Storing patterns in a spin-glass model of neural networks nears saturation

1987 J. Phys. A: Math. Gen. 20 2935

(http://iopscience.iop.org/0305-4470/20/10/036)

J. Phys. A: Math. Gen. **20** (1987) 2935-2947. Printed in the UK

Storing patterns in a spin-glass model of neural networks near saturation

D Grensing†, R Kühn† and J L van Hemmen‡

† Institut für Theoretische Physik und Sternwarte der Universität Kiel, D-300 Kiel, West Germany

‡ Sonderforschungsbereich 123 der Universität Heidelberg, D-6900 Heidelberg, West Germany

Received 22 August 1986, in final form 14 November 1986

Abstract. A Gaussian approximation for the synaptic noise and the $n \to 0$ replica method are used to study spin-glass models of neural networks near saturation, i.e. when the number p of stored patterns increases with the size of the network N as $p = \alpha N$. Qualitative features are predicted surprisingly well. For instance, at $T = 0$ the linear Hopfield network provides effective associative memory with errors not exceeding 0.05% for $\alpha \leq \alpha_c \approx 0.15$. In a network with clipped synapses, the number of patterns which can be stored with some given error tolerance is reduced by a factor of $2/\pi$ as compared with the linear Hopfield model. A simple learning within bounds algorithm is found to continuously interpolate between the linear Hopfield model and the network with clipped synapses.

1. Introduction

The recent theoretical interest in spin glasses as models of neural networks stems from the work of Little (1974) and Hopfield (1982), and a considerable number of research papers on the subject have appeared in the literature (see, for example, Peretto 1984, Amit *et al* 1985a, b, 1987, Kinzel 1985, Nadal *et al* 1986, Toulouse *et al* 1986, Sompolinsky 1986, van Hemmen and Kühn 1986). Several quantitative studies, analytical as well as numerical, have been based on the Hopfield model, where the states of the neurons are modelled by Ising spins S_i, $1 \leq i \leq N$ and the Monte Carlo dynamics of the network is governed by the Hamiltonian

$$H = -\sum_{(ij)} J_{ij} S_i S_j. \tag{1}$$

One stores the information, i.e. patterns $\{\xi_i^\mu, 1 \leq i \leq N\}$ with $1 \leq \mu \leq p$, in couplings according to

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{2}$$

and it is assumed that the ξ_i^μ $(= \pm 1)$ are quenched independent random variables. Information retrieval is defined as the existence of (meta-)stable states which have a non-zero overlap

$$m^\nu = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu \langle S_i \rangle \tag{3}$$

with one of the stored patterns. Here $\langle\ \rangle$ denotes a thermal average with respect to a specific ergodic component.

0305-4470/87/102935+13$02.50

For a finite number p of stored patterns the model has been solved exactly by Amit *et al* (1985a) in the thermodynamic limit: information retrieval is error-free, i.e. $m^{\nu} \to 1$ as $T \to 0$. Of particular interest is the question of the storage capacity of the network (Hopfield 1982, Weisbuch 1985, Amit *et al* 1985b, 1987, Kinzel 1985). From simulations and Gaussian noise arguments, Hopfield (1982) concluded that the system provides associative memory for $p \leq \alpha_c N$ with $\alpha_c \simeq 0.1$-0.2, but an abundance of errors precludes meaningful retrieval when $p \geq \alpha_c N$. Amit *et al* (1985b, 1987) calculated the storage capacity by studying the statistical mechanics of the network with $p = \alpha N$, $0 < \alpha < 1$, within replica theory. They found that the system exhibits associative memory for $p \leq \alpha_c N$ with $\alpha_c \simeq 0.14$ and that, as α passes through α_c, the retrieval overlaps (3) vanish discontinuously. In contrast, investigating the persistence of patterns under synchronous energy relaxation dynamics, Kinzel (1985) estimated that non-zero retrieval overlaps would exist up to $p = \alpha_0 N$, with $\alpha_0 = 2/\pi$, and that the overlaps would vanish continuously at α_0. Information retrieval with errors not exceeding 0.05% was found to be possible for $p \leq \alpha_c N$ with $\alpha_c \simeq 0.15$.

In this paper we turn to the question of storage capacity of generalised Hopfield memories, including the original linear Hopfield model, a model with clipped synapses and a family of models which continuously interpolates between these two. We study the statistical mechanics of these models, utilising the n-replica method. However, we use an approach which differs from that of Amit *et al.* It contains approximations concerning correlations between synaptic strengths J_{ij} which are partly ignored. While our calculations must be regarded as approximate as far as storage capacity is concerned, they represent an exact analysis of a model of learning in a pre-structured brain, as proposed by Toulouse *et al* (1986). Moreover, our method allows us to study also the non-linear models mentioned above. Most of our discussion will be within replica symmetric theory.

2. Statistical mechanics of the Hopfield model near saturation

We study the statistical mechanics of (1) and (2) in the limit $p = \alpha N$, $N \to \infty$. To this end, we evaluate

$$[Z^n] = \left[\mathrm{Tr}_{\{S_i^{\rho}\}} \exp\left(\frac{\beta}{N} \sum_{(ij)} \sum_{\mu=1}^{p} \xi_i^{\mu} \xi_j^{\mu} \sum_{\rho} S_i^{\rho} S_j^{\rho} \right) \right]. \tag{4}$$

Here (ij) denotes independent pairs of lattice sites, $\rho = 1, \ldots, n$ is a replica index and $[\ ]$ stands for an average over the ξ. Addressing the question of the storage capacity of the network, we are going to check whether the system allows (meta-)stable states which have non-zero overlaps (3) with one of the embedded patterns. To do this, we proceed as Amit *et al* did. We single out a finite number of patterns, say $\nu = 1, \ldots, s$, and calculate whether as $N \to \infty$ (3) retains non-zero values in the presence of the static (Gaussian) synaptic noise generated by the $p - s = \alpha N - s$ remaining patterns.

The evaluation of the configuration average in (4) is, therefore, performed in two steps. We decompose the couplings J_{ij} into two parts $J_{ij}^{(l)}$, $J_{ij}^{(h)}$ with

$$J_{ij}^{(l)} = \frac{1}{N} \sum_{\nu=1}^{s} \xi_i^{\nu} \xi_j^{\nu} \tag{5a}$$

$$J_{ij}^{(h)} = \frac{1}{p} \sum_{\mu > s} \xi_i^{\mu} \xi_j^{\mu}. \tag{5b}$$

Here (5*a*) depends on the first s 'low' ξ and (5*b*) on the remaining 'high' ξ. Then we have

$$J_{ij} = J_{ij}^{(l)} + \alpha J_{ij}^{(h)}. \tag{5c}$$

As N becomes large (s remaining finite), the $J_{ij}^{(l)}$ retain their discrete nature, whereas the $J_{ij}^{(h)}$ obey a Gaussian distribution

$$P(J_{ij}^{(h)}) \sim (2\pi/p)^{-1/2} \exp(-pJ_{ij}^{(h)2}/2) \tag{6}$$

by the central limit theorem. We now take

$$P(\{J_{ij}^{(h)}\}) = \prod_{(ij)} P(J_{ij}^{(h)}) \tag{7}$$

thereby ignoring correlations in the synaptic noise generated by the $p - s$ high ξ. Within this approximation, the elementary frustration loop $[J_{ij}^{(h)} J_{jk}^{(h)} J_{ki}^{(h)}]$ is zero whereas if the high-ξ average were performed exactly the result would be p^{-2}. Since we are interested in the saturation limit $p = \alpha N$, the difference between the two results vanishes when the sample size goes to infinity. It should however be noted that the number of elementary loops also increases with the system size, so that (7) is asymptotically exact only in the limit $\alpha \to \infty$. However, we have reason to believe that the results based on this approximation are reasonable also for finite α. First, for $\alpha \ll 1$ where we cannot *a priori* expect this approximation to be good, we obtain values for the retrieval overlaps which are exponentially close to unity, in agreement with previous analyses (Hopfield 1982, Amit *et al* 1985b, Kinzel 1985). Second, performing the $\alpha \to 0$ limit in the linear Hopfield model, we recover the exact finite-p equations of Amit *et al* (1985a).

Using the approximation (7) for the $\{J_{ij}^{(h)}\}$ distribution, we perform a partial average over the high ξ in (4) to obtain

$$\begin{aligned}[Z^n] &= \left[\mathrm{Tr}_{\{S_i^\rho\}} \prod_{(ij)} \exp\left(\frac{\alpha\beta^2}{2N} \left(\sum_\rho S_i^\rho S_j^\rho \right)^2 + \frac{\beta}{N} \sum_{\nu=1}^{s} \sum_\rho \xi_i^\nu \xi_j^\nu S_i^\rho S_j^\rho \right) \right]_{(l)} \\ &= C(N,n) \left[\mathrm{Tr}_{\{S_i^\rho\}} \exp\left(\frac{\alpha\beta^2}{2N} \sum_{(\rho,\sigma)} \left(\sum_i S_i^\rho S_i^\sigma \right)^2 + \frac{\beta}{2N} \sum_{\nu,\rho} \left(\sum_i \xi_i^\nu S_i^\rho \right)^2 \right) \right]_{(l)} \end{aligned} \tag{8}$$

where

$$C(N,n) = \exp[\alpha\beta^2(Nn - n^2)/4 - \beta sn/2] \tag{9}$$

and where $[\ ]_{(l)}$ denotes an average over the low ξ, which remains to be performed. We now linearise the exponential in (8):

$$\begin{aligned}[Z^n] = C(N,n) \int \prod_{(\rho,\sigma)} \frac{\mathrm{d}y_{\rho\sigma}}{(2\pi/N)^{1/2}} \prod_{\nu,\rho} \frac{\mathrm{d}z_{\nu\rho}}{(2\pi/N)^{1/2}} \exp\left(-\frac{N}{2} \sum_{(\rho,\sigma)} y_{\rho\sigma}^2 - \frac{N}{2} \sum_{\nu,\rho} z_{\nu\rho}^2 \right) \\ \times \left[\mathrm{Tr}_{\{S_i^\rho\}} \exp\left(\beta\sqrt{\alpha} \sum_{(\rho,\sigma)} y_{\rho\sigma} \sum_i S_i^\rho S_i^\sigma + \sqrt{\beta} \sum_{\nu,\rho} z_{\nu\rho} \sum_i \xi_i^\nu S_i^\rho \right) \right]_{(l)} \end{aligned} \tag{10}$$

so that the spin trace factorises with respect to the lattice sites.

Before proceeding, we shall briefly discuss the points where our approach differs from that of Amit *et al* and where their approach encounters difficulties which, we feel, one might want to avoid.

The principal difference between the approach of Amit *et al* and our approach is that Gaussian linearisation and high-ξ average are carried out in reverse order. While their procedure has the distinct advantage that the high-ξ average can be performed

exactly (at least for the linear model considered here), there is a serious drawback: the Gaussian linearisation in their approach leads to integrals whose evaluation requires expansion of a hyperbolic cosine to second order in its supposedly small argument. Unfortunately this procedure transforms integrals which are initially convergent into generalised Gaussian integrals which do not generally exist. For $\beta > \beta_c = 1$ the quadratic form involved is no longer positive definite, the diagonal elements of the matrix being $1-\beta$ (for all $1 \leq n \in \mathbb{N}$). However, it is conceivable that in the limit $n \to 0$ things become correct but there is no guarantee that one finds the proper saddle point, whence our motivation to take a second look at the problem.

We now proceed and rewrite equation (8) as

$$[Z^n] = C(N,n) \int \prod_{(\rho,\sigma)} \frac{dy_{\rho\sigma}}{(2\pi/N)^{1/2}} \prod_{\nu,\rho} \frac{dz_{\nu,\rho}}{(2\pi/N)^{1/2}} \exp\left(-\frac{N}{2}\sum_{(\rho,\sigma)} y_{\rho\sigma}^2 - \frac{N}{2}\sum_{\nu,\rho} z_{\nu\rho}^2\right)$$
$$\times \exp N\left(\frac{1}{N}\sum_i \ln \mathrm{tr}_{\{S_i^\rho\}} \exp\left(\beta\sqrt{\alpha}\sum_{(\rho,\sigma)} y_{\rho\sigma} S_i^\rho S_i^\sigma + \sqrt{\beta}\sum_{\nu,\rho} z_{\nu\rho}\xi_i^\nu S_i^\rho\right)\right). \quad (11)$$

Here $\mathrm{tr}_{\{S_i^\rho\}}$ denotes a trace over the 2^n states of the replicated spins on a single lattice site and we have exploited the fact that the low-ξ average $[\ ]_{(l)}$ can be obtained by self-averaging. For large N the integral in (11) is dominated by its saddle-point value. The physical interpretation of the parameters $y_{\rho\sigma}$ and $z_{\nu\rho}$ appearing in (11) is determined from the saddle-point equations. We find

$$y_{\rho\sigma} = \beta\sqrt{\alpha}\, q_{\rho\sigma} = \beta\sqrt{\alpha}\,\frac{1}{N}\sum_i \langle S_i^\rho S_i^\sigma\rangle_i \qquad \rho \neq \sigma$$
$$z_{\nu\rho} = \sqrt{\beta}\, m_{\nu\rho} = \sqrt{\beta}\,\frac{1}{N}\sum_i \xi_i^\nu \langle S_i^\rho\rangle_i \qquad \nu = 1,\ldots,s;\ \rho = 1,\ldots,n \quad (12)$$

with

$$\langle \ldots\rangle_i = \frac{\mathrm{tr}_{\{S_i^\rho\}}(\ldots)\exp(\alpha\beta^2 \Sigma_{(\rho,\sigma)} q_{\rho\sigma} S_i^\rho S_i^\sigma + \beta \Sigma_{\nu,\rho} m_{\nu\rho}\xi_i^\nu S_i^\rho)}{\mathrm{tr}_{\{S_i^\rho\}}\exp(\alpha\beta^2 \Sigma_{(\rho,\sigma)} q_{\rho\sigma} S_i^\rho S_i^\sigma + \beta \Sigma_{\nu,\rho} m_{\nu\rho}\xi_i^\nu S_i^\rho)}. \quad (13)$$

Thus $q_{\rho\sigma}$ is a spin-glass order parameter and $m_{\nu\rho}$ is related to the retrieval overlaps defined in (3). Unlike Amit *et al* we do not find, nor do we require, an order parameter analogous to their $r_{\rho\sigma}$.

The average free energy per spin, as obtained by the n-replica method†, is

$$-\beta f = \lim_{n\to 0}\lim_{N\to\infty}\frac{1}{nN}\ln[Z^n]. \quad (14)$$

Inserting (9), (11) and (12) we obtain

$$-\beta f = \frac{\alpha\beta^2}{4} + \lim_{n\to 0}\frac{1}{n}\max_{m,q}\left(-\frac{a\beta^2}{2}\sum_{(\rho,\sigma)} q_{\rho\sigma}^2 - \frac{\beta}{2}\sum_{\nu,\sigma} m_{\nu\rho}^2\right.$$
$$\left. + \frac{1}{N}\sum_i \ln \mathrm{tr}_{\{S_i^\rho\}}\exp\left(\alpha\beta^2\sum_{(\rho,\sigma)} q_{\rho\sigma} S_i^\rho S_i^\sigma + \beta\sum_{\nu,\rho} m_{\nu\rho}\xi_i^\nu S_i^\rho\right)\right). \quad (15)$$

† We prefer this formulation (van Hemmen and Palmer 1979) to the conventional one, because it allows limits to be taken in the correct order.

We shall now proceed to look for replica symmetric solutions of (12) with

$$\begin{aligned} q_{\rho\sigma} &= q \qquad \rho \neq \sigma \\ m_{\nu\rho} &= m^{\nu} \qquad \rho = 1, \ldots, n. \end{aligned} \tag{16}$$

It is then straightforward to evaluate the free energy. The result is

$$-\beta f = \frac{\alpha\beta^2}{4}(1-q)^2 - \frac{\beta}{2}\sum_{\nu}(m^{\nu})^2 + \frac{1}{N}\sum_{i}\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \ln(2\cosh(\beta\sqrt{\alpha q}z + \beta\sum_{\nu} m^{\nu}\xi_i^{\nu})) \tag{17}$$

where $\boldsymbol{m} = (m^{\nu})$ and q satisfy the fixed point equations

$$\begin{aligned} \boldsymbol{m} &= \langle\!\langle \boldsymbol{\xi} \tanh(\beta\sqrt{\alpha q}z + \beta \boldsymbol{m}\cdot\boldsymbol{\xi}) \rangle\!\rangle \\ q &= \langle\!\langle \tanh^2(\beta\sqrt{\alpha q}z + \beta \boldsymbol{m}\cdot\boldsymbol{\xi}) \rangle\!\rangle. \end{aligned} \tag{18}$$

Here $\langle\!\langle\ \rangle\!\rangle$ denotes the combination of a Gaussian zero-mean/unit-variance average with respect to z and a discrete average over the ξ which occurs in (17).

Note that by taking the limit $\alpha \to 0$ in equations (18) we recover the equations describing the statistical physics of the Hopfield model for a finite number (s) of embedded patterns as derived by Amit *et al* (1985a).

For non-zero α, equations (18) have two types of solution.

(i) A solution with $\boldsymbol{m} = 0$, $q \neq 0$. It represents a true spin-glass (SG) state which has no macroscopic overlap with any of the embedded patterns. Note that in the pure SG phase q obeys the fixed point equation

$$q = \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh(\beta\sqrt{\alpha q}z). \tag{19}$$

This is precisely the Sherrington–Kirkpatrick (1975) fixed point equation for replica symmetric solutions of the infinite range Edwards–Anderson spin glass having independent random couplings with a Gaussian distribution of zero mean and variance $\sqrt{\alpha/N}$.

(ii) The so-called retrieval solutions with $\boldsymbol{m} \neq 0$, $q \neq 0$. The solutions, which exist for sufficiently small α, are responsible for the functioning of the network as an associative memory.

In what follows, we shall be concerned with the nature and existence of the retrieval solutions. The most important retrieval states are those which have macroscopic overlaps with a single pattern, $m^{\mu} = m\delta_{\mu\nu}$. Following Amit *et al* (1985b), we discuss the nature of these solutions to resolve the issue of the storage capacity of the network.

For retrieval states with $m^{\mu} = m\delta_{\mu\nu}$, equations (18) are

$$\begin{aligned} m &= \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh(\beta\sqrt{\alpha q}z + \beta m) \\ q &= \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh^2(\beta\sqrt{\alpha q}z + \beta m). \end{aligned} \tag{20}$$

Again, these equations describe replica symmetric solutions of an SK spin glass having independent random couplings with a Gaussian distribution of common mean $1/N$ and variance $\sqrt{\alpha/N}$.

At low temperatures, equations (20) do not, however, give rise to a maximum of (15). Replica symmetry is known to be broken below the de Almeida-Thouless (1978) line

$$(k_B T)^2 = \alpha \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \operatorname{sech}^4(\beta\sqrt{\alpha q}z + \beta m) \tag{21}$$

which for small α takes the form

$$k_B T \sim \frac{4}{3}\left(\frac{\alpha}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2\alpha}\right). \tag{22}$$

For $\alpha \to 0$, replica symmetry breaking (RSB) occurs exponentially close to $T = 0$, where the system is already almost fully ordered and we expect the effects of RSB to be weak in this limit.

Ignoring the implications of RSB for a while, we discuss the solution of (20) at $T = 0$. As $T \to 0$, equations (20) yield

$$\begin{aligned} q &= 1 \\ m &= \operatorname{erf}(m/\sqrt{2\alpha}). \end{aligned} \tag{23}$$

Non-zero solutions for the amplitude m of the retrieval overlap exist only for $\alpha \leqslant \alpha_0 = 2/\pi$. Near α_0 we have

$$m = m(\alpha) \propto (\alpha_0 - \alpha)^{1/2}. \tag{24}$$

Since for $\alpha \leqslant \alpha_0$ the retrieval amplitude is small, the system does not provide useful associative memory in that regime. To give an estimate of the storage capacity, we must specify an error tolerance (thereby, of course, introducing an element of arbitrariness into the definition). If we require that the retrieval error does not exceed 0.05%, then efficient associative memory is provided for $\alpha \leqslant \alpha_c \simeq 0.15$.

These results were previously derived by Kinzel (1985) without using replicas on the basis of estimating the persistence of patterns under synchronous energy relaxation dynamics. The physical interpretation of α_0 in that approach is that, under completely parallel dynamics, an initial state which has some small macroscopic overlap with one of the embedded patterns will move away from that pattern if $\alpha \geqslant \alpha_0$.

Summarising, in the independent Gaussian approximation (7) the Hopfield memory provides efficient associative memory for $\alpha \leqslant \alpha_c \simeq 0.15$. As $p = \alpha N$ with $\alpha > \alpha_c$, the memory function rapidly deteriorates. Retrieval overlaps with a single pattern are non-zero for $\alpha < \alpha_0 = 2/\pi$ where they vanish continuously. There is no indication of a discontinuous transition, as found by Amit *et al* (1985b, 1987). However, as $\alpha \to 0$, we do recover the exact finite-p solution of Amit *et al* (1985a).

Before closing this section, we note that the calculations presented above are interesting in another context, namely they represent an exact analysis of learning in a pre-structured brain (Toulouse *et al* 1986). To establish this correspondence, we rewrite the couplings J_{ij} (equations (5*a*)-(5*c*)) in the form

$$J_{ij} = \sqrt{\alpha}\left(J_{ij}^0 + \frac{\varepsilon}{N}\sum_{\nu=1}^{s} \xi_i^\nu \xi_j^\nu\right)$$

with

$$\varepsilon = 1/\sqrt{\alpha} \qquad J_{ij}^0 = \sqrt{\alpha} J_{ij}^{(h)}.$$

The common prefactor $\sqrt{\alpha}$ can be absorbed by rescaling the temperature. The J_{ij}^0 are interpreted as inherent Gaussian distributed synaptic strengths, with zero mean and variance $1/N$, and they provide a spin-glassy initial state of the network. Patterns are then learnt according to the generalised Hebb rule. The retrieval quality is now a function of $\varepsilon = 1/\sqrt{\alpha}$ and non-zero retrieval overlaps exist only for $\varepsilon > \varepsilon_0 = \sqrt{\pi/2}$. Retrieval is practically perfect (with an error margin better than 0.05%) for $\varepsilon > 2.58$, in good agreement with results of Toulouse *et al* (1986) and Nadal *et al* (1986).

3. Networks with synapses of bounded strength

The model discussed in § 2 may be generalised in various ways (Hopfield 1982, 1984, Nadal *et al* 1986, Sompolinsky 1986). In the following two subsections we discuss the storage capacity of two simple generalisations of the Hopfield model which have synapses of bounded strength, a model with so-called clipped synapses (Hopfield 1982) and a family of models which implements a simple algorithm of learning within bounds. While a motivation to investigate clipped synapses is that they might be easier to construct in terms of electronic circuitry, the learning within bounds algorithm may be considered as a first and crude attempt to mimic the saturation effects observed in real biological systems. Other networks with synapses of bounded strength were recently studied by Mézard *et al* (1986), but these are linear generalisations of the Hopfield model, whereas in the present section we investigate non-linear models.

3.1. Clipped synapses

The genuine Hopfield model discussed above combines binary and analogue data processing elements, since the patterns presented to the network are given in a binary representation whereas the storing of these N-bit words via Hebb's learning rule is an analogue process. In this section we turn to a neural network which is wholly digitised, i.e. memory is located in so-called clipped synapses of the form

$$J_{ij} = \frac{\sqrt{p}}{N} \operatorname{sgn}\left(\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \right). \tag{25}$$

One important reason for considering clipped synapses is that they appear to be easier to realise in silicon versions of the Hopfield model, which are constructed of simple processing elements. For odd p, the flip-flop nature of the synapses permits an elementary estimate of the loading capacity. Obviously, only $N(N-1)/2$ bits of information can possibly be stored in this system of N neurons with its $N(N-1)/2$ clipped synapses (25). Thus, presenting more than $p = (N-1)/2$ patterns to this network, each representing an N-bit word, will overload the synaptic system. By this crude reasoning one can, of course, not decide whether overloading the system will mean a total or rather gradual loss of previously stored information. Moreover, the possibility cannot be excluded that retrieval of the patterns $\{\xi_i^\mu\}$ is significantly degraded before the storage capacity of the synapses is exhausted, i.e. for $\alpha < \frac{1}{2}$. The aim of this subsection is to demonstrate that (within our independent Gaussian approximation (7) and replica symmetric theory) the retrieval overlaps vanish continuously at $\alpha_0 = (2/\pi)^2 = 0.405$, which is remarkably close to our crude estimate $\frac{1}{2}$.

The starting point is again the averaged partition function of n replicas of the system denoted by $[Z^n]$. Following the procedure developed in the last section, the

synaptic efficacies appearing now in the argument of the non-linear sign function in equation (25) are decomposed into the high and the low ξ parts. Then the couplings given by equation (25) can be written in the form

$$J_{ij} = \frac{\sqrt{p}}{N} \operatorname{sgn}(J_{ij}^{(l)} + \alpha J_{ij}^{(h)}) \tag{26}$$

where we have used the definitions of equation (5). Again, we use the independent Gaussian approximation (7) for the $\{J_{ij}^{(h)}\}$ distribution, and we investigate whether the system has states of macroscopic overlap with the low ξ as defined in equation (3). The partial averaging over the high ξ leads to the multiple integral

$$[Z^n] = \left[\operatorname{Tr}_{\{S_i^\rho\}} \int \prod_{(ij)} \frac{\mathrm{d}J_{ij}^{(h)}}{(2\pi/p)^{1/2}} \exp\left(-\frac{p}{2} \sum_{(ij)} J_{ij}^{(h)2} + \frac{\beta\sqrt{p}}{N} \sum_{(ij)} \sum_{\rho} S_i^\rho S_j^\rho \operatorname{sgn}(J_{ij}^{(l)} + \alpha J_{ij}^{(h)}) \right) \right]_{(l)} . \tag{27}$$

Making the substitution $J_{ij}^{(h)} + \alpha^{-1} J_{ij}^{(l)} = y_{ij}$ and breaking the integration $\int \mathrm{d}y_{ij}$ from $-\infty$ to 0 and from 0 to ∞, the quenched average of Z^n in equation (27) is expressible in terms of the complementary error function

$$[Z^n] = \left[\operatorname{Tr}_{\{S_i^\rho\}} \prod_{(ij)} \tfrac{1}{2} \sum_{t_{ij}=\pm 1} \exp\left(-t_{ij} \frac{\beta\sqrt{p}}{N} \sum_{\rho} S_i^\rho S_j^\rho + \ln \operatorname{erfc}(t_{ij} N J_{ij}^{(l)} / \sqrt{2p}) \right) \right]_{(l)} . \tag{28}$$

Expanding ln erfc to second order in its $O(1/\sqrt{p})$ argument and performing the t_{ij} sums leads to a product of hyperbolic cosines. These are re-exponentiated to give

$$[Z^n] = \left[\operatorname{Tr}_{\{S_i^\rho\}} \exp\left(-\frac{1}{\pi p} \sum_{(ij)} (N J_{ij}^{(l)})^2 \right) \times \prod_{(ij)} \exp\left(\ln \cosh\left(\frac{\beta\sqrt{p}}{N} \sum_{\rho} S_i^\rho S_j^\rho + \left(\frac{2}{\pi p} \right)^{1/2} N J_{ij}^{(l)} \right) \right) \right]_{(l)} . \tag{29}$$

The details of the subsequent procedure are the same as for the linear Hopfield model described in the previous section. Expanding $\ln \cosh(x)$, retaining only terms of the order $1/N$, one ends up with the following result within the replica symmetric theory:

$$-\beta f = \tfrac{1}{4}\alpha\beta^2(1-q)^2 - \frac{1}{2}\left(\frac{2}{\pi}\right)^{1/2} \beta \sum_{\nu} (m^\nu)^2 + \frac{1}{N} \sum_i \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \ln\left\{ 2 \cosh\left(\beta\sqrt{\alpha q} z + \beta \left(\frac{2}{\pi} \right)^{1/2} \sum_{\nu} m^\nu \xi_i^\nu \right) \right\} \tag{30}$$

with $m^\nu (\nu = 1, \ldots, s)$ and q satisfying the transcendental equations

$$q = \left\langle\!\left\langle \tanh^2\left(\beta\sqrt{\alpha q} z + \beta \left(\frac{2}{\pi} \right)^{1/2} \boldsymbol{m} \cdot \boldsymbol{\xi} \right) \right\rangle\!\right\rangle \tag{31}$$

$$\boldsymbol{m} = \left\langle\!\left\langle \boldsymbol{\xi} \tanh\left(\beta\sqrt{\alpha q} z + \beta \left(\frac{2}{\pi} \right)^{1/2} \boldsymbol{m} \cdot \boldsymbol{\xi} \right) \right\rangle\!\right\rangle . \tag{32}$$

Here $\langle\!\langle\ \rangle\!\rangle$ again denotes the combined average over the low ξ and over the Gaussian noise z. These are the same type of equations as in the unclipped case, the only modification being that the ferromagnetic part in the local field appears with a factor $\sqrt{2/\pi}$. This entails, for instance, that the transcendental equation for the $T = 0$ retrieval solutions $m^\mu = m\delta_{\mu\nu}$ is now

$$\mathrm{m} = \operatorname{erf}(m/\sqrt{\alpha\pi}) \tag{33}$$

instead of (23) for linear synapses. This result has already been anticipated in the pioneering work of Hopfield (1982) who observed (without presenting the details): 'The signal-to-noise ratio can be evaluated analytically for this clipped algorithm and is reduced by a factor of $(2/\pi)^{1/2}$ compared with the unclipped case. For a fixed error probability, the number of memories must be reduced by $2/\pi$.' The present analysis, however, goes far beyond this Hopfield statement and leads to the remarkable relationship that all results obtained in § 2 and summarised in the phase diagram (figure 1) can be translated into the clipped case with the prescription that the inverse temperature β and the ratio $\alpha = p/N$ have to be rescaled by factors $\sqrt{\pi/2}$ and $2/\pi$, respectively. (This holds also for the occurrence and implications of replica symmetry breaking.) Sompolinsky (1986) also considered clipped synapses and reports a reduction of the storage capacity relative to the Hopfield model by a factor of about 1/1.4 which is to be compared with our result $2/\pi$.

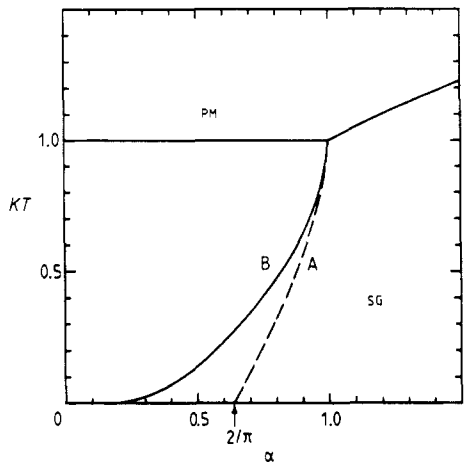

Figure 1. Phase boundaries of the linear Hopfield model. The broken curve A indicates where the macroscopic overlaps with a single pattern vanish and curve B is the AT line. PM and SG denote paramagnetic and spin-glass phases respectively.

3.2. Learning within bounds

In this subsection we study another neural network where the synaptic interconnections J_{ij} are bounded functions of the synaptic efficacies

$$T_{ij} = \frac{1}{\sqrt{p}} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{34}$$

but now the sgn function used in equation (25) is replaced by a function which is linear on the interval $[-a, a]$ and a constant otherwise, thus combining properties of both the linear Hopfield case and the digitising clipped version. To be precise, we shall work with

$$H = -\frac{\sqrt{p}}{N} \sum_{(ij)} \phi(T_{ij}) S_i S_j \tag{35}$$

where

$$\phi(x) = \begin{cases} bx & \text{for } |x| \leq a \\ ab\ \mathrm{sgn}(x) & \text{for } |x| > a. \end{cases} \tag{36}$$

This non-linear model takes account of saturation effects observed in neurobiology in a rather crude manner. More refined versions of learning within bounds were studied by Nadal *et al* (1986) and Parisi (1986), their intention being to avoid the total loss of the memory function for large α. The primitive version presented in this section is found to be unable to prevent this deterioration of the memory due to overloading.

Our remarkable result is that in the independent Gaussian approximation the phase diagram of the non-linear system described by equations (34)-(36) is again of the Hopfield type illustrated in figure 1, with the only modification being that the temperature T and the ratio $\alpha = p/N$ have to be rescaled by factors depending on the parameters a, b.

The steps towards this result are just the same as those in § 3.1.

Replicate the system and decompose the synaptic efficiencies (equation (34)) according to $T_{ij} = T_{ij}^{(l)} + T_{ij}^{(h)}$, the two contributions being composed of the low and high ξ, respectively. Finally, neglecting correlations in the synaptic noise generated by the high ξ by assuming that the probability distribution of the high T is (cf (6) and (7))

$$P(\{T_{ij}^{(h)}\}) = \prod_{(ij)} \frac{1}{\sqrt{2\pi}} \exp(-T_{ij}^{(h)2}/2) \tag{37}$$

one is able to perform the averaging over the high ξ. The definition of the non-linear function ϕ in equation (36) now suggests breaking the resulting integrals into three contributions from regions where ϕ is 'simply' defined, i.e. either a constant or a linear function of its argument. Then, as in the clipped case, the resulting integrals are expressible in terms of error functions. Exploiting the smallness of their arguments, simple algebra, which completely parallels that of § 3.1, yields the following result for the quenched average of the replicated partition sum:

$$[Z^n] = C(N, n; a, b)\ \mathrm{Tr}_{\{S_i^\rho\}}\ \exp\left(u \frac{\alpha\beta^2}{2N} \sum_{(\rho,\sigma)} \left(\sum_i S_i^\rho S_i^\sigma \right)^2 + v \frac{\beta}{2N} \sum_{\nu,\rho} \left(\sum_i \xi_i^\nu S_i^\rho \right)^2 \right) \tag{38}$$

where

$$C(N, n; a, b) = \exp(u\alpha\beta^2(Nn - n^2)/4 - v\beta sn/2) \tag{39}$$

and the parameters u and v are given by

$$u = b^2 \left(a^2 + (1 - a^2)\ \mathrm{erf}(a/\sqrt{2}) - a \left(\frac{2}{\pi} \right)^{1/2} \exp(-a^2/2) \right) \tag{40}$$

$$v = b\ \mathrm{erf}(a/\sqrt{2}). \tag{41}$$

Comparing this with the analogous formula (equation (8)) derived for the Hopfield model one can immediately conclude that the free energy per lattice site in the present case is related to the free energy f_0 in the Hopfield case via

$$f(\alpha, \beta) = f_0(uv^{-2}\alpha, v\beta). \tag{42}$$

The fixed point equations are readily found to be

$$\begin{aligned} \boldsymbol{m} &= \langle\!\langle \boldsymbol{\xi} \tanh(\beta\sqrt{u\alpha q}\, z + v\beta \boldsymbol{m} \cdot \boldsymbol{\xi}) \rangle\!\rangle \\ q &= \langle\!\langle \tanh^2(\beta\sqrt{u\alpha q}\, z + v\beta \boldsymbol{m} \cdot \boldsymbol{\xi}) \rangle\!\rangle \end{aligned} \tag{43}$$

from which it follows at once that the number of patterns which can be stored with some given error tolerance is reduced by a factor

$$\otimes(a) = v^2/u$$
$$= \frac{\operatorname{erf}^2(a/\sqrt{2})}{a^2 + (1-a^2)\operatorname{erf}(a/\sqrt{2}) - (2/\pi)^{1/2} a \exp(-a^2/2)} \tag{44}$$

as compared with the linear Hopfield model.

As illustrated in figure 2, $F(a) = v^2/u$ increases monotonically with the parameter a, remains finite in the whole interval $[0, \infty)$ and interpolates smoothly between the limiting cases of the linear Hopfield model ($a \to \infty$) and the clipped case ($a \to 0$).

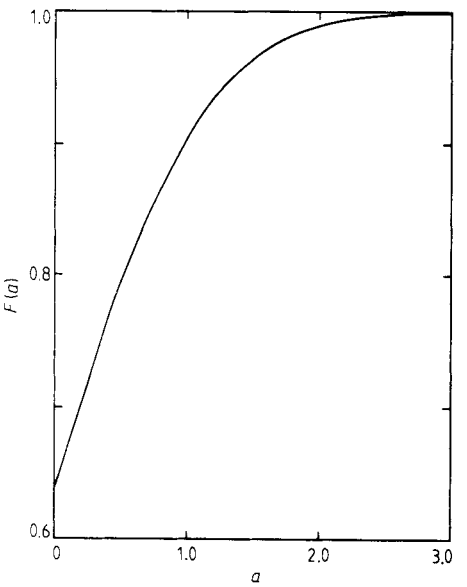

Figure 2. The reduction factor for the storage capacity $F(a) = v^2/u$ plotted against the parameter a which determines the region where the synaptic function ϕ increases linearly (cf equations (35) and (36)).

The fact that the parameter b does not occur in equation (44) stems from the fact that it can be absorbed in the inverse temperature and hence will not appear in zero-temperature results such as the ratio v^2/u given in equation (44).

Note that simply imposing bounds on the synaptic couplings while leaving the underlying Hebb rule unaltered appears not to be sufficient to prevent the deterioration of the memory function for large α. To achieve this, additional ingredients are required, like, for example, the iterative redefinition of the synaptic interconnections by learning (Hopfield 1982, Parisi 1986, Nadal *et al* 1986) or non-uniform learning intensities (Nadal *et al* 1986).

4. Discussion

We have studied generalised Hopfield memories near saturation where the number p of stored patterns increases with the size N of the system as $p = \alpha N$, using a Gaussian

approximation which neglects correlations in the synaptic noise. In the subspace of phases which have macroscopic (retrieval) overlaps with only one of the embedded patterns, the linear Hopfield model is found to be equivalent to an SK model with ferromagnetic anisotropy (see also Feigelman and Ioffe 1986). Surprisingly, the same conclusion also holds for the model with clipped synapses and the model with a simple learning within bounds algorithm, with inverse temperature and ferromagnetic anisotropy, however, appropriately rescaled. If α is small enough to permit reliable retrieval of information (i.e. with an error fraction $\leqslant 0.5\%$), the replica symmetry is unbroken down to very low temperatures where the system is already almost fully ordered and we expect the effects of replica symmetry breaking to be small. This, of course, deserves further investigation.

The complete analogy with the SK model breaks down when phases are studied which have macroscopic overlaps with several of the embedded patterns. In the linear Hopfield model, as $\alpha \to 0$, these phases approach the mixture states studied by Amit *et al* (1985a). For the non-linear models studied in § 3, however, the correct finite p equations (van Hemmen *et al* 1986, van Hemmen and Kühn 1986) cannot be recovered from our approximation.

This indicates that, within the Gaussian approximation, some information is lost, including possibly subtle details such as the order of the phase transition.

While at finite p the linear and non-linear versions of the Hopfield model appear to be quite different at very low temperatures (van Hemmen *et al* 1986, van Hemmen and Kühn 1986, Sompolinsky 1986), we find these differences to be smoothed out in the saturation limit. In both versions of the model, it is the synaptic noise generated by the embedded patterns which is ultimately responsible for the performance of the network. The non-linearities merely appear to be an additional source of noise.

Acknowledgments

The authors would like to thank the referees for helpful advice. Moreover, they are indebted to one of the referees for having drawn their attention to the correspondence that exists between the calculations presented in § 2 and the problem of learning in a pre-structured brain.

References

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev.* A **32** 1007
—— 1985b *Phys. Rev. Lett.* **55** 1530
—— 1987 *Ann. Phys., NY* **173** 30
de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
Feigelman M V and Ioffe L B 1986 *Europhys. Lett.* **1** 197
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
—— 1984 *Proc. Natl Acad. Sci. USA* **81** 3088
Kinzel W 1985 *Z. Phys.* B **60** 205
Little W A 1974 *Math. Biosci.* **19** 101
Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457
Nadal J P, Toulouse G, Changeux J P and Dehane S 1986 *Europhys. Lett.* **1** 335
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617
Peretto P 1984 *Biol. Cybern.* **50** 51
Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
Sompolinsky H 1986 *Phys. Rev.* A **34** 2571

Toulouse G, Dehane S and Changeux J P 1986 *Proc. Natl Acad. Sci. USA* **83** 1695
van Hemmen J L, Grensing D, Huber A and Kühn R 1986 *Preprint* Heidelberg
van Hemmen J L and Kühn R 1986 *Phys. Rev. Lett.* **57** 913
van Hemmen J L and Palmer R G 1979 *J. Phys. A: Math. Gen.* **12** 563
Weisbuch G 1985 *Proc. Les Houches Meeting, February 1985* ed E Bienenstock *et al* (Berlin: Springer)