interaction terms do appear when one considers the models of nonlinear holographic associative memories [4]. Modeling of the firing mechanism of neurons may also require introduction of the self-interaction in the dynamics of the neural network (see Ref. [5] for a detailed discussion of this aspect). Second, in contrast to the general belief, the presence of the self-interaction term in the synaptic matrix may even improve the performance of the network as an associative memory. Krauth *et al.* [6] have shown that a well-chosen value of the diagonal coupling can increase basins of attraction of stored patterns in the perceptron-like neural-network model. They have also discussed the positive role of the self-coupling for the retrieval properties of Hopfield-type models. They have demonstrated, by numerical simulations on a fully connected Hopfield net with pseudoinverse learning, the possibility of enhancing the basin of attraction substantially. Fontanari and Köberle [7] have carried out a detailed equilibrium analysis of the Little model (the synchronous version of the Hopfield model) in the presence of the self-coupling when the number of patterns p grows as $p=\alpha N$, where N is the number of neurons. It is shown that the self-coupling can be used to control the occurrences of cycles, thereby improving the performance of the model as an associative memory. It is also shown that the synchronous dynamics is much more stable to noise than the asynchronous one for the same memory loading level α, provided that neurons have a sufficiently large self-coupling. Athithan and Dasgupta [5] have performed Monte Carlo simulations for a Hopfield model with a negative self-coupling. They have shown that the self-coupling of an appropriate magnitude causes suppression of the spurious fixed-point attractors in the Hopfield model for $\alpha \leqslant 0.05$. Hence, the fractional volume of basins of spurious patterns goes to zero, while that of stored patterns remains finite. As explained below, the self-interaction term can also be considered as a nonzero threshold for the postsynaptic potential of neurons in the network. In the context of optimal neural networks, it is possible to select states with sizable basins of attraction by having a suitable threshold for the postsynaptic potential [8,9]. Treeves and Amit [10] have considered the effect of the positive threshold (equivalently the negative self-coupling) on the number of fixed points of an asymmetrically diluted Hopfield model in the limiting cases of very low and very high levels of memory loading. The general conclusion is that such a threshold causes an exponential reduction in the number of fixed points of the network.

In this paper, we present results of analytical studies of a Hopfield model in the presence of a self-interaction term. It is well known [1] that it is not possible to define the Lyapunov or energy function for such a system. As a result, one cannot use in the present case the methods of equilibrium statistical mechanics, which have been extensively used to analyze the Hopfield model. We concentrate, therefore, on the dynamics of the model, which is governed by the local alignment field. First, we look at the structure of the locally stable fixed points (i.e., states that are stable to all single spin flips) of the model. The counting of fixed points in the Hopfield model produces the following picture [11]. The fixed points, which are exponentially large in N, appear only in two distinct regions of phase space: in a narrow ''retrieval'' band where the fixed points are strongly correlated with a memory state, and in a wide ''spurious'' band which is centered around states having no macroscopic overlap with the chosen memory state. The two bands are disjoint from each other only below a certain critical value of $\alpha = 0.113$. It should be noted that the retrieval band is not exactly centered around the memory state. It happens as some of the neurons undergo spin-glass–type freezing in random individual states under the influence of conflicting synaptic inputs [1].

When the self-coupling d is increased from zero to a positive value, it results in the exponential growth of fixed points both in the retrieval and the spurious bands. These bands become wider with d and finally merge together. For moderate values of d, the retrieval band grows faster than the spurious band. It gives rise to the possibility of having an optimal value of the self-coupling where the positive effects due

to the growth of the retrieval band may take over the negative effects due to the growth the spurious band so far as the retrieval performance of the network is concerned. On the other hand, the presence of negative self-coupling results in making the two bands narrower by destabilizing exponentially many fixed points present in these bands. As a result, the two bands move farther apart in phase space. As is the case with the positive self-coupling, fixed points in the retrieval band are much more sensitive to the self-coupling than those in the spurious band. The retrieval band gets completely suppressed by a negative self-coupling of a very small magnitude compared to what is required to completely suppress the spurious band. This is not in agreement with the numerical simulation results of Ref. [5], namely that it is possible to suppress completely the spurious fixed points for $\alpha<0.05$. Further, the magnitude of the self-coupling which is required to suppress the retrieval band has a strong dependence on α. The higher the value of α, the smaller is the magnitude of the self-coupling required to completely suppress the retrieval band. Once again we are faced with a situation in which the positive effects due to suppression of the spurious band are accompanied by the negative effects that may arise due to suppression of the retrieval band. Naturally, the question arises whether there exists an optimal value of the negative self-coupling for the retrieval performance of the network. The question seems more relevant for smaller values of α, where the retrieval band survives for comparatively higher magnitude of the negative self-coupling.

At this point, it should be mentioned that the complete suppression of the retrieval band does not *necessarily* stop the function of the network as an associative memory [1,10]. A fixed point in the retrieval band is initially destabilized as the spin-glass freezing of the randomly aligned neurons is destroyed by a negative self-coupling of rather small magnitude, thereby causing a hopping around. This would, however, not affect the overlap with the stored pattern, which will remain fixed and large, as it is determined by the rest of the neurons. As long as α is small, the distinction between a fixed point and a trajectory spanning a small phase space corresponding to neurons with weak local fields is not very significant from an operational point of view. The network would work effectively provided that the dynamics draws the network to the neighborhood of a memory, even if a small fraction of neurons keep changing their states. If the time-averaged overlap is high enough, the memory would be properly recalled. From this point of view, it seems natural to study the time evolution of the overlap.

The foregoing discussion brings out the need for a dynamical theory. To this end, we generalize a dynamical theory due to Coolen and Sherrington [12,13], which has been used to study the dynamics of the Hopfield model with an extensive number of stored patterns on finite time scales. We find that the time taken by the network to converge to the desired memory state reduces in the presence of a positive self-coupling. It is also possible to enhance the basins of attraction of stored patterns by having a positive self-coupling of appropriate magnitude. The retrieval performance is found to deteriorate in the presence of a negative self-coupling—the convergence time increases, the basin of attraction decreases. One may expect enhancement in the storage capacity due to the presence of a negative self-coupling by looking at the structure of fixed points. However, this does not happen. Once again, contrary to the general expectation, we find that it is possible to enhance somewhat the storage capacity of the network by having a positive self-coupling. The positive self-coupling is also beneficial in different regions of the $(T-\alpha)$ phase diagram [1,14]. We observe faster retrieval and enhancement in the basin of attraction in the mixed phase, too. Moreover, retrieval becomes possible in some regions of the spin-glass phase.

The remaining part of this paper is organized as follows. Section II contains a description of the model we consider. In Sec. III, we present the results on the structure of fixed points of the deterministic dynamics. We study the dynamics of the model on finite time scales in Sec. IV. Section V contains a summary of the main results of this study and a few concluding remarks.

II. MODEL

The network under investigation consists of N two-state model neurons (''spins'') σ_i, each of which may assume the values $+1$ or -1. A configuration or state of the network is defined by giving specific values to all of its N spins. The off-diagonal elements of the synaptic interconnection matrix are given by the modified Hebb rule,

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu, \quad i\neq j, \tag{1}$$

where the $\{\xi_i^\mu\}$, $\mu=1,\ldots,p$ are the stored patterns or memories. Each ξ_i^μ may take the values ±1 with equal probability. The number of patterns stored in the network is p, and $\alpha=p/N$ is the memory-loading level of the network. Instead of taking the diagonal elements of the synaptic matrix $J_{ii}=0$, which is the case in the Hopfield model [15], we take them to be

$$J_{ii}=d, \quad \forall i. \tag{2}$$

The noise-free (zero temperature) dynamics of the network is given by the sequential update,

$$\sigma_i(t+\delta t)=\mathrm{sgn}[h_i(t)], \tag{3}$$

where the local field at the spin σ_i is given by

$$h_i(t)=\sum_j J_{ij}\sigma_j(t). \tag{4}$$

In order to compare this dynamics with that of the standard Hopfield model, we rewrite the local field $h_i(t)$ as

$$h_i(t)=h_i^0(t)+d\sigma_i(t), \tag{5}$$

where $h_i^0(t)=\Sigma_{j\neq i}J_{ij}\sigma_j(t)$ is the local field at the ith spin in the Hopfield model. It is possible to express the update rule for the ith spin as

$$\sigma_i(t+\delta t)=\mathrm{sgn}[h_i^0(t)+d\sigma_i(t)]. \tag{6}$$

Thus, a negative (positive) self-coupling in the synaptic matrix has the effect of introducing a positive (negative) threshold in the dynamics of the Hopfield model. It is clear from Eq. (6) that if $|d|\gg|\langle h_0\rangle|$, the initial state will not flow to the desired memory and the network will not act as an associative memory. Therefore, it is important to ensure by choosing carefully the magnitude of d that the dynamics of the network is not dominated by the self-coupling. It can be seen below that the average value of h_0 at any time is given by the overlap with the desired memory. Hence the magnitude of the self-coupling should be much smaller than the starting overlap with the memory state to be retrieved. We will now discuss the effect of the self-coupling on the structure of fixed points.

III. STRUCTURE OF FIXED POINTS

Fixed points are the states that remain unchanged under the single spin-flip dynamics given by Eq. (3). Accordingly, a state $\vec{\sigma}\equiv(\sigma_1,\ldots,\sigma_N)$ is a fixed point if it satisfies the following condition:

$$\sigma_i=\mathrm{sgn}[h_i],\quad i=1,\quad\ldots,N. \tag{7}$$

This can also be expressed as

$$h_i\sigma_i>0,\quad i=1,\ldots,N. \tag{8}$$

We follow closely the approach of Gardner [11] to calculate the average number of fixed-point attractors $\langle N_{fp}(N,\alpha,d,g)\rangle$ at a Hamming distance Ng from a stored pattern. We consider a state $\vec{\sigma}$ which is at a Hamming distance Ng from the νth stored pattern $\vec{\xi}^{\nu}$. According to Eq. (8), the state will be a fixed point if the quantity

$$R_i^{\nu}=\sigma_i\sum_j J_{ij}\sigma_j>0,\quad i=1,\ldots,N, \tag{9}$$

so that the average number of fixed points at a Hamming distance Ng from the νth stored pattern is given by

$$\langle N_{fp}(N,\alpha,d,g)\rangle=\int_0^{\infty}\prod_i d\lambda_i\mathrm{Tr}_{\{\sigma_i\}}\left\langle\prod_i\delta(\lambda_i-R_i^{\nu})\right\rangle. \tag{10}$$

Separation of the term coming from the νth pattern and the interference term coming from other patterns gives

$$R_i^{\nu}=1-2g+\frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu}\xi_i^{\mu}\xi_j^{\mu}\sigma_i\sigma_j+d \tag{11}$$

for $N(1-g)$ values of i for which $\xi_i^{\nu}=\sigma_i$, and

$$R_i^{\nu}=2g-1+\frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu}\xi_i^{\mu}\xi_j^{\mu}\sigma_i\sigma_j+d \tag{12}$$

otherwise. Using the integral representation for the δ functions in Eq. (10) and averaging over the patterns (see Ref. [16] for technical details of such a calculations), we get in the thermodynamic limit

$$\langle N_{fp}(N,\alpha,d,g)\rangle\approx e^{NF(\alpha,d,g)}. \tag{13}$$

$F(\alpha,d,g)$ is given in the terms of the saddle-point parameters a and b as

$$F(\alpha,d,g)=\alpha\left[b-\frac{1}{2}+\frac{(1-b)^2}{2a}+\frac{1}{2}\ln a\right]+(1-g)\ln\phi(t)+g\ln\phi(u)-g\ln g-(1-g)\ln(1-g), \tag{14}$$

where

$$t=\frac{2g-1+\alpha b-d}{\sqrt{\alpha a}}, \tag{15}$$

$$u=\frac{1-2g+\alpha b-d}{\sqrt{\alpha a}}, \tag{16}$$

and

$$\phi(x)=\frac{1}{\sqrt{2\pi}}\int_x^{\infty}dy\,\exp(-y^2/2). \tag{17}$$

The saddle-point parameters a and b satisfy the following equations:

$$\alpha\left[\frac{(1-b)^2}{a}-1\right]+(1-g)t\frac{\phi'(t)}{\phi(t)}+gu\frac{\phi'(u)}{\phi(u)}=0, \tag{18}$$

$$\alpha\left[1-\frac{(1-b)}{a}\right]+(1-g)\sqrt{\frac{\alpha}{a}}\frac{\phi'(t)}{\phi(t)}+g\sqrt{\frac{\alpha}{a}}\frac{\phi'(u)}{\phi(u)}=0. \tag{19}$$

We have solved Eqs. (18) and (19) numerically for a and b for different values of α and d to get corresponding $F(g)$. In Fig. 1, we have plotted $F(g)$ for $\alpha=0.05$. We note here that according to Eq. (13), fixed points exist only in regions of the phase space where $F(g)\geqslant 0$ as $N\to\infty$. As is well known, for $d=0$ we have fixed points in two distinct regions of phase space—in a narrow ''retrieval'' band where the fixed points are strongly correlated with the chosen memory state ($g\approx 0$), and in a wide ''spurious'' band which is centered around states having no macroscopic overlap with the memory state ($g=0.5$). These two bands are well separated from each other. For $d=0.02$, we have the same structure of fixed points. However, additional fixed points have appeared in both bands, making both of them broader. The retrieval band grows faster than the spurious band. For instance, at $d=0.02$, the peak value of F in the retrieval band is three to four orders of magnitude larger than that at $d=0$. On the other hand, the peak value of F in the spurious band at d

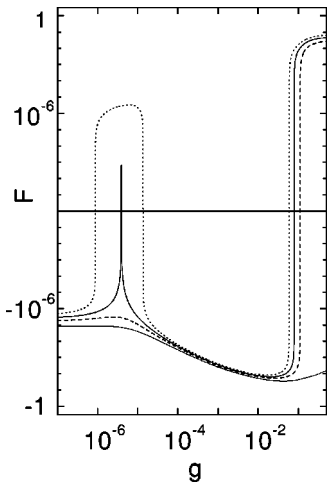

FIG. 1. $F(g)$ for various values of the self-coupling d. $\alpha = 0.05$. Full curve, $d=0$; dotted curve, $d=0.02$, dashed curve, $d=-0.02$; full curve at the bottom, $d=-0.06$.

$=0.02$ is only 1.3 times larger than that at $d=0$. When the self-coupling strength is reduced to a negative value $d=-0.02$, the retrieval band has become completely suppressed. The spurious band has still exponentially many fixed points in spite of a small fraction of fixed points becoming destabilized. When the strength of the self-coupling term is further reduced to $d=-0.06$, both of these bands become completely suppressed. The higher sensitivity of the fixed points in the retrieval band is due to the fraction of spins, which is not aligned to the chosen memory state and hence has very low stability.

In Fig. 2, we present the results for $\alpha=0.113$ where in the standard Hopfield model ($d=0$) we have the two bands nearly overlapping each other. As the self-coupling strength d is reduced to $d=-0.002$, both of these bands become narrower and more separated from each other. The retrieval band becomes completely suppressed at $d=-0.003$ whereas the spurious band becomes completely suppressed at $d=-0.09$. It is worth noticing here that the magnitude of the self-coupling $|d_{cr}|$ which is required to suppress the retrieval band has a strong dependence on the memory-loading level α. There is almost an order of magnitude variation in $|d_{cr}|$ as we go from $\alpha=0.05$ to $\alpha=0.113$. The origin of the higher sensitivity of the retrieval states at higher values of α lies in the fraction of spins which is not aligned to the chosen memory state. The higher the value of α, the higher is the probability of finding a spin with lower stability. However, there is little variation in the magnitude of the self-coupling $|d_{cs}|$ needed to destabilize the spurious band as we go from $\alpha=0.5$ to $\alpha=0.113$.

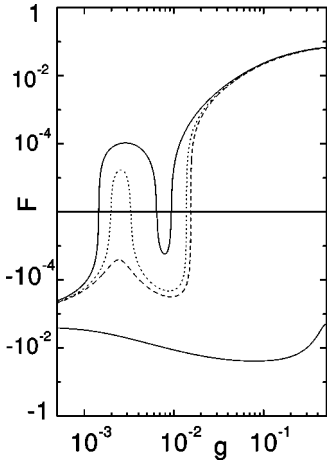

FIG. 2. Same as Fig. 1. $\alpha=0.113$. Full curve, $d=0$; dotted curve, $d=-0.002$; dashed curve, $d=-0.003$; full curve at the bottom, $d=-0.09$.

It is intriguing to investigate the effect of growth or suppression of fixed points on the retrieval properties of the network. In particular, it is interesting to study the behavior of the basin of attraction of stored patterns and the convergence time (time taken by an initial state to reach the desired state). This would be possible only in the framework of a dynamical theory, which will be discussed in the next section.

IV. DYNAMICS

To study the dynamics of the model, we generalize a dynamical theory due to Coolen and Sherrington [12,13]. The theory approximates reasonably well the dynamics of the fully connected Hopfield model near saturation on finite time scales. Here, we consider the stochastic dynamics of the network to account for the fast synaptic noise, which introduces undeterminism in the dynamics, e.g., given by Eq. (3). One is then compelled to carry out averages on ensembles of networks. It is convenient to start from a Glauber-type Markov chain description [1], where at each time step a randomly drawn spin is updated [and where the duration of each update δt is taken as $1/N$ so that on $O(N^0)$ time scales all spins have been updated once on average] such as

$$p_{t+\delta t}(\vec{\sigma}) = p_t(\vec{\sigma}) + \delta t \sum_{k=1}^{N} [p_t(F_k\vec{\sigma}) w_k(F_k\vec{\sigma}) - p_t(\vec{\sigma}) w_k(\vec{\sigma})]. \tag{20}$$

Here, $p_t(\vec{\sigma})$ is the probability of finding the system at time t in state $\vec{\sigma} \equiv (\sigma_1, \ldots, \sigma_N)$. F_k is a single spin-flip operator,

$$F_k \Phi(\vec{\sigma}) \equiv \Phi(\sigma_1, \ldots, -\sigma_k, \ldots, \sigma_N), \tag{21}$$

and the transition rates $w_k(\vec{\sigma})$ have the usual form,

$$w_k(\vec{\sigma}) \equiv \frac{1}{2}[1 - \sigma_k \tanh(\beta h_k(\vec{\sigma}))]. \tag{22}$$

The parameter β ($=T^{-1}$, the inverse of temperature) controls the degree of stochasticity. For $\beta=0$, the dynamics is completely random whereas for $\beta=\infty$ we recover the deter-

ministic update rule of Eq. (3). For $N\to\infty$, $\delta t\to 0$ and hence Eq. (20) results in the master equation:

$$\frac{d}{dt}p_t(\vec{\sigma})=\sum_{k=1}^{N}\left[p_t(F_k\vec{\sigma})w_k(F_k\vec{\sigma})-p_t(\vec{\sigma})w_k(\vec{\sigma})\right]. \tag{23}$$

We assume that the correlations $m_\mu(\vec{\sigma})\equiv(1/N)\Sigma_k\xi_k^\mu\sigma_k$ between system state and stored patterns are of order unity for $\mu=1$. The remaining $p-1$ correlations are assumed to be of order $1/\sqrt{N}$ (the condensed ansatz). Their cumulative impact on the system's dynamics is given by the order parameter $r(\vec{\sigma})$,

$$m(\vec{\sigma})\equiv\frac{1}{N}\sum_{k=1}^{N}\xi_k^1\sigma_k, r(\sigma)\equiv\frac{1}{\alpha}\sum_{\mu>1}^{p}\left[\frac{1}{N}\sum_{k=1}^{N}\xi_k^\mu\sigma_k\right]^2. \tag{24}$$

Local fields can now be expressed as

$$h_i(\vec{\sigma})=\xi_i^1[m(\vec{\sigma})+z_i(\vec{\sigma})]-\frac{1}{N}\sigma_i, \tag{25}$$

$$z_i(\vec{\sigma})\equiv\xi_i^1\sum_{\mu>1}^{p}\xi_i^\mu\frac{1}{N}\sum_{k\neq i}^{N}\xi_k^\mu\sigma_k+d\xi_i^1\sigma_i. \tag{26}$$

It is useful to define a distribution which gives the probability density in terms of the macroscopic order parameters (m,r):

$$P_t(m,r)\equiv\sum_{\vec{\sigma}}p_t(\vec{\sigma})\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma})). \tag{27}$$

Using Eq. (23), we can write the time derivative of the macroscopic distribution in the thermodynamic limit as

$$\frac{d}{dt}P_t(m,r)=\frac{\partial}{\partial m}\left\{P_t(m,r)\left[m-\int dzD_{m,r;t}[z]\tanh[\beta m+\beta z]\right]\right\}+2\frac{\partial}{\partial r}\left\{P_t(m,r)\left[r-1-\frac{1}{\alpha}\int dzD_{m,r;t}[z]z\tanh[\beta m+\beta z]\right.\right.$$
$$\left.\left.+\frac{d}{\alpha}\int dzD'_{m,r;t}[z]\tanh[\beta m+\beta z]\right]\right\}+\frac{1}{N}P_t(m,r)O\left[1,\int dzD_{m,r;t}[z]z,\int dzD_{m,r;t}[z]z^2,d,d^2\right], \tag{28}$$

where the intrinsic noise distributions $D_{m,r;t}[z]$ and $D'_{m,r;t}[z]$ are given by

$$D_{m,r;t}[z]\equiv\frac{\sum_{\vec{\sigma}}p_t(\vec{\sigma})\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))(1/N)\sum_i\delta(z-z_i(\vec{\sigma}))}{\sum_{\vec{\sigma}}p_t(\vec{\sigma})\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))}, \tag{29}$$

$$D'_{m,r;t}[z]\equiv\frac{\sum_{\vec{\sigma}}p_t(\vec{\sigma})\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))(1/N)\sum_i\sigma_i\xi_i^1\delta(z-z_i(\vec{\sigma}))}{\sum_{\vec{\sigma}}p_t(\vec{\sigma})\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))}. \tag{30}$$

The condensed ansatz allows us to neglect the last term on the right-hand side of Eq. (28) as the variance of $D_{m,r;t}[z]$ will remain finite for $N\to\infty$. Thus Eq. (28) takes the Liouville form on finite time scales in the limit $N\to\infty$. It therefore leads to deterministic evolution of the order parameter (m,r):

$$P_t(m,r)=\delta(m-m^*(t))\delta(r-r^*(t)),\quad N\to\infty, \tag{31}$$

where the deterministic trajectory $(m^*(t),r^*(t))$ is given by the solution of the coupled flow equations:

$$\frac{d}{dt}m=\int dzD_{m,r;t}[z]\tanh[\beta m+\beta z]-m, \tag{32}$$

$$\frac{1}{2}\frac{d}{dt}r=\frac{1}{\alpha}\int dzD_{m,r;t}[z]z\tanh[\beta m+\beta z]$$
$$-\frac{d}{\alpha}\int dzD'_{m,r;t}[z]\tanh[\beta m+\beta z]+1-r. \tag{33}$$

Although Eqs. (32) and (33) are exact within the condensed ansatz, the trouble with them is that they contain the distributions given by Eqs. (29) and (30), which depend on the microscopic distribution $p_t(\vec{\sigma})$, which in turn depends on the initial microscopic distribution $p_0(\vec{\sigma})$. In order to close these

equations, Coolen and Sherington [12] made two important assumptions: (i) the flow equations, and therefore the distributions (29) and (30), are self-averaging in the thermodynamic limit with respect to the microscopic realization of the stored patterns, and (ii) in calculating these distributions one can assume equipartitioning of probability within the macroscopic (m,r) subshells of the statistical ensemble. The first assumption has been supported by the numerical simulations [13]. As a result, the explicit time dependence and the dependence on microscopic initial conditions in the flow equations (32) and (33) are removed and the intrinsic noise distributions now become

$$D_{m,r;t}[z]\to D_{m,r}[z]\equiv\left\langle \frac{\sum_{\vec{\sigma}}\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))(1/N)\sum_i \delta(z-z_i(\vec{\sigma}))}{\sum_{\vec{\sigma}}\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))}\right\rangle_{\{\vec{\xi}\}}, \tag{34}$$

$$D'_{m,r;t}[z]\to D'_{m,r}[z]\equiv\left\langle \frac{\sum_{\vec{\sigma}}\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))(1/N)\sum_i \sigma_i\xi_i^1\delta(z-z_i(\vec{\sigma}))}{\sum_{\vec{\sigma}}\delta(m-m(\vec{\sigma}))\delta(r-r(\vec{\sigma}))}\right\rangle_{\{\vec{\xi}\}}. \tag{35}$$

The distributions $D_{m,r}[z]$ and $D'_{m,r}[z]$ are calculated by the replica method (see Ref. [13] for details). In the replica symmetric (RS) approximation, the results are

$$\begin{aligned}D^{\mathrm{RS}}_{m,r}[z]=&\frac{e^{-(\Delta+d+z)^2/2\alpha r}}{2\sqrt{2\pi\alpha r}}\\&\times\left\{1-\int Dy\tanh\left[\lambda y\sqrt{\frac{\Delta}{\alpha\rho r}}+(\Delta+d+z)\frac{\lambda^2}{\alpha\rho r}+\mu\right]\right\}+\frac{e^{-(\Delta+d-z)^2/2\alpha r}}{2\sqrt{2\pi\alpha r}}\\&\times\left\{1-\int Dy\tanh\left[\lambda y\sqrt{\frac{\Delta}{\alpha\rho r}}+(\Delta+d-z)\frac{\lambda^2}{\alpha\rho r}-\mu\right]\right\},\end{aligned} \tag{36}$$

$$\begin{aligned}D'^{\mathrm{RS}}_{m,r}[z]=&-\frac{e^{-(\Delta+d+z)^2/2\alpha r}}{2\sqrt{2\pi\alpha r}}\\&\times\left\{1-\int Dy\tanh\left[\lambda y\sqrt{\frac{\Delta}{\alpha\rho r}}+(\Delta+d+z)\frac{\lambda^2}{\alpha\rho r}+\mu\right]\right\}+\frac{e^{-(\Delta+d-z)^2/2\alpha r}}{2\sqrt{2\pi\alpha r}}\\&\times\left\{1-\int Dy\tanh\left[\lambda y\sqrt{\frac{\Delta}{\alpha\rho r}}+(\Delta+d-z)\frac{\lambda^2}{\alpha\rho r}-\mu\right]\right\},\end{aligned} \tag{37}$$

where Dy is the Gaussian measure, $Dy\equiv(dy/\sqrt{2\pi})e^{-y^2/2}$, $\Delta\equiv\alpha\rho r-\lambda^2/\rho$, and the parameters $\{q,\lambda,\rho,\mu\}$ are solutions of the following saddle-point equations:

$$r=\frac{1-\rho(1-q)^2}{[1-\rho(1-q)]^2}, \tag{38}$$

$$\lambda=\frac{\rho\sqrt{\alpha q}}{1-\rho(1-q)}, \tag{39}$$

$$m=\int Dy\tanh(\lambda y+\mu), \tag{40}$$

$$q=\int Dy\tanh^2(\lambda y+\mu). \tag{41}$$

Stability of the replica symmetric solution requires

$$\alpha-\rho^2(\alpha+\Delta)^2\int\frac{Dy}{\cosh^4(\lambda y+\mu)}\geqslant 0. \tag{42}$$

We numerically solve Eqs. (32) and (33) for various values of the memory loading level α, the self-coupling d, the temperature β^{-1}, and the initial conditions m_0 for the overlap m. We fix the initial condition for the order parameter r at $r_0=1$ for all the calculations. In Fig. 3, we plot the trajectory $m(t)$ for $m_0=0.22$, $\alpha=0.05$, and various values of the self-coupling d. It can be seen that retrieval is possible for $d=0$ with final overlap $m_f\approx 1$. The RS solution becomes unstable only after the retrieval has taken place, i.e., near $m\approx 1$ and $r\approx 1$. As d is increased to a moderate positive value, the retrieval becomes faster. If d is further increased, retrieval, as expected, becomes slower. Moreover, retrieval quality deteriorates and the RS solution becomes unstable

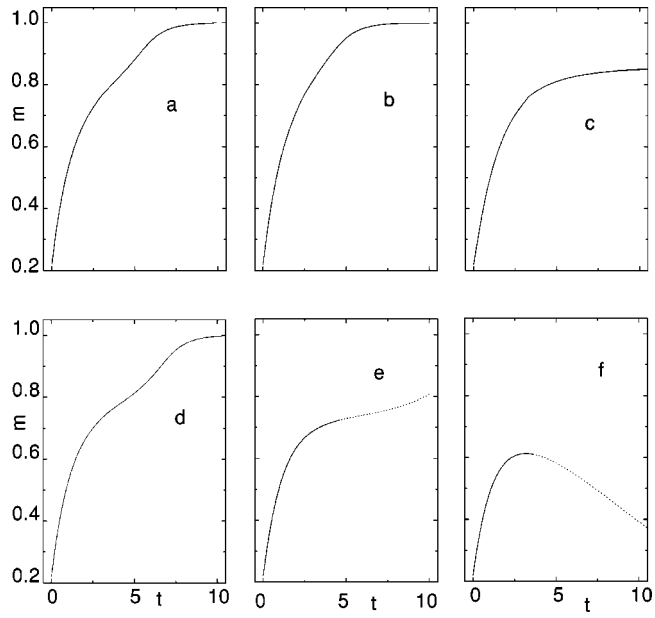

FIG. 3. Evolution of the overlap m of the network state $\vec{\sigma}$ with $\vec{\xi}^1$, the pattern under retrieval, for various values of the self-coupling d. $\alpha=0.05$, $T=0$, and $m_0=0.22$. Time is in units of iterations/spin. Full curve, solutions of the flow Eqs. (32) and (33) are stable against replica symmetry breaking (RSB); dotted curve, the solutions are unstable against RSB. (a) $d=0$. (b) $d=0.04$. (c) $d=0.25$. (d) $d=-0.02$. (e) $d=-0.04$. (f) $d=-0.08$.

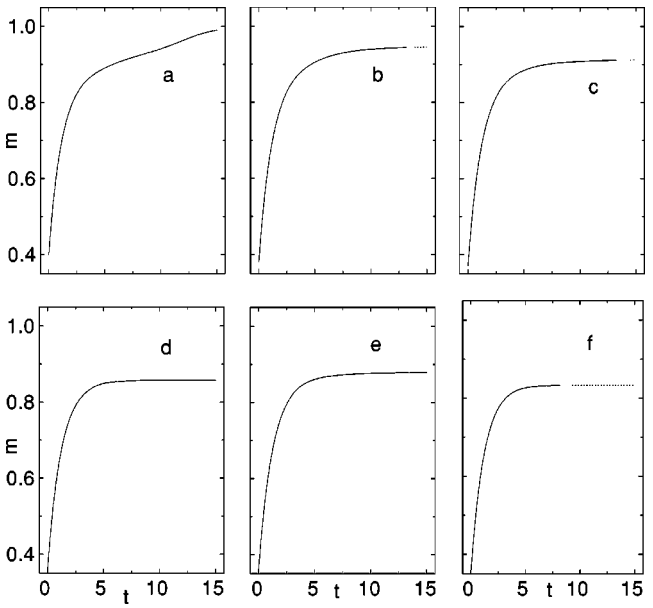

FIG. 4. Effect of self-coupling on the basin of attraction. $\alpha=0.1$ and $T=0$. Time is in units of iterations/spin. Full and dotted curves, same as Fig. 3. (a) $m_0=0.4$, $d=0.04$. (b) $m_0=0.38$, $d=0.12$. (c) $m_0=0.37$, $d=0.15$. (d) $m_0=0.36$, $d=0.15$. (e) $m_0=0.36$, $d=0.2$. (f) $m_0=0.35$, $d=0.2$.

before the retrieval. When d is negative, for $|d|\leqslant 0.02$ retrieval is possible with $m_f\approx 1$. The RS solution becomes unstable only after retrieval has been achieved. However, the retrieval becomes slower. As the magnitude of the negative self-coupling is increased to 0.04, the RS solution becomes unstable before the retrieval. Furthermore, the retrieval becomes very slow. For $d=-0.08$, the retrieval is not possible.

Next, we look at the effect of the self-coupling on the basin of attraction of a stored pattern. In the standard model ($d=0$), it is possible to retrieve the memory only when the initial overlap with the stored pattern $m_0\geqslant 0.22$ for $\alpha=0.05$. We find that it is possible to retrieve the memory with slightly lower values of the initial overlap $m_0\geqslant 0.21$ by having a positive self-coupling $d=0.04$. Similarly for $\alpha=0.1$ we observe improvement in the basin of attraction (Fig. 4). The minimum initial overlap $m_0=0.43$ is needed in order to achieve the retrieval at $d=0$. We find that with a positive self-coupling it can be accomplished with lower values of the m_0, e.g., 0.35 at $d=0.2$. The enhancement in the basin of attraction is much more prominent compared to that in $\alpha=0.05$. However, the retrieval quality becomes poor as we go for the lower values of the initial overlap and higher values of the self-coupling. We find reduction in the basin of attraction with the negative self-coupling.

Figure 5 shows the effect of self-coupling on the storage capacity of the network. We find enhancement of the storage capacity by having a positive self-coupling. It is possible to retrieve memory even for $\alpha=0.15$ with $d=0.15$, which is not possible otherwise.

The Hopfield model has a rich $(T-\alpha)$ phase diagram [1,14]. The spurious attractors affect the dynamics of the network differently in different regions of the phase diagram. We therefore solve the flow equations for finite temperatures to find out effect of the self-coupling in different regions of the phase diagram. First we look at the situation in which $\alpha=0.05$ and $T=0.5$. This point falls in a region that represents a mixed phase of the spin glass and the retrieval states. However, the retrieval states are not the global minima of the free energy. This results in the reduction of the basin of attraction of the stored patterns. The initial overlap m_0 should not be less than 0.44 in order to retrieve the memory. This should be compared with $m_0\geqslant 0.22$ in the zero-temperature case. Once again we find faster retrieval and enhancement in the basin of attraction of the stored memory

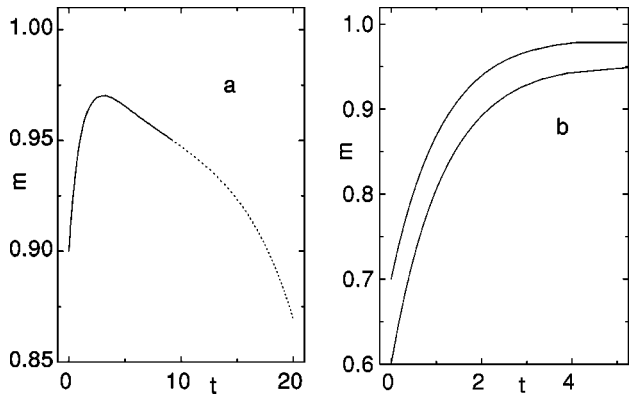

FIG. 5. Effect of self-coupling on the storage capacity of the network. $\alpha=0.15$ and $T=0$. Time is in units of iterations/spin. Full and dotted curves, same as Fig. 3. (a) $m_0=0.9$, $d=0$. (b) $m_0=0.6$ and 0.7, $d=0.15$.

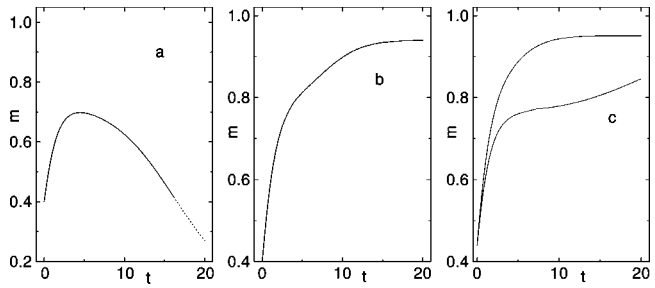

FIG. 6. Effect of self-coupling on the retrieval performance in the mixed phase region of the $(T-\alpha)$ phase diagram. $\alpha=0.05$ and $T=0.5$. Time is in units of iterations/spin. Full and dotted curves, same as Fig. 3. (a) $m_0=0.4$, $d=0$. (b) $m_0=0.4$, $d=0.08$. (c) $m_0=0.44$, $d=0$ (lower curve), and $d=0.12$ (upper curve).

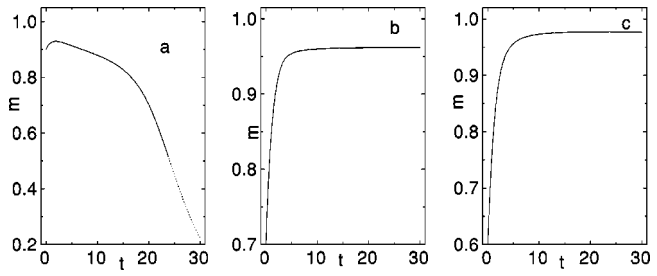

FIG. 7. Effect of self-coupling on the retrieval performance in the spin-glass region of the $(T-\alpha)$ phase diagram. $\alpha=0.09$ and $T=0.4$. Time is in units of iterations/spin. Full and dotted curves, same as Fig. 3. (a) $m_0=0.9$, $d=0$. (b) $m_0=0.7$, $d=0.15$. (c) $m_0=0.6$, $d=0.25$.

in the presence of a positive self-coupling (Fig. 6). Next we consider the case of $\alpha=0.09$ and $T=0.4$, which falls in the spin-glass phase (Fig. 7). In the standard model, it is not possible to retrieve the memory in such a situation. However, by having $d=0.15$ it is possible to retrieve the memory even in this case.

V. CONCLUSION

To summarize, we have analyzed the performance of a Hopfield model as an associative memory in the presence of a self-coupling term in the synaptic matrix by (i) counting the fixed points of the zero-temperature dynamics in the phase space and (ii) by applying a dynamical theory of Coolen and Sherrington, which has been used to study the dynamics of the Hopfield model on finite time scales. We have investigated the range of α from 0.01 to 0.15, which is the region of interest so far as the retrieval properties are concerned. We find that the spurious attractors are destabilized by a negative self-coupling, which depends very mildly on α. The magnitude of the self-coupling varies from $|d|=0.06$ in the case of $\alpha=0.05$ to $|d|=0.09$ for that in $\alpha=0.113$. Contrary to this, the magnitude of the self-coupling which causes suppression of the retrieval states has a rather strong dependence on α. A negative self-coupling of very small magnitude is enough to destabilize the retrieval states at higher values of α, compared to those at lower values of α. For example, the retrieval states for $\alpha=0.113$ become suppressed at $d=-0.003$, whereas those for $\alpha=0.05$ become suppressed at $d=-0.02$, a value that is an order of magnitude higher. In both cases, however, the suppression of retrieval states occurs much earlier than that of spurious states. As discussed above, the suppression of the retrieval states occurs because of the fraction of spins, which is not aligned to the stored pattern and hence has a very low value of local alignment fields.

There have been speculations [5] that the suppression of spurious fixed-point attractors by a negative self-coupling of appropriate magnitude will result in better performance of the network as an associative memory. The main reason behind such an expectation is the robustness of retrieval fixed points to the self-coupling. However, we did not find any direct correlation between the suppression of the large number of spurious metastable attractors and the improvement in the performance of the network as an associative memory. In fact, we find deterioration in the retrieval performance of the network in the presence of a negative self-coupling—the basin of attraction of stored memory decreases and the retrieval time increases. On the other hand, a positive self-coupling of appropriate magnitude has a positive effect on the retrieval performance of the network. We find some enhancement in the basin of attraction of the stored memory and the storage capacity of the network. The retrieval becomes faster. It also becomes possible to retrieve memory even in some regions of the spin-glass phase, which is not possible otherwise.

To what extent is the structure of fixed points relevant in understanding the behavior of the network in the presence of self-coupling? We can think of two possibilities. First, in the case of negative self-coupling, the positive effects of suppression of the spurious fixed-point attractors do not compensate for the negative effects that may arise due to suppression of the retrieval fixed-point attractors. Similarly, in the case of moderate positive self-coupling, growth of the retrieval band compensates well for the negative effects due to growth of the spurious band and leads to improvement in the retrieval performance of the network. Second, it may be possible that the effect of self-coupling is purely dynamic and the improvement or deterioration in the retrieval performance cannot be attributed to the structure of fixed points. Understanding this would be quite interesting. Numerical simulation results of Ref. [6] suggest that improvement in the retrieval performance of the network may also depend on the choice of the learning rule for the off-diagonal elements of the synaptic matrix. This issue will be addressed in our future publication.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor A. C. C. Coolen for helpful discussions and Dr. S. C. Mehendale for a critical reading of the manuscript.

[1] D. J. Amit, *Modeling Brain Functions* (Cambridge University Press, Cambridge, 1989).
[2] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[3] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).
[4] Y. Owechko, IEEE J. Quantum Electron. **25**, 619 (1989).
[5] G. Athithan and C. Dasgupta, IEEE Trans. Neural Netw. **8**, 1483 (1997).
[6] W. Krauth, M. Mezard, and J.-P. Nadal, Complex Syst. **2**, 387 (1988).
[7] J.F. Fontanari and R. Köberle, J. Phys. (France) **49**, 13 (1988).
[8] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).
[9] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[10] A. Treves and D.J. Amit, J. Phys. A **21**, 3155 (1988).
[11] E.G. Gardner, J. Phys. A **19**, L1047 (1986).
[12] A.C.C. Coolen and D. Sherrington, Phys. Rev. Lett. **71**, 3886 (1993).
[13] A.C.C. Coolen and D. Sherrington, Phys. Rev. E **49**, 1921 (1994).
[14] D.J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987).
[15] J.J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[16] M.P. Singh, C. Zhang, and C. Dasgupta, Phys. Rev. E **52**, 5261 (1995).