Learning noisy patterns in a Hopfield network

J. F. Fontanari* and R. Meir
Division of Chemistry, Mail Code 164-30 CH, California Institute of Technology, Pasadena, California 91125
(Received 24 March 1989; revised manuscript received 5 June 1989)

We study the ability of a Hopfield network with a Hebbian learning rule to extract meaningful information from a noisy environment. We find that the network is able to learn an infinite number of ancestor patterns, having been exposed only to a *finite* number of noisy versions of each. We have also found that there is a regime where the network recognizes the ancestor patterns very well, while performing very poorly on the noisy patterns to which it had been exposed during the learning stage.

Much of the recent work on spin-glass models for associative memory[1] has focused on the retrieval process[2] and on the process of learning itself.[3-5] In most of these studies however, it had been assumed that during the learning stage, the system is exposed to the very same patterns which it is later expected to recall correctly. An interesting question which arises is whether it is possible for a network, exposed to a noisy environment, to extract the meaningful information from that environment, using simple local learning rules.[6,7] In this paper we address this question, in the context of the Hopfield model, and obtain exact (replica symmetric) expressions for various quantities characterizing the performance of the network.

The model we study is the following. Consider a fully connected system of N spins, in which we wish to embed p *ancestor* patterns $\{\xi_i^\mu\}$, $\mu=1,\ldots,p$, $i=1,\ldots,N$. Assume that the system is exposed to noisy versions of these p patterns. We denote the noisy versions of the p ancestor patterns by $\{\xi_i^{\mu\nu}\}$, with $\nu=1,\ldots,s$, i.e., s noisy versions of each ancestor pattern are presented to the network. The connection J_{ij} between spins i and j is given by the Hebbian rule

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\sum_{\nu=1}^{s}\xi_i^{\mu\nu}\xi_j^{\mu\nu}\quad (i\neq j)\ . \tag{1}$$

We choose the probability distribution of the $\{\xi_i^{\mu\nu}\}$ to be

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

with $b\in[0,1]$ and $\xi_i^\mu=\pm 1$ with equal probability.

This probability distribution was used in a different context by Gutfreund[8] in his study of the storage of hierarchically correlated patterns. Our aim is to calculate the response of the system to the ancestor patterns $\{\xi_i^\mu\}$, given that the learning took place with the noisy patterns $\{\xi_i^{\mu\nu}\}$, and the Hebbian rule (1). The Hopfield model is obtained in the case $b=1$, i.e., the learned patterns are just the patterns to be stored. One also expects that for $s\gg 1$, i.e., learning many noisy versions of the same pattern, the noise should average out and we get good recall of the ancestor patterns. In this paper we study the case where $p=\alpha N$ and s is finite in the thermodynamic limit. We will see that even if the number of noisy versions is small, the network performs very well and retrieves the correct ancestor patterns $\{\xi_i^\mu\}$.

Before discussing the full solution of the model within the replica framework it is helpful to describe an approximation, first proposed for the Hopfield model by Kinzel,[9] and proven to be *exact* for the case of the strongly diluted model discussed by Derrida *et al.*[10] This approximation will be referred to below as the *diluted approximation.*

We assume that the network is in state $\mathbf{S}=\{S_i\}$, which has a nonzero correlation with the patterns $\xi_i^{1\nu}$ and ξ_i^1, but has negligible correlation $[O(1/\sqrt{N})]$ with all patterns with $\mu>1$. The field acting on S_i is

$$\begin{aligned}h_i&=\sum_{j\,(\neq i)}J_{ij}S_j\\&=\sum_\nu\xi_i^{1\nu}\frac{1}{N}\sum_{j\,(\neq i)}\xi_j^{1\nu}S_j+\sum_{\mu\,(>1)}\sum_\nu\xi_i^{\mu\nu}\frac{1}{N}\sum_{j\,(\neq i)}\xi_j^{\mu\nu}S_j\end{aligned} \tag{3}$$

and the state of the spin at the next time step S_i' is determined probabilistically according to the field h_i acting on it,

$$S_i'=\begin{cases}+1 \text{ with probability } [1+\exp(-2\beta h_i)]^{-1}\\ -1 \text{ with probability } [1+\exp(+2\beta h_i)]^{-1}\end{cases} \tag{4}$$

where $\beta^{-1}=T$ is a parameter measuring the noise level of the system (temperature). There are two quantities of interest. The first is the overlap of the state $\mathbf{S}$ with the pattern ξ_i^1 and the second is the overlap with the noisy patterns $\xi_i^{1\nu}$. Thus we define

$$M_1=\frac{1}{N}\sum_j\xi_j^1S_j,\quad M_{1\nu}=\frac{1}{N}\sum_j\xi_j^{1\nu}S_j\ . \tag{5}$$

We also define the single-site averaged quantities

$$m_1=\langle\!\langle\,\xi_j^1\langle S_j\rangle\,\rangle\!\rangle,\quad m_{1\nu}=\langle\!\langle\,\xi_j^{1\nu}\langle S_i\rangle\,\rangle\!\rangle\ , \tag{6}$$

where $\langle\ \rangle$ denotes a thermal average and $\langle\!\langle\ \rangle\!\rangle$ stands for an average over the distribution given in Eq. (2). Multiplying Eq. (3) by $\xi_i^{1\nu}$, and performing the thermal average, we obtain

$$m'_{1\nu}=\left\langle\!\left\langle \tanh\left[\beta\left[\xi_i^{1\nu}\sum_{\lambda}\xi_i^{1\lambda}M_{1\lambda}+\xi_i^{1\nu}\sum_{\mu\,(>1)}\sum_{\lambda}\xi_i^{\mu\lambda}\frac{1}{N}\sum_{j\,(j\neq i)}\xi_j^{\mu\lambda}S_j\right]\right]\right\rangle\!\right\rangle . \tag{7}$$

Without loss of generality, we set $\nu=1$ in the above equation. We also note that since we expect $M_{1\lambda}$ to be $O(1)$ we may replace $M_{i\lambda}$ for $\lambda>1$ by b^2M_{11}, since the fluctuations in this step are $O(1/\sqrt{N})$. In what follows we will be interested in the fixed point solutions, thus we take $m'_{1\nu}=m_{1\nu}$. In the diluted approximation, we take the second argument of the hyperbolic tangent to be a Gaussian random variable, neglecting its site dependence and the correlations between $\langle S_i\rangle$ and $\{\xi_i^{\mu\nu}\rangle$.

With these remarks in mind, we write the equation for m_{11} in the form

$$m_{11}=\langle\tanh\{\beta[(1+b^2z_s)m_{11}+\sqrt{\alpha r}z]\}\rangle , \tag{8}$$

where we have replaced M_{11} by m_{11}. The average in Eq. (8) is over the Gaussian random variable z, with zero mean and unit variance, and over the random variable z_s, with the probability distribution

$$P(z_s=k)=\left[\frac{s-1}{\frac{s-1-k}{2}}\right]\left[\frac{1+b^2}{2}\right]^{(s-1-k)/2}\times\left[\frac{1-b^2}{2}\right]^{(s-1+k)/2} . \tag{9}$$

The variable αr is given by

$$\alpha r=\left\langle\!\left\langle\left[\frac{1}{N}\sum_{\mu\,(>1)}\sum_{\nu}\xi_i^{\mu\nu}\sum_{j(\neq i)}\xi_j^{\mu\nu}S_j\right]^2\right\rangle\!\right\rangle . \tag{10}$$

Performing this average yields

$$\alpha r=\alpha s[1+(s-1)b^4] . \tag{11}$$

To obtain the equation for m_1 we proceed along similar lines. The only point to note is that, using arguments similar to those discussed above, we replace $M_{1\nu}$ appearing in Eq. (7) above by bm_1. We note that this procedure *cannot* be done with respect to the variables $M_{\mu\nu}$ and $M_{\mu}(\mu>1)$, since these variables themselves are $O(1/\sqrt{N})$. Thus we obtain

$$m_1=\langle\tanh[\beta(bx_sm_1+\sqrt{\alpha r}z)]\rangle . \tag{12}$$

Here αr is given by Eq. (11), and the probability distribution of x_s is

$$P(x_s=k)=\left[\frac{s}{\frac{s-k}{2}}\right]\left[\frac{1+b}{2}\right]^{(s-k)/2}\left[\frac{1-b}{2}\right]^{(s+k)/2} . \tag{13}$$

Taking the limit $\beta\to\infty$ in Eqs. (8) and (12), and expanding in powers of m_1 and m_{11} we find two critical values of α. For

$$\alpha>\alpha_c^{11}=(2/\pi s)[1+(s-1)b^4] \tag{14}$$

$m_{11}=0$, i.e., the network cannot retrieve the noisy versions of the ancestor patterns. For

$$\alpha>\alpha_c^{1}=(2/\pi)sb^4/[1+(s-1)b^4] \tag{15}$$

$m_1=0$, i.e., the network cannot retrieve the ancestor patterns.

At this stage we can already see that the network exhibits some interesting behavior. From Eqs. (14) and (15) we see that there is a region in the (b,s) space where $\alpha_c^1>\alpha_c^{11}$. In this region, we have $m_{11}=0$ and $m_1\neq0$. Thus, although the system is unable to recognize the patterns to which it has been exposed, it can still very well recognize the ancestor patterns. Similar behavior has been found by Virasorto[11] in a different context.

Having discussed the diluted approximation, we now turn to the fully connected model. To do that we introduce an energy function

$$E=-\tfrac{1}{2}\sum_i\sum_{j\,(\neq i)}J_{ij}S_iS_j \tag{16}$$

with J_{ij} given by Eq. (1).

Our goal is to characterize the states that minimize the quenched free-energy density

$$-\beta f=\lim_{N\to\infty}\lim_{n\to0}(\langle\!\langle Z^n\rangle\!\rangle-1)/nN , \tag{17}$$

where

$$Z=\mathrm{Tr}_{\{S_i\}}\exp(-\beta E) . \tag{18}$$

Using standard techniques[2] and assuming replica symmetry we obtain the following equations for the order parameter m_{11}:

$$m_{11}=\langle\tanh\{\beta[(1+b^2z_s)m_{11}+\sqrt{\alpha r_{11}}z]\}\rangle , \tag{19}$$

$$q_{11}=\langle\tanh^2\{\beta[(1+b^2z_s)m_{11}+\sqrt{\alpha r_{11}}z]\}\rangle , \tag{20}$$

$$r_{11}=sq_{11}\frac{\{1-C_{11}(1-b^2)[1+b^2(s-1)]\}^2+b^4(s-1)}{\{[1-(1-b^2)C_{11}]^2-sb^2C_{11}[1-(1-b^2)C_{11}]\}^2} , \tag{21}$$

where $C_{11}=\beta(1-q_{11})$ and z_s is distributed according to Eq. (9). In writing Eqs. (19)–(21) we have picked the solution with $m_{1\nu}=b^2m_{11}$, $\nu\neq1$, as in the diluted approximation. We have also assumed that $m_{\mu\nu}$ is nonzero only for $\mu=1$. We note that the expression for m_{11} is identical to the one we obtained in the diluted approximation, Eq. (8). However, the expression for r, the noise term, is much more complicated, since now we have taken into account the full correlations between $\{S_i\}$ and $\{\xi_i^{\mu\nu}\}$ for $\mu>1$.

In order to complete the calculation, we must compute the equation for the variable m_1. This equation cannot be obtained within the replica framework where the natu-

ral order parameter is m_{1v}. However, using the cavity method[12] in its simpler form proposed by Domany *et al.*,[13] we can show that m_1 satisfies the equations

$$m_1 = \langle \tanh[\beta(bx_s m_1 + \sqrt{\alpha r_1} z)] \rangle \ , \tag{22}$$

$$q_1 = \langle \tanh^2[\beta(bx_s m_1 + \sqrt{\alpha r_1} z)] \rangle \ , \tag{23}$$

where r_1 is given by Eq. (21) (with q_{11} replaced by q_1) and x_s has the probability distribution given in Eq. (13). In the remainder of the paper we discuss the solutions of the zero temperature limit of Eqs. (19)–(21) and Eqs. (22) and (23).

Let us focus first on the solutions of the equations for m_1. One finds three types of solutions. The first solution, $m=0$, corresponds to the spin-glass phase[2] and exists for all α. For $\alpha < \alpha_r(b)$, however, there appears an additional solution, the *retrieval* solution, with $m_1 \sim 1$. This solution appears through a first-order transition and corresponds to a state which is condensed into one of the ancestor patterns. In addition to these solutions, we find an additional *spurious* solution with $m_1 > 0$. This solution exists for $\alpha < \alpha_s(b)$, and emerges *continuously* from the spin-glass phase.

In Fig. 1 we plot α_r^1/α_0 versus b for several values of s. Here α_0 is the value obtained by Amit *et al.*[2] for the Hopfield model ($\alpha_0 \simeq 0.138$). As expected, for $s \to \infty$ the curve approaches the value $\alpha_r^1/\alpha_0 = 1$ for any *nonzero* b. The analytic expression for α_s^1, obtained by expanding

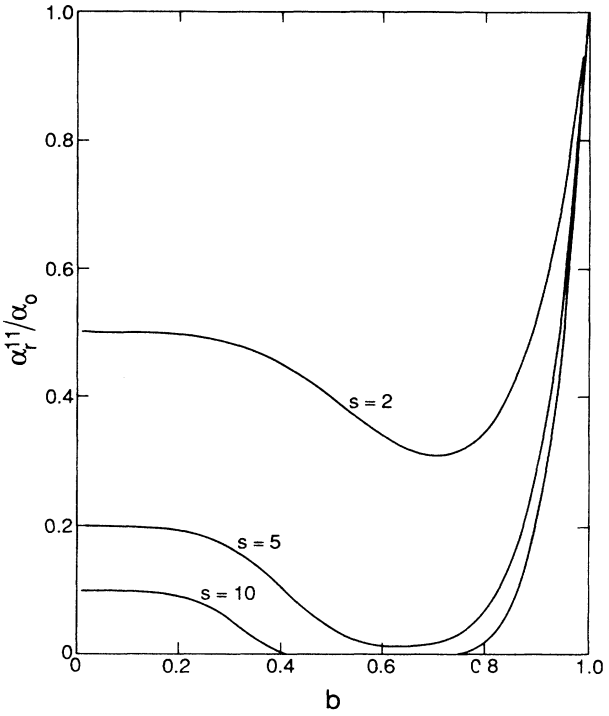

FIG. 2. Normalized critical α, α_r^{11} vs b, for the retrieval solution for the variable m_{11}, shown for $s=2$, 5, and 10. A retrieval solution with $m_{11} \sim 1$ exists only for $\alpha < \alpha_r^{11}$ and disappears discontinuously at α_r^1.

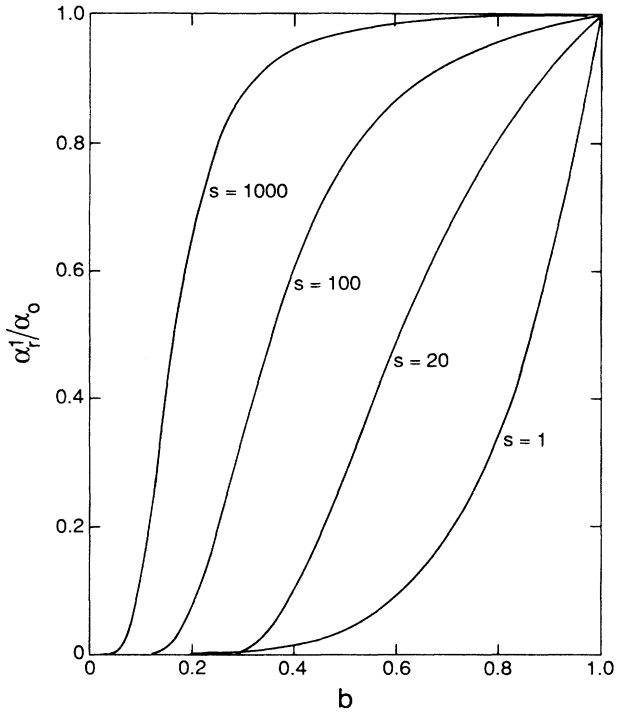

FIG. 1. Normalized critical α, α_r^1 vs b, for the retrieval solution for $s=1$, 20, 100, and 1000. A retrieval solution with $m_1 \sim 1$ exists only for $\alpha < \alpha_r^1$ and disappears discontinuously at α_r^1.

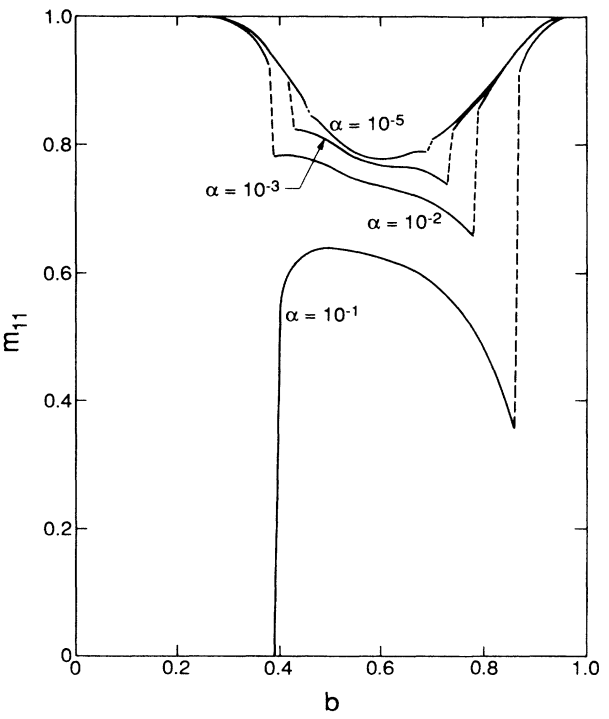

FIG. 3. Retrieval and spurious solution for m_{11}, plotted vs b for several values of α and $s=10$. The two solutions coincide at $\alpha=0$.

Eq. (22) for small m_1, is given by

$$\alpha_s^1=\frac{2}{\pi s}\frac{[(s+1)b^2-1]^2(1-b^2)^2}{[b^4(s-1)-1+2b^2]^2+b^8s^2(s-1)} . \quad (24)$$

Since $\alpha_s^1(b=0)\propto 1/s$ for $s\gg 1$, the region where the spurious solution exists diminishes with increasing s, as opposed to the retrieval solution.

The equations for m_{11} possess the same three types of solutions, as those for m_1 discussed above. However, the dependence of these solutions on b is different. In Fig. 2 we plot α_r^{11}, the critical value for the retrieval phase (we again normalize α with respect to α_0), versus b. It is easy to see that in the limit $b=0$, one has the result $\alpha_r^{11}/\alpha_0=1/s$, while for $b=1$ $\alpha_r^{11}/\alpha_0=1$. The analytic expression for α_s^{11} is given by

$$\alpha_s^{11}=\frac{2}{\pi}\frac{b^4}{s}\frac{[1+(s-1)b^2]^2(s-1)^2(1-b^2)^2}{[2b^2(s-1)-s+2]^2+(s-1)[1+(s-1)b^4]^2} . \quad (25)$$

To better understand the interplay between the retrieval and spurious solutions, we plot in Fig. 3 m_{11} versus b for several values of α near zero. For $\alpha=0$ these two solutions coincide. As α increases, however, there is an intermediate range in b where the retrieval solution does not exist, and m_{11} jumps discontinuously from the retrieval solution to the spurious one. As can be seen in the figure, the size of the jump increases with α. For $\alpha\geq 0.1$, only the spin-glass solution $m_{11}=0$ exists for small b. This solution undergoes a continuous transition to the spurious solution which in turn undergoes a discontinuous transition to the retrieval solution ($b\sim 1$). The same kind of behavior occurs in the solutions of the equations for m_1: the retrieval and spurious solutions coincide at $\alpha=0$ but split into two district solutions for any nonzero α. For any practical purposes the spurious solutions are not useful for the retrieval of either the ancestor patterns or the noisy patterns, since their retrieval quality cannot be distinguished from the remanent overlaps due to the nonequilibrium states.[14] We can easily see that for $s=1$ one should expect a solution with the same properties as the spurious states. In this limit we recover the Hopfield model with stored patterns $\{\xi_i^\mu\}$ and therefore there exist equilibrium states with a high overlap with these patterns. Since the ancestor patterns $\{\xi_i^\mu\}$ have an overlap b with the learned patterns, we expect a solution with $m_1\propto b$. In fact, for $b\ll 1$ we find for the spurious solution $m_1\approx b\ \mathrm{erf}(2/\pi\alpha)$. A similar argument in the limit of large s shows that m_{11} must also posses a similar kind of solution. Note that for $s\to\infty$, $\alpha_s^1\to 0$ and therefore the spurious solutions for m_1 disappear in this limit. From Eqs. (24) and (25) one can see that $\alpha_\infty^{11}=\alpha_1^1=(2/\pi)(1-b^2)^2$, in agreement with the intuitive idea that storing an infinite number of noisy versions of the ancestor states is equivalent to storing the ancestors themselves.

In summary, we have demonstrated analytically that a network using local Hebbian learning rules is able to learn an infinite number of patterns, having been exposed only to a *finite* number of noisy versions of each. We have also found that there is a regime where the network recognizes the ancestor patterns very well, while performing very poorly on the noisy patterns to which it had been exposed during the learning stage. However, if the number of noisy versions is small, and the correlation between them and their ancestor pattern is small, the network is unable to recognize the ancestor pattern although it may recognize the noisy patterns very well. We should emphasize that our results refer to the equilibrium states of the system, and do not tell us anything about the accessibility of these states by the dynamics.

The research at the California Institute of Technology was supported by Contract No. N00014-87-K-0377 from the U.S. Office of Naval Research. J. F. F. is partly supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico and R.M. is supported by the Weizmann Foundation.

*On leave of absence from Instituto de Fisica e Quimica de São Carlos, Universidade de São Paulo, 13560 São Carlos, São Paulo, Brazil.

[1]J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[2]D. J. Amit, H. Gutfreund, and H. Sompolinsky Ann. Phys. (N.Y.) **173**, 30 (1987).
[3]S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).
[4]W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).
[5]E. Gardner J. Phys. A **21**, 257 (1988).
[6]E. Gardner, N. Stroud, and D. J. Wallace (unpublished).
[7]B. M. Forrest J. Phys. A **21**, 245 (1988).
[8]H. Gutfreund, Phys. Rev. A **37**, 570 (1988).
[9]W. Kinzel, Z. Phys. B **60**, 205 (1985).
[10]B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[11]M. A. Virasoro, Europhys. Lett. **7**, 293 (1988).
[12]M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[13]E. Domany, R. Meir, and W. Kinzel, J. Phys. A (to be published).
[14]E. Gardner, J. Phys. A **19**, L1047 (1986).