Hierarchical neural network with high storage capacity

C. J. Perez-Vicente
Departament de Física, Divisió de Física Estadística, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

(Received 12 June 1989)

A recent method used to optimize biased neural networks with low levels of activity is applied to a hierarchical model. As a consequence, the performance of the system is strongly enhanced. The steps to achieve optimization are analyzed in detail.

I. INTRODUCTION

The wide knowledge of the Hopfield model has allowed important advances in the study of neural networks.[1-3] It has been the starting point for ambitious ideas aimed to build systems able to overcome some of its limitations. Models dealing with correlated patterns[4-8] or able to encode and retrieve configurations following particular structures[9-11] have been, among others, different approaches made lately.

The simplest way to introduce correlations in models with Hebbian learning rules is to assume that the patterns are biased, i.e., with levels of activity different than 50%, a property accepted from a neurophysiological point of view. Initial studies of these systems made by Amit *et al.*[12] showed not attractive results because they were characterized by a decrease of the storage capacity as the bias increases. The performance of the network was notably improved after the work of Buchmann *et al.*,[13] where the V model ($V=0,1$) is used rather than the S model ($S=-1,1$) because in their system the behavior of the network using one prescription or the other is completely different. Tsodyks and Feigelman[14] argued that such a difference comes from a possible lack of equivalence between both pictures.

Recently Perez and Amit,[15] working on the studies previously mentioned, have introduced a method to optimize biased neural networks. The point is to make certain assumptions about the local thresholds of the neurons.

Let us consider the local field acting on neuron i in terms of the V model

$$h_i=\sum_j J_{ij}V_j \ . \tag{1}$$

Since $V_i=(S_i+1)/2$ we have

$$h_i=\tfrac{1}{2}\sum_j J_{ij}S_j+\tfrac{1}{2}\sum_j J_{ij} \ . \tag{2}$$

For unbiased neural networks we know that if we add a local threshold $U_i=\frac{1}{2}\sum_j J_{ij}$ then $h_i=\frac{1}{2}\sum_j J_{ij}S_j$ and the storage capacity of the system doubles. For biased nets the best performance is achieved not by adding the previous value of U_i but by adding $U_i=(1+c)/2\sum_j J_{ij}$, where c ($c\subset[-1,1]$) is a parameter to be determined. Therefore

$$h_i=\tfrac{1}{2}\sum_j J_{ij}(S_j-c) \ . \tag{3}$$

Now, using the change $S_i=(2V_i-1)$ and introducing the variable $c'=(1+c)/2$, ($c'\subset[0,1]$), we have

$$h_i=\sum_j J_{ij}(V_j-c') \ . \tag{4}$$

In those terms there is a complete equivalence between the V and the S models solving the problem pointed out by Tsodyks and Feigelman.[14] For optimum c the performance of the network is enhanced over previous studies.

The main goal of this paper is to show that the method can be generalized to other models with local learning rules. As an example the procedure is applied to a neural network structured hierarchically studied by Gutfreund (Ref. 9, hereafter referred to as HG). The steps followed to achieve the optimization deserve special attention.

II. THE MODEL

The HG model stores a set of p patterns in k levels ($p=\prod_k p_k$) organized hierarchically. In the first level p_1 patterns are encoded. Every neuron ξ_i^μ ($\mu=1,\dots,p_1$, $i=1,\dots,N$) corresponding to such patterns takes values ±1 with a probability distribution given by

$$P(\xi_i^\mu)=\frac{1+a}{2}\delta(\xi_i^\mu-1)+\frac{1-a}{2}\delta(\xi_i^\mu+1) \ , \tag{5}$$

where the parameter a (the bias) measures the level of activity of the network.

The second level of the network is correlated with the previous one by means of a parameter b, being $0\le b\le 1$. Now, from each pattern $\{\xi_i^\mu\}$, p_2 patterns are encoded and its components $\xi_i^{\mu\nu}$ ($\nu=1,\dots,p_2$) follow the probability distribution given by

$$P(\xi_i^{\mu\nu})=\frac{1+\xi_i^\mu b}{2}\delta(\xi_i^{\mu\nu}-1)+\frac{1-\xi_i^\mu b}{2}\delta(\xi_i^{\mu\nu}+1) \ . \tag{6}$$

Generalization to k levels is straightforward. We consider now that we have only two levels. The synaptic matrix proposed in HG is

$$J_{ij}=\frac{1}{N}\sum_{\mu,\nu}(\xi_i^{\mu\nu}-\xi_i^\mu b)(\xi_j^{\mu\nu}-\xi_j^\mu b) \ , \qquad \mu=1,\dots,p_1, \quad \nu=1,\dots,p_2 \ . \tag{7}$$

With these ingredients we have enough information to apply the method indicated in the Introduction. The first step is the study of the signal-to-noise ratio from the local field [post-synaptic potential (PSP)] acting on neurons. Assuming that the network is in the state $\xi_i^{\eta\gamma}$ the local field acting on neuron i is

$$h_i=\frac{1}{N}\sum_j\sum_{\mu,\nu}(\xi_i^{\mu\nu}-\xi_i^{\mu}b)(\xi_j^{\mu\nu}-\xi_j^{\mu}b)(\xi_j^{\eta\gamma}-c)-U\ , \tag{8}$$

where U is a threshold. This last expression can be split in a signal plus a noise

$$\begin{aligned}h_i=&\frac{1}{N}\sum_j(\xi_i^{\eta\gamma}-\xi_i^{\eta}b)(\xi_j^{\eta\gamma}-\xi_j^{\eta}b)(\xi_j^{\eta\gamma}-c)-U\\&+\frac{1}{N}\sum_j\sum_{\substack{\mu,\nu\\ \mu,\nu\neq\eta,\gamma}}(\xi_i^{\mu\nu}-\xi_i^{\mu}b)(\xi_j^{\mu\nu}-\xi_j^{\mu}b)(\xi_j^{\eta\gamma}-c)\ .\end{aligned} \tag{9}$$

The effect of both new terms is the following. On one side the threshold U will be an external field whose effect is the optimization of the signal acting on neurons. On the other side the term $c\sum_j J_{ij}$ is a random field which because of its correlation with the patterns will reduce the destabilizing noise generated in the retrieval process. The constant c is chosen such that the noise becomes a minimum. The destabilizing noise is Gaussian with zero mean. Its square is

$$\langle R^2\rangle\simeq\alpha(1-b^2)^2(1+c^2-2abc)\ , \tag{10}$$

where $\alpha=p/N$ is the storage ratio. Therefore, the constant c is ab, and the noise is

$$\langle R^2\rangle\simeq\alpha(1-b^2)^2[1-(ab)^2]\ . \tag{11}$$

The signal is

$$S=\frac{N-1}{N}(\xi_i^{\eta\gamma}-\xi_i^{\eta}b)(1-b^2)-U\ . \tag{12}$$

U will be optimum when

$$U=-\xi_i^{\eta}b(1-b^2)\ . \tag{13}$$

Now, the signal acting on all the neurons in a given cluster is equalized to $|S|=(1-b^2)$. From this value of U one deduces that to start the dynamics in a certain level k the system needs to know its ancestors, i.e., the set $\{\xi_i^{\mu\nu,\ldots,k-1}\}$ has to be retrieved before and then the external field is applied. This type of dynamics is just the same obtained in HG. Therefore, the method to improve the performance of the network does not change the basic features of the system. Finally the signal-to-noise ratio is

$$\left|\frac{S}{R}\right|\simeq\frac{1}{\sqrt{\alpha[1-(ab)^2]}} \tag{14}$$

showing a divergence for $ab\rightarrow1$. Comparing this result with HG

$$\left|\frac{\left[\dfrac{S}{R}\right]_{\rm opt}}{\left[\dfrac{S}{R}\right]_{\rm HG}}\right|=\frac{1/\sqrt{\alpha[1-(ab)^2]}}{(1-|b|)/\sqrt{\alpha}}$$

$$=\frac{1}{(1-b)\sqrt{1-(ab)^2}}\ . \tag{15}$$

This analysis indicates that the optimized model always leads to a better performance of the network than the original model. Both terms (U and $c\sum_j J_{ij}$) are necessary to optimize the system. The lack of one of them still keeps a poor behavior characterized by a low storage capacity and an enormous amount of spurious states pervading the dynamics of the network.

III. MEAN-FIELD EQUATIONS

The mean-field equations give the collective behavior of the system in the $N\rightarrow\infty$ limit. They can be calculated from the study of the free energy per spin

$$f=-1/\beta N\langle\ln\mathrm{Tr}\exp(-\beta H)\rangle\ ,$$

where H is the Hamiltonian

$$H=-\tfrac{1}{2}\sum_{\substack{i,j\\ i\neq j}}J_{ij}(S_i-ab)(S_j-ab)+\sum_i U_iS_i\ . \tag{16}$$

The replica method[16] is used to average over the quenched variables $\{\xi^{\mu\nu}\}$. Following a standard calculation[10,3] and assuming that the solutions have only one macroscopic overlap with a single pattern, I have found in the replica symmetric framework that the free energy in the thermodynamic limit is

$$\begin{aligned}f=&\frac{\alpha}{2}(1-b^2)[1-(ab)^2]-\alpha abx(1-b^2)+a^2bh+\tfrac{1}{2}\sum_{\mu,\nu}(m^{\mu\nu})^2+\frac{ab\alpha\beta}{2}(2abr+y)\\&+\frac{\alpha}{2\beta}\left[\ln(1-\beta_1+\beta_2C)-\frac{\beta_2q}{1-\beta_1+\beta_2C}\right]+\frac{\alpha\beta}{2}(xy+r\{[1-(ab)^2]-[1+(ab)^2]q\})\\&-\frac{1}{\beta}\left\langle\!\!\left\langle\ln2\cosh\left[\beta[\sqrt{\alpha r}z+\frac{\alpha\beta}{2}(y+2abr)+m^{\mu\nu}(\xi_i^{\mu\nu}-\xi_i^{\mu}b)-h\xi^{\mu}]\right]\right\rangle\!\!\right\rangle\ ,\end{aligned} \tag{17}$$

where

$$\beta_1=\beta(1-b^2)[1-(ab)^2]\ ,$$

$$\beta_2=\beta(1-b^2)[1+(ab)^2]\ ,$$

and

$$C=\frac{2ab}{1+(ab)^2}x+q\ .$$

The double angular brackets mean the average over the patterns $\{\xi^{\mu\nu}\}$ and over a Gaussian variable z. Five order parameters are found: $m^{\mu\nu}$, x, q, y, and r. Their physical meaning comes from the saddle-point equations

$$m^{\mu\nu}=\langle\langle(\xi_i^{\mu\nu}-\xi_i^{\mu}b)\tanh\beta\phi\rangle\rangle\ , \tag{18}$$

where

$$\phi=\sqrt{\alpha r}\,z+\frac{\alpha\beta}{2}(y+2abr)+(\xi_i^{\mu\nu}-\xi_i^{\mu}b)m^{\mu\nu}-h\xi^{\mu}\ .$$

$m^{\mu\nu}$ measures indirectly the overlap between the state $\{S_i\}$ of the network and a stored pattern $\xi^{\mu\nu}$

$$m^{\mu\nu}=\frac{1}{N}\sum_i(\xi_i^{\mu\nu}-\xi_i^{\mu}b)\langle(S_i-ab)\rangle \tag{19}$$

where $\langle\ \rangle$ means thermal average. x gives the mean activity of the network

$$x=\langle\langle\tanh\beta\phi\rangle\rangle-ab\ , \tag{20}$$

q defined as

$$q=\frac{1}{N}\sum_i\langle(S_i-ab)(S_i-ab)\rangle \tag{21}$$

is a generalized Edward-Anderson parameter

$$q=\frac{1}{1+(ab)^2}[\langle\langle\tanh^2\beta\phi\rangle\rangle-2abx-(ab)^2]\ . \tag{22}$$

y and r are the mean-square fluctuations of the magnetization and of the overlaps between the thermodynamic state of the network and the patterns which are not condensed, respectively,

$$y=\frac{2ab}{\beta}(1-b^2)-\frac{(1-b^2)2ab}{1-\beta_1-\beta_2C}\left[1+\frac{\beta_2q}{1-\beta_1-\beta_2C}\right]\ , \tag{23}$$

$$r=\frac{q(1-b^2)^2[1+(ab)^2]}{(1-\beta_1+\beta_2C)^2}\ . \tag{24}$$

In the limit $T\rightarrow0$ these expressions reduce to

$$m=\tfrac{1}{4}(1-b^2)(1+a)(\mathrm{erf}\phi_1+\mathrm{erf}\phi_2)+\tfrac{1}{4}(1-b^2)(1-a)(\mathrm{erf}\phi_3+\mathrm{erf}\phi_4)\ , \tag{25}$$

$$x=\frac{1+a}{4}[(1+b)\mathrm{erf}\phi_1-(1-b)\mathrm{erf}\phi_2]+\frac{1-a}{4}[(1-b)\mathrm{erf}\phi_3-(1+b)\mathrm{erf}\phi_4]-ab\ , \tag{26}$$

$$C=\left[\frac{2}{\pi\alpha r}\right]^{1/2}\left[\frac{1+a}{4}[(1+b)\exp(-\phi_1^2)+(1-b)\exp(-\phi_2^2)]+\frac{1-a}{4}[(1-b)\exp(-\phi_3^2)+(1+b)\exp(-\phi_4^2)]\right]\ , \tag{27}$$

$$r=\frac{(1-b^2)[1-2abx-(ab)^2]}{[1-(1-b^2)C]^2}\ , \tag{28}$$

where

$$\phi_1=\frac{m(1-b)-h}{\sqrt{2\alpha r}}-\frac{ab\alpha(1-b^2)^2C}{\sqrt{2\alpha r}\,[1-(1-b^2)C]}\ ,$$

$$\phi_2=\frac{m(1+b)+h}{\sqrt{2\alpha r}}+\frac{ab\alpha(1-b^2)^2C}{\sqrt{2\alpha r}\,[1-(1-b^2)C]}\ ,$$

$$\phi_3=\frac{m(1+b)-h}{\sqrt{2\alpha r}}-\frac{ab\alpha(1-b^2)^2C}{\sqrt{2\alpha r}\,[1-(1-b^2)C]}\ ,$$

$$\phi_4=\frac{m(1-b)+h}{\sqrt{2\alpha r}}+\frac{ab\alpha(1-b^2)^2C}{\sqrt{2\alpha r}\,[1-(1-b^2)C]}\ ,$$

and where erf means the error function. In the limit $a=1$ the model simplifies and one recovers the set of equations obtained by Perez and Amit.[15]

IV. RESULTS

The numerical solution of the set of transcendental equations (25)–(28) gives the macroscopic behavior of the network at $T=0$. The interest is centered in the retrieval phase, solutions with $\alpha\neq0$ and $m\neq0$, because they are the only ones which have associative memory. The results can be observed in Fig. 1. The curves represent the variation of the storage capacity versus different values of

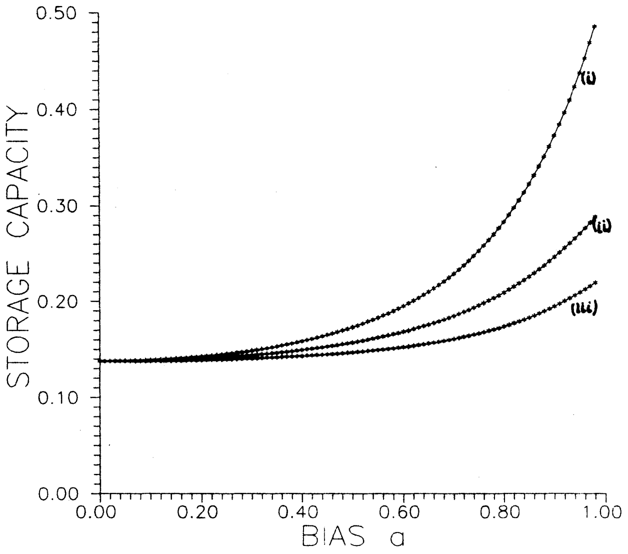

FIG. 1. Variation of storage capacity with respect to the parameter b for (i) $a=0.8$, (ii) $a=0.6$, and (iii) $a=0.4$.

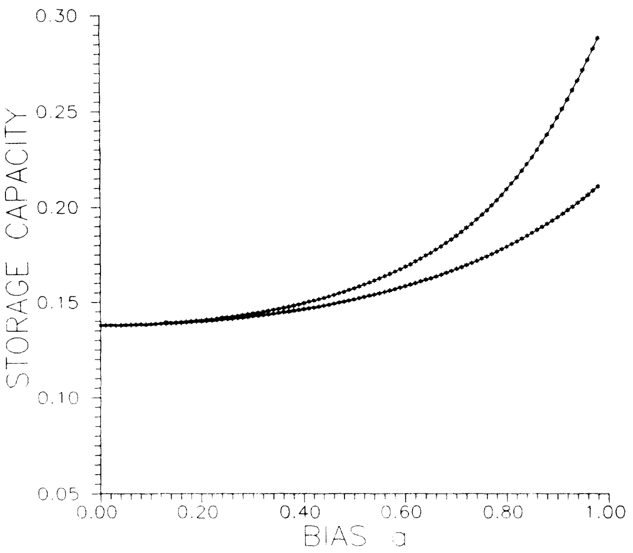

FIG. 2. Storage capacity vs b for $a=0.6$. Upper curve comes from MFT. Lower curve follows Eq. (29).

the parameters a and b, for an optimum external field. For small a,b the results obtained from the mean-field theory fit with those obtained from the signal-to-noise ratio. In this range α varies as

$$\alpha=\frac{\alpha_0}{1-(ab)^2} \tag{29}$$

with $\alpha_0=0.138$. However, both analyses differ in the opposite range. We observe in Fig. 2 that the behavior stated in (4.1) does not hold when a,b increases. For a equal to 1 the results are the same obtained by Ref. 15, consequently, when $b\rightarrow 1$ the capacity approaches the limit deduced by Gardner[17] and the content of information provided by the network is very high.

The variation of the storage capacity α with the external field U, for given a and b, depends on the product ab. For small ab, the capacity varies slightly for a wide band of values of U. As an example we see in Fig. 3(a) that for $ab=0.08$ a change of U of 35% respect its optimum value induces a change of only 5% in α. In contrast, when ab is close to 1 the dependence of the capacity with the field is so important in the neighborhood of the optimum value of the latter that small variations of U generate quite different answers in α. In Fig. 3(b) we observe that for $ab=0.72$ a change of 15% in U gives a change of almost 50% in the capacity. This effect shows the importance to constrain the dynamics of the system by means of an external field.

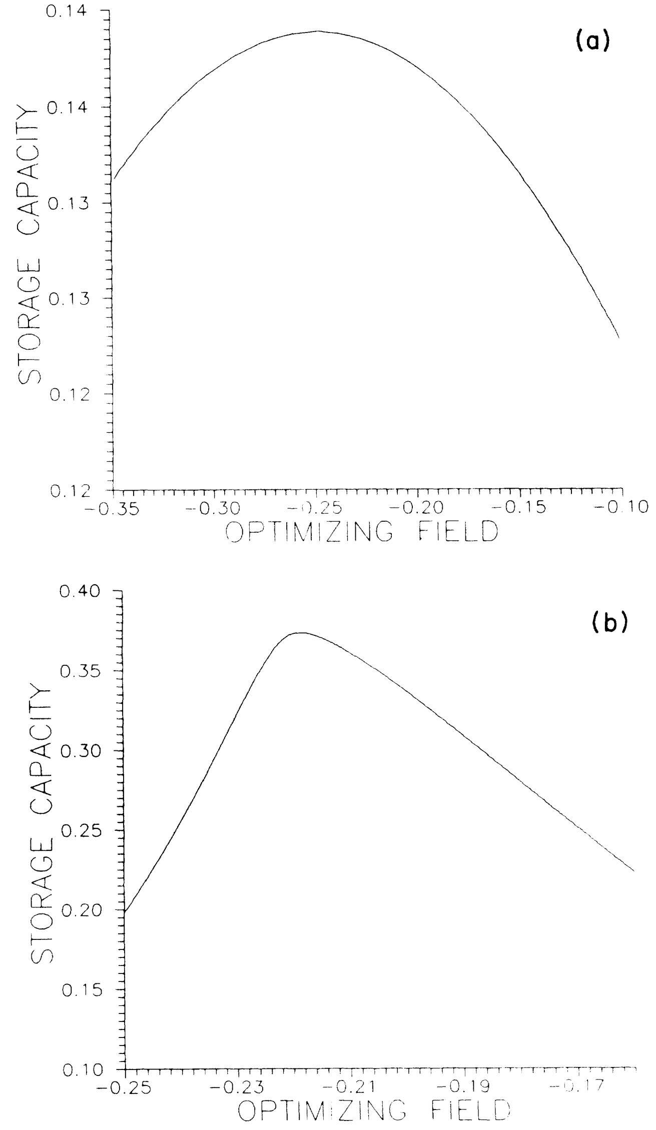

FIG. 3. (a) Dependence of the storage capacity on the external field for $a=0.2$ and $b=0.4$. (b) The same for $a=0.8$ and $b=0.9$.

V. CONCLUSIONS

In this paper I have shown that the method described in the Introduction to optimize neural networks with low levels of activity can be generalized to more complex systems, such as hierarchical models. The method consists of adding an external field whose effect is the equalization of the signal acting on neurons and to make certain assumptions about local thresholds whose effect is the reduction of the destabilizing noise. As a consequence the performance of the network is notably improved by increasing the storage capacity and by decreasing the spurious attractors. No characteristic features of the original model are modified.

ACKNOWLEDGMENTS

I want to express my gratitude to Dr. D. Jou for his interest in the author's work and for his careful reading of the text and Dr. D. Amit for useful discussions. This work has been supported by the grant of the Comision Interministerial de Ciencia y Tecnologia (CICYT) No. PB 86-0287.

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

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).

[3] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987).

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

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

[6] S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).

[7] G. Poppel and U. Krey, Europhys. Lett. **4**, 481 (1987).

[8] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).

[9] H. Gutfreund, Phys. Rev. A **37**, 570 (1988).

[10] A. Krogh and J. A. Hertz, J. Phys. A **21**, 2211 (1988).

[11] N. Parga and M. A. Virasoro, J. Phys. (Paris) **47**, 1857 (1986).

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

[13] J. Buhmann, R. Divko, and K. Schulten, Phys. Rev. A **39**, 2689 (1989).

[14] M. V. Tsodyks and M. V. Feigelman, Europhys. Lett. **6**, 101 (1988).

[15] C. J. Perez-Vicente and D. J. Amit, J. Phys. A **22**, 559 (1989).

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

[17] E. Gardner, J. Phys. A **21**, 257 (1988).