Capacity of neural networks with discrete synaptic couplings

H Gutfreund and Y Stein

The Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Abstract. We study the optimal storage capacity of neural networks with discrete local constraints on the synaptic couplings J_{ij}. Models with such constraints inlcude those with binary couplings $J_{ij} = \pm 1$ or $J_{ij} = 0, 1$, quantised couplings with larger synaptic range, e.g. $J_{ij} = \pm 1/L, \pm 2/L, \ldots, \pm 1$ and, in the limit, continuous couplings confined to the hypercube $|J_{ij}| \leq 1$ ('box confinement'). We find that the optimal storage capacity $\alpha(\kappa)$ is best determined by the vanishing of a suitably defined 'entropy' as calculated in the replica symmetric approximation. We also extend our results to cases with biased memories and make contact with sparse coding models.

1. Introduction

We study networks of N fully connected binary *neurons* $\{S_i\}_{i=1,\ldots,N}$, $S_i = \pm 1$, coupled by a matrix of *synaptic couplings* J_{ij}, having local thresholds Θ_i and obeying zero-temperature dynamics

$$S_i(t+1) = \text{sgn}\left(\sum_{j \neq i} J_{ij} S_j(t) - \Theta_i \sqrt{N}\right). \tag{1}$$

We are interested in the network's functioning as an associative memory, in which p random *memories* $\{\xi_i^\mu\}_{i=1,\ldots,n}^{\mu=1,\ldots,p}$, $\xi = \pm 1$, are stored as fixed points of the dynamics (1). We actually require

$$\xi_i^\mu\left(\sum_{j \neq i} J_{ij}\xi_j^\mu - \Theta_i\sqrt{N}\right) > \kappa\sqrt{N} \qquad \text{for all } \mu = 1, \ldots, p \text{ and } i = 1, \ldots, N \tag{2}$$

since, although the memories are fixed points of the dynamics even for $\kappa = 0$, positive κ is needed to ensure large basins of attraction. When $\kappa > 0$, its value is meaningful only when one specifies the normalisation of the coupling matrix J_{ij}, due to the possibility of an overall rescaling of the inequalities (2). One commonly used normalisation is the spherical normalisation:

$$\sum_{j \neq i} J_{ij}^2 = N \tag{3}$$

which is a *global* constraint on the rows of the coupling matrix. Provided that solutions to (2) subject to (3) exist, one can be found, for example, by applying the 'perceptron algorithm' (Rosenblatt 1962, Minsky and Papert 1969, Gardner 1988, Diederich and Opper 1987).

A significant contribution to this problem was achieved by Gardner (1988) who showed that the probability of existence of solutions can be deduced from the fractional volume in the phase space of the parameters J_{ij} within which equations (2) and (3)

are satisfied. Gardner's work effectively decouples the question of the existence of solutions and the theoretical storage capacity $\alpha_c = p/N$, from the problem of actually producing such solutions using a specific learning algorithm.

It is of great interest to study models in which the global constraint (3) is replaced by *local* constraints on individual J_{ij}. One important class of models of this nature is distinguished by J_{ij} which are only allowed to assume a discrete set of values. In other cases the J_{ij} values can be chosen from continuous intervals, such as for the 'box confinement' $|J_{ij}| \leq 1$, interval constraints of the type $0 < c \leq |J_{ij}| \leq 1$, and constraints which impose *a priori* probability distributions on the J_{ij}.

The study of neural networks with local constraints on the coupling strengths is well motivated from both biological and applications points of view. It is very implausible to assume a biological mechanism which preserves infinite precision of truly continuous J_{ij} and it is therefore interesting to study the effect of some coarse graining of synaptic efficacies, for example, by encoding the information using a finite number of discrete values. Likewise, in hardware implementations it may prove simpler to realise networks wherein the couplings are restricted to discrete values such as $J_{ij} = 0, 1$ (connected or not), $J_{ij} = \pm 1$ (direct or inverted), or more generally J_{ij} restricted to digital values.

In the present paper we discuss a class of models for which the J_{ij} are restricted to a discrete set of values. Networks with $J_{ij} = \pm 1$ and $J_{ij} = 0, 1$ have been studied previously in the context of models with specified dependence of the couplings on the stored memories, namely the clipped Hebb rule (Hopfield 1982, Sompolinsky 1986) and the Willshaw model (Willshaw *et al* 1969, Golomb *et al* 1990). In the context of studies of optimal storage capacity, the Ising interaction case $J_{ij} = \pm 1$ was considered by Gardner and Derrida (1988) in the replica symmetric approximation, who found for $\kappa = 0$ a critical storage capacity $\alpha_c = 4/\pi$. This result exceeds the information theoretic bound of $\alpha_c = 1$, indicating, as they explicitly show, that the replica symmetry must be broken. This adds an additional motivation to study the Ising interaction case, as a problem of basic interest for the understanding of the replica method. With mainly this goal in mind, this problem has been investigated recently by Krauth and Mézard (1989). They found a one-step replica symmetry breaking solution which gives $\alpha_c = 0.83$, precisely the value at which the replica symmetric entropy vanishes. In addition this value is in good agreement with numerical evidence (Gardner and Derrida 1989, Krauth and Opper 1989).

In section 2 we calculate the expectation of the logarithm of the number of solutions to the inequalities (2) subject to general local constraints and derive the replica symmetric saddle-point equations. In section 3 we define three lines of interest in the κ against α plane: (*a*) the GD (Gardner-Derrida) line, which defines an $\alpha_c(\kappa)$ similar to that of Gardner and Derrida (1988); (*b*) the AT (de Almeida-Thouless) line below which the replica symmetric solution is stable (de Almeida and Thouless 1978); (*c*) the ZE (zero-entropy) line, on which the replica symmetric entropy vanishes. In sections 4 and 5 we discuss specific cases. Finally, in section 6 we consider biased memories, with particular emphasis on extreme bias.

Let us first summarise our basic results.

(*a*) We extend Gardner's formalism (Gardner 1988, Gardner and Derrida 1988) to the general case of local discrete constraints, including the effects of biased memories.

(*b*) We find that for all cases considered here the GD line is unstable with respect to replica symmetry breaking (RSB), while the ZE line is stable. Thus the latter gives our best estimate of the optimal storage capacity.

(*c*) We calculate the optimal storage capacity and connectivity for the case $J_{ij} = 0$, 1, and verify the results by simulations.

(*d*) We determine optimal capacities for the multivalued discrete cases such as $J_{ij} = \pm 1/L, \pm 2/L, \ldots, \pm 1$, and find the $\kappa = 0$ values to be in consonance with a simple estimate when $L \geq 3$.

(*e*) We contrast the optimal storage capacity for the case $|J_{ij}| \leq 1$ ('box confinement constraint') which is a limiting case of discrete couplings, with that of the spherical constraint.

(*f*) We verify the theoretical predictions for the box confinement case by performing simulations based on the simplex linear programming algorithm.

(*g*) We extrapolate the information capacity of the binary-valued cases to the extremely biased case, thereby making contact with sparse coding models.

A short version of our results has been previously presented (Gutfreund and Stein 1989).

2. Replica symmetric theory

In the first subsection we will assume the stored memories to be unbiased $\langle \xi_i^\mu \rangle = 0$ and uncorrelated $\langle \xi_i^\mu \xi_j^\mu \rangle = 0$ and take the local thresholds to be zero, $\Theta_i = 0$. In the following subsection we will lift these restrictions.

2.1. Unbiased memories

The calculation commences with the observation that given the set of memories $\{\xi_i^\mu\}_{i=1,\ldots,N}^{\mu=1,\ldots,p}$, the function

$$I(J_{ij}, \xi_i^\mu) = \prod_\mu \theta\left(\xi_i^\mu \sum_{j \neq i} J_{ij}\xi_j^\mu - \kappa\sqrt{N} \right) \tag{4}$$

serves as an indicator function, i.e. equals one if J_{ij} obeys (2) and is zero otherwise. Thus, for the spherical normalisation (3), the fractional volume in the coupling space of the properly normalised J_{ij} matrices which obey (2) is given by

$$V = \frac{\Pi_i \int dJ_i \, \delta(\Sigma_j J_{ij}^2 - N) I(J_{ij}, \xi_i^\mu)}{\Pi_i \int dJ_i \, \delta(\Sigma_j J_{ij}^2 - N)} \tag{5}$$

where $dJ_i = \Pi_{j \neq i} dJ_{ij}$. In the absence of any restriction on the correlation between J_{ij} and J_{ji}, the neurons i are decoupled and the fractional volume can be calculated for each site separately. As we are interested in the typical case for all possible realisations of the memories, we should average $\ln V$ over the probability distribution of ξ_i^μ. This is the basis of Gardner's approach (Gardner 1988) to the calculation of the storage capacity. In the case of discrete J_{ij}, restricted to a finite set of values, the number of solutions Ω to (2) and (3) is finite. This number of solutions Ω replaces the fractional volume V and is given by

$$\Omega = \operatorname*{Tr}_{J_{ij}} I(J_{ij}, \xi_i^\mu) \tag{6}$$

where the trace stands for summation over all allowed discrete values of J_{ij}.

The typical value of Ω is given by $\exp\{\langle\ln\Omega\rangle\}$, and performing the average of $\ln\Omega$ over the different sets of memories requires the use of the 'replica trick':

$$\langle\ln\Omega\rangle = \lim_{n\to 0}\frac{\langle\Omega^n\rangle - 1}{n} \tag{7}$$

whereby averaging of the logarithm is replaced by averaging of the power. This latter is accomplished by introducing an ensemble of n identical replicas of the system in terms of which

$$\langle\Omega^n\rangle = \left\langle\prod_\alpha\mathop{\mathrm{Tr}}_{J_{ij}^\alpha}\prod_\mu\theta\left(\xi_i^\mu\sum_{j\neq i}J_{ij}^\alpha\xi_j^\mu - \kappa\sqrt{N}\right)\right\rangle_{\xi_i^\mu} \tag{8}$$

where α is the replica index.

We next introduce two pairs of conjugate order parameters. The *overlap* between two solutions labelled α and β is represented by the order parameter

$$q^{\alpha\beta} = \frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta \qquad -1\leq q^{\alpha\beta}\leq 1 \tag{9}$$

while the *normalisation* (self-overlap) of a solution α is specified by a second order parameter

$$Q^\alpha = \frac{1}{N}\sum_j J_{ij}^{\alpha 2} \qquad 0\leq Q^\alpha\leq 1. \tag{10}$$

This latter is absent in the treatment of the spherical normalisation since it equals one identically. The role of the conjugate order parameters $\hat{q}^{\alpha\beta}$ and $\hat{Q}^\alpha$ is to enforce (9) and (10).

The, by now standard, procedure (Gardner 1988) gives

$$\langle\Omega^n\rangle = \int\int\int\int\frac{\mathrm{d}Q^\alpha\,\mathrm{d}\hat{Q}^\alpha\,\mathrm{d}q^{\alpha\beta}\,\mathrm{d}\hat{q}^{\alpha\beta}}{(2\pi)^2}\exp[NG(Q^\alpha,\hat{Q}^\alpha,q^{\alpha\beta},\hat{q}^{\alpha\beta})] \tag{11}$$

$$G = \alpha G_1 + G_2 + \mathrm{i}\sum_\alpha\hat{Q}^\alpha Q^\alpha + \mathrm{i}\sum_{\alpha<\beta}\hat{q}^{\alpha\beta}q^{\alpha\beta} \tag{12}$$

where

$$G_1 = \ln\int_\kappa^\infty\prod_\alpha\frac{\mathrm{d}\lambda^\alpha}{\sqrt{2\pi}}\int_{-\infty}^\infty\prod_\alpha\frac{\mathrm{d}x^\alpha}{\sqrt{2\pi}}\exp\left(\mathrm{i}\sum_\alpha x^\alpha\lambda^\alpha - \tfrac{1}{2}\sum_\alpha Q^\alpha(x^\alpha)^2 - \sum_{\alpha<\beta}q^{\alpha\beta}x^\alpha x^\beta\right)$$

$$G_2 = \ln\mathop{\mathrm{Tr}}_{J_{ij}^\alpha}\exp\left(-\mathrm{i}\sum_\alpha\hat{Q}^\alpha(J_{ij}^\alpha)^2 - \mathrm{i}\sum_{\alpha<\beta}\hat{q}^{\alpha\beta}J_{ij}^\alpha J_{ij}^\beta\right). \tag{13}$$

We now adopt the replica symmetric ansatz:

$$q^{\alpha<\beta} = q \qquad q^{\alpha\alpha} = 0$$
$$Q^\alpha = Q$$
$$\hat{q}^{\alpha<\beta} = \hat{q} \qquad \hat{q}^{\alpha\alpha} = 0 \tag{14}$$
$$\hat{Q}^\alpha = \hat{Q}$$

and further perform a change of variables to

$$q_0 \equiv Q - q$$
$$F_1 \equiv -\mathrm{i}\hat{q} \tag{15}$$
$$F_2 \equiv \tfrac{1}{2}F_1 + \mathrm{i}\hat{Q}$$

with Q retaining its former status.

When N increases to the thermodynamic limit we can perform the integrals by saddle-point integration:

$$\langle \Omega^n \rangle = \int \int \int \int \ldots \exp[Nng(Q, q_0, F_1, F_2)] = \exp\left[Nn\left(\underset{Q,q_0,F_1F_2}{\text{extr}}\, g + \mathrm{O}(1/N)\right)\right]f \qquad (16)$$

$$g = G/n = \alpha g_1 + g_2 - F_1 q_0/2 + F_2 Q. \qquad (17)$$

The integrations are carried out over all four order parameters, and

$$g_1 = \int_{-\infty}^{\infty} \mathrm{D}t \, \ln H(A(t))$$

$$g_2 = \int_{-\infty}^{\infty} \mathrm{D}u \, \ln \underset{J}{\mathrm{Tr}} \exp(u\sqrt{F_1}\, J - F_2 J^2) \qquad (18)$$

where

$$\mathrm{D}x = \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$H(x) = \int_x^{\infty} \mathrm{D}z$$

$$A(t) = \frac{\kappa + (Q - q_0)^{1/2} t}{\sqrt{q_0}} \qquad (19)$$

$$B(t) = \kappa + \frac{Qt}{(Q - q_0)^{1/2}}$$

$$C(t) = \frac{\exp(-A^2(t)/2)}{H(A(t))}$$

the last two variables being defined for the following. The saddle-point equations are obtained by differentiating (17) with respect to all four order parameters, resulting in

$$F_1 = \frac{\alpha}{\sqrt{2\pi}\, q_0^{3/2}} \int_{-\infty}^{\infty} \mathrm{D}t \, B(t) C(t) \qquad (20)$$

$$F_2 = \frac{\alpha}{2\sqrt{2\pi}[q_0(Q - q_0)]^{1/2}} \int_{-\infty}^{\infty} \mathrm{D}t \, t \, C(t) \qquad (21)$$

$$Q = \int_{-\infty}^{\infty} \mathrm{D}u \, \overline{J^2} \qquad (22)$$

$$q_0 = \frac{1}{\sqrt{F_1}} \int_{-\infty}^{\infty} \mathrm{D}u \, u \, \bar{J} \qquad (23)$$

where

$$\overline{J^k} \equiv \frac{\mathrm{Tr}_J J^k \exp(u\sqrt{F_1}\, J - F_2 J^2)}{\mathrm{Tr}_J \exp(u\sqrt{F_1}\, J - F_2 J^2)} \qquad k = 1, 2. \qquad (24)$$

This constitutes a set of self-consistent equations for any given κ and α which determine all variables needed to calculate the function G and thus the typical number of solutions $\exp(\langle \ln \Omega \rangle)$. The saddle-point value of G, which is the extremum of equation (12) with respect to all order parameters, will henceforth be called the '*entropy*'. This is consistent with the usual definition in statistical physics where the logarithm of the number of states typically available to a system is proportional to the entropy.

2.2. Biased memories

We now consider the effect of allowing biased memories and a non-zero threshold. The bias is introduced by the following non-trivial distribution for the memories:

$$p(\xi_i^\mu) = \frac{1+m}{2} \delta(\xi_i^\mu - 1) + \frac{1-m}{2} \delta(\xi_i^\mu + 1). \tag{25}$$

Equation (8) is supplemented by a threshold term and the average is now interpreted to be over this distribution. Furthermore, we augment our set of order parameters (defined in (9) and (10)) with the *ferromagnetic bias*

$$M^\alpha = \frac{1}{\sqrt{N}} \sum_j J_{ij}^\alpha \tag{26}$$

and its conjugate variable $\hat{M}$, which turns out to be of negligible importance in the thermodynamic limit. The replica symmetric ansatz (14) is extended to include $M^\alpha = M$ and M can be replaced by an equivalent order parameter

$$v \equiv M - \Theta/m \tag{27}$$

thus eliminating explicit reference to the threshold.
    We thus obtain

$$\langle \Omega^n \rangle = \int \int \int \int \int \ldots \exp[Nng(v, Q, q_0, F_1, F_2)] \tag{28}$$

$$= \exp[Nn(\operatorname*{extr}_{vQq_0F_1F_2} g + O(1/N))]$$

$$g = \alpha g_1 + g_2 - \tfrac{1}{2} F_1 q_0 + F_2 Q \tag{29}$$

where

$$g_1 = \frac{1+m}{2} \int_{-\infty}^{\infty} Dt \ln H(A_+(t)) + \frac{1-m}{2} \int_{-\infty}^{\infty} Dt \ln H(A_-(t))$$

$$g_2 = \int_{-\infty}^{\infty} Du \ln \operatorname*{Tr}_J \exp(u\sqrt{F_1} J - F_2 J^2) \tag{30}$$

and defining

$$A_\pm(t) = \frac{\kappa \mp mv + [(Q - q_0)(1 - m^2)]^{1/2} t}{[q_0(1 - m^2)]^{1/2}}$$

$$B_\pm(t) = \frac{\kappa \mp mv}{[q_0(1 - m^2)]^{1/2}} + \frac{Qt}{(Q - q_0)^{1/2}} \tag{31}$$

$$C_\pm(t) = \frac{\exp(-A_\pm^2(t)/2)}{H(A_\pm(t))}$$

the saddle-point equations are

$$\frac{1+m}{2} \int_{-\infty}^{\infty} Dt \, C_+(t) = \frac{1-m}{2} \int_{-\infty}^{\infty} Dt \, C_-(t) \tag{32}$$

$$F_1 = \frac{\alpha}{\sqrt{2\pi} \, q_0^{3/2}} \left(\frac{1+m}{2} \int_{-\infty}^{\infty} Dt \, B_+(t) C_+(t) + \frac{1-m}{2} \int_{-\infty}^{\infty} Dt \, B_-(t) C_-(t) \right) \tag{33}$$

$$F_2 = \frac{\alpha}{2\sqrt{2\pi} \, [q_0(Q - q_0)]^{1/2}} \left(\frac{1+m}{2} \int_{-\infty}^{\infty} Dt \, t C_+(t) + \frac{1-m}{2} \int_{-\infty}^{\infty} Dt \, t C_-(t) \right). \tag{34}$$

These equations, along with (22) and (23), constitute a set of self-consistent equations for any given κ, α and m which determine all variables needed to calculate the entropy and thus the typical number of solutions $\exp(\langle \ln \Omega \rangle)$.

3. The three lines of interest

3.1. The GD line

For the spherical normalisation constraint studied by Gardner (equations (2) and (3)) the subspace of solutions is continuous and convex and is embedded in a continuous and unbounded surroundings of candidate matrices. When, for fixed κ, α is increased from zero towards its critical value, this subspace shrinks until it contains only a single point in the limit $q \to 1$. For the case of local constraints this corresponds to $q \to Q$ or equivalently $q_0 \to 0$; we are, however, no longer confident that at the end of this shrinking process, we remain with a matrix which satisfies the local constraints. Thus, while we expect this process to provide the actual capacity for the spherical normalisation constraint, it will in general supply an upper bound, which we call the GD line $\alpha_{(\kappa)}^{\text{GD}}$. This line is the upper boundary of meaningful phase space.

For unbiased memories, in the limit $q_0 \to 0$, the expression for g_1 in (17) simplifies to

$$g_1 = -\frac{1}{2q_0} \int_{-\kappa/\sqrt{Q}}^{\infty} Dt(\kappa + \sqrt{Q}t)^2 \tag{35}$$

and the saddle-point equations can be expressed as

$$F_1 = \frac{\alpha Q}{q_0^2} \int_{-\kappa/\sqrt{Q}}^{\infty} Dt \left(t + \frac{\kappa}{\sqrt{Q}} \right)^2 \tag{36}$$

$$F_2 = \frac{\alpha}{2q_0} \int_{-\kappa/\sqrt{Q}}^{\infty} Dt \, t \left(t + \frac{\kappa}{\sqrt{Q}} \right) \tag{37}$$

where, once again, Q and q_0 are as in (22) and (23) respectively.

For biased memories

$$g_1 = -\frac{1}{2} \left(\frac{1+m}{2} \int_{-\tau_+}^{\infty} Dt \, A_+^2(t) + \frac{1-m}{2} \int_{-\tau_-}^{\infty} Dt \, A_-^2(t) \right) \tag{38}$$

and the saddle-point equations can be expressed as

$$\frac{1+m}{2} \int_{\tau_+}^{\infty} Dt(t - \tau_+) = \frac{1-m}{2} \int_{\tau_-}^{\infty} Dt(t - \tau_-) \tag{39}$$

$$F_1 = \frac{\alpha Q}{q_0^2} \left(\frac{1+m}{2} \int_{\tau_+}^{\infty} Dt(t - \tau_+)^2 + \frac{1-m}{2} \int_{\tau_-}^{\infty} Dt(t - \tau_-)^2 \right) \tag{40}$$

$$F_2 = \frac{\alpha}{2q_0} \left(\frac{1+m}{2} \int_{\tau_+}^{\infty} Dt \, t(t - \tau_+) + \frac{1-m}{2} \int_{\tau_-}^{\infty} Dt \, t(t - \tau_-) \right) \tag{41}$$

where

$$\tau_\pm \equiv \frac{-\kappa \mp mv}{[Q(1-m^2)]^{1/2}}$$

and, again, the equations for Q and q_0 are unchanged.

These equations can be solved numerically using an iterative procedure (equation (39) being viewed as a non-linear function of v, whose root is being sought), and then repeating with decreasing values of q_0 allows extrapolation $q_0 \to 0$, thereby determining the storage capacity α_c. If we are interested only in α_c a simpler procedure is available, which we will demonstrate for the unbiased case. Since $\sqrt{F_1}$ and F_2 are proportional to q_0^{-1}, the traces in (24) are dominated by a single term when $q_0 \to 0$. This term contains J_{opt}, the allowed value of J closest to

$$\mathscr{J} = \frac{\sqrt{F_1}}{2F_2} u. \tag{42}$$

As long as $J_{opt} \neq 0$, the traces in both numerator and denominator of the equations for Q and q_0 can be approximated by a single term, and the integrands reduce to J_{opt}^2 and uJ_{opt}, respectively. For the special but important case $\kappa = 0$ the saddle-point equations simplify even further, because then

$$\mathscr{J} = \left(\frac{2Q}{\alpha}\right)^{1/2} u. \tag{43}$$

3.2. The AT line

Formally the GD line is unsatisfactory because it is not stable with respect to replica symmetry breaking. We are thus lead to study the AT line, below which the replica symmetric solution is stable with respect to local variations of $q^{\alpha\beta}$, Q^α and their conjugate order parameters. Being the limit of replica symmetry stability we might expect the AT line $\alpha^{AT}(\kappa)$ to reliably predict the true capacity.

The onset of RSB is signalled by a change of sign in at least one of the eigenvalues of the matrix of second derivatives of G with respect to all the order parameters. This matrix can be represented schematically, in view of equation (12), in block form

$$\begin{pmatrix} \alpha\partial^2 G_1 & i \\ i & \partial^2 G_2 \end{pmatrix}.$$

For the unbiased case, the upper left block contains derivatives with respect to $q^{\alpha\beta}$ and Q^α, while the lower right contains those with respect to $\hat{q}^{\alpha\beta}$ and $\hat{Q}^\alpha$. Requiring all of the eigenvalues (and thus the determinant) to be positive leads to the replica symmetry stability criterion (Gardner 1988, Gardner and Derrida 1988)

$$\alpha\gamma_1\gamma_2 < 1 \tag{44}$$

where γ_1 and γ_2 are the transverse eigenvalues of $\partial^2 G_1$ and $\partial^2 G_2$, respectively:

$$\gamma_1 = \int_{-\infty}^{\infty} Dt (\overline{x^2} - \bar{x}^2)^2$$
$$\gamma_2 = \int_{-\infty}^{\infty} Dt (\overline{J^2} - \bar{J}^2)^2 \tag{45}$$

where $\overline{J^k}$ is given by (24) and

$$\overline{x^k} \equiv \frac{\int_\kappa^\infty d\lambda \int_{-\infty}^\infty dx \exp[-w(\tilde{\lambda}, x)]x^k}{\int_\kappa^\infty d\lambda \int_{-\infty}^\infty dx \exp[-w(\tilde{\lambda}, x)]} \tag{46}$$

where

$$w(\tilde{\lambda}, x) = \frac{\tilde{\lambda}^2 + (q_0 x - i\tilde{\lambda})^2}{2q_0}$$

and $\tilde{\lambda} = \lambda + \sqrt{q}t$.

For the biased case, $\partial^2 G_2$ remains unchanged, while $\partial^2 G_1$ involves the additional order parameter M^α. Using the same techniques as in de Almeida and Thouless (1978) and then proceeding as above (Gardner 1988, Gardner and Derrida 1988) we find that the stability criterion has the same form as (44), but now γ_1 is given by

$$\gamma_1 = (1 - m^2) \left\langle \left(\int_{-\infty}^{\infty} Dt(\overline{x^2} - \bar{x}^2)^2 \right) \right\rangle_\xi \tag{47}$$

where $\overline{x^k}$ has the same form as (46), except that

$$w(\tilde{\lambda}, x) = \frac{(1 - m^2)^{1/2}\tilde{\lambda}^2 + (\tilde{q}_0 x - i\tilde{\lambda})^2}{2\tilde{q}_0}$$

with $\tilde{\lambda} = \lambda - mv\xi + [q(1 - m^2)]^{1/2}t$ and $\tilde{q}_0 = q_0(1 - m^2)$.

The α at which (44) breaks down defines $\alpha^{AT}(\kappa, m)$. For all the discrete cases discussed in this paper $\alpha^{AT}(\kappa, m) < \alpha^{GD}(\kappa, m)$.

3.3. The ZE line

In the previous subsection we presented an argument for accepting the AT line as an estimate of the actual capacity. However, for the cases studied here the entropy is negative on the AT line. From the very definition of $G = \ln \Omega$, positive entropy indicates that the number of J_{ij} that obey the constraints is exponential in N, while negative entropy implies that there will be no valid solutions in the thermodynamic limit. Thus we expect the value of α for which the entropy becomes zero to determine the storage capacity. This criterion defines the ZE line $\alpha^{ZE}(\kappa)$. We find that for all the discrete constraints studied here

$$\alpha^{ZE}(\kappa) < \alpha^{AT}(\kappa) < \alpha^{GD}(\kappa) \tag{48}$$

which, in particular, means that the solutions on the ZE line are stable with respect to RSB. Thus, $\alpha^{ZE}(\kappa)$ is a consistent estimate of the storage capacity; one should, however, keep in mind that due to a possible first-order transition, the true solution may actually be somewhere else in replica space (Krauth and Mézard 1989).

4. Specific unbiased cases

4.1. The $J_{ij} \pm 1$ case

This case, frequently referred to as the 'Ising' interaction, was considered by Gardner and Derrida (1988). The saddle-point equations simplify considerably in this case, since J^2 and Q are replaced by unity. In particular the GD line can be found analytically, since from (36) and (23) one obtains

$$\left(\alpha \int_{-\kappa}^{\infty} Dt(t + \kappa)^2 \right)^{1/2} = \int_{-\infty}^{\infty} Du\, u \tanh(u\sqrt{F_1}).$$

In the small q_0 limit the tanh becomes the sign function and the right-hand side integral gives $(2/\pi)^{1/2}$. After slight reshuffling we find

$$\alpha^{GD}(\kappa) = \frac{2}{\pi}\,\alpha^S(\kappa) \tag{49}$$

(where $\alpha^S(\kappa)$ is the optimal storage capacity obtained by Gardner for the spherical normalisation case) which is a result of Gardner and Derrida (1988).

To check the stability of the GD line with respect to RSB, we note that, as $q_0 \to 0$, γ_1 simplifies to

$$\gamma_1 = \frac{1}{q_0^2} \int_{-\kappa/\sqrt{Q}}^{\infty} Dt. \tag{50}$$

This is true in general, while for the Ising case γ_2 simplifies as well, one finds

$$\gamma_2 = \int_{-\infty}^{\infty} Dt[1 - \tanh^2(\sqrt{F_1}\,t)]^2.$$

Since γ_1 is proportional to q_0^{-2} while γ_2 behaves as q_0, the left-hand side of (44) diverges on the GD line. Hence, the AT line must lie below the GD line. Gardner and Derrida noted that the entropy on the Ising GD line is negatively infinite—implying that at this storage level, there are no solutions at all rather than a unique solution.

We present the three lines of interest for the Ising case as compared with Gardner's spherical constraint line in figure 1. We note that the GD line is indeed proportional to Gardner's line as implied by (49). This turns out to be the case due to the fact that the spherical constraint (3) is trivially obeyed for the Ising case—although it is not for any other discrete case. In addition we see, as indicated above, that the GD line is unstable with respect to RSB while the ZE line, lying beneath the AT line, is stable. In fact, we will show that relations (48) and (49) turn out to be correct for memories with bias m as well.
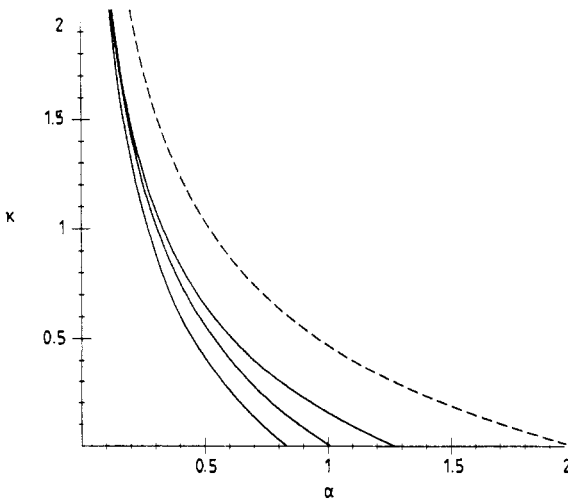
Figure 1. Phase diagram for the Ising case. The continuous curves, from top to bottom, are the GD, AT and ZE curves, respectively. The broken curve gives, for comparison, the storage capacity for the spherical constraint.

The AT line touches the axis at approximately $\alpha = 1.01$ while the ZE line gives $\alpha^{ZE}(\kappa = 0) = 0.832$. Krauth and Mézard (1989) find that, with one-step replica symmetry breaking, α_c is determined precisely by the vanishing of the replica symmetric entropy. In addition, they point out that if one terminates the aforementioned shrinking process for the spherical constraint in such a fashion that one corner of the hypercube must still be included, then $\alpha_c(\kappa = 0) \approx 0.85$, which is just slightly higher than the value of $\alpha^{ZE}(\kappa = 0)$.

4.2. The $J_{ij} = 0, 1$ case

One can similarly handle the binary-valued case $J_{ij} = 0, 1$, breaking the positive–negative symmetry of allowed couplings. This seems to be the simplest case for hardware implementation and it characterises the 'Willshaw' model (Willshaw *et al* 1969, Golomb *et al* 1990) wherein, given the p memories V_i^μ ($V_i^\mu = 0, 1$), a specific form of the synaptic couplings

$$J_{ij} = \theta\left(\sum_\mu V_i^\mu V_j^\mu\right) \tag{51}$$

is assumed. Simply stated, if neurons i and j are active in at least one stored pattern, then the Willshaw model takes $J_{ij} = 1$. These couplings are appropriate for the sparse coding problem, when the fraction of active neurons is of the order $\ln(N)/N$. When the activity is extensive the coupling matrix is saturated after storing a finite number of memories.

In the Willshaw model, however, the neural activity is taken to be $V_i = 0, 1$ as well. This $(0, 1)$ representation of neural activity is well known to be equivalent to our $(-1, +1)$ representation, to within a local threshold. It is indeed possible to modify our basic inequalities (2) to get

$$(2V_i^\mu - 1)\left(\sum_j J_{ij} V_j^\mu - \Theta\sqrt{N}\right) > \kappa\sqrt{N} \tag{52}$$

and one sees that allowing thresholds, the results for capacity are completely independent of which representation is used.

We find that it is possible to store an extensive number of memories with extensive activity in a network with binary couplings $J_{ij} = 0, 1$. The saddle-point equations in the $q_0 \to 0$ limit reduce to

$$Q = \int_{\sqrt{\alpha/8Q}}^\infty Du = \frac{2}{\alpha}\left(\int_{\sqrt{\alpha/8Q}}^\infty Du\, u\right)^2. \tag{53}$$

From this one may deduce that $Q \to \frac{1}{2}$ when $\alpha \to 0$ and that the GD prediction is $\alpha^{GD}(\kappa = 0) = 0.81$ and $Q = 0.27$. The zero-entropy calculation gives in this case a storage capacity $\alpha^{ZE}(\kappa = 0) = 0.59$ and self-overlap (which for $J_{ij} = 0, 1$ bonds means the fraction of bonds present) $Q = 0.32$. Due to the separability of the post-synaptic neurons (as mentioned after equation (5) above), all our results are equally valid for a single layer perceptron with one binary ± 1 output unit. Thus we deduce that one can implement an association of $p = 0.59N$ randomly chosen patterns (where N is the number of input units) with a single binary ± 1 output by connecting only 32% of the input units to the output unit. This implies that about one third of the elementary units serve to classify a set of patterns stored in such a perceptron.

We have verified this result by simulations on small systems (up to $N = 14$) by performing exhaustive search. We start with a randomly chosen initial pattern and find all possible vectors J_j which satisfy the required inequalities. We then chose a second pattern at random (only requiring that it be different from the first), and determine which of the J_j vectors found in the previous step stabilise this new pattern as well. We thus continue adding new patterns until no J_j vectors remain, at which point we have found the number of random patterns that can be stored. We then update the statistics of the number of non-zero bonds in the vectors remaining in the final step, in order to calculate the typical Q. This process is repeated for many sets of randomly chosen patterns and for various sizes of N (we found 20 000 different pattern sets for $N = 4, 6, 8$ and 10 000 for $N = 10, 14$ to give sufficiently accurate statistics). Our results are presented in figure 2 where average α_c and typical Q are plotted against $1/N$. The broken curve is the best quadratic fit to the experimental points. We find that extrapolation to $N \to \infty$ gives $\alpha_c = 0.55 \pm 0.04$ and $Q = 0.32 \pm 0.03$.
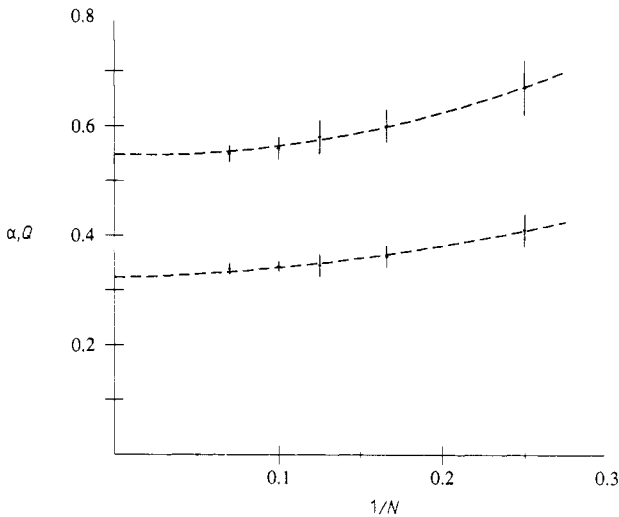
Figure 2. Simulation results for the $J_{ij} = 0, 1$ case. The upper points are the empirical α values, while the lower points represent Q values. The broken curves are best quadratic fits.

4.3. The digital case

We now treat several discrete levels of synaptic strengths (the synaptic depth will be denoted by L in the following), which, in order to facilitate comparison, will be taken as follows:

$$J_{ij} = \pm 1/L, \pm 2/L, \ldots, \pm 1. \tag{54}$$

For all finite L, the GD line lies above the AT line (and is thus unstable and does not provide a good estimate for the capacity) which in turn lies above the ZE line (which is thus stable and should provide a good estimate). We plot the ZE lines for various depths in figure 3.

One can estimate the $\kappa = 0$ storage capacity by the following approximate argument. Assume that, for given α, J_{ij} is a solution obeying the spherical constraint corresponding
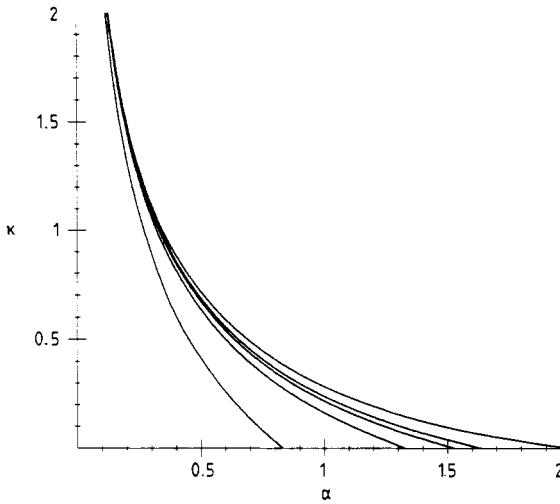
Figure 3. ZE lines for the digital case. The curves from bottom to top are for $L = 1, 2, 3, 4, \infty$.

to the maximal κ for this α. Then

$$\xi_i^\mu \sum_{j \neq i} J_{ij} \xi_j^\mu > \kappa_c^S(\alpha)\sqrt{N} \qquad \text{for all } \mu = 1, \dots, p \qquad (55)$$

where the function $\kappa_c^S(\alpha)$ is the inverse function of $\alpha_c^S(\kappa)$. We can rescale this inequality by L after which most of the rescaled J_{ij} values will be in the interval $[-L, \dots, L]$ and thus can be represented by

$$L J_{ij} = \overline{J_{ij}} + \Delta_{ij} \qquad (56)$$

where $\overline{J_{ij}}$ are integers and $|\Delta_{ij}| \leq \frac{1}{2}$. The inequalities (55) become

$$\xi_i^\mu \sum_{j \neq i} \overline{J_{ij}} \xi_j^\mu > L\sqrt{N} \, \kappa_c^S(\alpha) - \xi_i^\mu \sum_{j \neq i} \Delta_{ij} \xi_j^\mu. \qquad (57)$$

Estimating the second term on the right-hand side by $\frac{1}{2}\sqrt{N}$ one finds that the critical α for which the left-hand side is positive is determined by

$$\kappa_c^S(\alpha) = 1/2L. \qquad (58)$$

The α thus determined is an estimate for the $\kappa = 0$ storage capacity of the discrete model with synaptic depth L. In figure 4 we plot for $L = 1, \dots, 10$ this estimate along with the results of the GD (triangles) and ZE (squares) calculations described above. We see that the estimate is very close to the ZE result, at least for $L \geq 3$ and that for large L one approaches the limit $\alpha = 2$.

Another interesting variation on the discrete synapse theme is to allow for absent bonds as well, i.e. to permit the value $J_{ij} = 0$. One can repeat the calculations with equal facility for this case and the effects prove not to be drastic. In all cases the capacity as determined by the ZE line for $J_{ij} = 0, \pm 1/L, \dots, \pm 1$ is between the depth L and depth $L+1$ capacities. For example, with all bonds present, we obtain $\alpha^{ZE}(\kappa = 0) = 0.832, 1.331, 1.529$ for $L = 1, 2, 3$ respectively, while allowing for absent bonds we find storage capacities $\alpha^{ZE}(\kappa = 0) = 1.174$ and 1.477 for $L = 1, 2$.

Finally, one can extend the $J_{ij} = 0, 1$ binary case to $J_{ij} = 0, 1/L, \dots, 1$; for example, when $J_{ij} = 0, \frac{1}{2}, 1$ one finds $\alpha^{ZE}(\kappa = 0) = 0.74$. As L increases in this model, α approaches unity as expected for networks with continuous J_{ij} restricted to positive values (Amit *et al* 1989).
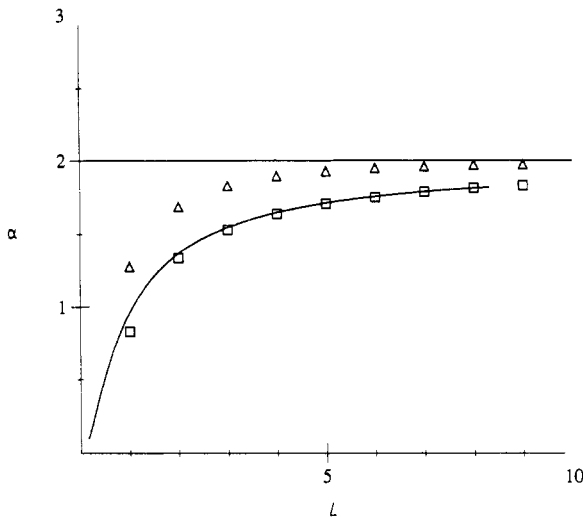
Figure 4. Storage capacity for the digital case as a function of synaptic depth. The full curve represents the estimate (see text), while the squares and triangles represent the ZE and GD results, respectively.
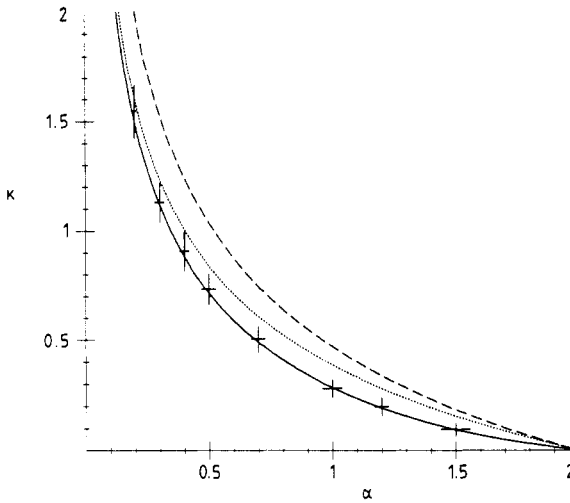
5. The box confinement constraint

If we allow the synaptic depth to grow indefinitely, we obtain the continuous 'box confinement' constraint

$$|J_{ij}| \leq 1. \tag{59}$$

As $L \to \infty$ the three lines of interest approach one another, and for the truly continuous case one can show, by substituting (42) into (24) for (59) and explicitly performing the integrations in (18), (22) and (23), that the entropy is zero on the GD line. One finds that this line lies well below the capacity for the spherical constraint $\alpha^S(\kappa)$, except at $\kappa = 0$, when both give $\alpha = 2$. To compare the two cases, let us, for a given α and κ, scale the couplings which satisfy (2) (with $\Theta_i = 0$) and (59), by the corresponding $1/\sqrt{Q}$. The matrix J'_{ij} thus obtained is normalised as in (3) and satisfies the inequalities (2) with $\kappa' = \kappa/\sqrt{Q}$. Nevertheless, the curve $\alpha^{ZE}(\kappa')$ is still below $\alpha^S(\kappa)$, due to the further restriction on the individual matrix elements $J_{ij} \leq 1/\sqrt{Q}$. This indicates that the optimal solutions obtained with the spherical constraint, for $\kappa > 0$, have large dynamic range, i.e. some matrix elements are very large while the majority remain small.

One of the advantages of the box confinement constraint (59) is that, in this case, we can cast the calculation of J_{ij} satisfying (2), with the largest possible κ, for a given set of memories, into a 'linear programming' (LP) problem, for which efficient methods of solution are known (e.g. the simplex method (Krauth and Mézard 1987)).

We have performed simulations using the simplex method for $N = 100$. Given many realisations of $p = \alpha N$ patterns we found in each case the optimal $\kappa(\alpha)$ and then averaged over the values of κ thus obtained. This procedure is repeated for various values of α and the average $\kappa(\alpha)$ is presented in figure 5 superposed on the box constraint ZE line. The agreement is seen to be quite good. What has been achieved here is just what was done by Krauth and Mezard (1987) using the 'minover' algorithm, who showed that $\alpha^S(\kappa)$ is indeed the optimal storage capacity for the spherical constraint.

Figure 5. Phase diagram for the box confinement case. The full curve is the theoretical capacity and the results from simplex simulations are the overlaid points. The broken curve is the storage capacity for the spherical constraint, while the dotted curve is the box confinement capacity modified so as to obey the spherical constraint (see text).

6. Biased memories and the limit $m \to 1$

We have shown in section 2.2 that the entire program outlined in the present paper can be extended to the treatment of biased memories. In particular, we have investigated the two binary cases of sections 4.1 and 4.2 and verified that for all m the inequalities (48) hold. We present in figure 6 the three lines of interest for the Ising interaction, taking $\kappa = 0$ and plotting the critical capacity as a function of the bias. For comparison we present the spherical constraint line as well.
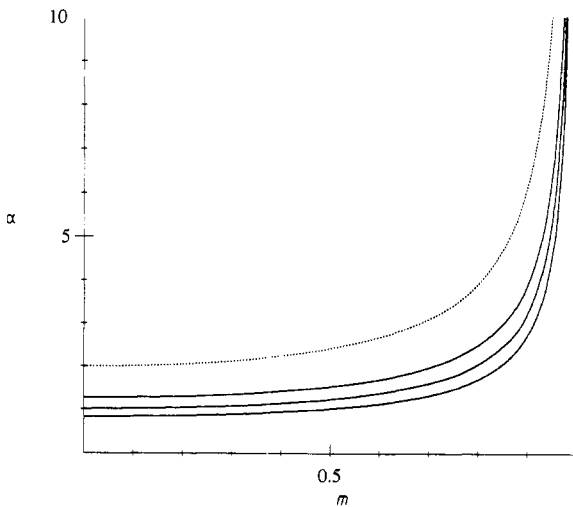
Figure 6. The $\alpha-m$ plane of the phase diagram for the biased Ising case. The full curves, from top to bottom, are the GD, AT and ZE curves, respectively. The dotted curve gives, for comparison, the storage capacity for the spherical constraint.

We now wish to discuss the limit $m \to 1$, which is of interest due to its relation with sparse coding networks, wherein only a small fraction of the neurons are active. The connection can be made since, in view of the $(+1, -1)$ symmetry our neurons, the limit $m \to 1$ is equivalent to the $m \to -1$ limit, for which indeed only a very small fraction of the neurons is active. For the $(0, 1)$ representation, mentioned before in connection with the Willshaw model and which is equivalent to our $(-1, +1)$ to within a local threshold, this is the limit of sparse coding $a \to 0$ (a being the fraction of active neurons).

The $m \to 1$ limit was discussed by Gardner (1988) who compared the information capacity per bond for optimal storage spherical normalisation networks to that achieved by the Willshaw model. The information capacity is the total number of bits stored in the memories divided by the number of bonds, and is thus given by

$$I = -\frac{\alpha_c(m)}{\ln 2}\left[\frac{1-m}{2}\ln\left(\frac{1-m}{2}\right) + \frac{1+m}{2}\ln\left(\frac{1+m}{2}\right)\right]. \tag{60}$$

For the spherical constraint, in the limit $m \to 1$:

$$\alpha_c^S(m) \to -[(1-m)\ln(1-m)]^{-1} \tag{61}$$

one obtains an optimal information capacity of $I(m \to 1) = 0.72$. The Willshaw model, with its specific couplings, gives the surprisingly good $I(m \to 1) = 0.69$. The information capacity of networks with sparsely coded memories has been recently discussed (Nadal 1989a, b, Nadal and Toulouse 1990, Perez Vicente 1989).

From equations (60) and (61) we find that $I(m)$, when $m \to 1$, has the form

$$I(m) = a_1 - \frac{a_2}{\ln(1-m)} \tag{62}$$

where a_1 and a_2 are constants. Assuming this form to be more general, we have plotted in figure 7 the information capacity for the two binary cases as a function of $\mu \equiv 1/\ln(1-m)$, for values of m approaching unity. The uppermost curve in the figure
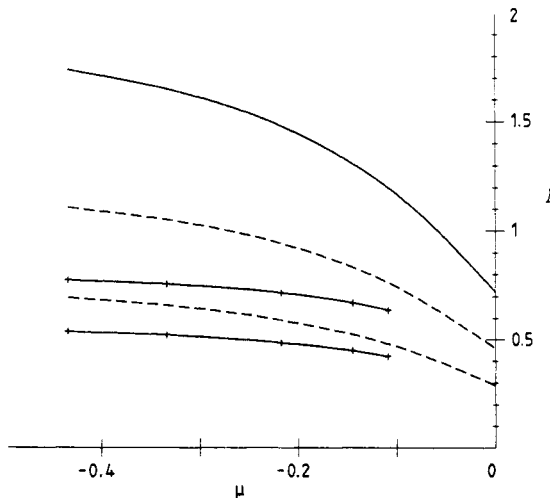

Figure 7. Information capacity for spherical and binary cases, in the limit $m \to 1$. The different curves are explained in the text.

represents the information capacity of the spherical model with the limiting value (at $m \to 1$ which corresponds to $\mu \to 0$) of 0.72. It is noteworthy to point out that, even at $\mu \approx -0.1$ (or $m \approx 0.9999$), this curve is still not a straight line, which indicates that the limiting form (61) is approached very slowly. The broken curves represent results based on GD line calculations, for the Ising interaction (upper) and the $J_{ij} = 0, 1$ case (lower). The remaining full curves are the ZE results for these two cases (in the same order). The numerical accuracy of our ZE calculation breaks down beyond $\mu \approx -0.1$; even so we observe that the $m \to 1$ information content of the $J_{ij} = 0, 1$ case is certainly less than 0.29 (the limiting value of the GD result) which is much lower than the value obtained in the Willshaw model. One would expect that the information content of the optimal storage capacity connectivity matrix would surpass that of the Willshaw model which imposes a specific form of J_{ij}. We believe that the resolution of this apparent paradox lies in one or both of the following points: (a) the Willshaw model corresponds to $a = c \ln N / N$ with $N \to \infty$ while the $a \to 0$ limit explored in the present calculation is $a = c\,N$ with $c \to 0$ and thus the result of the former is not necessarily obtained as the limit of the latter; (b) the results we obtain are based on the condition that the stored memories are exact fixed points, while the Willshaw result follows from the requirement of zero noise-to-signal ratio. If one requires strictly zero retrieval error instead, one gets a lower information capactity (Nadal 1989a).

7. Concluding remarks

We have extended Gardner's work to a new class of constraints on the coupling matrix. In the process we have established that the best and most consistent estimate of storage capacity in networks with discrete couplings is given by the criterion of zero entropy. This criterion can be applied in all these cases where the entropy can be unambiguously defined as the logarithm of the number of available configurations. This conclusion is supported by simultations for the binary and box confinement cases. Moreover, an approximate estimate of the critical storage capacity for finite L, which is expected to improve as L increases, gives very good agreement with the ZE result.

The basic problem impeding implementation of models with discrete constraints is that we do not have general algorithms for finding optimal matrices for most cases. In fact, we do not even know whether efficient algorithms exist. One possible approach is to find a continuous valued solution using, for example, the perceptron algorithm, and then to discretise the couplings as in section 4.3. In general, the resulting J_{ij} will not satisfy (2). However, they may furnish starting points for further learning sessions. This procedure may be repeated until the discretised matrix becomes a solution. Although there is no guarantee that this process converges to a solution, preliminary results indicate that, sufficiently far from saturation, it does. This method deserves more systematic study.

References

Amit D J, Wong K Y M and Campbell C 1989 *J. Phys. A: Math. Gen.* **22** 2039

de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983

Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949

Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

—— 1989 *J. Phys. A: Math. Gen.* **22** 1983

Golomb D, Rubin N and Sompolinsky H 1990 *Phys. Rev. A* **41** 1843

Gutfreund H and Stein Y 1989 *Proc. STATPHYS 17 Workshop on Neural Networks and Spin Glasses, Porto Alegre* (Singapore: World Scientific) in press

Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554

Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745

—— 1989 *J. Physique* **50** 3057

Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** 1983

Minsky M and Papert S 1969 *Perceptrons. An Introduction to Computational Geometry* (Cambridge, MA: MIT Press)

Nadal J-P 1989a *Proc. STATPHYS 17 Workshop on Neural Networks and Spin Glasses, Porto Alegre* (Singapore: World Scientific) in press

—— 1989b private communication

Nadal J-P and Toulouse G 1990 *Network* **1** 61

Perez Vicente C J 1989 *Proc. STATPHYS 17 Workshop on Neural Networks and Spin Glasses, Porto Alegre* (Singapore: World Scientific) in press

Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan Books)

Sompolinsky H 1986 *Phys. Rev. A* **34** 2571

Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960