Three unfinished works on the optimal storage capacity of networks

E. Gardner, Bernard Derrida

J. Phys. A: Math. Gen. **22** (1989) 1983-1994. Printed in the UK

Three unfinished works on the optimal storage capacity of networks

E Gardner and B Derrida

The Institute for Advanced Studies, The Hebrew University of Jerusalem, Jerusalem, Israel and Service de Physique Théorique de Saclay†, F-91191 Gif-sur-Yvette Cedex, France

Received 13 December 1988

Abstract. The optimal storage properties of three different neural network models are studied. For two of these models the architecture of the network is a perceptron with $\pm J$ interactions, whereas for the third model the output can be an arbitrary function of the inputs. Analytic bounds and numerical estimates of the optimal capacities and of the minimal fraction of errors are obtained for the first two models. The third model can be solved exactly and the exact solution is compared to the bounds and to the results of numerical simulations used for the two other models.

1. Introduction

The problem of the optimal storage capacity of a network (Gardner 1987a, 1988 and references therein) can be formulated in its simplest version as follows. Consider a set of P patterns, each pattern μ consisting of N input bits $S_i^\mu = \pm 1$ for $1 \leqslant i \leqslant N$ and one output bit $R^\mu = \pm 1$. Is it possible to find a Boolean function F of N variables such that

$$R^\mu = F(S_1^\mu, S_2^\mu, \ldots, S_N^\mu) \tag{1}$$

is satisfied for each pattern μ?

Each pattern μ can be viewed as an example ($\{S_i^\mu\}$ is the input and R^μ is the right answer to that input) and the question is whether there exists a Boolean function F which gives the right answer for all the P examples.

It is in general possible to find such a Boolean function F (unless there are contradictory examples, i.e. examples with the same inputs and opposite output: see model C below) since a Boolean function F of N variables is fully determined by 2^N bits corresponding to its 2^N possible inputs. For each input $\{S_i^\mu\}$ which belongs to the P examples, one chooses for the output the right answer R^μ whereas one can choose the output at random for the remaining $2^N - P$ possible inputs.

The problem becomes more difficult when restrictions are imposed on the possible functions F, in particular if one assumes that the inputs and the output belong to a neural network with a complex architecture (for example, with layers of hidden units). The simplest network for which the role of the architecture is essential is the perceptron

† Laboratoire de l'Institut de Recherche Fondamentale du Commissariat à l'Énergie Atomique.

0305-4470/89/121983+12$02.50

(Rosenblatt 1962, Minsky and Papert 1969) (figure 1). In that case, the possible functions F are characterised by N real numbers J_i, $1 \leqslant i \leqslant N$ (the synapses):

$$R^\mu = F(S_1^\mu, S_2^\mu, \ldots, S_N^\mu) = \operatorname{sgn}\left(\sum_{i=1}^{N} J_i S_i^\mu\right). \tag{2}$$

Because equation (2) is unchanged when all the J_i are multiplied by the same positive constant ($\{J_i\} \to \{\lambda J_i\}$), one can impose a normalisation constraint on the J_i:

$$\sum_{i=1}^{N} J_i^2 = N. \tag{3}$$

The typical questions one can ask about such a problem are the following.

(i) What is the maximum number P_c of examples that the network can store?

(ii) For $P < P_c$, how many different functions F solve the P equations (2)?

(iii) For $P > P_c$, what is the minimal fraction $f_{\min}$ of errors of this network: i.e. what is the function F which minimises the number $Pf_{\min}$ of patterns for which equation (2) is not satisfied?

(iv) How do these results depend on the statistical properties of the P patterns and on their correlations?

(v) Are there algorithms able to find the optimal choice of the J_i (i.e. the set of J_i which satisfies the maximum number of equations (2))?

All these questions have been studied recently (Gardner 1987a, 1988, Gardner and Derrida 1988 and references therein) for the case of the spherical constraint (3).

The only technique which has been used so far to study questions (i)-(iv) analytically is the replica method. For the spherical model (defined by (3)) it was found that if the P patterns are chosen at random and are not correlated, there exists a maximal storage capacity (Venkatesh 1986)

$$\alpha_c = 2 \tag{4}$$

where

$$\alpha = P/N. \tag{5}$$

The volume of phase space (i.e. of the set of the J_i for which all, or more precisely almost all, of the equations (2) are satisfied) was obtained for $\alpha < \alpha_c$ (Gardner 1987a, 1988). The minimal fraction $f_{\min}$ of wrong patterns was calculated for $\alpha > \alpha_c$ (Gardner and Derrida 1988). All these calculations having been done with the replica approach assuming a replica symmetric solution, it is generally believed that the results should be trusted only when the replica symmetric solution is stable. For the spherical model,

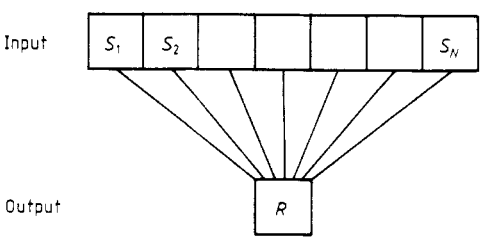

Figure 1. The network for models A and B. The input consists of N bits $\{S_i = \pm 1\}$ and the output is a single bit $R = \pm 1$, expressed as $R = \operatorname{sgn}(\sum_{i=1}^{N} J_i S_i)$.

the replica symmetric solution is stable for $\alpha \leqslant \alpha_c = 2$ and unstable for $\alpha > 2$. So the calculation of the volume of phase space for $\alpha < 2$ and the threshold $\alpha_c = 2$ are in principle right whereas the expression of $f_{\min}$ obtained for $\alpha > 2$ should be incorrect.

In this paper, some results on three similar problems will be presented. For the first two problems (models A and B), even the value of α_c is not known exactly. Upper bounds for α_c, for the volume of phase space (for $\alpha < \alpha_c$) and a lower bound for $f_{\min}$ (for $\alpha > \alpha_c$) will be calculated. Also numerical estimates of α_c will be obtained. Lastly, the third model C is a kind of random energy limit of the problem which allows one to calculate $f_{\min}$ exactly. These three models are defined as follows.

1.1. Model A

The definition is the same as for the spherical model (2) except that the spherical constraint (3) is replaced by an Ising constraint

$$J_i = \pm 1. \tag{6}$$

So, if P uncorrelated random patterns $(\{S_i^\mu\}, R^\mu)$ are given, the first question is whether there exists a choice of $J_i = \pm 1$ such that the P equations

$$R^\mu = \operatorname{sgn}\left(\sum_{i=1}^{N} J_i S_i^\mu \right) \tag{7}$$

are satisfied simultaneously.

The next questions are: what is α_c; how many choices of the J_i solve all these equations (7) for $\alpha < \alpha_c$; what is the minimal fraction $f_{\min}$ of wrong patterns for $\alpha > \alpha_c$; etc.

Using the replica symmetric ansatz for this model A, all these quantities have been calculated (Gardner and Derrida 1988). In particular, the replica symmetric solution (which was unstable for $\alpha > \alpha_c$) gave

$$\alpha_c = 4/\pi \simeq 1.273. \tag{8}$$

This prediction is certainly incorrect because α_c must be less than 1 (Gardner and Derrida 1988). For each pattern μ, one bit R^μ is stored by the network. The network can store up to $N\alpha_c$ bits. This information is encoded in the N bits J_i. Therefore $N\alpha_c$ should be less than N (see §§ 2 and 3).

1.2. Model B

For model B, the network is the same as for model A: a perceptron with Ising interactions ($J_i = \pm 1$). The only difference is that for each pattern the output R^μ is correlated with the input $\{S_i^\mu\}$ in the following sense. One considers that the P examples, $(\{S_i^\mu\}, R^\mu)$ are given by a teacher which is itself a perceptron with N interactions $K_i = \pm 1$ chosen at random. For each example, a pattern μ is constructed by choosing the input $\{S_i^\mu\}$ at random and by asking the teacher what is the right answer for R^μ:

$$R^\mu = \operatorname{sgn}\left(\sum_{i=1}^{N} K_i S_i^\mu \right). \tag{9}$$

The question one can then ask is: how many different choices of the J_i give for the P examples the same answers as the teacher? For $N \to \infty$, one expects a critical value α_c of $\alpha = P/N$. For $\alpha < \alpha_c$, there exist several choices of $\{J_i\}$ which give the same R^μ

as $\{K_i\}$, whereas for $\alpha > \alpha_c$ this choice is unique ($\{J_i\} = \{K_i\}$). So α_c gives the size of the learning set. Once the network has learnt more than $N\alpha_c$ examples, it knows how to generalise perfectly and it can answer correctly any new question posed by the teacher.

1.3. Model C

For model C, the P patterns are random (random input $\{S_i^\mu\}$ and random output R^μ), and no constraint is imposed on the function F. So F is an arbitrary Boolean function of N variables. The only problem in this model C comes from contradictory patterns (patterns with the same input but opposite output). For this model C, the minimal fraction $f_{\min}$ of errors can be calculated exactly as a function of the number P of patterns. The main interest of this model C is that $f_{\min}$ can be calculated exactly. This exact result should be a good test for the replica method: all the replica calculations of $f_{\min}$ which have been done so far are not acceptable because the replica symmetric solution is always unstable.

2. Bounds for models A and B

Let us call $\mathscr{C} = \{J_i\}$ a configuration of the $\{J_i\}$. For model A, the probability $Q(\mathscr{C})$ that a configuration $\mathscr{C}$ gives the right answer R^μ for the P random patterns ($\{S_i^\mu\}$, R^μ) is

$$Q(\mathscr{C}) = (\tfrac{1}{2})^P. \tag{10}$$

So if Ω is the number of configurations which give the right answer for all the P random patterns, the average $\langle\Omega\rangle$ (which is the average over the P random patterns) is

$$\langle\Omega\rangle = \sum_{\mathscr{C}} Q(\mathscr{C}) = 2^{N-P} = 2^{N(1-\alpha)}. \tag{11}$$

For $\alpha > 1$, $\langle\Omega\rangle$ is much less than 1. Since Ω can only take integer values (Derrida 1980, Gardner 1987b), it is clear that the probability that $\Omega \neq 0$ is less than $\langle\Omega\rangle$. This implies that $\Omega \neq 0$ with a probability which vanishes as $N \to \infty$. So for $\alpha > 1$, $\Omega = 0$ with probability 1 when $N \to \infty$. This implies that

$$\alpha_c \leqslant 1. \tag{12}$$

Notice that for $\alpha < 1$, one cannot deduce anything from (10) except bounds: when $\langle\Omega\rangle$ is much larger than 1, it is either possible that $\Omega \neq 0$ with a finite probability or that $\Omega = 0$ with probability 1 (when $N \to \infty$). This last case is possible if there are only rare events for which $\Omega \neq 0$ and if for some of these rare events Ω is very large.

The bound (12) agrees with the simple argument given in the introduction (that α_c cannot exceed 1) and gives another way of showing that the replica prediction $\alpha_c = 4/\pi$ is incorrect.

In most cases in disordered systems, the typical values of quantities which increase exponentially with the system size N are given by averaging the logarithm. Since from the convexity of the log, one always has

$$\langle \log z \rangle < \log\langle z \rangle$$

for any random variable z, one obtains from (11) a bound (Gardner 1986):

$$\frac{1}{N}\langle \log \Omega \rangle < \frac{1}{N} \log\langle\Omega\rangle = (1-\alpha)\log 2. \tag{13}$$

This upper bound is not a very profound result because one expects that $\langle \log \Omega \rangle = -\infty$. This is due to the events with contradictory patterns ($\{S_i^\mu\} = \{S_i^\nu\}$ but $R^\mu = -R^\nu$) which give $\Omega = 0$. Although these events are very rare, they force $\langle \log \Omega \rangle$ to be $-\infty$. To avoid this problem, one can consider $\langle \log(1+\Omega) \rangle$ and one gets, using again the convexity of the log,

$$\langle \log(1+\Omega) \rangle < \log(1+\langle \Omega \rangle) \tag{14}$$

which implies the following bound for $\alpha < 1$ (model A):

$$\frac{1}{N} \langle \log(1+\Omega) \rangle < (1-\alpha) \log 2. \tag{15}$$

For $\alpha > 1$, it is also possible to obtain a lower bound for the minimal fraction $f_{\min}$ of errors. If one defines $Q(\mathscr{C}, f)$ to be the probability that a configuration $\mathscr{C}$ gives the right answer (the right R^μ) for $P(1-f)$ patterns and the wrong one for Pf patterns, then one has

$$Q(\mathscr{C}, f) = \frac{P!}{(Pf)![P(1-f)]!} \left(\frac{1}{2}\right)^P. \tag{16}$$

If $\Omega(f)$ is the number of configurations which give $P(1-f)$ examples right and Pf wrong, then one has

$$\langle \Omega(f) \rangle = \frac{P!}{(Pf)!(P(1-f))!} 2^{N-P}. \tag{17}$$

Using the same argument as above (if $\langle \Omega \rangle \ll 1$, then $\Omega = 0$ with probability 1), one gets a lower bound $\tilde{f}$ for $f_{\min}$:

$$f_{\min} > \tilde{f} \tag{18}$$

where $\tilde{f}$ is such that $\langle \Omega \rangle \sim 1$, i.e. is the (smaller) solution of

$$-\tilde{f} \log \tilde{f} - (1-\tilde{f}) \log(1-\tilde{f}) = (1-1/\alpha) \log 2. \tag{19}$$

The α dependence of the lower bound $\tilde{f}$ for model A is shown in figure 2 (it is non-zero only for $\alpha > 1$).

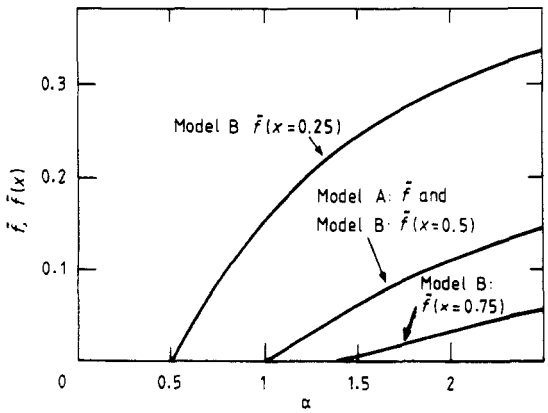

Figure 2. The α dependence of the lower bounds of the minimal fraction $f_{\min}$ of errors expressed as $\tilde{f}$ for model A and as $\tilde{f}(x)$ for model B, with $x = 0.25$, 0.50 and 0.75. The two curves $\tilde{f}$ for model A and $\tilde{f}$ ($x = 0.50$) for model B coincide.

All the previous arguments can easily be generalised to the case of model B.

For model B, it is convenient to classify the configurations $\mathscr{C}=\{J_i\}$ according to their projections on the teacher configuration $\{K_i\}$

$$\frac{1}{N}\sum_{i=1}^{N} J_i K_i = 2x-1 \tag{20}$$

meaning that for Nx interactions $J_i = K_i$ and for $N(1-x)$ interactions $J_i = -K_i$. The total number of configurations having a projection (20) is given by

$$\frac{N!}{(Nx)!(N(1-x))!} \sim \exp\{-N[x\log x+(1-x)\log(1-x)]\}. \tag{21}$$

One can calculate the probability $Q(\mathscr{C}, x)$ that a configuration $\{J_i\}$ gives the same R^μ as $\{K_i\}$ for all the P random patterns μ; one can write

$$\tilde{R}^\mu = \operatorname{sgn}(Y+Z) \qquad R^\mu = \operatorname{sgn}(Y-Z) \tag{22}$$

where $\tilde{R}^\mu$ is the output obtained by the configuration $\{K_i\}$ and R^μ is the output obtained by $\{J_i\}$ for the same random input $\{S_i^\mu\}$. Y is the sum of Nx random numbers ± 1 whereas Z is the sum of $N(1-x)$ such random numbers. The probability that $R^\mu = \tilde{R}^\mu$ for a given random pattern μ is the probability that $|Z|<|Y|$. When N increases, Y and Z become Gaussian variables of widths Nx and $N(1-x)$ and therefore

$$\begin{aligned} Q(\mathscr{C}, x) &= \left[\frac{1}{2\pi N\sqrt{x(1-x)}}\int_{-\infty}^{\infty} dY \int_{-|Y|}^{|Y|} dZ \exp\left(-\frac{Y^2}{2Nx}-\frac{Z^2}{2N(1-x)}\right)\right]^P \\ &= \left[\frac{2}{\pi}\tan^{-1}\left(\sqrt{\frac{x}{1-x}}\right)\right]^P. \end{aligned} \tag{23}$$

If $\Omega(x)$ is the number of configurations $\{J_i\}$ having an overlap (20) with $\{K_i\}$ which give the right answer for the P patterns, one has

$$\begin{aligned} \langle\Omega(x)\rangle &= \sum_{\mathscr{C}} Q(\mathscr{C}, x) = \frac{N!}{(Nx)!(N(1-x))!}\, Q(\mathscr{C}, x) \\ &\sim \exp\left[\!\left[N\left\{-x\log x-(1-x)\log(1-x)+\alpha\log\left[\frac{2}{\pi}\tan^{-1}\left(\sqrt{\frac{x}{1-x}}\right)\right]\right\}\right]\!\right]. \end{aligned} \tag{24}$$

For any value of x, one can get an upper bound for $\alpha_c(x)$, the critical number of examples above which there is no configuration $\{J_i\}$ which gives the same output R^μ as $\{K_i\}$ for all the patterns μ. The argument is again that if $\langle\Omega(x)\rangle \ll 1$, then $\Omega(x)=0$ with probability 1:

$$\alpha_c(x) < [-x\log x-(1-x)\log(1-x)]\left\{-\log\left[\frac{2}{\pi}\tan^{-1}\left(\sqrt{\frac{x}{1-x}}\right)\right]\right\}^{-1}. \tag{25}$$

One can calculate the maximal value (over x) of the right-hand side of (25) and one finds 1.448. This implies that

$$\alpha_c = \max_x(\alpha_c(x)) < 1.448\ldots. \tag{26}$$

Below α_c, there exist some configurations $\{J_i\} \neq \{K_i\}$ which give the right answer for all the P examples. Above α_c, the only configuration is $\{J_i\} = \{K_i\}$. Therefore $N\alpha_c$ is the number of examples that the system has to learn before it is able to generalise without any error.

As for model A, it is possible to obtain for model B an upper bound on the volume of phase space, i.e. on the number of configurations $\{J_i\}$ which give the same answer as $\{K_i\}$ for the P patterns:

$$\frac{1}{N}\langle \log \Omega \rangle = \max_x \left(\frac{1}{N} \langle \log \Omega(x) \rangle \right)$$

$$\leqslant \max_x \left\{ -x \log x - (1-x) \log(1-x) + \alpha \log \left[\frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{x}{1-x}} \right) \right] \right\}. \quad (27)$$

One can obtain also a lower bound $\tilde{f}(x)$ for the minimal fraction $f_{\min}(x)$ of errors for $\alpha > \alpha_c(x)$. For each configuration $\{J_i\}$ which has an overlap given by (20), the probability $Q(\mathscr{C}, x, f)$ that it gives the right answer to $P(1-f)$ patterns and the wrong one to Pf patterns is

$$Q(\mathscr{C}, x, f) = \frac{P!}{(P(1-f))!(Pf)!} \left[\frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{x}{1-x}} \right) \right]^{P(1-f)} \left[1 - \frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{x}{1-x}} \right) \right]^{Pf}. \quad (28)$$

Using exactly the same argument as for model A, one can calculate the lower bound $\tilde{f}(x)$:

$$f_{\min}(x) \geqslant \tilde{f}(x) \quad (29)$$

where $\tilde{f}(x)$ is the smaller solution of

$$\frac{1}{\alpha}[-x \log x - (1-x) \log(1-x)] + [-\tilde{f} \log \tilde{f} - (1-\tilde{f}) \log(1-\tilde{f})]$$

$$= -(1-\tilde{f}) \log \left[\frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{x}{1-x}} \right) \right] - \tilde{f} \log \left[1 - \frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{x}{1-x}} \right) \right]. \quad (30)$$

This result is valid as long as $x \neq 0$ and $x \neq 1$ because we have used in (22) and (23) the fact that both Y and Z are sums of a large number of random variables. (If $\{J_i\} = \{K_i\}$ except for a finite number of interactions, the calculation should be done in a different way). This is why from the very beginning one knows that $\alpha_c(1) = \infty$ since the solution $\{J_i\} = \{K_i\}$ gives the right output for an arbitrary number P of patterns whereas (25) would give $\alpha_c(1) = 0$.

The lower bound $\tilde{f}(x)$ for model B is plotted as a function of α for $x = 0.25$, 0.50 and 0.75 in figure 2.

3. Numerical simulations of models A and B

In this section, numerical estimates of α_c are obtained for models A and B.

For model A, the procedure is the following. For finite N ($N = 3, 5, 7, \ldots, 15$) one calculates by a Monte Carlo sampling the average number $\langle P \rangle$ of examples that the system can store without error: the first example is chosen at random (i.e. the

$S_i^1 = \pm 1$ and $R^1 = \pm 1$ are chosen at random). Then one goes through the 2^N possible choices of the $\{J_i\}$ and one keeps only the $\{J_i\}$ which give the right answer R^1. Then one chooses at random the second pattern and one goes through the remaining configurations $\{J_i\}$ in order to eliminate again the $\{J_i\}$ which give a wrong answer for R^2, and so on. If there exist some choices of the $\{J_i\}$ which give the right answer for the first P patterns and if no choice exists for $P+1$ patterns, this means that for this sample the system can store exactly P patterns. This value of P depends on the sample since it depends on the random choices of the S_i^μ and R^μ. One averages P over many samples (here the number of samples will always be 10 000) and one defines an estimate $\alpha_c(N)$ of α_c for a system of size N by

$$\alpha_c(N) = \frac{\langle P \rangle}{N}. \tag{31}$$

$\alpha_c(N)$ for model A is plotted as a function of $1/N$ by the open circles in figure 3. Error bars are not given because they would be smaller than the symbols. Only odd values of N are used in order to avoid cases for which the sum $\Sigma_i J_i S_i^\mu$ would vanish.

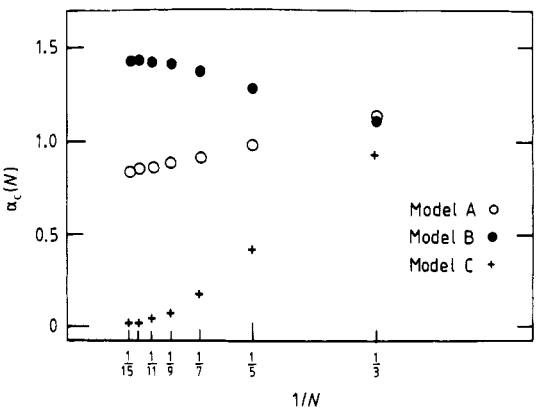

Figure 3. The maximal storage capacity $\alpha_c(N)$ as a function of $1/N$ obtained by Monte Carlo sampling (10^4 samples) for models A and B and by the method discussed in § 4 for model C.

We see in figure 3 that $\alpha_c(N)$ has a regular N^{-1} dependence and, from the data of figure 3, one can estimate that in the limit $N \to \infty$

$$\alpha_c = 0.75 \pm 0.05. \tag{32}$$

The same procedure can be repeated for model B. One chooses first a fixed set of random interactions $\{K_i\}$ (for example $K_i = +1$ for all i). Then if there exist some choices of $\{J_i\} \neq \{K_i\}$ which give the right R^μ for P patterns and if the only choice is $\{J_i\} = \{K_i\}$ for $P+1$ patterns, this means that the system has perfectly learnt after P patterns. $\alpha_c(N)$ defined by (31) is shown for model B by the full circles in figure 3.

The N^{-1} dependence of $\alpha_c(N)$ for model B is not linear and seems to reach a maximum at $N \sim 13$. This makes the extrapolation difficult. However, from the data shown in figure 3, it seems reasonable that for model B

$$\alpha_c \simeq 1.35 \pm 0.10. \tag{33}$$

We see that this estimate satisfies rather well the bound $\alpha_c < 1.448$ obtained in § 2 (equation (26)). If the extrapolated value (33) is right, then model B is a case for which the bound (26) gives a rather good approximation of α_c.

4. Solution of model C

For model C, the inputs $\{S_i^\mu\}$ and the outputs R^μ of each pattern are chosen at random and the Boolean function F is arbitrary. This would correspond to the limit of a multiconnected network with p spin interactions in the limit $p \to \infty$ (Gardner 1985, 1987b, Derrida 1980).

For a given set of P patterns chosen at random, one can define for each Boolean function F, an energy $W(F)$ as the number of patterns μ for which the function F gives the wrong output R^μ. There are 2^{2^N} possible Boolean functions F and for each function the energy $W(F)$ is an integer between 0 and P.

By analogy with Gardner and Derrida (1988), one can define a partition function $Z(h)$ by

$$Z(h) = \sum_F \exp(-hW(F)) \tag{34}$$

where the sum runs over the 2^{2^N} possible Boolean functions. It is possible to derive the following exact result (equations (36)-(40) below): in the limit $N \to \infty$

$$\frac{1}{P}\langle \log Z(h)\rangle = \frac{1}{\alpha}\sum_{n=1}^{\infty}\frac{e^{-\alpha}\alpha^n}{2^n}\sum_{q=0}^{n}\frac{\log(e^{-qh}+e^{-(n-q)h})}{q!(n-q)!} \tag{35}$$

where $\langle\cdot\rangle$ denotes the average over the random choices of the P patterns and where α is defined by

$$\alpha = P/2^N \tag{36}$$

(one should notice that this new definition of α is different from (5): since the number of possible Boolean functions is much larger (2^{2^N}) than for models A and B, P can also be much larger).

The derivation of (35) is rather easy: there are $M = 2^N$ possible input configurations $\{S_i\}$. When one chooses P patterns at random, some input configurations $\{S_i\}$ never occur among these P patterns, some input configurations $\{S_i\}$ occur only once, some occur twice, and so on.

Let M_n be the number of input configurations $\{S_i\}$ such that n (among the P random patterns) have this configuration $\{S_i\}$ as input. One has of course

$$\sum_{n=0}^{\infty} M_n = M = 2^N \qquad \sum_{n=0}^{\infty} nM_n = P = \alpha 2^N. \tag{37}$$

Because the P patterns are chosen at random and independently, one can easily show that for $M \to \infty$, one has

$$M_n = \frac{\alpha^n}{n!}e^{-\alpha}M \tag{38}$$

(P points are chosen at random among M points; M_0 is the number of points which are never chosen, M_1 is the number of points which are chosen once, ..., M_n is the

number of points which are chosen n times). Among these M_n input configurations, which occur n times among the P patterns,

$$\frac{1}{2^n}\frac{n!}{q!(n-q)!}M_n \tag{39}$$

are such that q patterns have the output $R^\mu = +1$ and $n-q$ patterns have the output $R^\mu = -1$. The contribution of these patterns to $W(F)$ is either q or $(n-q)$ and therefore $\log Z(h)$ is given by

$$\log Z(h) = \sum_{n=0}^{\infty} M_n \sum_{q=0}^{n} \frac{n!}{2^n q!(n-q)!} \log(e^{-qh} + e^{-(n-q)h}) \tag{40}$$

and this expression leads obviously to (35).

In the limit $h \to \infty$ (Gardner and Derrida 1988), one can obtain the exact expression of $f_{\min}$ for model C:

$$f_{\min} = \lim_{h\to\infty} -\frac{1}{h}\frac{1}{P}\langle \log Z(h)\rangle = \frac{1}{\alpha}\sum_{n=0}^{\infty}\frac{e^{-\alpha}\alpha^n}{2^n}\sum_{q=0}^{n}\frac{\min(q, n-q)}{q!(n-q)!}. \tag{41}$$

The curve $f_{\min}$ is plotted as a function of α in figure 4. It is clear from this curve and from (41) that $f_{\min} > 0$ if $\alpha > 0$ and therefore

$$\alpha_c = 0. \tag{42}$$

For model C, it is also possible to redo what was done in §§ 2 and 3 for models A and B. To obtain bounds as in § 2, one can start with the expression of the probability $Q(F)$ that a given Boolean function F gives the right answer for P patterns:

$$Q(F) = (\tfrac{1}{2})^P. \tag{43}$$

Then the average number $\langle\Omega\rangle$ of functions which give the right answer for all the P patterns is

$$\langle\Omega\rangle = \sum_F (\tfrac{1}{2})^P = 2^{2^N - P} = 2^{2^N(1-\alpha)}. \tag{44}$$

This gives as in § 2 an upper bound for α_c:

$$\alpha_c \leqslant 1. \tag{45}$$

Of course, the exact value $\alpha_c = 0$ (42) satisfies this bound (45). However, we see that, for model C, the bound gives a rather poor estimate of the exact value.

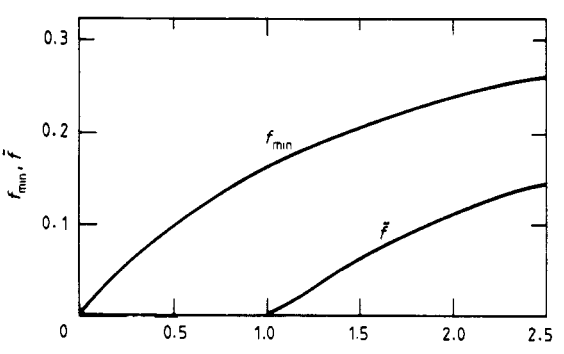

Figure 4. The exact expression of $f_{\min}$ and the lower bound $\tilde{f}$ as a function of α for model C.

For $\alpha > 1$, the calculation of the lower bound $\tilde{f}$ of $f_{\min}$ can be repeated. The probability $Q(F,f)$ that a given function F gives the right answer for $P(1-f)$ patterns and the wrong answer for Pf patterns is:

$$Q(F,f) = \frac{P!}{(Pf)!(P(1-f))!}\left(\frac{1}{2}\right)^P. \tag{46}$$

Exactly in the same way as in § 2, this leads to the following expression of the lower bound $\tilde{f}$ of $\tilde{f}_{\min}$:

$$\tilde{f} < f_{\min} \tag{47}$$

with

$$-\tilde{f}\log\tilde{f} - (1-\tilde{f})\log(1-\tilde{f}) = (1-1/\alpha)\log 2 \tag{48}$$

$\tilde{f}$ is shown in figure 4. We see again that this bound is a rather bad estimate for $f_{\min}$.

For model C, it is also possible to obtain the $\alpha_c(N)$ as in § 3. It would be too difficult to use the same Monte Carlo sampling and to go through all the possible functions (there are 2^{2^N} possible functions for model C instead of 2^N for models A and B). However, because model C is exactly soluble, one can obtain exact expressions for the $\alpha_c(N)$.

One can first calculate the probability $Q_K(P)$ that choosing P input patterns at random among $M = 2^N$ possible input patterns, one gets K different input patterns. Of course one has

$$Q_1(1) = 1 \qquad \text{and} \qquad Q_K(1) = 0 \qquad \text{for } K \geqslant 2. \tag{49}$$

For $P = 2$, one can easily check that

$$Q_1(2) = \frac{1}{M} \qquad Q_2(2) = \frac{M-1}{M} \qquad Q_K(2) = 0 \qquad \text{for } K \geqslant 3 \tag{50}$$

and one can show that

$$Q_K(P+1) = \frac{K}{M}Q_K(P) + \frac{M-K+1}{M}Q_{K-1}(P). \tag{51}$$

This recursion relation allows one to obtain all the $Q_K(P)$ (at least on the computer).

If all the patterns have different input patterns, i.e.: if $K = P$, there always exists a function F which gives the right answer for all the P patterns (there is no contradiction).

If $K = P-1$, one input pattern $\{S_i\}$ is repeated twice. There is a probability $\frac{1}{2}$ that the two outputs corresponding to this input pattern are contradictory. Therefore, there exists a function F only with probability $\frac{1}{2}$.

For arbitrary K and P, the probability that there is no contradiction for the patterns which have identical inputs is 2^{K-P}. It follows that for P random patterns, the probability $H(P)$ that there exists a Boolean function F which gives the right output for all the P patterns is

$$H(P) = \sum_{K=1}^{P} 2^{K-P}Q_K(P). \tag{52}$$

The probability that there exists such a function for P random patterns and not for $P+1$ patterns is $H(P) - H(P+1)$ and $\alpha_c(N)$ is given by

$$\alpha_c(N) = \frac{1}{M}\sum_{P=1}^{\infty} P(H(P) - H(P+1)) = \frac{1}{M}\sum_{P=1}^{\infty} H(P). \tag{53}$$

The values of $\alpha_c(N)$ calculated using equations (49)-(53) are shown in figure 3. We see that $\alpha_c(N)$ converges quickly to its exact limit $\alpha_c = 0$. This convergence seems faster for model C than for models A and B. This is probably due to the fact that phase space is much larger (2^{2^N}) for model C than for models A and B (2^N).

This good convergence to the exact limit for model C gives more confidence in the numerical extrapolations (32) and (33) of § 3.

5. Conclusion (BD alone)

It is with a great deal of pain and emotion that I come to the end of this paper. For many years, Elizabeth has been both a very close collaborator and a very faithful friend. During all these years, by her exceptional scientific talent and by her kindness, she made all our work together most pleasant and fruitful. It is hard to accept that our collaboration, which I found both enjoyable and stimulating, should have ended so prematurely.

This work is unfinished. What is written here corresponds to what was discussed with Elizabeth between January 1987 and May 1988. We started to work on models A and C in the spring of 1987 in Paris. The work on model B and the calculations of § 3 were started when both of us were visiting the Hebrew University of Jerusalem in the spring of 1988. This paper should have included another part: Elizabeth wanted, and had started, to extend to these models the replica approach she had developed in her previous works.

References

Derrida B 1980 *Phys. Rev. Lett.* **45** 79
Gardner E 1985 *Nucl. Phys.* B **257** [FS14] 747
—— 1986 *J. Phys. A: Math. Gen.* **19** L1047
—— 1987a *Europhys. Lett.* **4** 481
—— 1987b *J. Phys. A: Math. Gen.* **20** 3453
—— 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
Venkatesh S 1986 *PhD thesis* California Institute of Technology