J. Phys. A: Math. Gen. 24 (1991) 4907–4940. Printed in the UK

Finite-size effects and bounds for perceptron models

B Derrida, R B Griffiths† and A Prügel-Bennett‡
Service de Physique Théorique§, CEN-Saclay, F-91191 Gif-sur-Yvette Cedex, France.

Received 26 March 1991

Abstract. In this paper we consider two main aspects of the binary perceptron problem: the maximal capacity when random patterns are stored (model A), and its generalization ability (model B). We have extended previous numerical estimates of critical capacities and studied thermal properties of systems of small sizes to test recent replica predictions. We have also considered some simpler versions of these models. The discrete spherical versions can be solved exactly using Gardner's replica calculation for the spherical model and are shown to give a rigorous upper bound and lower bound on the capacities of models A and B, respectively. Toy versions of models A and B are solved in detail and provide information which is useful for interpreting the finite-size effects present in the numerical studies of models A and B.

1. Introduction

Ever since it was realized that tools employed for studying disordered systems could be applied to models of neural networks, the investigation of such models has become a part of theoretical physics [1, 2]. An important question is whether a given neural network architecture can perform tasks such as storing information, classifying it, and generalizing from examples. The simplest architecture which has been proposed is the perceptron [3, 4]. In this paper we are concerned with two aspects of a perceptron: its maximal storage capacity (model A) and its ability to generalize (model B).

The problem of the optimal capacity of the perceptron can be formulated in its simplest version as follows: given P input patterns S_i^μ, $1 \leqslant \mu \leqslant P$ and $1 \leqslant i \leqslant N$, and P outputs T^μ, is there a configuration of synaptic weights or couplings J_i, $1 \leqslant i \leqslant N$, such that each input pattern gives the desired output, i.e.

$$T^\mu = \operatorname{sgn}\left(\sum_{i=1}^{N} J_i S_i^\mu\right) \tag{1.1}$$

for every μ? As the number P of patterns increases while the number of couplings N is fixed, it becomes harder and harder to find a choice $\{J_i\}$ which satisfies (1.1). The maximal capacity P_c is the maximum number of patterns which can be stored, in the sense that for $P > P_c$ there is no solution to (1.1).

† Also at Institut des Hautes Etudes Scientifiques, 91440 Bures-sur-Yvette, France. Permanent address: Physics Department, Carnegie-Mellon University, Pittsburgh, PA 15213, USA.

‡ Now at Department of Computer Science, University of Manchester, UK.

§ Laboratoire de la Direction des Sciences de la Matière du Commisariat à l'Energie Atomique.

0305-4470/91/204907+34$03.50

This maximal capacity depends on the method used to choose the patterns $\{S_i^\mu\}$ and the targets $\{T^\mu\}$. If they are chosen at random in the sense that S_i^μ and T^μ are $+1$ or -1 with equal probability and have no correlations, Cover [5] showed that the maximal capacity is

$$P_c \simeq \alpha_c N$$

for large N, with $\alpha_c = 2$. This result was rederived by Gardner [6, 7], who developed a replica approach by which the volume in the space of $\{J_i\}$ with $\sum_i J_i^2 = N$ of those couplings satisfying (1.1) could be calculated for $P < 2N$, as well as the minimum fraction of errors [8] for $P > 2N$. This replica approach was later extended to several other situations [9–17].

A serious difficulty arose when this replica procedure was applied to the case of binary couplings $J_i = \pm 1$, which we shall call 'model A'. It gave $\alpha_c = 4/\pi \simeq 1.27$, whereas one can prove [8, 18, 19] that $\alpha_c \leqslant 1$, and numerical simulations [18–22] yield estimates for α_c in the range 0.7 to 0.85. Since this calculation assumes replica symmetry, a possible source for this erroneous α_c was replica symmetry breaking. However, it was shown that the replica symmetric solution was stable for $\alpha \leqslant 1.015$ [19], excluding the possibility that the true α_c, which cannot exceed 1, occurs at the limit of stability of this solution. The sole remaining possibility within the replica approach was a discontinuous transition from a replica symmetric to a replica non-symmetric saddle point. Krauth and Mézard [19] found such a transition [23] and realized that it was similar to the one which occurred in the random energy model [24], in which the entropy vanishes at the transition temperature at the same time as the replica symmetric solution remains locally stable [25]. They thereby obtained a value of $\alpha_c = 0.833$. One of the motivations of the present paper was to use numerical simulations to test this prediction of the Krauth and Mézard theory and to improve previous estimates of the critical capacity [18].

Besides the maximal capacity, another property of interest for neural network models is their ability to generalize [18, 26–32]. For the binary perceptron, the problem of generalization (model B) can be formulated as follows [18]: the couplings J_i are again ± 1 and the input patterns $\{S_i^\mu\}$ are chosen at random. However, the $\{T^\mu\}$ are now by definition the values given by a 'teacher' $\{\tilde{J}_i\}$, which is simply one possible choice of couplings (e.g. $\tilde{J}_i = +1$ for all i). Just as in model A, the number of configurations satisfying (1.1) for this choice of $\{T^\mu\}$ decreases as P increases until a critical value $P_c \simeq \alpha_c N$ is reached, above which the only solution is $\{J_i\} = \{\tilde{J}_i\}$. For $P < P_c$ the generalization of the network is imperfect, because there are solutions $\{J_i\} \neq \{\tilde{J}_i\}$ which give the same results as the teacher for the P specified patterns which have been 'learnt', but different results for patterns not yet learnt. By contrast, for $P > P_c$ the only solution to (1.1) is the teacher itself, which thus will (by definition!) give correct results in all cases. A rigorous upper bound of $\alpha_c \leqslant 1.448$ as well as numerical extrapolations to $\alpha_c \simeq 1.35$ have been obtained previously [18]; both substantially exceed the value of $\alpha_c = 1.245$ obtained [26, 27] using the replica approach, which also predicts a freezing transition at finite temperature for $\alpha > \alpha_c$.

The goal of the research reported here was to improve the existing estimates for the critical capacity for models A and B, and to use numerical simulations to test the Krauth–Mézard [21], and Györgyi [26, 27] predictions for the nature of the freezing transitions. To this end, it turns out to be very useful to introduce several variants

of models A and B, as described in §2: the S_i^μ are allowed to have a continuous Gaussian distribution ('Gaussian patterns'), and the 2^N possible couplings $\{J_i\}$ can be chosen at random on the sphere in $\mathbb{R}^N$ ('discrete spherical couplings'). For all these models, we show that the problem of maximal capacity can be formulated in terms of the distribution of points or objects in various cells: for $\alpha > \alpha_c$ of model A, most of the cells are empty, and for $\alpha > \alpha_c$ of model B, almost all occupied cells are occupied by a single point. From this perspective, models A and B are different aspects of a single problem. The discrete spherical models, while rather artificial from the point of view of neural networks, are interesting in that the A and B transitions can be found numerically from the results of Gardner's replica calculations, and these values are then upper and lower bounds respectively, as shown in §5, of the critical capacities of the model with binary couplings and Gaussian patterns.

In addition we introduce, in §3, two simplified or 'toy' models corresponding to A and B, in which the energies of the different configurations (defined as the number of errors occurring in (1.1)) are independent random variables. These models can then be solved exactly, and the finite-size effects turn out to be useful in interpreting the numerical results presented in §4. The latter are of two types: estimates of critical capacities which extend previous results to larger values of N, and calculations of thermal properties for systems of finite size which, in conjunction with the results of §3, allow a test of the replica prediction for models A and B. Our conclusions are summarized in §6.

2. Transitions in the discrete spherical models

The perceptron model considered here can be characterized as follows. The P patterns correspond to a $P \times N$ matrix S_i^μ with $\mu = 1, 2, \ldots, P$ and $i = 1, 2, \ldots, N$. A set of couplings J is a vector with N components J_i, and for the μth pattern gives rise to an output

$$R^\mu = \operatorname{sgn}\left(\sum_{i=1}^{N} S_i^\mu J_i\right) \tag{2.1}$$

which takes values ± 1. Let the target T be a P-component vector $T^\mu = \pm 1$. The energy E attributed to a set of couplings J is the number of patterns for which the output differs from the target, that is

$$E(J) = \frac{1}{4}\sum_{\mu=1}^{P}(R^\mu - T^\mu)^2. \tag{2.2}$$

Using these energies, a partition function [8] analogous to that used in statistical mechanics can be defined:

$$Z = \sum_{\{J\}} \mathrm{e}^{-\beta E(J)}. \tag{2.3}$$

We shall assume that the elements of the patterns are chosen randomly: the $N \times P$ numbers S_i^μ are independent identically distributed random variables. For *Ising*

patterns, each S_i^μ is $+1$ or -1 with equal probability. For *Gaussian patterns*, each S_i^μ is chosen from a Gaussian distribution with zero mean.

The term *binary couplings* will refer to the case in which each J_i is $+1$ or -1 and *spherical couplings* when the J_i are real numbers satisfying

$$\sum_{i=1}^{N} J_i^2 = N. \tag{2.4}$$

A third possibility is that of *discrete spherical couplings* in which a set of e^{Nb} couplings are chosen randomly as vectors on the sphere defined by (2.4). Typically we shall assume that $b = \ln 2$, giving the same number (2^N) of possibilities as the binary case.

The two types of patterns and the three kinds of couplings give rise to a total of six distinct but related models which can be studied for large N as a function of the parameter

$$\alpha = P/N. \tag{2.5}$$

In the case of Ising patterns and binary couplings we shall consider only the case where N is odd so that the sign in equation (2.1) is well defined. It has been asserted [21], on the basis of replica studies, that Ising and Gaussian patterns give rise to identical results as $N \to \infty$. Nonetheless, it is useful to distinguish them because the behaviour for finite N will be different, and because a certain inequality (see §5) relating the binary and discrete spherical couplings can be proved for Gaussian patterns but not (at least by the same methods) for Ising patterns.

The following perspective in these models is sometimes helpful. Formula (2.1) assigns each set of couplings J, each of which can be thought of as an object or 'particle', to one of 2^P categories or 'boxes' labelled by the P numbers R^μ. Thus a particular set of patterns S_i^μ gives rise to a histogram consisting of the number of particles in each box. The target T is a particular box which has zero energy, while the other boxes are assigned energies relative to T by means of (2.2). The partition function (2.3) is a sum over particles, each assigned the energy of the box in which it is located. Note that as long as the distribution for each S_i^μ is symmetrical about zero the choice of the target is arbitrary in that any choice will yield the same statistics. This fact can be used to speed up numerical studies, because once a histogram has been constructed, sets of energies can be computed for various targets. (The results are correlated, but not biased, and by using several histograms corresponding to independent choices of patterns, it is possible to make error estimates by standard procedures.)

Both models A and B are determined by the same histogram, and thus can be thought of as two aspects of a single model. The difference is that for model A, any one of the 2^P boxes may be chosen at random to define the zero of energy, whereas for model B, one of the particles is chosen at random (it is, by definition, the 'teacher'), and the box containing it is assigned the energy zero. The critical α for model A, denoted by α_A, is the value such that for $\alpha < \alpha_A$ a box chosen at random will contain at least one particle, whereas for $\alpha > \alpha_A$ the typical box will be empty. The critical α for model B, denoted α_B, has the property that, for $\alpha > \alpha_B$, almost every particle is alone in the box which it occupies, whereas for $\alpha < \alpha_B$ it almost certainly has the company of at least one other particle in the same box, i.e. a 'student' is able to yield the same result as the 'teacher'. These definitions can be

made more precise by appropriate uses of the phrase 'with a probability approaching one in the limit as N tends to infinity'. Unfortunately there are no proofs (known to us) that the desired limits exist, or that α_A and α_B are well defined in the sense that there exists no intermediate phases: for example, a range of α values for model A in which the probability that a typical box is empty is neither zero nor one.

A geometrical representation of the 'boxes' introduced above is obtained by imagining that for each μ, the S_i^μ are the components of a vector normal to a hyperplane in an N-dimensional space $\mathbb{R}^N$. The P hyperplanes corresponding to the P patterns cut this space into a set of 2^P convex regions (some may have zero volume), and these regions intersect the sphere (2.4) in a corresponding number of *cells*, with the property that all the points in a particular cell are, by (2.1), mapped into the same set of outputs. Thus these cells correspond to the 'boxes', while the 'particles' in a particular box are those sets of couplings J corresponding to points on the sphere (2.4) inside the cell in question.

In the case of the discrete spherical model, transitions A and B can be discussed in terms of the volumes of these cells, where 'volume' denotes the appropriate rotationally invariant measure on the sphere (2.4). (For example, for $N = 3$ and $P = 3$ the cells are spherical triangles, and the 'volume' is the corresponding area on the surface of the sphere.) It will be convenient to assume this measure is normalized, so that if v_R is the volume of the cell labelled by $R = \{R^\mu\}$,

$$\sum_R v_R = 1. \tag{2.6}$$

A cell R will be said to be of 'size' k provided its volume lies in the interval

$$e^{Nk} \leqslant v_R \leqslant e^{N(k+\Delta k)} \tag{2.7}$$

where Δk is a small positive number, and k takes on a discrete set of values determined by Δk. Let $\exp\{N\,c(k)\}$ be the number of cells with volumes in the interval (2.7). Of course, $c(k)$ is a random variable which depends on the choice of hyperplane normals $\{S_i^\mu\}$. We shall assume that when N is large, the typical $c(k)$—the one occurring with high probability—approaches some limit, of which a plausible form is sketched in figure 1. The normalization condition (2.6) then reads

$$\sum_k e^{N[c(k)+k]} = 1 \tag{2.8}$$

and, assuming the sum is dominated by its maximum term, we conclude that

$$0 = c(k_B) + k_B \tag{2.9}$$

where (see figure 1) k_B is the value of k which maximizes $c(k) + k$, the point where $dc/dk = -1$.

Since in the discrete spherical model a set of e^{Nb} points or 'particles' are chosen randomly on the surface of the sphere, the average number $\nu(k)$ falling into a cell of size k will be proportional to its volume:

$$\nu(k) = e^{Nb}e^{Nk} = e^{N(k-\kappa)} \tag{2.10}$$

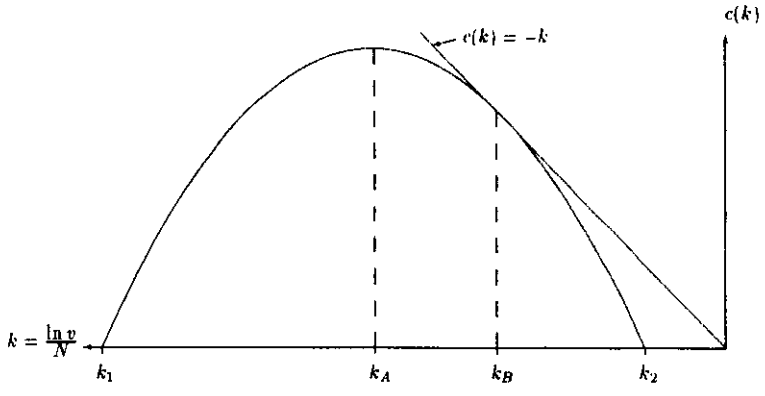

Figure 1. Sketch of the distribution of cell sizes.

where $\kappa = -b$ is a sort of negative chemical potential. Its significance is that cells of size $k < \kappa$ are essentially all empty, while those with $k > \kappa$ contain a large number of the e^{Nb} particles. Note that the vast majority of cells have a size $k \approx k_A$, corresponding to the maximum of $c(k)$, figure 1, whereas almost all of the volume is associated with cells of size $k \approx k_B$. Consequently the vast majority of particles (whatever the value of κ) will be in cells with $k \approx k_B$.

Now consider what happens if, for a given set of cells, κ increases continuously (corresponding to a decreasing number of particles) from a value less than k_1 to a value greater than k_2 (figure 1). For $\kappa < k_1$, all cells (with a probability approaching 1) will contain at least one particle, whereas for $\kappa > k_1$ there will be a large number (though a negligible fraction of the total) which are empty. For $\kappa > k_A$, almost all the cells are empty. Thus $\kappa = k_A$ corresponds to the A transition for this model. Similarly, as the vast majority of particles are in cells with $k \approx k_B$, for $\kappa < k_B$ most particles are in cells containing other particles, whereas for $\kappa > k_B$ most are isolated in separate cells. Thus $\kappa = k_B$ corresponds to transition B. Finally, as κ passes k_2 the last cases of more than one particle in a cell disappear.

The foregoing discussion needs minor modifications if some of the cells have zero volume, and are thus not represented by $c(k)$. If their number is a small fraction of the total, the only modification is that when κ passes k_1 the 'transition' involves only those cells with non-zero volume. (Of course it is also possible that k_1 is at $-\infty$, so there is no transition of this type in any case.) However, if the majority of cells have zero volume, there can obviously be no A transition as κ varies, even if $c(k)$ has a maximum. There can still be a B transition, described in the same way as previously.

The same considerations apply if b is held fixed and α varied, for changing α will change the (typical) distribution $c(k)$. The A transition occurs when $\kappa = -b = k_A$, which means that the volume of the typical cell is e^{-Nb}, or 2^{-N} in the case $b = \ln 2$. Gardner's replica calculation indicates that this value occurs at $\tilde{\alpha}_A = 0.847$. (This value is obtained by finding the value of α for which the $G(q)$ in [7, equation (23)] is equal to $-\ln 2$. This has been previously calculated by Krauth and Mézard [19].) The B transition will occur when $\kappa = -b = k_B = -c(k_B)$, which is to say, for that α for which $c(k_B) = \ln 2$. The corresponding α can be determined from Gardner's replica calculation of the moments

$$\langle v^n \rangle = 2^{-P} \sum_R \langle v_R^n \rangle \tag{2.11}$$

of the cell-size distribution. Here v is the volume of some specific cell (by symmetry it does not matter which cell is chosen), and the angular brackets $\langle\cdots\rangle$ indicate an average over all possible sets of the P patterns.

Let us define $g(n)$ by

$$e^{Ng(n)} = \sum_R v_R^n = \sum_k e^{N[c(k)+nk]} \tag{2.12}$$

for a typical distribution $c(k)$. Replacing the sum by its maximum term, we have $g(n)$ related to $c(k)$ by a Legendre transform,

$$g(n) = \max_k [c(k) + nk]. \tag{2.13}$$

Assuming $c(k)$ is differentiable, this tells us that

$$g(n) = c(k) + nk \tag{2.14}$$

where k is the solution of

$$n = -c'(k) \tag{2.15}$$

and, by differentiating (2.14) with respect to n,

$$k = g'(n). \tag{2.16}$$

Now k_B is the value of k where $c'(k) = -1$. This corresponds, by (2.15), to $n = 1$, and hence, by (2.9) and (2.16), to

$$c(k_B) = -k_B = -g'(1). \tag{2.17}$$

Consequently the α corresponding to the transition B for the discrete spherical model is the one for which $g'(1) = -\ln 2$.

Note that $g(n)$ is defined, (2.12), for the typical case, whereas the replica calculation yields $\hat{g}(n)$ defined by

$$e^{N\hat{g}(n)} = \left\langle \sum_R v_R^n \right\rangle = 2^P \langle v^n \rangle. \tag{2.18}$$

Since $\sum_R v_R^n$ is non-negative, its typical value (that achieved with high probability) cannot be larger than its average value times a factor very close to one, although it might be much smaller.

As a consequence one has

$$\hat{g}(n) \geqslant g(n). \tag{2.19}$$

In addition, the normalization condition (2.6) tells us that

$$\hat{g}(1) = g(1) = 0. \tag{2.20}$$

As a consequence of (2.19) and (2.20)

$$\hat{g}'(1) = g'(1) \tag{2.21}$$

assuming the derivatives exist.

The nth moment of the volume $\langle v^n \rangle$ was calculated by Gardner [7, equations (14)–(17)]. Assuming the replica symmetric ansatz, which allows analytic continuation of $\langle v^n \rangle$ to non-integer n, yields the formula

$$\hat{g}(n) = \frac{\log\langle v^n \rangle}{N} = \frac{1}{2}(n-1)\log(1-q) + \frac{1}{2}\log(1+(n-1)q) + \alpha \log\left\{ \int Dz \left[H\left(\frac{z\sqrt{q}}{\sqrt{1-q}} \right) \right]^n \right\} \tag{2.22}$$

where

$$Dz = e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \qquad H(x) = \int_x^\infty Dz \tag{2.23}$$

and where q is the Edwards–Anderson parameter given by the fixed point equation

$$q = \frac{n(n-1)q}{2[(n-1)q^2 + (2-n)q - 1]} - \alpha n \int Dz\, z \left[H\left(\frac{z\sqrt{q}}{\sqrt{1-q}} \right) \right]^{n-1} \exp\left(\frac{qz^2}{2(1-q)} \right) \times \left\{ 2[2\pi q(1-q)^3]^{1/2} \int Dz \left[H\left(\frac{z\sqrt{q}}{\sqrt{1-q}} \right) \right]^n \right\}^{-1}. \tag{2.24}$$

Using this equation, one finds that $\hat{g}'(1) = -\ln 2$ when $\tilde{\alpha}_B = 1.197$, which is therefore the critical capacity of the discrete spherical model B.

The replica calculation applies equally to Ising or Gaussian patterns, and thus (assuming it is correct), the values of $\tilde{\alpha}_A$ and $\tilde{\alpha}_B$ given above for the discrete spherical model are also valid for both cases.

The notion of cell-size distribution can also be employed in the case of binary couplings by defining the 'volume' of a cell as the number of hypercube vertices (J with $J_i = \pm 1$) which it contains, dividing by 2^N to ensure normalization (2.6). This new definition gives rise to a larger number of empty cells than in the preceding case. The A transition occurs again, at the value of α at which k_A, the value of k where $c(k)$ has its maximum, is equal to $-\ln 2$ and the B transition when $c(k_B)$ is equal to $\ln 2$.

3. Toy models A and B

In this section we introduce and solve two simplified models: toy models A and B. These models are simplified versions of the models A and B with binary couplings and Ising patterns introduced in the previous section. Their main simplification is that the energies of the different configurations are independent random variables. This

simplification allows one to calculate the free energy exactly in the thermodynamic limit ($N \to \infty$) as well as finite-size corrections in approaching the limit.

These solutions can serve as a guide to interpreting the numerical results obtained for other models (defined in §2) which are too complicated to be solved exactly. These toy models also provide bounds on the free energy of the true models A and B defined in §2.

The reason for these bounds is the same as for other random energy models [18, 24] and can be explained as follows: the partition function Z is always given by

$$Z = \sum_E \mathcal{N}(E) e^{-E/T} \tag{3.1}$$

where $\mathcal{N}(E)$ is the number of configurations at energy E. However, one is interested in the value Z_{typ} of Z in a typical case, one which occurs with a high probability, say $1 - \epsilon$, where ϵ is a small number. To calculate this one needs to know the typical value, $\mathcal{N}_{typ}(E)$, rather than the average $\langle \mathcal{N}(E) \rangle$. Because $\mathcal{N}(E)$ is non-negative, its average can obviously not be much smaller than its typical value

$$\langle \mathcal{N}(E) \rangle \geqslant \mathcal{N}_{typ}(E) \tag{3.2}$$

where, if we want to be precise, the right-hand side can be multiplied by $(1 - \epsilon)$. And since $\mathcal{N}(E)$ is an integer, $\mathcal{N}_{typ}(E)$ will be 0 if $\langle \mathcal{N}(E) \rangle$ is small compared to 1.

In the corresponding toy model, the 2^N configurations are independently assigned energies at random (as described below) in a manner which makes the average $\mathcal{N}_{toy}(E)$ equal to $\langle \mathcal{N}(E) \rangle$ for the corresponding real model. However, the typical $\mathcal{N}_{toy}(E)$ is close to its average value [24] when the latter is large compared to 1, making (3.2) an approximate equality, whereas in the true model the two sides can be very different, with, for example, the average $\langle \mathcal{N}(E) \rangle$ quite large even when $\mathcal{N}_{typ}(E)$ is zero. As a consequence

$$\mathcal{N}_{typ}(E) \leqslant \mathcal{N}_{toy}(E) \tag{3.3}$$

in the large-N limit. (The seeming contradiction with the fact the sum over E is 2^N in both cases is disposed of by noting that $\mathcal{N}_{typ}(E)$ can exceed $\mathcal{N}_{toy}$ by a small fraction at energies where both quantities are exponentially large and essentially equal.) Hence one has

$$N^{-1} \ln Z_{typ} \leqslant N^{-1} \ln Z_{toy} \tag{3.4}$$

in the large-N limit, which means that the free energy (multiply both sides of the inequality by $-T$) of the toy model is a lower bound for the true free energy and the toy ground state, a lower bound for the true (typical) ground-state energy.

Another consequence of (3.3) is that

$$\alpha_c(\text{true}) \leqslant \alpha_c(\text{toy}) \tag{3.5}$$

because α_c is the largest value of α for which $\mathcal{N}(0) \geqslant 1$ in model A, or $\mathcal{N}(0) \geqslant 2$ in model B.

3.1. Toy model A

3.1.1. Definition and thermodynamic properties. In toy model A the output R^μ for each pattern and for each of the 2^N configurations $\{J_i\}$ is chosen at random, so that the energy E of each configuration is a random number chosen with probability

$$\Pr(E) = \frac{1}{2^N}\langle \mathcal{N}(E)\rangle = \frac{1}{2^P}\binom{P}{E} \tag{3.6}$$

where

$$\langle \mathcal{N}(E)\rangle = 2^{N-P}\binom{P}{E} \tag{3.7}$$

is the average for the true model A with binary couplings and Ising or Gaussian patterns. As explained above, the typical value of $\mathcal{N}_{\text{toy}}(E)$ is 0 for $\langle \mathcal{N}(E)\rangle$ substantially less than 1 and equal to $\langle \mathcal{N}(E)\rangle$ when the latter is large compared to 1. One can show that $\mathcal{N}_{\text{toy}}(E)$ has a distribution which is approximately Poisson [33], and the values for different E are approximately independent. (They are not completely independent, because their sum is 2^N.)

Now let

$$\epsilon = E/N \tag{3.8}$$

be the energy per coupling, and define

$$s(\epsilon) = \lim_{N\to\infty} \frac{\ln \mathcal{N}_{\text{typ}}(N\epsilon)}{N} \tag{3.9}$$

which, by using (3.7) and applying Stirling's approximation, gives

$$s(\epsilon) = (1-\alpha)\ln 2 + \alpha \ln \alpha - \epsilon \ln \epsilon - (\alpha - \epsilon)\ln(\alpha - \epsilon). \tag{3.10}$$

In view of the preceding remarks, we see that $s(\epsilon)$ is the (microcanonical) entropy per coupling whenever it is non-negative; when the right-hand side of (3.10) is negative, $\mathcal{N}_{\text{typ}}$ for the corresponding energy is 0 and s is $-\infty$. In particular, for $\alpha < 1$, $s(\epsilon)$ is positive for all $\epsilon \geqslant 0$, implying that the $\epsilon = 0$ ground state has a finite entropy. This entropy vanishes linearly in $(1-\alpha)$ as α approaches its critical value, 1, from below, similar to the Krauth and Mézard prediction [19] for the behaviour of the true model A (where the critical value of α is, of course, less than 1). For $\alpha > 1$, on the other hand, there is a ground-state energy in the range $0 < \epsilon_c < \alpha/2$ where the right-hand side of (3.10) vanishes.

The temperature T is given by the usual formula

$$1/T = \mathrm{d}s(\epsilon)/\mathrm{d}\epsilon \tag{3.11}$$

as long as $s(\epsilon) \geqslant 0$. Using this one finds that, for $\alpha > 1$, the entropy vanishes at a critical temperature T_c related to α through

$$\alpha = \frac{\ln 2}{\ln 2 - \ln(1 + \mathrm{e}^{1/T_c}) + T_c^{-1}(1 + \mathrm{e}^{-1/T_c})^{-1}}. \tag{3.12}$$

Alternatively the thermodynamic properties can be obtained from the (typical) partition function for the toy model—we hereafter omit the subscript 'toy'—which is

$$\frac{\ln Z}{N} = (1-\alpha)\ln 2 + \alpha \ln(1+\mathrm{e}^{-1/T}) \tag{3.13}$$

for $T > T_c$ (including all T for $\alpha \leqslant 1$), and

$$\frac{\ln Z}{N} = \frac{1}{T}\left(\frac{-\alpha}{1+\mathrm{e}^{1/T_c}}\right) \tag{3.14}$$

for $T < T_c$ and $\alpha > 1$. At $T = T_c$, the specific heat drops discontinuously to zero and for $T < T_c$ the system is in a frozen state with constant (ground-state) energy and zero macroscopic entropy and specific heat. This is the same behaviour observed in the random energy model [24], which differs from the toy model A (and toy model B) mainly in the fact that in the former E is a continuous variable.

3.1.2. The critical capacity α_c of toy model A. The critical capacity α_c for a perceptron of finite size N can be defined in the following way [18]. Assume the target is $T^\mu = 1$ for all μ. (Any other choice of target will, of course, give the same result.) One starts with 2^N configurations, and as each pattern is added, those configurations which do not give the correct output, $R^\mu = 1$, are discarded. If there are still some configurations remaining after P_c patterns, but none after P_c+1 patterns, the number of stored patterns is defined as P_c. Clearly P_c is a random variable depending on the patterns S_i^μ, and we define the critical capacity

$$\alpha_c(N) = \frac{\langle P_c \rangle}{N} \tag{3.15}$$

in terms of its average over the corresponding probability distribution.

An equivalent expression can be obtained by introducing random variables $y_1, y_2, \ldots$ defined as follows: $y_P = 1$ if, after P patterns have been added, there are still some allowed configurations, and 0 if there are no allowed configurations. If the number of stored configurations is P_c, then, obviously, $y_P = 1$ for $1 \leqslant P \leqslant P_c$, and $y_P = 0$ for $P > P_c$. In other words

$$P_c = \sum_{P=1}^{\infty} y_P \tag{3.16}$$

and, consequently,

$$\alpha_c(N) = \frac{1}{N}\sum_{P=1}^{\infty}\langle y_P \rangle. \tag{3.17}$$

Note that, because the patterns are all chosen independently, $\langle y_P \rangle$ is simply the probability that given *any* P patterns (not necessary the first P), there is at least one of the 2^N configurations which gives $R^\mu = 1$ for each of these patterns.

For toy model A, the output for each pattern and for each configuration is chosen at random. Thus the probability that a particular configuration gives $R^\mu = 1$ for each

of P patterns is 2^{-P}, and hence the probability that *none* of the 2^N configurations gives $R^\mu = 1$ for each of these patterns is

$$1 - \langle y_P \rangle = (1 - 2^{-P})^{2^N}. \tag{3.18}$$

If this expression is inserted in (3.17), the result when N is large is

$$\alpha_c(N) = 1 + \frac{1}{N}\left(-e + \sum_{n=1}^{\infty}(1 - e^{-2^{-n}} - e^{-2^n})\right) + O(N^{-1}2^{-N}). \tag{3.19}$$

Thus for this model, $\alpha_c(N)$ converges to 1 with a $1/N$ correction. This will be compared with more complicated models in §4 below.

3.1.3. The low-temperature phase in toy model A. In the low-temperature $T < T_c$ phase, only a small number of configurations near the ground state need to be considered. The number of configurations at each energy E is given by a Poisson distribution with average $\langle \mathcal{N}(E) \rangle$ and the explicit dependence of this average on E, (3.7), can be approximated by an exponential,

$$\langle \mathcal{N}(E) \rangle \simeq A^{(E-E_c)} \tag{3.20}$$

where

$$A = e^{1/T_c} = (\alpha - \epsilon_c)/\epsilon_c \tag{3.21}$$

and E_c, the value of E where $\langle \mathcal{N}(E) \rangle$ is 1, is given by

$$E_c = N\epsilon_c + \frac{\ln[2\pi N \epsilon_c(\alpha - \epsilon_c)/\alpha]}{2\ln[(\alpha - \epsilon_c)/\epsilon_c]}. \tag{3.22}$$

Of course $\mathcal{N}(E)$ is only defined when E is an integer, and the value of E_c, (3.22), which makes (3.20) a good approximation for E near E_c will, in general, not be integer. The tendency of the fractional part of E_c to oscillate as N and P vary thus gives rise to oscillations in certain properties of the ground state, as we shall see.

The fact that the $\mathcal{N}(E)$ are independent Poisson variables means that the probability that the ground state has an energy E and is n-fold degenerate is given by

$$\frac{\langle \mathcal{N}(E) \rangle^n}{n!} \exp\left(-\sum_{E' \leqslant E} \langle \mathcal{N}(E') \rangle\right) \tag{3.23}$$

from which it follows that the average ground-state entropy is given by

$$S = \sum_{E=-\infty}^{\infty} \sum_{n=1}^{\infty} \frac{A^{(E-E_c)n}}{n!} \exp\left(-\frac{A^{(E-E_c)}}{1-A}\right) \ln n. \tag{3.24}$$

Here the lower limit $-\infty$ rather then 0 for E produces a negligible error (exponentially small in N). Changing E_c by an integer obviously leaves this expression unchanged (alter the dummy variable E by the same amount). On the other hand,

as is easily seen if one does the sum numerically, S depends on the fractional part of E_c, and thus S oscillates as a function of N and P. The average ground-state energy shows similar oscillations superimposed on a smooth dependence on N and P. We have looked for similar effects in our numerical studies of the true model A (§4) but have not seen anything definite.

The exponential approximation, (3.20), is also useful in an analysis of the low-temperature properties of the random energy model [33, 34]. The main difference is that in the latter E can take on any real value, and is not restricted to integers. As a consequence the random energy model always (with probability 1) has a non-degenerate ground state, whereas toy model A can have a degenerate ground state leading to a finite (order 1, not order N) average ground-state entropy. Both models also exhibit macroscopic fluctuations in the magnetic susceptibility, as defined and discussed in appendix 1.

3.1.4. Finite-size effects in toy model A near the freezing temperature. Finite-size effects round the singular behaviour at a phase transition. These can be computed explicitly for toy model A for T near the freezing temperature T_c, with $\alpha > 1$, and the results will be of use in discussing our simulations of the real model A in §4 below.

In the low-temperature phase, the decrease of $e^{-E/T}$ dominates the growth of $\langle \mathcal{N}(E) \rangle$ as E increases, which justifies the exponential approximation (3.20). These two effects become comparable for $T \simeq T_c$, so that (3.20) is no longer adequate. However, it suffices to expand $\langle \mathcal{N}(E) \rangle$ to second order in $E - E_c$ yielding a Gaussian approximation:

$$\langle \mathcal{N}(E) \rangle = \exp\left(\frac{E - E_c}{T_c} - \frac{(E - E_c)^2 b}{N} \right) \tag{3.25}$$

with

$$\frac{1}{T_c} = \ln\left(\frac{\alpha - \epsilon_c}{\epsilon_c} \right) \qquad \text{and} \qquad b = \frac{1}{2}\left(\frac{1}{\alpha - \epsilon_c} + \frac{1}{\epsilon_c} \right). \tag{3.26}$$

Once again, we make use of the fact that $\mathcal{N}(E)$ can be treated as independent random variables with a Poisson distribution corresponding to the average (3.25). The average of the logarithm of the partition function can be obtained using

$$\langle \ln Z \rangle = \int_0^\infty \frac{e^{-t} - \langle e^{-tZ} \rangle}{t} \, dt \tag{3.27}$$

where independence of the $\mathcal{N}(E)$ (see also appendix 2) leads to the expression

$$\begin{aligned} \langle e^{-tZ} \rangle &= \exp\left(\sum_E \langle \mathcal{N}(E) \rangle \left[\exp\left(-te^{-E/T} \right) - 1 \right] \right) \\ &= \exp\left(\sum_E e^{(E-E_c)/T_c - (E-E_c)^2 b/N} \left[\exp\left(-te^{-E/T} \right) - 1 \right] \right). \end{aligned} \tag{3.28}$$

When $T - T_c$ is of order $1/\sqrt{N}$, one can replace the sum by an integral over $\varphi = (E - E_c)/\sqrt{N}$, and, as shown in appendix 2, for the value of t of importance

for the integral (3.27), only the $\varphi > 0$ part gives a significant contribution, and the term in the final parentheses in (3.28) can be replaced with $-te^{-E/T}$, leading to

$$\langle e^{-tZ} \rangle = \exp\left\{-te^{-E_c/T} \int_0^\infty \exp\left[-b\varphi^2 + \varphi\sqrt{N}\left(\frac{1}{T_c} - \frac{1}{T}\right)\right] d\varphi\right\} \tag{3.29}$$

and hence

$$\langle \ln Z \rangle = -\frac{E_c}{T} + \log\left\{\int_0^\infty \exp\left[-b\varphi^2 + \varphi\sqrt{N}\left(\frac{1}{T_c} - \frac{1}{T}\right)\right] d\varphi\right\}. \tag{3.30}$$

The result is identical to that obtained in the random energy model with a continuous distribution of energies [34], due to the fact that contributions come from energies within the order of $\sqrt{N}$ of E_c so the effects of discreteness are washed out.

The specific heat per coupling can be computed from $\langle \ln Z \rangle$ and is given, again for $T - T_c$ of order $1/\sqrt{N}$, by

$$c = \frac{1}{T^2}[I_2/I_0 - (I_1/I_0)^2] \tag{3.31}$$

where

$$I_m = \int_0^\infty \varphi^m \exp\left[-b\varphi^2 + \varphi\sqrt{N}\left(\frac{1}{T_c} - \frac{1}{T}\right)\right] d\varphi. \tag{3.32}$$

Because this is a function of $\sqrt{N}(T_c^{-1} - T^{-1})$, the curves of $c(T)$ corresponding to different values of N all cross at $T = T_c$ at a value

$$c(T_c) = \frac{\pi - 2}{\pi}\frac{1}{2bT_c^2} \tag{3.33}$$

which is just $(\pi - 2)/\pi$ times the discontinuity in c at T_c in the $N \to \infty$ limit, where, using (3.13) and (3.14), one finds

$$c(T_c^+) = \frac{1}{2bT_c^2} \qquad c(T_c^-) = 0. \tag{3.34}$$

Thus for the toy model, the crossing of the specific heat curves for different sizes gives a good criterion for T_c.

3.2. Definition and solution of toy model B

In model B with Ising patterns and binary couplings, let the 'teacher' be the configuration $\{\tilde{J}_i\}$, and consider a 'student' $\{J_i\}$ which has n of the N couplings identical with the teacher, that is,

$$\sum_{i=1}^{N} J_i \tilde{J}_i = 2n - N. \tag{3.35}$$

Assuming for convenience that the first n components are identical, one has for any pattern S_i^μ the result

$$\sum_i \tilde{J}_i S_i^\mu = Y + Z \qquad \sum_i J_i S_i^\mu = Y - Z \tag{3.36}$$

where Y represents the sum over $1 \leqslant i \leqslant n$, and Z that for $n+1 \leqslant i \leqslant N$. The probability p_n that the teacher and the student give the same output, (2.1), for a randomly chosen pattern is the same as the probability that $|Y| > |Z|$, which is easily shown to be

$$p_n = 2^{-N} \sum_{m_1=0}^{n} \binom{n}{m_1} \sum_{m_2=0}^{N-n} \binom{N-n}{m_2} \theta[(n-2m_1)^2 - (N-n-2m_2)^2] \tag{3.37}$$

where $\theta[\cdots]$ is 1 when its argument is positive and 0 otherwise; the argument is never 0 when, as we shall assume, N is odd. Note that $p_0 = 0$, $p_N = 1$ and

$$p_{N-n} = 1 - p_n. \tag{3.38}$$

One can also show that the p are equally in pairs: $p_1 = p_2$, $p_3 = p_4$, etc, which follows from the fact that $\binom{M}{m} = \binom{M-1}{m} + \binom{M-1}{m-1}$. When N is large and

$$n = xN \tag{3.39}$$

one finds [18]

$$p_n \simeq \gamma(x) = \frac{2}{\pi} \tan^{-1} \left(\frac{x}{1-x} \right)^{1/2}. \tag{3.40}$$

With P randomly chosen patterns and n given by (3.35), the probability $P_n(E)$ that the configuration $\{J_i\}$ has energy E is given by

$$P_n(E) = \binom{P}{E} p_n^{P-E} (1-p_n)^E. \tag{3.41}$$

In toy model B, one divides the 2^N configurations into sets of $\binom{N}{n}$ configurations, $0 \leqslant n \leqslant N$, and the configurations in each set are randomly assigned energies according to the distribution (3.41). Consequently the average number of configurations with energy E is given by

$$\langle \mathcal{N}_{\text{toy}}(E) \rangle = \sum_{n=0}^{N} \binom{N}{n} P_n(E) \tag{3.42}$$

which is identical to $\langle \mathcal{N}(E) \rangle$ for the true model. For each E in the interval $0 < E < P$, $\mathcal{N}_{\text{toy}}(E)$ has, approximately, a Poisson distribution, and this is also the case for $\mathcal{N}_{\text{toy}}(0) - 1$ and $\mathcal{N}_{\text{toy}}(P) - 1$. And, unlike the true model, $\mathcal{N}_{\text{toy}}(E)$ coincides with $\langle \mathcal{N}(E) \rangle$ whenever the latter is large. This makes it possible to compute the partition function in the large-N limit,

$$\frac{\ln Z_{\text{toy}}}{N} = \max_x \{-(1-x)\ln(1-x) - x \ln x + \alpha \ln[\gamma(x) + (1-\gamma(x))e^{-1/T}]\} \tag{3.43}$$

where $\alpha = P/N$ and $\gamma(x)$ is defined in (3.40). For $T \geqslant 0$ the right-hand side always has a local maximum at $x = 1$ (corresponding to the ground state $E = 0$) and as T decreases this eventually becomes the absolute maximum at a first-order phase transition where the entropy and the energy fall discontinuously to zero, and the overlap with the teacher, $2x - 1$, jumps discontinuously to one, where they remain for the whole low-temperature phase. For finite N these discontinuities will be rounded out, and one expects the energy (or entropy) against temperature curves to intersect in a manner similar to the heat capacity curves for model A.

3.2.1. Critical capacity of toy model B. The critical capacity for model B can be determined in a manner analogous to that for model A: one considers all $2^N - 1$ 'student' configurations and as each pattern is added, those configurations which do not give the same answer as the teacher are eliminated, until with $P_s + 1$ patterns there are none left. Equations (3.15) and (3.17) apply once again where $\langle y_P \rangle$ is the probability that for P randomly chosen patterns, at least one student gives the same answer as the teacher.

For toy model B, the probability that all the students fail,

$$1 - \langle y_p \rangle = \prod_{n=0}^{N-1} [1 - p_n^P]^{\binom{N}{n}} \tag{3.44}$$

is just the product of the probability of failure for each student. Thus one has

$$\alpha_c(N) = \frac{1}{N} \sum_{P=1}^{\infty} \langle y_P \rangle \simeq \frac{1}{N} \sum_{P=1}^{\infty} (1 - e^{-x_P}) \tag{3.45}$$

where

$$x_P = \sum_{n=0}^{N-1} \binom{N}{n} (p_n)^P. \tag{3.46}$$

For a plot of $\alpha_c(N)$ against $1/N$ see figure 5 below. As N tends to infinity, $\alpha_c(N)$ tends to 1.448, the upper bound for model B derived in [18].

4. Numerical studies

4.1. Model A

For the various versions of model A introduced in §2 we have performed some numerical simulations which are presented in this section. Just as in spin-glass systems, the energy landscape is rugged [37]. This makes it difficult to use standard Monte Carlo techniques [22] since the system gets trapped in local minima. To avoid these problems, we have used exact enumeration—that is, we calculate the energy for each of the 2^N configurations $\{J_i\}$. The obvious disadvantage of this approach is that one is limited to studying very small sizes so that analysis of finite-size effects is unavoidable. We have tested the replica theory [19] by looking at two different aspects: the critical capacity α_c and the phase transition at finite temperature.

To study the critical capacity we have measured the capacity $\alpha_c(N)$ as a function of N as explained in §2. For the binary perceptron with Ising patterns, N took on the values $5, 7, \ldots, 19$, while for the other models the simulations were carried out at all integer values of N between 5 and some maximum value. The method, in brief, is to choose a pattern at random and discard the configurations $\{J_i\}$ which do not give the correct response. The process is repeated until no configurations are left; this occurs, by definition, after $P_c + 1$ patterns have been used. The value $\alpha_c(N)$ is then the average of P_c/N for many samples. The number of samples used varied with size from 100 000 for the smaller systems to 4000 for the largest systems. Figure 2

shows $\alpha_c(N)$ against $1/N$ for (*a*) the binary perceptron with Ising and Gaussian patterns and for (*b*) the discrete spherical model with Ising and Gaussian patterns as well as the analytic calculation of $\alpha_c(N)$ for the toy model (shown on both graphs (*a*) and (*b*)). The error bars show the statistical errors in the mean calculated in the standard way. The data points are fitted with a best quadratic fit in $1/N$ as a guide for the reader. The simulations were performed using a combination of bit encoding, Gray code [21] and vectorization where appropriate.

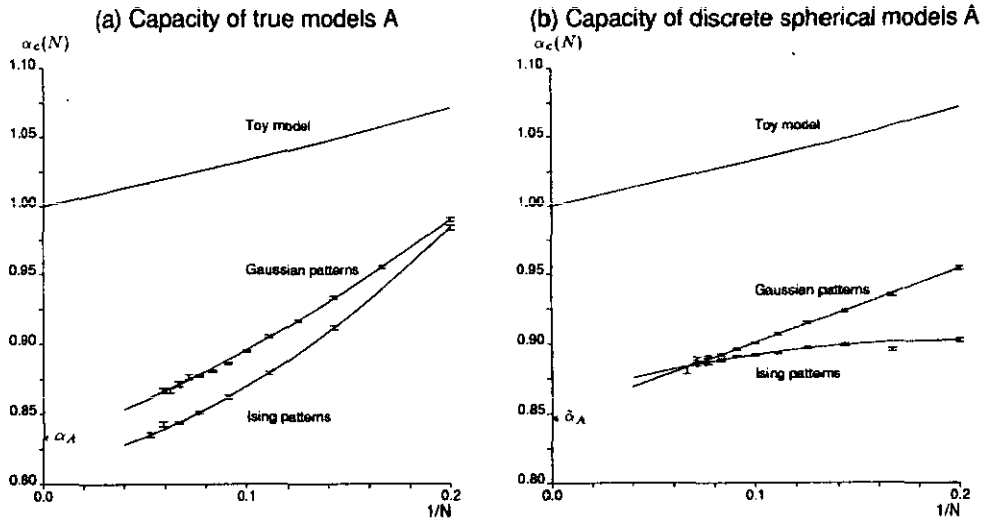

Figure 2. Capacity curves $\alpha_c(N)$ against $1/N$ calculated for toy model A, and for (*a*) the binary and (*b*) the discrete spherical versions of model A with both Ising and Gaussian patterns. The details of the methods used to obtain these curves are described in the text.

The convergence of $\alpha_c(N)$ as a function of $1/N$ seems to resemble that of toy model A in all cases. In the case of Gaussian patterns the results look as if they will extrapolate rather well to the predicted value of 0.833 for binary couplings and 0.847 for the discrete spherical case. These values are indicated as α_A and $\hat{\alpha}_A$, respectively, on the ordinate. The situation is more worrisome for the case of Ising patterns. For the discrete spherical couplings, the data cross those for the Gaussian patterns and look as if they could extrapolate to a higher value. For the binary couplings, it is hard to imagine that the data will extrapolate to a value as large as 0.833, although it may be closer to this figure than to the 0.75 estimated previously by the same method [18]. If one accepts the idea, based on the replica calculation, that Ising and Gaussian patterns yield the same α_c in the large-N limit, there must be very substantial finite-size effects for N larger than 20, and these cannot, of course, be excluded by our calculation. Another possible source of difficulty in the extrapolation is that oscillations with N could be present (see §3) but there is no firm evidence for this in our data for $\alpha_c(N)$.

Another prediction of the Krauth–Mézard theory [19] is that, for $\alpha > \alpha_c$, as one lowers the temperature, there is a complete freezing at a transition temperature T_c. At that temperature the specific heat jumps discontinuously to zero and remains there for the whole low-temperature phase.

In figure 3(*a*) we show numerical results obtained by exact enumeration of the

specific heat per coupling at $\alpha = 2$, for model A with Ising patterns, together with the replica prediction (broken curve). Because the enumeration method prohibits the study of large systems, our results are limited to $N \leqslant 21$. For each sample we compute the degeneracy of energy levels $\mathcal{N}(E)$ and use these histograms to produce the specific heat curves for different temperatures. For the true model 5000 samples were used. The data points show the calculated average values with the usual sample error estimates (the curve through the points are fitted by a cubic spline).

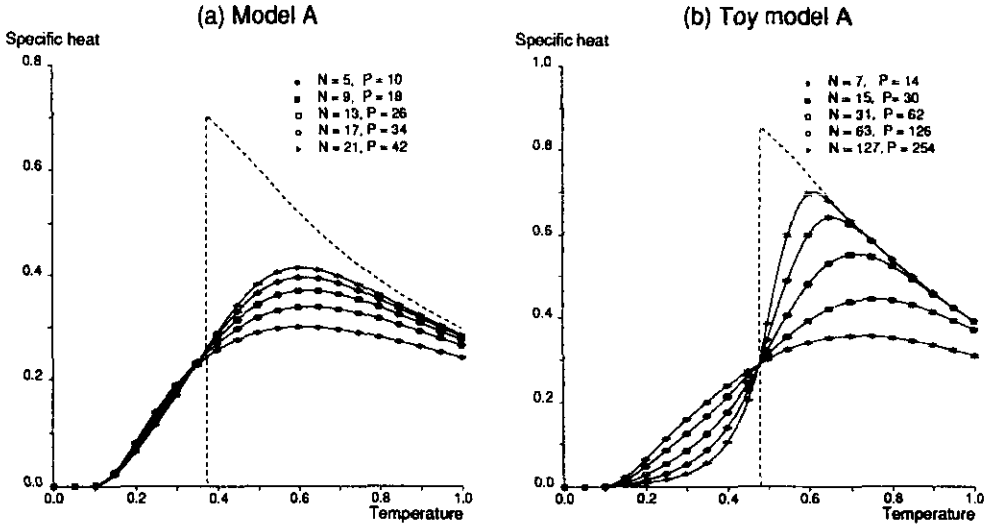

Figure 3. The specific heat per coupling against temperature at $\alpha = 2$ for (a) the true model A and (b) the toy model A for finite sizes and with the exact solution (broken curve).

While at first sight our results might appear inconsistent with the replica predictions, a comparison with the simulation of toy model A in figure 3(b) shows a similar behaviour for systems of comparable size. In the toy model one can simulate much larger systems, because if the $\mathcal{N}(E)$ are treated as random variables (§3), and since E is discrete, the time needed to simulate a sample of size N is proportional to N rather than 2^N. Thus in figure 3(b), each curve is for a system twice as large as for the previous curve, whereas in (a) the largest system is only four times the size of the smallest.

The finite-size specific heat curves for toy model A all cross at the transition temperature of the infinite-N limit. Those of the true model A go through a common point at a temperature slightly less than the transition predicted by the replica calculation, although closer examination suggests a crossing point moving to slightly higher temperatures with increasing system size. Hence it is not implausible that were the simulations of much larger systems possible, the results would eventually approach the replica prediction.

If for both models the specific heat at T_c is discontinuous, as predicted by the exact solution for toy model A and by the replica calculation for the true model A, one expects the specific heat to satisfy the following finite-size scaling form:

$$c_N(T) = F(N^{-x}(T - T_c)) \tag{4.1}$$

valid for N large and T close to T_c. It seems possible that the finite-size scaling is actually the same in both models (the same exponent $x = \frac{1}{2}$ and the same function F) corresponding to the same universality class. This would be consistent with the fact that the crossing point in both cases is close to $(1-2/\pi)$ times the total discontinuity in the specific heat.

However, there is at least one feature which seems to be different in the low-temperature behaviour of model A and toy model A, reflecting the existence of correlations of the energies which are present in the true model. In the case of the toy model A we have seen that to describe the low-temperature phase one could replace the distribution of energies by an exponential distribution. As a consequence one expects that the sample to sample fluctuations—defined as $(\langle E^2\rangle - \langle E\rangle^2)^{1/2}$, where the angular brackets denote average over samples—of the ground-state energy do not increase as $N \to \infty$. If the same picture were valid for the true model A one would expect the sample to sample fluctuations not to increase with N. We observe, however, that these fluctuations do in fact increase with N typically like $N^{1/3}$ (this behaviour is similar to that found in the Sherrington–Kirkpatrick spin-glass model [36, 37]). This is shown in figure 4: the curves are best fits of the form aN^x. This means that the picture of an exponential distribution has to be modified, one possibility being that the position of the exponential distribution shifts from sample to sample.

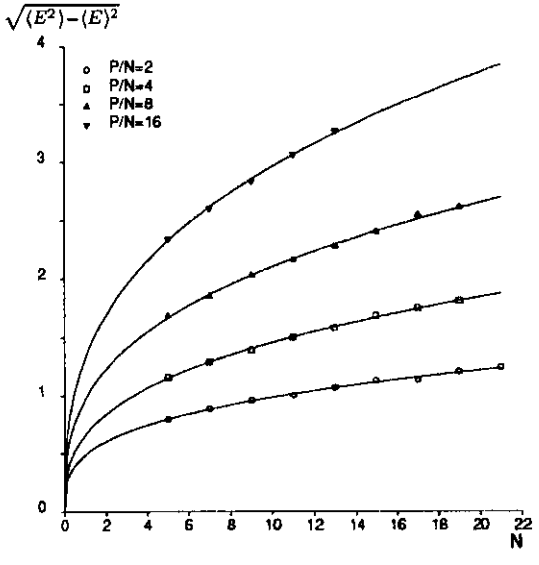

Figure 4. Sample to sample fluctuations of the energy plotted against N for model A.

The susceptibility and its sample to sample fluctuations were also studied. Unfortunately due to the limited sizes of systems that could be examined and the subtle nature of the transition (i.e. a cusp), these quantities did not prove to be good indicators of the transition, although they were consistent with a spin-glass-like transition.

4.2. Model B

We have repeated the numerical calculation of the critical capacity and of the thermal

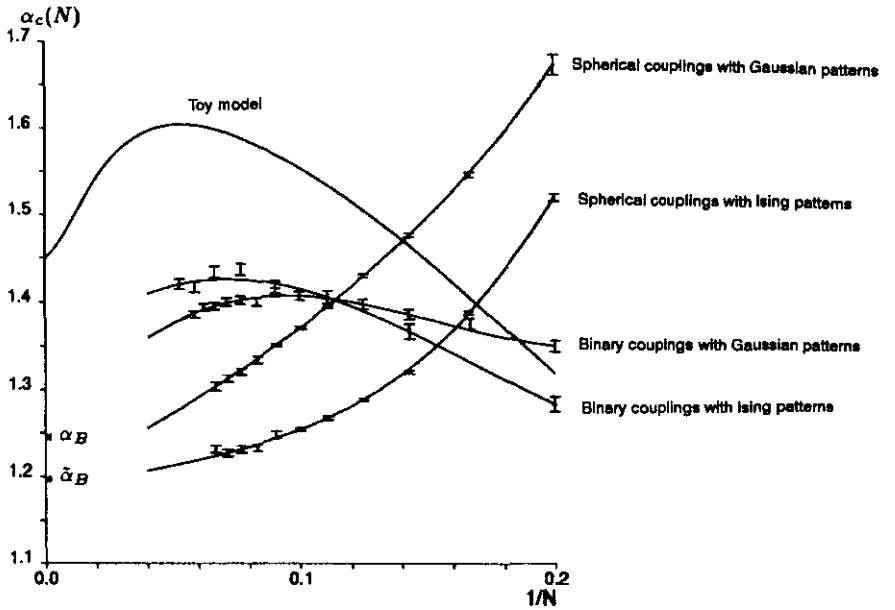

Figure 5. Capacity curves $\alpha_c(N)$ against $1/N$ calculated for toy model B, and for the binary and the discrete spherical versions of model B with both Ising and Gaussian patterns. The details of the methods used to obtain these curves are described in the text.

properties in the case of model B.

In figure 5 we show the critical value $\alpha_c(N)$ against $1/N$ for various versions of model B. The critical capacity was calculated using the method described in §3, (3.45). In the case of the discrete spherical model with Ising patterns it is possible to have a situation in which two configurations are never separated by any of the 2^N possible hyperplanes, and consequently the sum in (3.45) diverges. To avoid this problem we replace the definition of $\langle P_c \rangle$, in this case, by the value of P where $\langle y_P \rangle = \frac{1}{2}$, as determined by linear interpolation between successive integer values of P. As in figure 2, the error bars show the statistical errors in the mean.

In the case of the discrete spherical model with both Ising and Gaussian patterns, it is easy to imagine an extrapolation of the curves to the value of $\tilde{\alpha}_B = 1.197$ obtained in §2 on the basis of Gardner's replica calculation, and indicated on the ordinate. On the other hand, the results with binary couplings look as if they will extrapolate to distinct values substantially above the replica prediction of $\alpha_B = 1.245$. However, the curve for toy model B, (3.45), shows large finite-size effects, and were an extrapolation based on the part corresponding to $N \leqslant 20$, the estimated α_c would be near 1.6 instead of the correct 1.448. If a similar effect is present for binary couplings, it is not difficult to imagine an extrapolation of both the Ising and the Gaussian cases to a value near the replica prediction.

Turning to thermal properties we note that, for $\alpha > \alpha_c$, the calculation for the toy model B as well as the replica calculation [26, 27] for the true model B predict a first-order phase transition. For toy model B, the first-order phase transition appears as a jump in the energy curve, for example, as a function of the temperature. For finite systems this jump is rounded and the curves corresponding to different sizes all cross at the transition temperature T_c as seen in figure 6(*b*). The curves for the

true model B are shown in figure 6(a). We observe that the same crossing behaviour seems to hold (although once again we are limited to examining small systems).

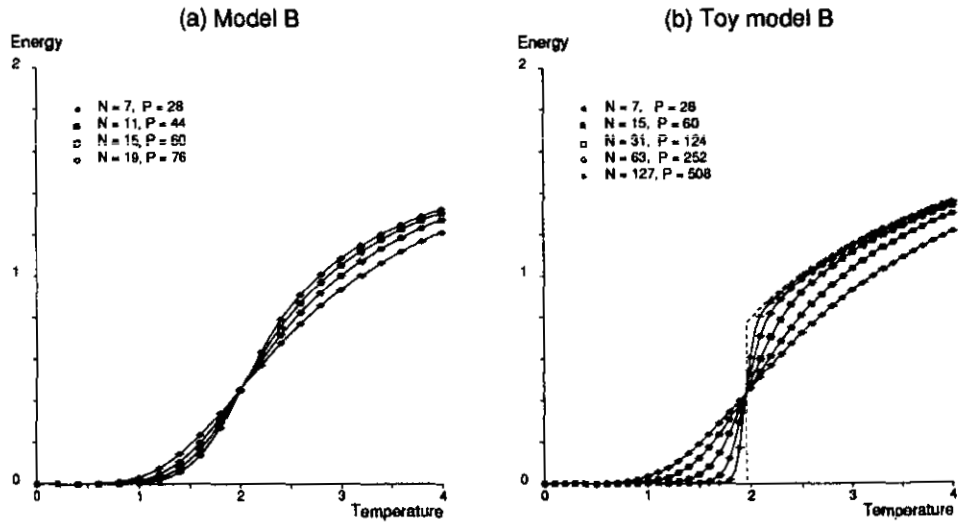

Figure 6. The energy per coupling against temperature with $\alpha = 4$ for (a) the true model B and (b) the toy model B.

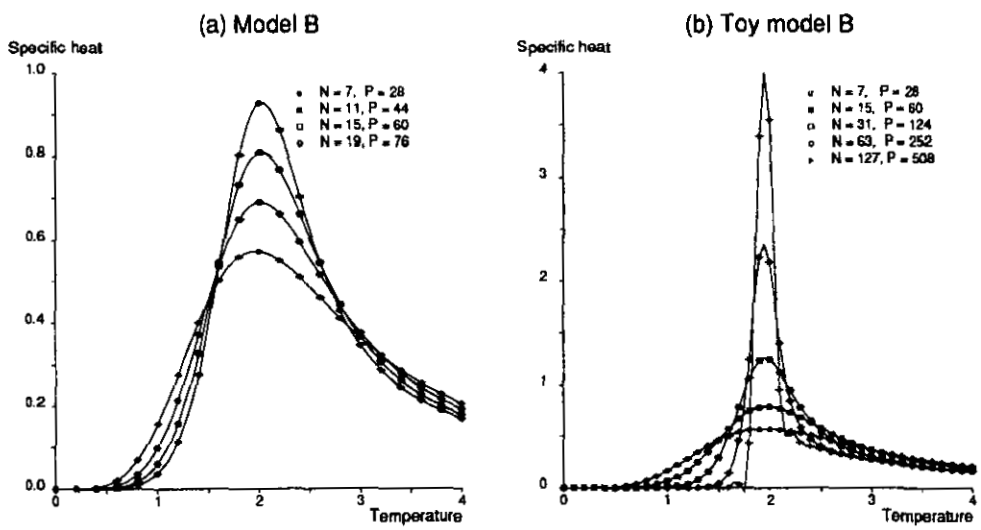

Figure 7. The specific heat per coupling against temperature with $\alpha = 4$ for (a) the true model B and (b) the toy model B. Note the difference in scales. The apparent negative specific heat is an artefact of the cubic spline curve fitting.

As well as a jump in the energy a first-order transition gives rise to a delta peak in the specific heat. For small systems this peak becomes rounded. Figure 7 shows the curves for the specific heat for various different system sizes for (a) the true model and (b) the toy model. For the true model the peak is apparent but not very pronounced. In the toy model where much larger sized systems can be studied the

peak is very much clearer. Notice that there is an apparent common crossing point for the true model around $T = 1.5$, but comparison with the toy model B—where the crossing points are observed to move towards higher temperatures—suggests that this is unlikely to be a true fixed point but is an artefact of studying systems very close to each other in size.

5. Bounds for α_A and α_B

Let α_A and α_B be the values of α for the A and B transitions in the case of binary couplings. As the number of particles is 2^N and the number of cells is 2^P, it follows that

$$\alpha_A \leqslant 1 \leqslant \alpha_B \tag{5.1}$$

(for $\alpha > 1$, the number of cells is much larger than the number of particles, and so most cells must be empty, whereas for $\alpha < 1$ the reverse is true, and thus most particles must be in cells which contain other particles). The first inequality in (5.1) also follows, as noted in §3, from the fact that the critical α for toy model A is an upper bound on α_A. An additional bound

$$\alpha_B \leqslant 1.448 \tag{5.2}$$

was obtained by Gardner and Derrida in [18, equation (26)]. It is also a consequence of the fact, also pointed out in §3, that the right-hand side of (5.2) is the critical α for toy model B. Note that (5.1) and (5.2) hold both for Ising and Gaussian patterns.

In the case of Gaussian (but not Ising) patterns it is possible to establish two additional inequalities:

$$\alpha_A \leqslant \tilde{\alpha}_A \tag{5.3}$$

$$\alpha_B \geqslant \tilde{\alpha}_B \tag{5.4}$$

where $\tilde{\alpha}_A$ and $\tilde{\alpha}_B$ are the values of α for the A and B transitions for discrete spherical couplings. In fact, the values of $\tilde{\alpha}_A$ and $\tilde{\alpha}_B$ are determined by properties of the distribution of cell sizes (§2), and Gardner's calculations [6, 18] using replica methods, yielding the values

$$\tilde{\alpha}_A = 0.847 \qquad \tilde{\alpha}_B = 1.197. \tag{5.5}$$

To be sure, one must always be cautious regarding the results of replica calculations, but assuming (5.5) is correct, the bounds (5.3) and (5.4) are a substantial improvement over (5.1). A comparison with our numerical estimates has already been made in §4 above. That $\tilde{\alpha}_A$ is close to the value α_A obtained by alternative methods was noted in [19].

In addition, the arguments leading to (5.3) and (5.4) are of interest because of their general character: they employ no property of the binary couplings apart from the fact that there are 2^N of them. That is, $\tilde{\alpha}_A$ is an upper bound on α_A for any model with 2^N configurations and Gaussian patterns, and similarly $\tilde{\alpha}_B$ is a general

lower bound for α_B. This generality is reflected in the statement of the two theorems, A and B below, in which (5.3) and (5.4) are established. Furthermore, in order to emphasize their geometrical character, the theorems are expressed in terms of two hypotheses, HA and HB, the latter in two versions, related to the distribution of cell sizes introduced in §2.

In HA and HB, and in the theorems, we assume that a sequence of values P and N is given, tending to infinity, with P/N tending to a specified value α. The notation $\Pr(\mathcal{E})$ stands for the probability of the event $\mathcal{E}$. Volumes of cells, denoted by v, correspond to the normalization (2.6). The hypotheses and theorems apply equally well to any cell $\boldsymbol{R} = \{R^\mu\}$, by symmetry of the probability distribution or measure for S, denoted by $\nu(S)$, even though it may be convenient to think of a particular cell, say $\boldsymbol{R} = \mathbf{1}$, which means $R^\mu = 1$ for all μ. Hypothesis HB is stated in two versions which are actually equivalent because of the spherical symmetry of $\nu(S)$; the reason for doing so is to make explicit the point where this symmetry enters the argument, and thus the gap which has thus far prevented us from extending the argument (5.4) to the case of Ising patterns.

Hypothesis HA. Let v be the volume of the cell $\boldsymbol{R} = \mathbf{1}$. Then there are positive numbers ϵ and η, depending on N and tending to zero as $N \to \infty$, such that

$$\Pr(v \geqslant \epsilon 2^{-N}) < \eta. \tag{5.6}$$

Hypothesis HB1. Given any point specific $\boldsymbol{J}$ on the sphere (2.4), let v be the volume of the cell which it occupies. (The probability that $\boldsymbol{J}$ falls on the boundary between two or more cells is zero.) Then there are positive numbers ϵ and η, depending on N and tending to zero as $N \to \infty$, such that

$$\Pr(v \leqslant \epsilon^{-1} 2^{-N}) < \eta. \tag{5.7}$$

Hypothesis HB2. There is for each N at least one specific point $\boldsymbol{J}$ on the sphere (2.4) for which (5.7) holds.

Note that in both HB1 and HB2 the point $\boldsymbol{J}$ remains fixed while S varies. Also note that HA and the pair HB are independent in the sense that the values of α where one holds have no necessary relationship with the values where the other holds. Clearly HB1 implies HB2, and the spherical symmetry of $\nu(S)$ for Gaussian planes implies that the two are equivalent, for the probability is independent of the point chosen (see appendix 3).

Theorem A. (a) If for some α, HA is satisfied, and if for each N, $\mathcal{C}_N$ is some arbitrary collection of 2^N points on the sphere (2.4), then, given that $\nu(S)$ is invariant under rotations, the probability that the cell $\boldsymbol{R} = \mathbf{1}$ (or any other particular cell) contains at least one point from $\mathcal{C}_N$ tends to zero as $N \to \infty$.
(b) For $\alpha > \tilde{\alpha}_A$ (transition A for the discrete spherical couplings), condition HA is satisfied.
(c) In the case of Gaussian patterns, $\tilde{\alpha}_A$ is the infimum of those α for which HA holds, and consequently $\alpha_A \leqslant \tilde{\alpha}_A$.

Note that part (b) of the theorem does not require the spherical symmetry of $\nu(S)$, so it is also valid for the case of Ising patterns.

In stating theorem B, the following terminology is helpful. A member of a collection $\mathcal{C}_N$ of points on the sphere (2.4) is said to be *isolated* for a given set of hyperplanes S if it falls in a cell which contains no other points from $\mathcal{C}_N$.

Theorem B. (a) If for some α, HB1 is satisfied, and $\mathcal{C}_N$ is for each N some arbitrary collection of 2^N points on the sphere (2.4), then the fraction f of isolated points in $\mathcal{C}_N$ tends to zero with probability one as $N \to \infty$. That is there are positive numbers $\tilde{\epsilon}$ and $\tilde{\eta}$ tending to zero as $N \to \infty$ such that

$$\Pr(f \geqslant \tilde{\epsilon}) < \tilde{\eta}. \tag{5.8}$$

(b) For $\alpha < \tilde{\alpha}_B$ (transition B for discrete spherical couplings), condition HB2 is satisfied.
(c) In the case of Gaussian patterns, $\tilde{\alpha}_B$ is the supremum of those α for which HB1 (equivalent to HB2) is satisfied, and consequently $\alpha_B \geqslant \tilde{\alpha}_B$.

Note that neither (a) nor (b) require the spherical symmetry of $\nu(S)$, whereas it is needed for (c). We shall now indicate the intuitive ideas, which are actually quite simple, underlying the technical proofs of theorems A and B, which are in appendix 3.

To begin with (a) of theorem A, HA tells us that the volume of a typical cell—or the typical volume of a particular cell—is smaller than $\epsilon 2^{-N}$. This means that the *average* number of points from the collection $\mathcal{C}_N$ of 2^N points which fall in the cell is ϵ or less; the actual location of the points in $\mathcal{C}_N$ is irrelevant, because the distribution $\nu(S)$ is spherically symmetrical. The only way that this average can be small is if the typical cell is empty, which is the same (by symmetry among the cells) as saying that a particular cell is typically empty. For part (b), note that $\alpha > \tilde{\alpha}_A$ means that if $\mathcal{C}_N$ is a collection of 2^N points chosen at random on the sphere, a typical cell will be empty. Now because the points are chosen at random, the average number in a cell of volume v is $2^N v$, and because the distribution of the number of points in a cell of a given size is (essentially) Poisson, the only way a typical cell can be empty is if the average number of points from $\mathcal{C}_N$ which it contains is small, meaning that its volume is small. Hence $\alpha > \tilde{\alpha}_A$ implies HA.

Finally, part (c) is a consequence of noting that $\alpha > \tilde{\alpha}_A$ implies HA, but HA, by part (a), implies $\alpha \geqslant \tilde{\alpha}_A$, making $\tilde{\alpha}_A$ the lower limit of the α for which HA holds, and thus—applying (a) to the case where $\mathcal{C}_N$ is the collection of hypercube vertices—an upper bound on α_A.

The intuitive idea behind part (a) of theorem B is as follows. Given any collection $\mathcal{C}_N$ of 2^N points, hypothesis HB1 implies that most of them fall in cells which are relatively large, of volume greater that $\epsilon^{-1}2^{-N}$. But as the total volume of all the cells is 1, there are at most $\epsilon 2^N$ of these large cells, and it is obvious that among the points falling in the large cells, at most $\epsilon 2^N - 1$ of them, a small fraction of the total, can be isolated. (Given a hotel with m rooms and $M \gg m$ guests, it is clear that only a small fraction of the guests can be in rooms by themselves.)

For part (b), the reasoning is analogous to that of the same part of theorem A. For $\alpha < \tilde{\alpha}_B$, an extremely large fraction of the 2^N points chosen at random on the sphere fall in cells containing other points, a situation which is only possible (given Poisson statistics) if the average number of points in those cells containing at least one point is large, corresponding to the fact that the cells themselves are large. One thereby establishes a result which is actually somewhat stronger than required for HB2. However, to obtain HB1, and thus conclusion (c) of the theorem by reasoning entirely analogous to that employed at the corresponding point in theorem A, the spherical symmetry of $\nu(S)$ is important, which is why the argument fails for Ising patterns.

6. Conclusion

In this paper we have extended previous work on binary perceptron models in two directions. First, we have carried out numerical studies on systems of finite sizes using exact enumeration of states in order to extend previous estimates of the critical capacity α_A and α_B of models A and B, and to find the thermodynamic properties as a function of temperature, for both models. Second, we have introduced and solved a set of simplified models: the discrete spherical versions of models A and B, and the 'toy' models A and B. These simplified models are interesting because they provide bounds on the critical capacities of the original binary perceptron models and in addition because the finite-size effects can be computed in the toy models and compared with our numerical studies of the real models.

While the upper bounds on α_A and α_B provided by the toy models were known previously [18], those given by the discrete spherical models, an upper bound $\tilde{\alpha}_A$ for α_A and a lower bound $\tilde{\alpha}_B$ on α_B, are new. (In fact $\tilde{\alpha}_A$ was calculated previously [19]; the fact that it is an upper bound is new.) They are also a noticeable improvement on previous values, assuming that the actual numerical values provided by Gardner's replica calculation are correct. It seems likely that there are other contexts in which such discrete spherical bounds might be useful, in particular in cases where the energy function would be different (more complicated cost functions [38], other architectures [39, 40]). So far as we know, there is no rigorous lower bound (greater than zero) on α_A, and our efforts in this direction have proved futile. Finding a rigorous lower bound seems surprisingly difficult.

The finite-size properties of the toy models turn out to be extremely useful in interpreting our numerical results for the temperature dependence of thermodynamic properties of models A and B in systems of finite size. By comparing the toy and real models, one can make a very plausible argument for phase transitions into a 'frozen' low-temperature phase with vanishing entropy occurring at a finite temperature: a continuous (second-order) transition for model A and a first-order transition for model B, provided α exceeds the corresponding critical capacity. Such transitions have been proposed on the basis of replica calculations by Krauth and Mézard [19] for model A and by Györgyi [26] and Sompolinsky, Tishby and Seung [27] for model B, but given the usual uncertainty about the limits of validity of replica studies, we think that our numerical calculations as interpreted with the help of the corresponding toy models provide an important confirmation.

Our estimates of the critical capacities α_A and α_B based on numerical studies of small systems, while they have been extended to larger systems than previously studied (up to $N \simeq 21$), are somewhat disappointing in that the finite-size corrections are still not very well understood and make it difficult to obtain any precise extrapolation to $N = \infty$. The numerical evidence for model A with Ising patterns taken by itself suggests an α_A which is significantly less than that for Gaussian patterns, and although we cannot exclude the possibility that the difference is solely a consequence of finite-size effects, it is worth pointing out that there are no completely compelling arguments for their equality in the $N \to \infty$ limit.

The situation in the case of model B is, if anything, even worse. If one had only numerical evidence, one might plausibly suppose that Ising and Gaussian patterns give different values for α_B, and these well in excess of the replica estimate. However, the toy model B shows extremely large finite-size effects of a type which would make a reliable extrapolation based on a maximum N of only 21 out of the question. If

the real model has a similar behaviour, it is easy to imagine the curves of figure 5 bending over for larger N and reaching the replica value.

By contrast, the corresponding extrapolation for the discrete spherical versions of model A and B are consistent with a smooth extrapolation as $N \to \infty$ to the $\tilde{\alpha}_A$ and $\tilde{\alpha}_B$ expected from the replica calculations, and with results identical (in the same limit) for Ising and Gaussian planes.

It seems clear that further progress in the direction of estimating critical capacities from the study of finite systems would benefit from a better understanding of finite-size effects.

Acknowledgments

The research of one of the authors (RBG) has received financial support from the US National Science Foundation under grant DMR–9009474.

Appendix 1. Fluctuations in magnetic susceptibility

In this appendix we consider fluctuations in the magnetic susceptibility χ (suitably defined) for toy model A and for the random energy model [24], in their low-temperature phases.

To begin with, we suppose that 2^N configurations, labelled by subscripts a and b, have energies E_a which depend on the sample, and thus the Boltzmann weights

$$W_a = \frac{e^{-E_a/T}}{\sum_b e^{-E_b/T}} \tag{A1.1}$$

are random variables. The configurations are then independently assigned random magnetizations M_a between $-N$ and N according to the probability distribution

$$\Pr(M_a) = 2^{-N} \binom{N}{(N+M_a)/2}. \tag{A1.2}$$

For large N, $M_a/\sqrt{N}$ is, to a good approximation, a Gaussian random variable with zero mean and unit variance, which means that

$$\langle M_a \rangle = \langle M_a^3 \rangle = 0$$
$$\langle M_a^2 \rangle = N \qquad \langle M_a^4 \rangle = 3N^2 \tag{A1.3}$$

while M_a and M_b for $a \neq b$ are uncorrelated. Note that M_a and W_a are independent random variables.

The susceptibility χ is defined in terms of the thermal fluctuations in the magnetization,

$$\chi = \frac{1}{NT}\left[\sum_a M_a^2 W_a - \left(\sum_a M_a W_a\right)^2\right] \tag{A1.4}$$

and is, of course, sample dependent. Its average and variance are easily calculated using (A1.3); one finds:

$$\langle \chi \rangle = \frac{1}{T} \sum_a [\langle W_a \rangle - \langle W_a^2 \rangle] = \frac{1}{T}(1 - \langle Y_2 \rangle) \tag{A1.5}$$

$$\langle \chi^2 \rangle - \langle \chi \rangle^2 = 2\langle Y_2 \rangle - 4\langle Y_3 \rangle + 3\langle Y_2^2 \rangle - \langle Y_2 \rangle^2 \tag{A1.6}$$

where the variables Y_k are defined by

$$Y_k = \sum_a W_a^k. \tag{A1.7}$$

Note that

$$Y_k = \frac{\sum_E \mathcal{N}(E) \mathrm{e}^{-kE/T}}{\left[\sum_E \mathcal{N}(E) \mathrm{e}^{-E/T}\right]^k} \tag{A1.8}$$

where $\mathcal{N}(E)$ is the number of configurations of energy E. Using the integral representation

$$z^{-k} = \frac{1}{\Gamma(k)} \int_0^\infty t^{k-1} \mathrm{e}^{-tz} \mathrm{d}t \tag{A1.9}$$

along with the fact that in the case of toy model A and the random energy model the $\mathcal{N}(E)$ are independent Poisson variables, one arrives at the expression

$$\langle Y_k \rangle = \frac{1}{\Gamma(k)} \int_0^\infty t^{k-1} \left(\sum_E \langle \mathcal{N}(E) \rangle \exp\left(- kE/T - t\mathrm{e}^{-E/T} \right) \right)$$
$$\times \exp\left(- \sum_{E'} \langle \mathcal{N}(E') \rangle \left[1 - \exp\left(-t\mathrm{e}^{-E'/T} \right) \right] \right). \tag{A1.10}$$

A similar procedure can be applied to obtain averages of moments of Y^k or the products $Y_k Y_l$, etc.

We now consider the low-temperature phase $T < T_c$, where the exponential approximation (3.20) for $\langle \mathcal{N}(E) \rangle$ can be employed for toy model A. The resulting expressions are rather complicated. They simplify considerably if one considers the random energy model, in which E is a continuous variable not restricted to integer values, and where at low temperature the $\mathcal{N}(E)$ are again independent Poisson variables, with the average number of configurations with an energy between E and $E + \Delta E$ given by

$$\langle \mathcal{N}(E) \rangle = \Delta E \, \mathrm{e}^{(E - E_c)/T_c}. \tag{A1.11}$$

In this case one finds

$$\langle Y_k \rangle = \frac{\Gamma(k - T/T_c)}{\Gamma(k)\Gamma(1 - T/T_c)}. \tag{A1.12}$$

and

$$\langle Y_2^2 \rangle = \frac{1}{3}\left[3 - 5\frac{T}{T_c} + 2\left(\frac{T}{T_c}\right)^2\right] \tag{A1.13}$$

and thus

$$\langle \chi \rangle = \frac{1}{T_c} \tag{A1.14}$$

$$\langle \chi^2 \rangle - \langle \chi \rangle^2 = \frac{T}{T_c}\left(1 - \frac{T}{T_c}\right). \tag{A1.15}$$

Hence for $T < T_c$ in the random energy model there are macroscopic (order 1) fluctuations in the susceptibility from sample to sample, while the average susceptibility remains constant. In toy model A, there are also macroscopic fluctuations in χ for $T < T_c$. However, in addition $\langle \chi \rangle$ diverges as T goes to zero, due to the fact (not true for the random energy model) that the ground state can be degenerate.

Appendix 2. Discussion of (3.28) and (3.29)

If one inserts (3.28) in (3.27), the result is a value of $-\infty$ for $\langle \ln Z \rangle$ due to the fact that as $t \to \infty$ the right-hand side of (3.28) does not go to zero, but tends to an exceedingly small constant. This is a spurious effect arising from the fact that the Poisson approximation allows $\mathcal{N}(E)$ to be simultaneously 0 for all E (yielding $Z = 0$ because there are no configurations) with an astronomically small probability. The cure is simply to cut the integral off for some large value of t; as shown in the analysis below, the result is essentially independent of the cut-off in a suitable range.

The justification of (3.29) is subtler because of the fact that one must show that it is adequate for all the t values which dominate the integral (3.27), as well as for derivatives of this integral with respect to temperature, for the latter are used in §3 for calculating the heat capacity. For this purpose it is convenient to replace t with the variable

$$\tau = te^{-\beta E_c} \tag{A2.1}$$

where $\beta = 1/T$ and $\beta_c = 1/T_c$ in what follows. Then (3.27) becomes

$$\langle \ln Z \rangle + \beta E_c = \int_0^\infty \frac{d\tau}{\tau}\left[e^{-\tau} - e^{-z(\tau)}\right] \tag{A2.2}$$

where

$$z(\tau) = \sqrt{N}\int_{-\infty}^{\infty} d\varphi\left[1 - \exp\left(-\tau e^{-\sqrt{N}\beta\varphi}\right)\right]e^{\sqrt{N}\varphi\beta_c - b\varphi^2} \tag{A2.3}$$

results from replacing the sum over E in (3.28) with a corresponding integral. Note that $z(\tau)$ is monotone increasing, $z(0) = 0$, and its derivative

$$z'(0) = \sqrt{\frac{N\pi}{b}}\exp\left(\frac{N(\beta - \beta_c)^2}{4b}\right) \tag{A2.4}$$

at $\tau = 0$ is an upper bound on $z'(\tau)$ for all $\tau > 0$. In the following analysis *we will always assume that* $|\beta - \beta_c|$ *is at most of order* $1/\sqrt{N}$. Given this assumption, $z'(0)$ is of order $\sqrt{N}$.

It will be convenient to split the integration interval in (A2.3) into two pieces, $\varphi \geqslant \varphi_0(\tau)$ and $\varphi \leqslant \varphi_0(\tau)$, where

$$\varphi_0(\tau) = \frac{\ln \tau}{\beta\sqrt{N}} \tag{A2.5}$$

i.e. the value of φ where

$$\tau e^{-\sqrt{N}\beta\varphi} = 1. \tag{A2.6}$$

We then use the approximation

$$1 - \exp(-\tau e^{-\sqrt{N}\beta\varphi}) \simeq \begin{cases} \tau e^{-\sqrt{N}\beta\varphi} & \varphi \geqslant \varphi_0(\tau) \\ 1 & \varphi \leqslant \varphi_0(\tau) \end{cases} \tag{A2.7}$$

to obtain $z(\tau)$ as (approximately) the sum of

$$\bar{z}_+(\tau) = \tau\sqrt{N} \int_{\varphi_0(\tau)}^{\infty} d\varphi \exp[\sqrt{N}(\beta_c - \beta)\varphi - b\varphi^2] \tag{A2.8}$$

and

$$\bar{z}_-(\tau) = \sqrt{N} \int_{-\infty}^{\varphi_0(\tau)} d\varphi \exp(\sqrt{N}\beta_c\varphi - b\varphi^2). \tag{A2.9}$$

In fact, this sum is an upper bound on $z(\tau)$ because the right-hand side of (A2.7) is always larger than the left-hand side for $\tau \geqslant 0$. Setting $b = 0$ in (A2.9) produces an upper bound

$$z_-(\tau) = \beta_c^{-1} e^{\sqrt{N}\varphi_0(\tau)\beta_c} = \beta_c^{-1} \tau^{\beta_c/\beta} \tag{A2.10}$$

for $\bar{z}_-(\tau)$, and we will later show that this is small compared with (A2.8).

To estimate the error involved in using the approximation (A2.7) in the integral (A2.8), note that as soon as φ exceeds φ_0 by $(\ln N)/\sqrt{N}$, $\tau e^{-\sqrt{N}\beta\varphi}$ will be at most $N^{-\beta}$. Hence the fractional error from this source is of order $(\ln N)/\sqrt{N}$; remember that $\beta_c - \beta$ is at most $1/\sqrt{N}$. A further error of this same magnitude will occur if the lower limit in (A2.8) is set to zero yielding

$$z_+(\tau) = \tau\sqrt{N} J_0 \tag{A2.11}$$

with

$$J_0 = \int_0^{\infty} d\varphi \exp[\sqrt{N}(\beta_c - \beta)\varphi - b\varphi^2] \tag{A2.12}$$

provided $|\varphi_0(\tau)|$ is at most of order $(\ln N)/\sqrt{N}$, a condition which holds provided

$$N^{-q} \leqslant \tau \leqslant N^{Q} \tag{A2.13}$$

for some q and Q positive and not too large.

In order that (A2.10) be small compared to (A2.11), noting that J_0 is of order 1, τ must not be too small, assuming $\beta > \beta_c$, or too large assuming $\beta < \beta_c$. The crossover where the two are comparable comes at

$$\tau \simeq N^{\beta_c/2(\beta_c-\beta)} \tag{A2.14}$$

which means that we can always choose $q > 1/2$ and $Q > 0$ in (A2.13). We conclude that in the range (A2.13), $z(\tau)$ can be replaced by $z_+(\tau)$, (A2.11), with fractional errors which go to zero with increasing N, and hence the right-hand side of (A2.2) is given by $\ln(J_0\sqrt{N})$ plus correction terms going to zero with N, provided the contribution from τ outside the limits (A2.13) also go to zero with N.

The bounds

$$0 \leqslant z(\tau) \leqslant \tau z'(0) \tag{A2.15}$$

(see (A2.4)), together with the inequality

$$0 \leqslant (e^{-a\tau} - e^{-b\tau}) \leqslant (b-a)\tau \tag{A2.16}$$

valid for any

$$0 \leqslant a \leqslant b \qquad \tau \geqslant 0 \tag{A2.17}$$

can be used to bound the part of the integral (A2.2) corresponding to $0 \leqslant \tau \leqslant N^{-q}$ by a quantity of order $N^{-q-1/2}$, which goes to zero for $q > 1/2$. As for $\tau > N^Q$, note that the contribution from $e^{-\tau}$ is of order $\exp(-N^Q)$, and that from $e^{-z(\tau)}$ is even smaller for any reasonable cut-off (needed for reasons noted in the introductory paragraph).

Bounds on the errors in calculating $\langle \ln Z \rangle$ using the approximation (A2.11) for $z(\tau)$ are not necessarily valid for its derivatives; note that $\ln J_0$, (A2.12), is of order 1, and its second derivative with respect to β is of order N. However, the integrals obtained by differentiating the right-hand side of (A2.2) one or more times always contains an $e^{-z(\tau)}$ in the integrand, and thus the key estimates needed to show that the error terms are relatively small are already contained in the preceding discussion.

Appendix 3. Proof of results in §5

We present here the proofs that HB2 implies HB1 if the probability distribution $\nu(S)$ has spherical symmetry, as well as the detailed proofs of theorems A and B.

A3.1. Equivalence between HB1 and HB2

By spherical symmetry of ν we mean the following. Let R be an element of the rotation group in $\mathbb{R}^N$, i.e. an $N \times N$ orthogonal matrix, and let

$$RS = \{RS^\mu\} \tag{A3.1}$$

be the set of hyperplane normals obtained by applying R to each S^μ for $1 \leqslant \mu \leqslant P$. Then ν is an invariant probability measure in the sense that

$$\nu(RS) = \nu(S). \tag{A3.2}$$

Note that in the case of Gaussian patterns, (A3.2) holds for an arbitrary rotation, while for Ising patterns it is only true for the subgroup of rotations which map all hypercube vertices into other hypercube vertices.

Let $\chi_R(J, S)$ be 1 if the point J on the sphere (2.4) falls in the cell $R = \{R^\mu\}$ for the set of hyperplanes S, and 0 otherwise. Clearly the property of being in a given cell is preserved if both J and the hyperplanes are subject to the same rotation,

$$\chi_R(RJ, RS) = \chi_R(J, S) \tag{A3.3}$$

as is also the volume of the cell,

$$v_R(S) = \int \mu(J)\chi_R(J, S) = v_R(RS) \tag{A3.4}$$

where $\mu(J)$ denotes the spherically symmetric normalized measure.

Thus if hypothesis HB2 holds, so that (5.7) is correct for one point J on the sphere, the invariance of ν, (A3.2), means that (5.7) holds for any other point J', as there is always some rotation R such that $J' = RJ$. Consequently (A3.2) implies the equivalence of HB1 and HB2.

A3.2. Proof of theorem A

Theorem A is proved as follows. For part (a), let $\mathcal{E}$ be the event that some J from the collection $\mathcal{C}_N$ falls in the cell R and that this cell has a volume $v_R < v_0$. Its probability is

$$\begin{aligned} \Pr(\mathcal{E}) &= \int \nu(S)\chi_R(J, S)\theta(v_0 - v_R(S)) \\ &= \int \nu(S)\chi_R(RJ, S)\theta(v_0 - v_R(S)) \end{aligned} \tag{A3.5}$$

where $\theta(x)$ is 1 for $x \geqslant 0$ and 0 for $x < 0$, and the second equality follows from (A3.3), (A3.4) and (A3.2). If (A3.5) is integrated over all rotations R using normalized Haar measure, the effect is the same as integrating over J with the measure $\mu(J)$, see (A3.4):

$$\Pr(\mathcal{E}) = \int \nu(S)v_R(S)\theta(v_0 - v_R(S)) \leqslant v_0 \Pr(v_R \leqslant v_0). \tag{A3.6}$$

If n_R is the number of points of $\mathcal{C}_N$ which fall in the cell R, we can write

$$\Pr(n_R > 0, v_R \leqslant v_0) \leqslant 2^N \Pr(\mathcal{E}) \leqslant 2^N v_0 \Pr(v_R \leqslant v_0) \tag{A3.7}$$

where the first inequality comes about by noting that the probability that n_R is positive, is certainly not larger than the average value of n_R, and the latter, with $v_R \leqslant v_0$, is $2^N \Pr(\mathcal{E})$ since (A3.6) is independent of the point J in the collection $\mathcal{C}_N$. What we are interested in is, of course,

$$\Pr(n_R > 0) = \Pr(n_R > 0, v_R \leqslant v_0) + \Pr(n_R > 0, v_R > v_0). \tag{A3.8}$$

If $v_0 = \epsilon 2^{-N}$, then (A3.7) tells us that the first term on the right-hand side of (A3.8) is no bigger than ϵ, while (5.6) gives η as an upper bound on the second term. Thus $\Pr(n_R > 0)$ is less than $\epsilon + \eta$, which tends to zero as $N \to \infty$, by hypothesis HA.

For part (b) of theorem A, note that $\alpha > \tilde{\alpha}_A$ means that for any given cell R, $\Pr(n_R > 0)$ for the discrete spherical model goes to zero as $N \to \infty$, and thus given some $\epsilon > 0$ we can be sure that

$$\Pr(n_R > 0) < \epsilon^2/2 \tag{A3.9}$$

for all N sufficiently large. Since for this model the 2^N points are chosen at random on the sphere, we have

$$\Pr(n_R > 0 | v_R) = 1 - (1 - v_R)^{2^N} \tag{A3.10}$$

for the conditional probability of a non-empty cell given a cell volume v_R.

Noting that the right-hand side of (A3.10) is monotone increasing in v_R and ignoring cases with $v_R < \epsilon 2^{-N}$, we obtain the inequality

$$\Pr(n_R > 0) \geqslant [1 - (1 - \epsilon 2^{-N})^{2^N}] \Pr(v_R \geqslant \epsilon 2^{-N}). \tag{A3.11}$$

Combining (A3.9) with (A3.11), and assuming N is large enough so that

$$(1 - \epsilon 2^{-N})^{2^N} \simeq e^{-\epsilon} \tag{A3.12}$$

and that ϵ is less than 1, we obtain

$$\Pr(v_R \geqslant \epsilon 2^{-N}) < \epsilon \tag{A3.13}$$

which is to say, (5.6) with $\eta = \epsilon$.

Part (c) of theorem A is a consequence, as noted in §5, of first choosing C_N in part (a) to be the 2^N discrete spherical couplings, and then the 2^N hypercube vertices.

A3.3. Proof of theorem B

For theorem B, part (a), it is convenient to introduce random variables I_j and K_j, $1 \leqslant j \leqslant 2^N$, associated with the points of the collection C_N in the following way: I_j is 1 if point j is isolated and zero otherwise: K_j is 1 if j is in a cell with volume $v \leqslant \epsilon^{-1} 2^{-N}$, and zero otherwise. The inequality

$$2^N f = \sum_j I_j \leqslant \sum_j K_j + \epsilon 2^N \tag{A3.14}$$

where f is the fraction of isolated points, comes about by noting that either j is in a 'small' cell where $K_j = 1$, whence $I_j \leqslant K_j$, or it is in a 'large' cell with $K_j = 0$. As the total volume of all cells is 1, there are at most $\epsilon 2^N$ isolated particles in large cells.

Upon averaging (A3.14) and using (5.7) we conclude that

$$\langle f \rangle < \epsilon + \eta. \tag{A3.15}$$

As f is non-negative, for any $\zeta > 0$,

$$\Pr(f > \zeta) \leqslant (\epsilon + \eta)/\zeta \tag{A3.16}$$

and thus (5.8) holds with

$$\tilde{\epsilon} = \tilde{\eta} = \zeta = \sqrt{\epsilon + \eta}. \tag{A3.17}$$

Note that if $\mathcal{C}_N$ is the collection of hypercube vertices and $\nu(S)$ is invariant under all rotations, or at least a subgroup large enough to map any hypercube vertex into any other (as is the case for Ising patterns), the fact that $\langle I_j \rangle$ does not depend on j yields the additional result that the probability that any hypercube vertex is isolated does not exceed $\epsilon + \eta$.

For part (b) of theorem B, note that $\alpha < \tilde{\alpha}_B$ means that with $\mathcal{C}_N$ a set of 2^N points chosen at random on the sphere, the probability that a particular one, say $j = 1$, is isolated goes to zero as $N \to \infty$, and thus for some small $\epsilon > 0$ we can be sure that

$$\langle I_1 \rangle < \tfrac{1}{2}\epsilon e^{-1/\epsilon} \tag{A3.18}$$

for all N sufficiently large. Given that this point falls in a cell of volume v, the probability that it is isolated is

$$(1 - v)^{2^N - 1} \tag{A3.19}$$

as the remaining $2^N - 1$ random points lie outside the cell. Since (A3.19) is monotone decreasing in v, a lower bound for $\langle I_1 \rangle$ is

$$\langle I_1 \rangle > (1 - \epsilon^{-1}2^{-N})^{2^N - 1} \Pr(v \leqslant \epsilon^{-1}2^{-N}) \tag{A3.20}$$

if we simply ignore cases with $v > \epsilon^{-1}2^{-N}$. Combining (A3.18) and (A3.20), and making the appropriate exponential approximation (always assuming N is not too small), yields the inequality (5.7), and thus the hypothesis HB2, with $\eta = \epsilon$.

Part (c) of theorem B is established by the same type of argument as for part (c) of theorem A.

References

[1] Amit D 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)
[2] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison Wesley)
[3] Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[4] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[5] Cover T 1965 *IEEE Trans. Electron. Comput.* EC-14 326
[6] Gardner E 1987 *Europhys. Lett.* **4** 481
[7] Gardner E 1987 *J. Phys. A: Math. Gen.* **21** 257
[8] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[9] Abbott L and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** 2031
[10] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[11] Gardner E, Gutfreund H and Yekutieli I 1989 *J. Phys. A: Math. Gen.* **22** 1995
[12] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659

[13] Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **24** 3361
[14] Franz S, Amit D J and Virasoro M A 1990 *J. Physique* **51** 387
[15] Del Giudice P, Franz S and Virasoro M A 1989 *J. Physique* **50** 121
[16] Wong K Y M and Sherrington D 1989 *Europhys. Lett.* **10** 419
[17] Fontanari J F and Meir R 1989 *J. Phys. A: Math. Gen.* **22** L803
[18] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983
[19] Krauth W and Mézard M 1989 *J. Physique* **50** 3056
[20] Amaldi E and Nicolis S 1989 *J. Physique* **50** 2333
[21] Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519
[22] Köhler H 1990 *J. Phys. A: Math. Gen.* **23** L1265
[23] Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **24** 2613
[24] Derrida B 1981 *Phys. Rev.* B **24** 2613
[25] Gross D J and Mézard M 1984 *Nucl. Phys.* **B240** [FS12] 431
[26] Györgyi G 1990 *Phys. Rev.* A **41** 7097
[27] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[28] Vallet F 1989 *Europhys. Lett.* **8** 747
[29] Vallet F, Cailton J and Refregier R 1989 *Europhys. Lett.* **9** 315
[30] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581
[31] Hansel D and Sompolinsky H 1990 *Europhys. Lett.* **11** 687
[32] Györgyi G and Tishby N 1989 Preprint, to appear in *Proc. STATPHYS-17 Workshop* ed W K Theumann and R Köberle
[33] Ruelle D 1987 *Commun. Math. Phys.* **108** 225
[34] Cook J and Derrida B 1991 *J. Stat. Phys.* **63** 505
[35] Fontanari J F and Köberle R 1990 *J. Physique* **51** 1403
[36] Young A P, Bray A J and Moore M A 1984 *J. Phys. C: Solid State Phys.* **17** L149
[37] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[38] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715
[39] Barkai E and Kanter I 1991 *Europhys. Lett.* **14** 107
[40] Bouten M, Komoda A and Serneels R 1991 Preprint