Storing extensively many weighted patterns in a saturated neural network

1987 J. Phys. A: Math. Gen. 20 3989

(http://iopscience.iop.org/0305-4470/20/12/043)

J. Phys. A: Math. Gen. **20** (1987) 3989-3999. Printed in the UK

Storing extensively many weighted patterns in a saturated neural network

J L van Hemmen† and V A Zagrebnov‡

† Sonderforschungsbereich 123, Universität Heidelberg, D-6900 Heidelberg, Federal Republic of Germany

‡ Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, 141980 Dubna, USSR

Received 21 October 1986

Abstract. The performance of the Hopfield model of a neural network with *extensively* many weighted patterns is analysed. If the system size is N, then N patterns, each provided with a suitable weight, are stored. The weights may be associated with a temporal order and, if appropriately chosen, they allow a *gradual* fading out of the extensively many stored patterns. Particular emphasis is put on the underlying mathematical structure.

1. Introduction

At a formal level, a neural network is a set of two-state neurons connected by synapses and equipped with a certain dynamics that has the memorised patterns as stable attractors. One of the main problems of the theory of neural networks is to describe the way in which the information should be stored and to find a mechanism which allows both retrieval (recollection) and forgetting of the stored patterns.

One usually considers a *fully* interconnected network of, say, N neurons. According to McCulloch and Pitts [1], a neuron can be in only one of two states (firing and non-firing) and it, therefore, can be described by an Ising spin $S(i)$, $1 \leqslant i \leqslant N$, with $+1$ corresponding to firing and -1 to quiescent. In this context, a pattern is a specific Ising spin configuration.

It is generally expected [2-6] that the essential characteristics of the temporal behaviour of the network are captured by a Monte Carlo dynamics with a Hamiltonian of the form

$$H_N = -\tfrac{1}{2} \sum_{i \neq j} J_{ij} S(i) S(j). \tag{1.1}$$

Then the dynamics of the network is reduced to a downhill motion in the (free) energy landscape associated with H_N and the asymptotic stability is governed by its equilibrium statistical mechanics. Below a critical temperature T_c the ergodicity is broken [7] and the stored patterns are associated with attractive sets (equilibrium states or ergodic components) in the phase space of the underlying Ising spin glass.

Following Hebb [8] one locates memory *in the synapses*, i.e. more precisely, in the distribution of values of the synaptic efficacies, which are then mapped onto the exchange couplings J_{ij} of the Ising spin Hamiltonian (1.1). For suitable couplings, the network operates as a fault-tolerant content-addressable (associative) memory; see for example [2-6]. Additional patterns may be learnt by appropriately modifying the

0305-4470/87/123989+11$02.50

J_{ij}. To facilitate the modelling one assumes the patterns $\{\xi_{i\alpha}; 1 \leqslant i \leqslant N\}$, say with $1 \leqslant \alpha \leqslant p$, to be random. That is, the $\xi_{i\alpha} = \pm 1$ are independent, identically distributed random variables which assume the values ± 1 with equal probability.

The Hopfield model of associative memory [2] is defined by

$$J_{ij} = N^{-1} \sum_{\alpha=1}^{p} \xi_{i\alpha}\xi_{j\alpha}. \tag{1.2}$$

In a recent paper [9], Amit *et al* analysed this network near saturation, i.e. when $p = \alpha N$ and $\alpha > 0$. Through an ingenious mean-field analysis they showed that at zero temperature ($T = 0$) the system can efficiently retrieve information if $\alpha < \alpha_c = 0.14$. At α_c, however, the retrieval states disappear *dis*continuously and above α_c no useful retrieval is possible. The system performs, so to speak, a first-order transition at α_c and has more or less forgotten *all* information thereafter. From a physiological point of view this does not seem very plausible. Therefore, other learning rules have been proposed but, as yet, their status is not very satisfying. For instance, Parisi [10] has constructed a memory 'which forgets' but in so doing the memory forgets everything but the very last patterns.

In this paper two things are done. First we rederive the main result of Amit *et al* [9] in a simpler and more transparent way, paying due attention to the underlying mathematical structure. Second, we extend the Hopfield model by storing N patterns in a system of size N (so it is fully saturated) and giving each pattern a weight ε_ν. If the labelling of the patterns corresponds to the temporal order in which they arrived and $\varepsilon_\nu \to 0$ as $\nu \to \infty$, which seems reasonable, then for suitable weights an extensive number of patterns may be stored but, as ν proceeds, they are *gradually* faded out.

In § 2 the model itself is defined and its mean-field treatment is analysed. In the next section the replica symmetric solution to the free energy is presented. The associated storage capacity at $T = 0$ is obtained in § 4. A discussion of our results can be found in the final section.

2. Hopfield model with weighted patterns

We consider the Hamiltonian (1.1) with exchange couplings

$$J_{ij} = N^{-1} \sum_{n=1}^{N} \varepsilon_n \xi_{in}\xi_{jn}. \tag{2.1}$$

Each pattern n has a weight ε_n. For the time being the weights are arbitrary except for the requirement that $0 \leqslant \varepsilon_n \leqslant 1$. In fact, we could replace the upper limit of the sum in (2.1) by a number proportional to but larger than N. We recover the Hopfield model by putting $\varepsilon_n = 1$ for $1 \leqslant n \leqslant \alpha N$ and $\varepsilon_n = 0$ for $n > \alpha N$.

Following Amit *et al* [9] we use the replica method and determine the stability of a certain pattern by selecting a *finite* number of patterns, to be denoted by ν, and integrating out the remaining ones, whose labels are indicated by μ. For later purposes it may be convenient to add an external field that singles out the ν-patterns,

$$H_{\text{ext}} = -\sum_{\nu} h_\nu \sum_{i=1}^{N} \xi_{i\nu} S(i). \tag{2.2}$$

In the context of the replica method [11], one first determines

$$\phi_N(n) = N^{-1} \ln\langle Z_N^n \rangle \tag{2.3}$$

for positive integer n, takes the thermodynamic limit $N \to \infty$ so as to arrive at $\phi(n)$, and obtains an extension (usually the replica-symmetric one) to a neighbourhood of $n = 0$. Then $\phi'(0)$ is supposed to give $-\beta f(\beta)$ where $f(\beta)$ is the free energy per spin at inverse temperature β. As usual, $Z_N = \mathrm{Tr} \exp(-\beta H_N)$ is the partition function, a sum over all Ising spin configurations.

The angular brackets in (2.3) denote an average over the disorder, here the N patterns $\xi_{i\mu}$. Since we first integrate out the μ-patterns, we will leave aside the rest of the Hamiltonian and instead of $\langle Z_N^n \rangle$ concentrate on

$$\left\langle \exp\left[\frac{\beta}{2N} \sum_{\mu,\rho} \varepsilon_\mu \left(\sum_{i=1} \xi_{i\mu} S_\rho(i) \right)^2 - \tfrac{1}{2}\beta n \sum_\mu \varepsilon_\mu \right] \right\rangle. \tag{2.4}$$

Here $1 \leqslant \rho \leqslant n$ labels the n replicas. Until the next section (equation (3.5)) we drop the constant term $-\frac{1}{2}\beta n \sum_\mu \varepsilon_\mu$. Using the relation

$$\exp(\tfrac{1}{2}\lambda a^2) = \int_{-\infty}^{+\infty} \frac{dz}{(2\pi)^{1/2}} \exp(-\tfrac{1}{2}z^2 + \sqrt{\lambda}\, az) \tag{2.5}$$

we linearise the squares in the exponent of (2.4),

$$\left\langle \int \prod_{\mu,\rho} \frac{dm_{\mu\rho}}{(2\pi)^{1/2}} \exp\left\{ -\frac{1}{2} \sum_{\mu,\rho} m_{\mu\rho}^2 + \sum_{\mu,\rho} m_{\mu\rho} \left[\left(\frac{\beta\varepsilon_\mu}{N} \right)^{1/2} \sum_{i=1}^{N} \xi_{i\mu} S_\rho(i) \right] \right\} \right\rangle \tag{2.6}$$

and perform the average with respect to $\xi_{i\mu}$ so as to find

$$\int \prod_{\mu,\rho} \frac{dm_{\mu\rho}}{(2\pi)^{1/2}} \exp\left[\left[-\frac{1}{2} \sum_{\mu,\rho} m_{\mu\rho}^2 + \sum_{i,\mu} \ln\left\{ \cosh\left[\left(\frac{\beta\varepsilon_\mu}{N} \right)^{1/2} \sum_{\rho=1}^{n} m_{\mu\rho} S_\rho(i) \right] \right\} \right] \right]. \tag{2.7}$$

One now imagines that the term between the square brackets in (2.7) is 'small' and replaces $\ln(\cosh(x))$ for 'small' x by $\frac{1}{2}x^2$. This gives

$$\int \prod_{\mu,\rho} \frac{dm_{\mu\rho}}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2} \sum_{\mu,\rho} m_{\mu\rho}^2 + \sum_{\substack{i,\mu \\ \rho,\sigma}} \frac{\beta\varepsilon_\mu}{2N} m_{\mu\rho} m_{\mu\sigma} S_\rho(i) S_\sigma(i) \right) \tag{2.8}$$

where $1 \leqslant \rho, \sigma \leqslant n$. Fixing μ the integration with respect to $m_{\mu\rho}$ can be performed exactly and we obtain

$$\det(Q_\mu)^{-1/2} \tag{2.9}$$

where Q_μ is a *symmetric* $n \times n$ matrix with elements

$$(Q_\mu)_{\rho\sigma} = \delta_{\rho\sigma} - \beta\varepsilon_\mu \left(N^{-1} \sum_{i=1}^{N} S_\rho(i) S_\sigma(i) \right). \tag{2.10}$$

Here and elsewhere $\delta_{\rho\sigma}$ is the Kronecker delta. Collecting terms we obtain

$$\prod_\mu \det(Q_\mu)^{-1/2} = \exp\left(-\frac{1}{2} \sum_\mu \mathrm{Tr}(\ln Q_\mu) \right). \tag{2.11}$$

Stepping back for a first overview we see [12] that something must be wrong. For (2.9) to make sense the matrix Q_μ has to be positive-definite and, therefore, its diagonal elements *must* be positive. However, $(Q_\mu)_{\rho\rho} = 1 - \beta\varepsilon_\mu$ is negative for β large enough. The reason for this can be traced back to the transition from (2.7) to (2.8). Whereas (2.7) is well defined for all β (since $\ln(\cosh(x)) \sim |x|$ as $|x| \to \infty$), (2.8) is not (fix μ and take the diagonal elements of the quadratic form in the integrand), and there is no

way out. We do need the quadratic approximation of $\ln(\cosh(x))$ since otherwise no analytic evaluation of (2.7) is possible. Furthermore, as we will see shortly, Q_μ becomes positive-definite in the replica limit $n \to 0$.

The term (2.11), which still depends on the spins but contains no randomness anymore, represents the *noise* produced by the other patterns and has to be added to the remaining part of the Hamiltonian, which contains the ν-patterns. This complex can be treated exactly, modulo the difficulty we just noted. We have, by (2.3) and (2.11),

$$\phi_N(n) = N^{-1} \ln\left\langle \mathrm{Tr}_{S_\rho} \exp\left\{ N\left[\frac{1}{2}\beta \sum_{\nu,\rho} \varepsilon_\nu \left(N^{-1} \sum_{i=1}^{N} \xi_{i\nu} S_\rho(i) \right)^2 \right.\right.\right.$$
$$\left.\left.\left. + \beta \sum_{\nu,\rho} h_\nu \left(N^{-1} \sum_{i=1}^{N} \xi_{i\nu} S_\rho(i) \right) - \frac{1}{2} N^{-1} \sum_{\mu} \mathrm{Tr}(\ln Q_\mu) \right] \right\} \right\rangle. \tag{2.12}$$

The first trace is a sum over all 2^{nN} Ising spin configurations of the n replicas and the second one is an ordinary trace. Let us define the order parameters

$$\begin{aligned} m_{\nu\rho} &= N^{-1} \sum_{i=1}^{N} \xi_{i\nu} S_\rho(i) \qquad 1 \leq \rho \leq n \\ q_{\rho\sigma} &= N^{-1} \sum_{i=1}^{N} S_\rho(i) S_\sigma(i) \qquad 1 \leq \rho < \sigma \leq n. \end{aligned} \tag{2.13}$$

Then the expression between the square brackets in (2.12) may be written

$$F(\boldsymbol{m}, \boldsymbol{q}) = \beta \sum_{\nu,\rho} (\tfrac{1}{2}\varepsilon_\nu m_{\nu\rho}^2 + h_\nu m_{\nu\rho}) - \tfrac{1}{2} N^{-1} \sum_{\mu} \mathrm{Tr} \ln Q_\mu(\boldsymbol{q}) \tag{2.14}$$

and it therefore seems natural to perform a coordinate transformation from the $S_\rho(i)$, $1 \leq i \leq N$ and $1 \leq \rho \leq n$, to $m_{\nu\rho}$ and $q_{\rho\sigma}$ as new 'integration' variables. To this end we only need the 'Jacobian' $\mathscr{D}_N(\boldsymbol{m}, \boldsymbol{q})$. Indeed, it can be shown [13-15] that, as $N \to \infty$, the coordinate transformation which we referred to is possible and, with probability one,

$$\mathscr{D}_N(\boldsymbol{m}, \boldsymbol{q}) = \exp(-N c^*(\boldsymbol{m}, \boldsymbol{q})) \tag{2.15}$$

where

$$c^*(\boldsymbol{m}, \boldsymbol{q}) = \sup_{(\boldsymbol{x},\boldsymbol{y})} (\boldsymbol{m} \cdot \boldsymbol{x} + \boldsymbol{q} \cdot \boldsymbol{y} - c(\boldsymbol{x}, \boldsymbol{y})) \tag{2.16}$$

is the Legendre transform [16] of a (strictly) convex c-function [13-15],

$$c(\boldsymbol{x}, \boldsymbol{y}) = \left\langle \ln \mathrm{Tr} \exp\left(\sum_{\nu,\rho} x_{\nu\rho} \xi_\nu S_\rho + \dot{\sum_{(\rho,\sigma)}} y_{\rho\sigma} S_\rho S_\sigma \right) \right\rangle. \tag{2.17}$$

The second sum in (2.17) is over pairs (ρ, σ) only. The trace refers to n Ising spins S_ρ, $1 \leq \rho \leq n$, and in the outer average each ξ_ν appears only once; there are finitely many of them. In addition [15], as is already implicit in its formulation, (2.15) does not depend on the specific random configuration as $N \to \infty$. For that reason we may drop the angular brackets from (2.12). (This is *not* a consequence of the self-averaging property of the free energy as was asserted by Amit *et al* [9]. The averaging has to be done *inside* the logarithm. In passing we also note that the order parameters $r_{\rho\sigma}$ as introduced by these authors simply can be dispensed with.)

Combining (2.12)-(2.17) and writing $\boldsymbol{\mu}$ instead of the pair $(\boldsymbol{m}, \boldsymbol{q})$ we find

$$\phi(n) = \lim_{N \to \infty} \frac{1}{N} \ln \int \mathrm{d}\boldsymbol{\mu} \exp[N(F(\boldsymbol{\mu}) - c^*(\boldsymbol{\mu}))] \tag{2.18}$$

which is, by a Laplace argument,

$$\phi(n) = \sup_{\mu}(F(\mu) - c^*(\mu)). \tag{2.19}$$

For small enough β this expression is exact but it becomes purely formal as soon as $\beta \max_\mu(\varepsilon_\mu) > 1$. In spite of that we proceed and obtain, as shown in the appendix,

$$\phi(n) = \max_{\mu}[F(\mu) - \mu \cdot \nabla F(\mu) + c(\nabla F(\mu))] \tag{2.20}$$

where μ satisfies the fixed point equation

$$\mu = \nabla c(\nabla F(\mu)). \tag{2.21}$$

Equation (2.20) has to be understood in the same sense as (2.19).

The matrix $Q_\mu(q)$ has elements $(\rho < \sigma)$

$$(Q_\mu)_{\rho\sigma} = \delta_{\rho\sigma}(1 - \beta\varepsilon_\mu) - \beta\varepsilon_\mu q_{\rho\sigma} = (Q_\mu)_{\sigma\rho}. \tag{2.22}$$

Using the relation

$$\frac{\partial}{\partial Q_{\rho\sigma}} \operatorname{Tr} \ln Q = 2(Q^{-1})_{\rho\sigma} \tag{2.23}$$

one easily verifies that, with $\mu = (m, q)$,

$$\nabla F(\mu) = \begin{pmatrix} \beta(\varepsilon_\nu m_{\nu\rho} + h_\nu) \\ N^{-1} \Sigma_\mu \beta\varepsilon_\mu (Q_\mu^{-1})_{\rho\sigma} \end{pmatrix}. \tag{2.24}$$

We now perform the extension of (2.20) to $n = 0$ by assuming replica symmetry.

3. Replica symmetry

We require that all the replicas be equal so that $m_{\nu\rho} = m_\nu$ and $q_{\rho\sigma} = q$ $(\rho \neq \sigma)$. This requirement is consistent with the fixed point equation (2.21). Moreover,

$$Q_\mu(q) = (1 - \beta\varepsilon_\mu + \beta\varepsilon_\mu q)\mathbb{1} - \beta\varepsilon_\mu qn \left| \frac{1}{\sqrt{n}} \mathbf{1} \right\rangle \left\langle \frac{1}{\sqrt{n}} \mathbf{1} \right| \equiv a(n) - nb(n)P \tag{3.1}$$

where $\mathbb{1}$ is the unit matrix, $\mathbf{1}$ is the vector $(1, 1, \ldots, 1) \in \mathbb{R}^n$, and $P = P^2$ is a projection operator. Through the ansatz $Q^{-1} = c - dP$ the inverse of Q is easily obtained and $(\rho \neq \sigma)$

$$(Q^{-1})_{\rho\sigma} = -b(n)[a(n)(nb(n) - a(n))]^{-1}. \tag{3.2}$$

Hence we can write

$$N^{-1} \sum_\mu \beta\varepsilon_\mu (Q_\mu^{-1})_{\rho\sigma} \equiv \beta^2 q r(n), \tag{3.3}$$

where $r(n)$ does not depend explicitly on β, and by (2.24)

$$\nabla F(\mu) = \begin{pmatrix} \beta(\varepsilon_\nu m_\nu + h_\nu) \\ \beta^2 q r(n) \end{pmatrix}. \tag{3.4}$$

Using (3.1) one directly verifies that the eigenvalues of $Q_\mu(q)$ are $1 - \beta\varepsilon_\mu(1 - q) + \beta\varepsilon_\mu qn$, which is simple, and $1 - \beta\varepsilon_\mu(1 - q)$, which is $(n - 1)$-fold degenerate. In the limit $n \to 0$ one is left with $1 - \beta\varepsilon_\mu(1 - q)$, which has to be positive; cf (3.7) below. In the next section and in the Hopfield model as considered by Amit *et al* [9] one can verify that this limit value is indeed positive.

Combining (2.17), (2.20) and (3.1)-(3.4), and adding the constant we dropped from (2.4) we obtain

$$\phi(n) = -\tfrac{1}{2}\beta n\Big(N^{-1}\sum_{\mu}\varepsilon_{\mu}\Big) - \tfrac{1}{2}\beta n\Big(\sum_{\nu} m_{\nu}^2\Big) - \frac{1}{2}N^{-1}\sum_{\mu}\{\ln[1-\beta\varepsilon_{\mu}(1-q)+\beta\varepsilon_{\mu}qn] + (n-1)\ln[1-\beta\varepsilon_{\mu}(1-q)]\} - \tfrac{1}{2}n(n-1)(\beta q)^2 r(n) + \Big\langle \ln \mathrm{Tr}_{S_\rho} \exp\Big(\beta\sum_{\nu,\rho}(\epsilon_{\nu}m_{\nu}+h_{\nu})\xi_{\nu}S_{\rho} + \tfrac{1}{2}\beta^2\sum_{\rho\neq\sigma} qr(n)S_{\rho}S_{\sigma}\Big)\Big\rangle. \tag{3.5}$$

If we use the linearisation trick (2.5) and carry out the trace, we can rewrite the last term of (3.5) in the form

$$-\tfrac{1}{2}n\beta^2 qr(n) + n\ln 2 + \Big\langle \ln\int_{-\infty}^{+\infty}\frac{\mathrm{d}z}{(2\pi)^{1/2}}\mathrm{e}^{-z^2/2}\cosh^n\{\beta[(\varepsilon\boldsymbol{m}+\boldsymbol{h})\cdot\boldsymbol{\xi}+(qr(n))^{1/2}z]\}\Big\rangle \tag{3.6}$$

where $\varepsilon = \mathrm{diag}(\varepsilon_{\nu})$ is a diagonal matrix. Taking the 'evident' real variable extension of $\phi(n)$ we then obtain

$$-\beta f(\beta) = \lim_{n\to 0} n^{-1}\phi(n) = -\tfrac{1}{2}\Big(N^{-1}\sum_{\mu}\varepsilon_{\mu}\Big) - \tfrac{1}{2}\beta\Big(\sum_{\nu}\varepsilon_{\nu}m_{\nu}^2\Big) - \tfrac{1}{2}N^{-1}\sum_{\mu}\{\ln[1-\beta\varepsilon_{\mu}(1-q)] - \beta\varepsilon_{\mu}q[1-\beta\varepsilon_{\mu}(1-q)]^{-1} - \tfrac{1}{2}\beta^2 qr(1-q) + \Big\langle\int\frac{\mathrm{d}z}{(2\pi)^{1/2}}\mathrm{e}^{-z^2/2}\ln[\![2\cosh\{\beta[(\varepsilon\boldsymbol{m}+\boldsymbol{h})\cdot\boldsymbol{\xi}+\sqrt{qr}z]\}]\!]\Big\rangle \tag{3.7}$$

with N very large and

$$r = \lim_{n\to 0} r(n) = N^{-1}\sum_{\mu}\varepsilon_{\mu}^2[1-\beta\varepsilon_{\mu}(1-q)]^{-2} \geqslant 0. \tag{3.8}$$

Furthermore, one should choose that solution of the fixed point equations,

$$\boldsymbol{m} = \langle\!\langle \boldsymbol{\xi}\tanh[\beta((\varepsilon\boldsymbol{m}+\boldsymbol{h})\cdot\boldsymbol{\xi}+(qr)^{1/2}z)]\rangle\!\rangle \tag{3.9a}$$

$$q = \langle\!\langle \tanh^2[\beta((\varepsilon\boldsymbol{m}+\boldsymbol{h})\cdot\boldsymbol{\xi}+(qr)^{1/2}z)]\rangle\!\rangle \tag{3.9b}$$

which maximises the right-hand side of (3.7). The double angular brackets in (3.9) denote an average with respect to both the finite number of ξ_{ν} and the Gaussian distribution of z. If $\varepsilon_{\mu}=1$ for $1\leqslant\mu\leqslant\alpha N$ and $\varepsilon_{\mu}=0$ for $\mu>\alpha N$, then (3.7)-(3.9) reproduce the result of Amit *et al* [9]. Throughout what follows we put $\boldsymbol{h}=0$. In view of (3.8) and (3.9) we interpret r as a renormalisation constant that rescales the order parameter q. It is a consequence of the noise generated by the 'infinitely many' other patterns ($N\to\infty$).

4. Storage capacity

Given a collection of weights ε_{μ}, the storage capacity is the (maximum) number of patterns which have not lost their own stability completely. In this section we study the storage capacity at $T=0$ of a network with $\varepsilon_{\mu}=\mu^{-x}$ and $x>0$. For suitable x there is a gradual fading out of the patterns as μ proceeds. One easily verifies that for $x>0$ the average $N^{-1}\Sigma_{\mu}\mu^{-x}$ converges to zero (Césaro convergence) and,

therefore, that $N^{-1}\Sigma_\mu \operatorname{Tr}(\ln Q_\mu)$, the background noise, converges to zero too. This does not mean, however, that it can be neglected completely, since the weight ε_μ of a pattern μ also converges to zero as $\mu \to \infty$. There is a trade off between the two, which we now want to determine.

We have to take two limits, $N \to \infty$ and $\beta \to \infty$, which have to be performed in a specific order. First we have to take the thermodynamic limit $N \to \infty$, then the zero-temperature limit $\beta \to \infty$. The two limits cannot be interchanged. Since it is evident that q will approach one we start with (3.9*a*). In the limit $\beta \to \infty$, the $\tanh\{\ldots\}$ in (3.9*a*) reduces to $\operatorname{sgn}\{\ldots\}$ and the integral over the Gaussian distribution gives up to a term of order T an error function

$$\int_{-\infty}^{+\infty} \frac{dz}{(2\pi)^{1/2}} e^{-z^2/2} \tanh\{\ldots\} \approx \left(\frac{2}{\pi}\right)^{1/2} \int_0^{\varepsilon \boldsymbol{m}\cdot\boldsymbol{\xi}/\sqrt{r}} dz\, e^{-z^2/2} = \operatorname{erf}(\varepsilon \boldsymbol{m}\cdot\boldsymbol{\xi}/(2r)^{1/2}) \tag{4.1}$$

where

$$\operatorname{erf}(x) = \frac{2}{\pi^{1/2}} \int_0^x dy\, e^{-y^2} \tag{4.2}$$

is the error function, so that (3.9*a*) reduces to

$$\boldsymbol{m} = \langle \boldsymbol{\xi} \operatorname{erf}(\varepsilon \boldsymbol{m}\cdot\boldsymbol{\xi}/(2r)^{1/2})\rangle_{\boldsymbol{\xi}}. \tag{4.3}$$

We are interested in the behaviour of a *specific* pattern, say ν. Then $\boldsymbol{m}$ is assumed to have only one component m and

$$m = \langle \xi_\nu \operatorname{erf}(m\varepsilon_\nu \xi_\nu/(2r)^{1/2})\rangle = \operatorname{erf}(m\varepsilon_\nu/(2r)^{1/2}). \tag{4.4}$$

As is evident from (2.13), the closer m is to one, the better is the retrieval. Because the weights ϵ_ν tend to zero in Césaro mean, r approaches zero as $N \to \infty$ and, for *fixed* ν, m tends to one. In the same sense we get for a finite group of fixed indices ν and for finite β,

$$\boldsymbol{m} = \langle \boldsymbol{\xi} \tanh\{\beta\varepsilon \boldsymbol{m}\cdot\boldsymbol{\xi}\}\rangle_{\boldsymbol{\xi}} \tag{4.5}$$

and for a single non-zero component

$$m = \tanh(\beta\varepsilon_\nu m). \tag{4.6}$$

The larger ν, the smaller the critical temperature $T_c(\nu)$ at which a non-zero m branches off into the direction of the pattern ν.

We now turn to (3.9*b*). In the thermodynamic limit it reduces to

$$q = \langle \tanh^2\{\beta\varepsilon \boldsymbol{m}\cdot\boldsymbol{\xi}\}\rangle_{\boldsymbol{\xi}} \tag{4.7}$$

and, as $\beta \to \infty$, it then converges to one at an *exponential* rate. Hence $C(\beta) = \beta(1-q)$ converges to zero. As before, $\varepsilon = \operatorname{diag}(\varepsilon_\nu)$.

What happens at finite but very large N? How many patterns can we store at $T = 0$? The noise factor r is not completely zero yet. It can be estimated by using the above observation in tandem with (3.8),

$$r = N^{-1} \sum_\mu (\varepsilon_\mu^{-1} - C(\beta))^{-2} \xrightarrow[\beta\to\infty]{} N^{-1} \sum_\mu \mu^{-2x} \tag{4.8}$$

which we approximate by

$$r = N^{-1} \int_1^N d\mu\, \mu^{-2x} = (1-2x)^{-1}(N^{-2x} - N^{-1}) \tag{4.9}$$

if $x \neq \frac{1}{2}$ and by

$$r = N^{-1} \int_1^N d\mu\, \mu^{-1} = N^{-1} \ln N \tag{4.10}$$

if $x = \frac{1}{2}$. We discern three cases: (a) $0 < x < \frac{1}{2}$, (b) $x = \frac{1}{2}$ and (c) $x > \frac{1}{2}$. In the first case the tail of the integral in (4.9) dominates and we find

$$r = (1-2x)^{-1} N^{-2x}. \tag{4.11a}$$

In the second case, when $x = \frac{1}{2}$, we obtain the same behaviour but the divergence is logarithmic,

$$r = N^{-1} \ln N \tag{4.11b}$$

whereas in the last case the region $\mu \approx 1$ dominates so that

$$r = (2x-1)^{-1} N^{-1}. \tag{4.11c}$$

With these values of r return to (4.4).

The function $\psi(x) = \operatorname{erf}(\eta x)$ is concave ($\eta > 0$) and monotonically increasing for $x \geqslant 0$. It starts at zero and converges to one. In short, it more or less behaves like $\tanh(\eta x)$ were it not that its slope at zero is $2\eta/\pi^{1/2}$; cf (4.2). Therefore, the equation

$$m = \operatorname{erf}(\eta m) \tag{4.12}$$

has only one solution, $m = 0$, if $2\eta/\pi^{1/2} < 1$, i.e. if $\eta < \eta_c = \frac{1}{2}\pi^{1/2}$. For $\eta > \eta_c$ there also exists a non-trivial solution that converges to zero *continuously* as η approaches η_c from above.

In view of (4.4) and (4.11) there exists a critical ν_c such that for $\nu > \nu_c$ no pattern can be stored. We have

$$\left.\frac{\varepsilon_\nu}{(2r)^{1/2}}\right|_{\nu=\nu_c} = \tfrac{1}{2}\sqrt{\pi} \Rightarrow \sqrt{r}\,\nu_c^x = \left(\frac{2}{\pi}\right)^{1/2}. \tag{4.13}$$

In case (a) we then find

$$(1-2x)^{-1/2}\left(\frac{\nu_c}{N}\right)^x = \left(\frac{2}{\pi}\right)^{1/2} \tag{4.14}$$

so that ν_c is *extensive*,

$$\nu_c = [(1-2x)2/\pi]^{1/2x} N. \tag{4.15a}$$

In case (b) we simply obtain

$$\nu_c = (2/\pi) N/\ln N \tag{4.15b}$$

whereas in case (c), where $2x > 1$,

$$\nu_c = [(2x-1)2/\pi]^{1/2x} N^{1/2x}. \tag{4.15c}$$

That is, neither case (b) nor case (c) gives rise to an extensive ν_c.

One can optimise the prefactor $\alpha \equiv [(1-2x)2/\pi]^{1/2x}$ in (4.15a) by varying x. This gives $\alpha_{\max} = 0.103$ for $x = 0.280$. However, for this value of x one has to decrease ν to about $0.01\,N$ to get an error percentage less than 0.5%. Alternatively, one can fix the error percentage, say 0.5%, and look for the value of x that maximises the corresponding ν. This gives $\nu_{\max} = 0.013\ N$ for $x = 0.386$, which is not a great improvement.

The rationale of the above optimisation is simple. If $x = 0$, the memory is in a state of total confusion [9] and no pattern can be stored. On the other hand, if $x \geqslant \frac{1}{2}$ and $\nu \to \infty$, a pattern is too readily lost in the background noise and an extensive storage capacity is not possible either. The optimal x is between these two situations.

5. Discussion

For the weight $\varepsilon_\nu = \nu^{-x}$ a critical ν_c exists such that all correlations with the stored patterns are lost if the index ν exceeds ν_c: the patterns have disappeared in the background noise. For $\nu < \nu_c$ there is a definite correlation, which approaches zero *continuously* as ν tends to ν_c from below. The patterns are gradually faded out and there is a continuous transition at ν_c. If $0 < x < \frac{1}{2}$, then ν_c is an extensive quantity, proportional to the size of the system.

A qualitative understanding of these results, in particular of (4.15), is easy to obtain. At zero temperature a pattern ν should be a stable fixed point of the dynamics

$$S(i) := \operatorname{sgn}\left(\sum_j J_{ij} S(j)\right). \tag{5.1}$$

The J_{ij} are given by (2.1) and, thus,

$$J_{ij} \propto \sum_n \varepsilon_n \xi_{in} \xi_{jn}. \tag{5.2}$$

Hence $(N - 1 \approx N)$

$$\sum_j J_{ij} \xi_{j\nu} \propto \sum_{n,j} \varepsilon_n \xi_{in} \xi_{jn} \xi_{j\nu} = \varepsilon_\nu \xi_{i\nu} N + \sum_{\substack{\mu \neq \nu \\ j}} \varepsilon_\nu \xi_{i\mu} \xi_{j\mu} \xi_{j\nu}. \tag{5.3}$$

The sum over $\mu (\neq \nu)$ and $j (\neq i)$ is a sum over independent identically distributed random variables whose order of magnitude is given by the square root of its variance, $(N \Sigma_\mu \varepsilon_\mu^2)^{1/2}$. Then ν_c is determined by the condition that the two terms in (5.3) be of equal magnitude,

$$(\varepsilon_\nu N)^2 \approx N \sum_\mu \varepsilon_\mu^2. \tag{5.4}$$

At $\nu = \nu_c$ we then find

$$\varepsilon_\nu^2 \approx N^{-1} \sum_\mu \varepsilon_\mu^2 \tag{5.5}$$

where, by definition, $\varepsilon_\nu^2 = \nu^{-2x}$. This, combined with (4.8)-(4.10), reproduces (4.15)—apart from the (essential) $2/\pi$, which comes from the error function in (4.4). Without this factor $2/\pi$ the optimal value of the proportionality constant α would have been given by the maximum $e^{-1} = 0.368$ of the function $(1 - 2x)^{1/2x}$, which is attained at $x = 0^+$.

Summarising, we have presented a simple, straightforward, and careful derivation of the thermodynamics of the Hopfield model near saturation. In so doing we have confirmed the key results of Amit *et al* [9] and extended the model by introducing weights ε_ν. For suitable ε_ν, there is a gradual fading out of the patterns as ν proceeds (or time runs backwards). Further evidence is needed to decide whether forgetting is an inherent property of the memory, or a matter of time (or both).

Acknowledgments

J L van Hemmen gratefully acknowledges the hospitality of the Joint Institute for Nuclear Research in Dubna, where this work was done. He also thanks the Deutsche Forschungsgemeinschaft (Bonn) for financial support.

Appendix

We want to show that if $c(t)$ is a strictly convex function with Legendre transform $c^*(m)$ and $F(m)$ is smooth, then

$$\sup_m (F(m) - c^*(m)) \tag{A1}$$

may be written

$$\max_\mu (F(\mu) - \mu \cdot \nabla F(\mu) + c(\nabla F(\mu))) \tag{A2}$$

where μ satisfies the fixed point equation

$$\mu = \nabla c(\nabla F(\mu)). \tag{A3}$$

The proof is simple. We note that $t \to \nabla c(t)$ is a mapping of $\mathbb{R}^n$, say, into $\mathbb{R}^n$. Its inverse exists and equals [16] ∇c^*. Since the supremum in (A1) is realised among those m which satisfy the relation $\nabla F(m) = \nabla c^*(m)$, we immediately obtain the fixed point equation

$$m = \nabla c(\nabla F(m)). \tag{A4}$$

Its solutions are denoted by μ.

We now evaluate $c^*(\mu)$. By definition,

$$c^*(\mu) = \sup_t (\mu \cdot t - c(t)). \tag{A5}$$

To obtain the supremum we have to find a t so well behaved that

$$\nabla c(t) = \mu. \tag{A6}$$

However, μ satisfies (A4). By comparison we see that $t = \nabla F(\mu)$. If we substitute this into (A5) and return to (A1), then (A2) follows directly.

Note added. After this paper was completed in Dubna (August 1986) we learned that a parallel work has been performed by Mézard *et al* [17]. Since the intentions of these authors are rather different, as are the conclusions, there are in our opinion enough reasons to warrant separate publication of the present paper.

References

[1] McCulloch W S and Pitts W A 1943 *Bull. Math. Biophys.* **5** 115
[2] Hopfield J J 1982 *Proc. Natl. Acad. Sci. USA* **79** 2554; 1984 *Proc. Natl. Acad. Sci. USA* **81** 3088
[3] Little W A 1974 *Math. Biosci.* **19** 101
Little W A and Shaw G L 1978 *Math. Biosci.* **39** 281
[4] Peretto P 1984 *Biol. Cybern.* **50** 51
[5] Toulouse G, Dehaene S and Changeux J-P 1986 *Proc. Natl. Acad. Sci. USA* **83** 1695
[6] van Hemmen J L and Kühn R 1986 *Phys. Rev. Lett.* **57** 913

[7] van Enter A C D and van Hemmen J L 1984 *Phys. Rev.* A **29** 3555
[8] Hebb D 1949 *The Organization of Behavior* (New York: Wiley)
[9] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530; 1987 *Ann. Phys., NY* **173** 30
[10] Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617
[11] van Hemmen J L and Palmer R G 1979 *J. Phys. A: Math. Gen.* **12** 563
[12] Grensing D, Kühn R and van Hemmen J L 1987 *J. Phys. A: Math. Gen.* **20** 2935
[13] van Hemmen J L 1982 *Phys. Rev. Lett.* **49** 409
[14] van Hemmen J L, van Enter A C D and Canisius J 1983 *Z. Phys.* B **50** 311
[15] van Hemmen J L 1983 *Heidelberg Colloquium on Spin Glasses* ed J L van Hemmen and I Morgenstern (*Lecture Notes in Physics* **192**) (Berlin: Springer) pp 203-233
[16] Roberts A W and Varberg D E 1973 *Convex Functions* (New York: Academic)
[17] Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457