Optimised network for sparsely coded patterns

1989 J. Phys. A: Math. Gen. 22 559

(http://iopscience.iop.org/0305-4470/22/5/018)

J. Phys. A: Math. Gen. 22 (1989) 559-569. Printed in the UK

Optimised network for sparsely coded patterns

Conrado J Perez Vicente† and Daniel J Amit‡§

† Departament de Fisica, Universitat Autonoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

‡ Institute for Advanced Studies, Hebrew University, Jerusalem, Israel

Received 23 August 1988

Abstract. The performance of attractor neural networks storing sparsely coded patterns has been shown to be greatly improved on shifting from the −1, +1 representation of neural states to the 0, 1 representation. Here we show that when this shift is considered as a special case of the transformation of the dynamical variables which depends on a continuous parameter, the value of the parameter can be chosen to improve the performance of the network even further for every value of the bias in the patterns.

1. Introduction

Sparse coding is entering the centre stage of attractor neural network (ANN) research (Amit *et al* 1987a (hereafter referred to as AGS), Gutfreund 1988). It is intended to describe networks characterised by mean spatial levels of activity much lower than 50%, typical of the original proposals (Hopfield 1982, Amit *et al* 1985, 1987b, Personnaz *et al* 1985, Kanter and Sompolinsky 1987). One pressure in this direction comes from neurophysiological observations which find that typically no more than 5% of the neurons will fire within an absolute refractory period. The imaginative studies of Whilshaw *et al* (1969) have concentrated on memories of sparsely coded patterns because their technique—a feed-forward network—found it to be natural. Networks with feedback, ANNs, have made their initial advances storing patterns with 50% of the neurons active.

Early studies of these systems (AGS) extended the Hopfield model so as to make it capable of storing and retrieving patterns with activities of $(1+a)/2$, where the bias (a) can take arbitrary values in the interval, $-1 \leqslant a \leqslant +1$. In the magnetic context a is the magnetisation per spin. The standard model (Hopfield 1982, Amit *et al* 1985, 1987b) stores and retrieves patterns with $a=0$, by prescribing a synaptic matrix

$$J_{ij} = \frac{1}{N} \sum_{\mu}^{p} \zeta_i^\mu \zeta_j^\mu \tag{1.1}$$

where ζ^μ are patterns of randomly chosen +1 and −1, with equal probability, N is the number of neurons in the network and p is the number of stored patterns. The first step on the extension is to write

$$J_{ij} = \frac{1}{N} \sum_{\mu}^{p} (\zeta_i^\mu - a)(\zeta_j^\mu - a) \tag{1.2}$$

for bias a.

§ On leave from the Racah Institute of Physics, Hebrew University, Jerusalem.

This simple extension, which deals with the elimination of the finite mean noise projected by one pattern on another, leads to an extensive storage capacity for the network, but this capacity is always lower than that of the unbiased network. Recall that the storage capacity is defined as the maximal value of α $(=p/N)$, for which retrieval is possible. Moreover, this prescription gives rise to a severe proliferation of spurious states even when the storage level is low, $\alpha \to 0$, unless the configuration space of the network is constrained to states with a given level of activity (AGS). When the constraint is introduced the spurious states are suppressed; there is some mild increase in storage capacity for most of the values of the bias, with a final decrease to zero as $a \to 1$; monotonic decrease to zero of the storage capacity per neuron.

In a seminal paper Gardner (1988) has investigated the general space of connectivities (synaptic efficacies) which can have a given set of patterns as attractors. As a by-product she had found that ANNs can store biased patterns more effectively than unbiased ones. In fact, the storage capacity diverges as $|a| \to 1$. In the process a uniform optimising field (threshold) is introduced, which remains finite as $|a| \to 1$. There is some reduction in the information content per spin, but its limit as $a \to 1$ is finite. The divergence in the storage capacity of the network was corroborated (Buhmann *et al* 1987 (hereafter referred to as BSD), Tsodyks and Feigelman 1988 (hereafter referred to as TF)) in an explicit modification of the standard model.

The main ingredient in the latter proposals (BSD, TF) has been a shift to $V_i = (0, 1)$ designations for natural activity states, from $S_i = (-1, +1)$ used previously. This shift is, of course, a mere formal step and can be effective only if it is accompanied by some assumption about local threshold. The natural description of the local field (post-synaptic potential—PSP) of a neuron i is

$$h_i = \sum_j J_{ij} V_j \tag{1.3}$$

where J_{ij} are synaptic efficacies and $V_j = (0, 1)$ determines whether neuron j has been active and will contribute J_{ij} to the PSP of the neuron i, or is passive and will not contribute. The question of whether neuron i will emit a spike is decided by the variable

$$U_i = h_i - T_i$$

where T_i is the threshold of neuron i. If $U_i > 0$ the firing probability is high and if $U_i < 0$ the probability is low.

For the Hopfield model the reformulation in terms of the $S = (+1, -1)$ can be done by making the substitution

$$V_i = (S_i + 1)/2 \tag{1.4}$$

which implies

$$h_i = \tfrac{1}{2} \sum_j J_{ij} S_j + \tfrac{1}{2} \sum_j J_{ij}. \tag{1.5}$$

In these terms the equivalence between both pictures V and S is complete. If one adds a neuronal threshold

$$T_i = \tfrac{1}{2} \sum_j J_{ij} \tag{1.6}$$

the capacity associated with the 'S' model, with unbiased patterns, becomes double that of the 'V' model (Weisbuch and Fogelman-Soulie 1985, Bruce *et al* 1987).

To improve the network's performance for sparsely coded patterns one chooses

$$T_i = U$$

where U is determined by optimising the retrieval properties. When this optimising field is appropriately fixed (TF, BSD), α can arrive close to the bound derived by Gardner.

Clearly the transformation, equation (1.4), is a particular instance of

$$S_i \to V_i = (S_i - b)/2 \tag{1.7}$$

with $b = -1$. But the transformation can be extended to allow b to take values in $[+1, -1]$. Accompanying the general change of dynamical variables will be a shift of local thresholds

$$h_i \to h_i - (b/2) \sum_j J_{ij} \tag{1.8}$$

on going from V variables to S variables. The main motivation of this paper is to show that there is a choice of the parameter b which optimises the performance of the network.

Following the transformation the network is described by a Hamiltonian

$$H = \tfrac{1}{2} \sum_j J_{ij}(S_i - b)(S_j - b) \tag{1.9}$$

where a factor of $\frac{1}{4}$ will be systematically absorbed into J_{ij}. The minima of H are the stationary states of the following dynamics:

$$S_i(t+1) = \mathrm{sgn}(h_i) \tag{1.10}$$

where h_i is

$$h_i = \sum_j J_{ij}(S_j - b). \tag{1.11}$$

Our main result is that the network performs best when $b = a$ compared with other explicit storage prescriptions (AGS, TF, BSD). It coincides with the Hopfield model for $a = 0$.

The paper is structured as follows. In § 2 we study the signal-to-noise ratio of the uniform external field and of the shift in the dynamical variables. In § 3 we derive the mean-field equations at finite noise T as well as in the limit $T = 0$. The study of the order parameters allows us to describe the behaviour of the simple memories and the spurious states as functions of the bias. In § 4 we analyse the influence of the uniform field on the performance of the network and discuss the role and form of the optimising field. In § 5 we analyse the asymptotic behaviour of the solutions of the mean-field equations to find an explicit expression for the divergence in the storage capacity. The results are confirmed by numerical computation.

2. Signal-to-noise ratio

A useful preliminary guide to the performance of the network can be obtained from the analysis of the signal-to-noise ratio in the local field (PSP). Equation (1.11), supplemented by the uniform field, gives for the PSP of neuron i

$$h_i = \sum_j J_{ij}(S_j - b) - U \tag{2.1}$$

where the synaptic efficacies are given by the modified Hebb rule (1.2) and U is the value of a uniform threshold. The random varuable ζ_i^μ takes the value ± 1 with a probability distribution given by

$$P(\zeta_i^\mu)=\frac{1+a}{2}\delta(\zeta_i^\mu-1)+\frac{1-a}{2}\delta(\zeta_i^\mu+1). \tag{2.2}$$

When the network state coincides with pattern ν ($S_j=\zeta_j^\nu$), the local field is

$$h_i=\frac{1}{N}\sum_{j\neq i}^{N}\sum_{\mu}^{p}(\zeta_i^\mu-a)(\zeta_j^\mu-a)(\zeta_j^\nu-b)-U \tag{2.3}$$

which may be divided into a signal and a noise, namely

$$h_i=\frac{1}{N}(\zeta_i^\nu-a)\sum_{j\neq i}(\zeta_j^\nu-a)(\zeta_j^\nu-b)-U+\frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu}(\zeta_i^\mu-a)(\zeta_j^\mu-a)(\zeta_j^\nu-b). \tag{2.4}$$

In a large system one has for the signal

$$S\simeq(1-a^2)(\zeta_i^\nu-a)-U \tag{2.5}$$

which is optimised when (BSD)

$$U=-a(1-a^2). \tag{2.6}$$

For this value of U the magnitude of the signals on all neurons is equalised to be

$$|S|=(1-a^2). \tag{2.7}$$

The noise—the last term in (2.4)—is Gaussian with zero mean. Its mean square, evaluated with (2.2), is

$$\langle R^2\rangle\simeq\alpha(1-a^2)^2(1+b^2-2ab). \tag{2.8}$$

The main qualitative observation of BSD and TF can be generalised to the fact that the random term $-b\sum J_{ij}$ in (2.3) reduces the noise.

The value of b which minimises R is $b=a$. In this case the signal-to-noise ratio is

$$|S|/R=[\alpha(1-a^2)]^{-1/2} \tag{2.9}$$

which diverges as $a\to 1$, i.e. the coding becomes sparse (Whilshaw *et al* 1969). The amplification of the ratio $|S|/R$ when the bias goes to ± 1 suggests an enhancement of the storage capacity. An estimation of this enhancement of α is

$$\alpha_c\simeq(1-a^2)^{-1} \tag{2.10}$$

arriving close to the bound derived by Gardner (1988). This approach is analysed in detail in the next section. In the unbiased case ($a=0$) the optimal b is 0 and the standard result is recovered (AGS). If one takes $b=1$ then the denominator in (2.9) becomes $\sqrt{2\alpha}$ and the capacity is reduced by a factor of 2 (Weisbuch and Fogelman-Soulie 1985, Bruce *et al* 1987).

3. Mean-field equations

The transformation $S_i\to(S_i-b)/2$ implies a dynamics governed by the Hamiltonian

$$H=-\tfrac{1}{2}\sum_{i\neq j}J_{ij}(S_i-b)(S_j-b)+U\sum_i S_i \tag{3.1}$$

with synaptyic efficacies J_{ij} given by (1.2). The models of biased neural networks studied previously are particular cases of this generalised expression. Concretely, when

$b=-1$ we recover the model of TF and BSD, when $b=+1$ that of Bruce *et al* (1987) and when $b=0$ that of AGS.

To study the statistical mechanics of the system we apply the replica method (Kirkpatrick and Sherrington 1978) to average over the quenched variables ζ. The free energy in the replica symmetric approximation is

$$f=\frac{\alpha}{2}(1-a^2)(1-b^2)+(b-x)U+\frac{\alpha}{2\beta}\left(\ln[1-\beta_1-\beta_2]-\frac{\beta_2 a}{1-\beta_1-\beta_2 C}\right)+\frac{\alpha}{2}(1-a^2)bx$$
$$+\tfrac{1}{2}\sum_\mu (m^\mu)^2+\frac{\alpha\beta}{2}(xy+r[(1-b^2)-(1+b^2)q])-\frac{b\alpha\beta}{2}(y-2br)$$
$$-\frac{1}{\beta}\left\langle\!\!\left\langle \ln 2\cosh\left[\beta\left(\sqrt{\alpha r}\,z-\frac{\alpha\beta}{2}(y-2br)+\sum_\mu m^\mu(\zeta^\mu-a)\right)\right]\right\rangle\!\!\right\rangle \tag{3.2}$$

where β is the inverse temperature, $\beta_1=\beta(1-a^2)(1-b^2)$, $\beta_2=\beta(1-a^2)(1+b^2)$ and $\langle\!\langle\ldots\rangle\!\rangle$ stands for the quenched average over the ζ and

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

The various order parameters appearing in the free energy are defined at the saddle point: m^μ_ρ is a generalised overlap of a state with a learned pattern. The order parameter x, the mean activity of the network or the magnetisation, is

$$x_\rho=-\frac{1}{N}\sum_i\langle(S_i^\rho-b)\rangle\to x \tag{3.3}$$

and the Edwards–Anderson order parameter is

$$q_{\rho\sigma}=\frac{1}{N(1+b^2)}-\sum_i\langle(S_i^\rho-b)(S_i^\sigma-b)\rangle\to q \qquad \text{for} \qquad \rho\neq\sigma \tag{3.4}$$

where single brackets denote thermal or temporal averages. The variable y is the Lagrange multiplier conjugate to x and r is the Lagrange multiplier conjugate to q (see e.g. AGS). They express the mean-square fluctuations of the magnetisation and of the overlaps between the thermodynamic state and the 'non-condensed' patterns, respectively.

Note that the order parameters x and y, which previously (BSD, TF, Bruce *et al* 1987) were considered as a particular feature of the V prescription, appear here in a natural way, which expresses the consequences of the formal transformation (1.7).

Broken ergodicity in the dynamics of the network is parametrised by the solutions of the mean-field equations, which are just the equations for the saddle points (Amit *et al* 1985, 1987b). These equations are obtained by varying the free energy f in (3.2) with respect to the order parameters. One finds five equations

$$m^\mu=\langle\!\langle(\zeta^\mu-a)\tanh(\beta\varphi)\rangle\!\rangle \tag{3.5}$$

$$x=b-\langle\!\langle\tanh(\beta\varphi)\rangle\!\rangle \tag{3.6}$$

$$q=\frac{1}{(1+b^2)}(2bx-b^2+\langle\!\langle\tanh^2(\beta\varphi)\rangle\!\rangle) \tag{3.7}$$

$$r=\frac{q(1-a^2)(1+b^2)}{(1-\beta_1-\beta_2 C)^2} \tag{3.8}$$

$$y=\frac{2U}{\alpha\beta}-\frac{(1-a^2)2b}{\beta}+\frac{2}{\beta^2(1-\beta_1-\beta_2 C)(1+b^2)}\left(1+\frac{q\beta_2}{1-\beta_1-\beta_2 C}\right) \tag{3.9}$$

where

$$\varphi = \sqrt{\alpha r}\, z - \frac{\alpha\beta}{2}(y - 2br) + \sum_{\mu} m^{\mu}(\zeta^{\mu} - a).$$

3.1. Mean-field equations at $T = 0$

To study the retrieval states we assume that the network has a macroscopic overlap with a single stored pattern and microscopic overlap with the result. In the noiseless limit ($\beta \to \infty$) the expressions for the mean-field equations (3.5)-(3.9) reduce to

$$m = \frac{(1-a^2)}{2}[\operatorname{erf}(\phi_1) + \operatorname{erf}(\phi_2)] \tag{3.10}$$

$$x = b - \frac{(1+a)}{2}\operatorname{erf}(\phi_1) + \frac{(1-a)}{2}\operatorname{erf}(\phi_2) \tag{3.11}$$

$$C = \frac{1}{\sqrt{2\pi\alpha r}}\{(1+a)\exp[-\phi_1^2] + (1-a)\exp[-\phi_2^2]\} \tag{3.12}$$

$$r = \frac{(1-a^2)^2(1+2bx-b^2)}{[1-(1-a^2)C]^2} \tag{3.13}$$

where

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

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

and

$$\operatorname{erf}(x/\sqrt{2}) = \sqrt{2/\pi}\int_0^x \exp(-z^2/2)\,dz.$$

The retrieval properties of the noiseless system are derived from a numerical solution of (3.10)-(3.13). We analyse separately the effect of each of the new terms, emphasising their respective influence on the properties of stored and spurious attractors.

4. Properties of the network

4.1. $U = 0$; $b = a$

Even when $U = 0$ the introduction of $b = a$ improves the signal-to-noise ratio relative to the case $b = 0$, studied in AGS. We therefore consider this case first. There is a $T = 0$ phase transition at a storage capacity $\alpha_c = \alpha_c(a)$. $\alpha_c(a)$ decreases monotonically in the entire range of the bias, as can be observed in the lower curve in figure 1. Below this value the retrieval states are dynamically stable attractors with an overlap m bigger than $m_c(a)$, whereas above α_c the network is in a pure spin-glass phase characterised by $m = 0$.

Spurious states are attractors with macroscopic overlaps with several patterns. For $U = 0$ these spurious attractors pervade the dynamics of the system, much as in AGS.

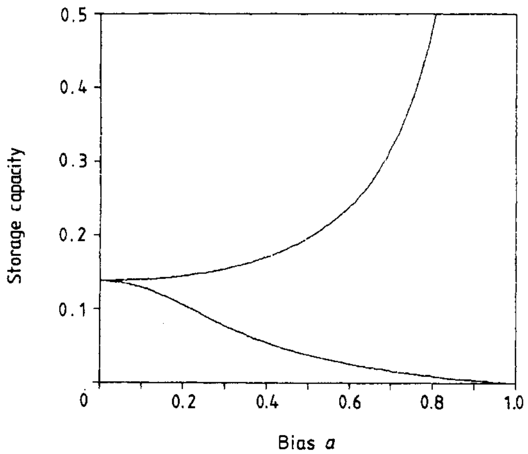

Figure 1. Storage capacity α against the bias a. There is a continuous decrease in α when $U=0$ and $|a|\rightarrow 1$ (lower curve), and a divergence when U is optimum (upper curve).

As an illustration we present in figure 2 the storage capacity associated with attractors which are symmetric mixtures of two, three and five stored patterns. As $|a|\rightarrow 1$ one finds a sequence of peaks corresponding to spurious states mixing an increasing number of patterns. Comparing these results with those of AGS we observe in figure 3(a) and 3(b) a slight improvement in the performance of the network characterised by a small increase of the capacity of the single retrieval states, $\alpha_c(b=a) > \alpha_c(b=0)$, and a small decrease of the capacity associated with spurious attractors, but it looks rather pitiful compared with the results of Gardner (1988).

4.2. *The effect of an optimising threshold: $U \neq 0$; $b=a$*

The behaviour changes radically when the optimising field is added to the system. We still find the same type of phase transition in the α-a plane, but now $\alpha_c(a)$ increases

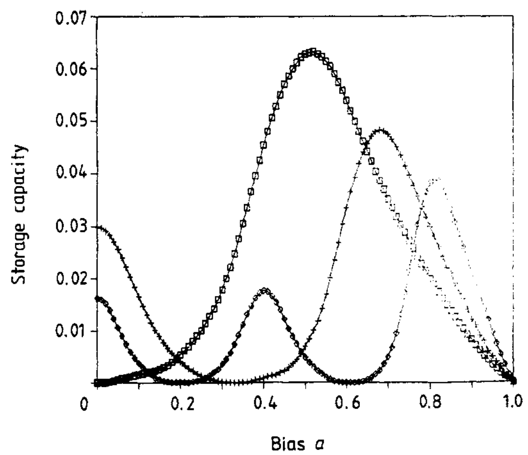

Figure 2. Storage capacity of the spurious states mixing symmetrically two (□), three (+) and five (◇) patterns.

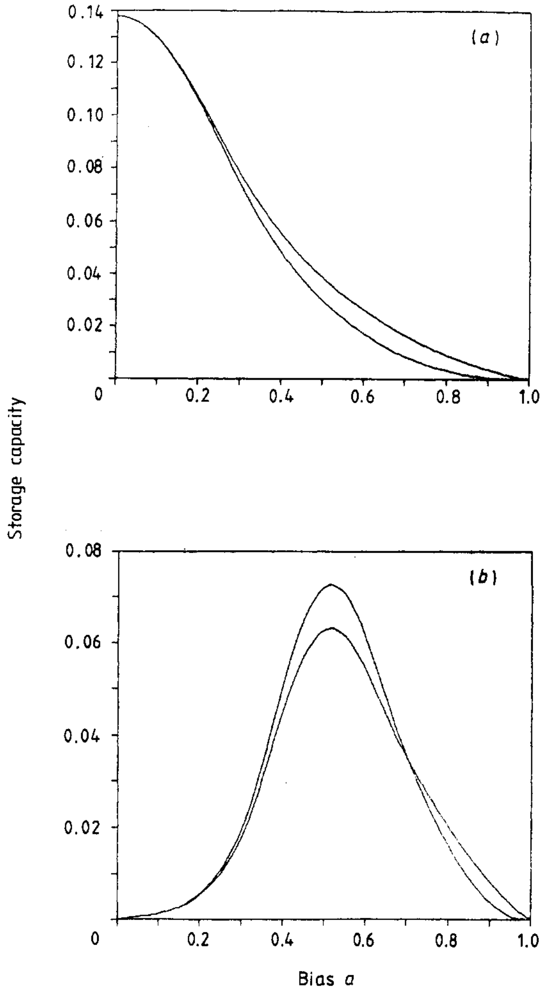

Figure 3. (a) Storage capacity for single retrieval states for $U=0$ and $b=a$ (higher curve), and for $U=0$ and $b=0$ (lower curve). (b) Storage capacity of spurious states mixing two patterns for $U=0$ and $b=a$ (lower peak) compared with $U=0$ and $b=0$ (higher peak).

monotonically for all a, as we can see in the upper curve in figure 1. In fact α_c goes as

$$\alpha_c \simeq k(a)[(1-a)\ln(1-a)]^{-1} \tag{4.1}$$

when $|a| \to 1$. The proportionality factor $k(a)$ is studied in the limit of high bias in the next section. The value of U which optimises the network, i.e. the value which provides the highest α_c for a given bias, has the same parabolic shape deduced in the signal-to-noise analysis (figure 4) and agrees with the results obtained by Gardner (1988). When U is chosen at its optimum value the spurious attractors are eliminated. This is to be expected since the presence of a uniform field is a soft way of introducing a constraint on the states available to the network (AGS). In the thermodynamic limit it prefers states with a given magnetisation. The other effect of the optimising field is that α_c is higher than in other schemes storing biased patterns by explicit storage prescriptions (AGS, BSD, TF).

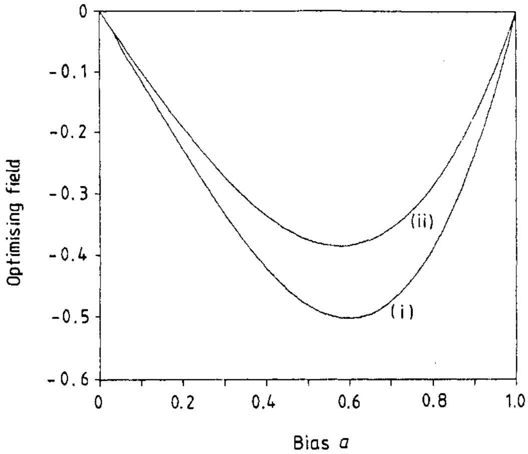

Figure 4. Optimising field U against a, for $T=0$: (i) from the mean-field equations; (ii) from the analysis of the signal-to-noise ratio.

5. Asymptotic behaviour

In this section we examine analytically the asymptotic performance of the network in the limit $|a|\to 1$, at $T=0$ (AGS). When $a=1-\varepsilon$, ($\varepsilon \ll 1$) the error function can be expanded as

$$\operatorname{erf}(x)=1-\frac{\exp(-x^2)}{\sqrt{\pi}\,x}. \tag{5.1}$$

Substituting in the set of equations (3.10)-(3.13) we find a divergence for the storage capacity close to the bound deduced by Gardner. The solutions when the parameter b is equal to the bias a are

$$\alpha_c(a)\simeq\frac{A[1+\sqrt[3]{\ln(1-a)}]}{(1-a)|\ln(1-a)|} \qquad A\simeq 0.144 \tag{5.2}$$

$$m_c(a)\simeq 2(1-a)\left(1-\frac{B}{\sqrt{\ln(1-a)}}\right) \qquad B\simeq 0.307 \tag{5.3}$$

$$r\simeq 8(1-a)^3\left(1-\frac{D}{\sqrt{\ln(1-a)}}\right) \qquad D=\frac{1}{2\sqrt{\pi}}\simeq 0.282 \tag{5.4}$$

$$x\simeq\frac{-D(1-a)}{\sqrt{\ln(1-a)}} \tag{5.5}$$

$$C\simeq\frac{F}{[\ln(1-a)]^{1/8}} \qquad F\simeq 0.522 \tag{5.6}$$

$$U\simeq -2(1-a)\{1+G[\ln(1-a)]^{2/9}\} \qquad G\simeq\frac{1}{2\sqrt{\pi}}. \tag{5.7}$$

The constants $(A-G)$ have been computed from the numerical analysis of the solutions for $\varepsilon<10^{-5}$. As $\varepsilon\to 0$ $(a\to 1)$ our result merges with that of TF and BSD. This is accounted for by the asymptotic expansion which for general b and a close to 1 gives

$$\alpha_c(a,b)\simeq\frac{A[1+\sqrt[3]{\ln(1-a)}]}{(1-a)|\ln(1-a)|}\left(1-\frac{(b-a)^2}{a}\right). \tag{5.8}$$

The capacity is maximised at $b = a$ and for $b = 1$ the results merge when $a \to 1$. The agreement between the asymptotical expressions (5.2)-(5.7) and the results obtained from the numerical study of the equations (3.10)-(3.13) in the $|a| \to 1$ limit is reflected in figure 5, where, as an example, the storage capacity is plotted against the bias.

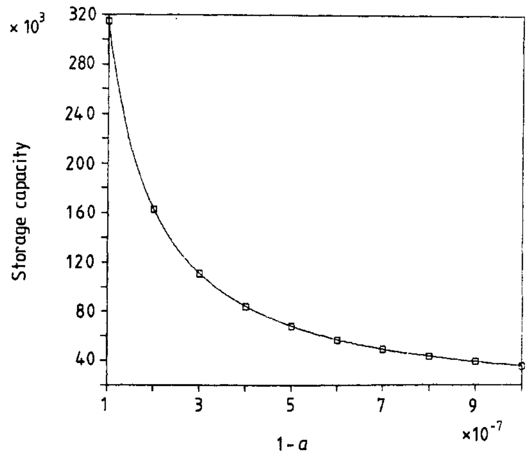

Figure 5. Storage capacity in the asymptotic limit. The full curve corresponds to (5.2). The squares correspond to numerical computation.

6. Conclusions

In this paper we have shown that a transformation, equation (1.7), of the dynamical variables—neural states or spins—with its accompanying thresholds, interpolates between models of $+1, -1$ states and those of $0, 1$. In the process a whole family of models is generated. When the continuous family of models is examined one finds that there is a natural way of optimising the performance of the network. Two contributions bring about the improvement. One is the appearance of a random local field, proportional to the $\Sigma_j J_{ij}$, and hence correlated with the patterns. This term reduces the destabilising noise. The second is the addition of a uniform external field which optimises the signal. Both terms depend on the continuous parameter which characterises the transformation and hence their effect can be optimised by searching in the range of values of this parameter. As a consequence, the retrieval properties of the network are improved, for all values of the bias, even relative to the 0, 1 model, which has produced values close to the bound found by Gardner.

Acknowledgments

CJPV is very grateful to the members of the Institute of Advanced Studies in Jerusalem for their hospitality. The work of CJPV has been supported by a grant from the CIRIT. The work of DJA has been partially supported by a grant from the US-Israel Binational Science Foundation.

References

Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
—— 1987a *Phys. Rev.* A **35** 2293
—— 1987b *Ann. Phys., NY* **173** 30
Bruce A D, Gardner E J and Wallace D J 1987 *J. Phys. A: Math. Gen.* **20** 2909
Buhmann J, Divko R and Schulten K 1987 *Associative memory with high information content* Munich Technical University preprint
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gutfreund H 1988 *Phys. Rev.* A **37** 570
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kirkpatrick S and Sherrington D 1978 *Phys. Rev.* B **17** 4384
Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
Tsodyks M V and Feigelman M V 1988 *Europhys. Lett.* **6** 101
Weisbuch G and Fogelman-Soulie F 1985 *J. Physique Lett.* **46** L623
Whilshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960