Perceptrons above saturation

1993 J. Phys. A: Math. Gen. 26 7405

(http://iopscience.iop.org/0305-4470/26/24/015)

J. Phys. A: Math. Gen. 26 (1993) 7405–7416. Printed in the UK

Perceptrons above saturation

P Majer, A Engel and A Zippelius

Institut für Theoretische Physik, Georg-August-Universität, 37073 Göttingen, Federal Republic of Germany

Received 19 August 1993

Abstract. We study the storage of random patterns by a perceptron above its storage capacity α_c, i.e. in the region where perfect storage becomes impossible. We determine the minimal fraction of learning errors and the distribution of stabilities for different learning rules in one-step replica symmetry breaking. Thereby we not only extend the known replica symmetric results to values of the storage capacity beyond the AT line but also show that, depending on the learning rule, replica symmetry may be globally unstable already well below the AT line. As an example for possible implications we compare the results for the typical basins of attraction of an extremely diluted attractor neural network as given by replica symmetry and one-step replica symmetry breaking.

1. Introduction

In the storage problem for a perceptron one tries to find a synaptic vector $J \in \mathbb{R}^N$ that implements p random input–output mappings $\xi^\mu \to \sigma^\mu$ according to $\sigma^\mu = \mathrm{sign}(J \cdot \xi^\mu)$, $\mu = 1, \ldots, p$. One of the central results is that for $N \to \infty$ there is a sharp threshold α_c of the ratio $\alpha = p/N$ such that for $\alpha < \alpha_c$ the problem can be solved, whereas for $\alpha > \alpha_c$ there is no vector J realizing all mappings [1, 2]. For $\alpha < \alpha_c$, it is easy to show that the space of all possible solutions J is connected (even convex). For $\alpha > \alpha_c$, since no vector J correctly realizes all the mappings, the interesting task is to find the J which minimizes the possible fraction of errors [3]. Although difficult to prove exactly, it seems reasonable, that the solution space can now be disconnected, since different solutions make errors for different patterns.

To average the various quantities of interest over the distribution of the random patterns forming the training set, one usually employs the replica trick. Eventually, this results in saddle-point integrals which necessitate the minimization of non-trivial functions with respect to appropriate order parameters carrying one or two replica indices. To find these extrema in the limit where the number of replicas goes to zero, one has to make an ansatz for the replica structure of these order parameters. It is generally believed that a connected solution space (being the analogue of an ergodic dynamics in the related problem of spin-glasses) implies that replica symmetry (RS) holds. Hence a replica-symmetric ansatz is justified for $\alpha \leqslant \alpha_c$. To test the correctness of the replica-symmetric ansatz for $\alpha \geqslant \alpha_c$ one can determine the local stability of the corresponding saddle point. One finds that for some $\alpha_{AT} \geqslant \alpha_c$ the replica-symmetric saddle point becomes locally unstable [3, 4]. This is, however, not a sufficient criterion, since other saddle points with lower values of the function to be minimized may exist. Hence it remains unclear whether a replica-symmetric ansatz is correct for $\alpha_c \leqslant \alpha \leqslant \alpha_{AT}$. This question is of particular interest because replica-symmetric

0305-4470/93/247405+12$07.50

results have been used for $\alpha > \alpha_c$ to calculate, for example, the basins of attraction of extremely diluted models [4,5], the generalization ability [6] and the storage capacity of multilayer networks [7].

In the present paper we investigate the *global* stability of the replica-symmetric ansatz for perceptrons beyond the saturation limit α_c by considering the alternative ansatz of one-step replica symmetry breaking (RSB). This also provides approximate results for the region $\alpha > \alpha_{AT}$. This has also been done recently for a special cost function by Erichsen and Theumann [8]. We extend the analysis to other cost functions, and calculate the minimal possible fraction of errors and the distribution of stabilities. The results show that *depending* on the cost function, replica symmetry breaking may or may not occur for $\alpha < \alpha_{AT}$ and hence the results using the replica-symmetric ansatz in this region are not reliable. As an example, we study the basins of attraction of an extremely diluted attractor network and compare the replica-symmetric results with those from the solution using one-step replica symmetry breaking.

2. Model

The network we want to consider is a perceptron consisting of N input neurons ξ_i connected to a single output neuron σ by synaptic couplings J_i $i = 1, \ldots, N$. All neurons can take on binary values $+1, -1$. Given an input pattern $\boldsymbol{\xi}$, the output is determined by the dynamics

$$\begin{aligned} \sigma &= \operatorname{sign}(h) \\ h &= \boldsymbol{J}\boldsymbol{\xi} = \sum_{i=1}^{N} J_i \xi_i . \end{aligned} \tag{1}$$

We say that a network stores the pattern $\boldsymbol{\xi}^\mu$ if it satisfies the *a priori* given input–output relation $(\boldsymbol{\xi}^\mu, \sigma^\mu)$, which is equivalent to $\Delta^\mu = \sigma^\mu \boldsymbol{J}\boldsymbol{\xi}^\mu/\sqrt{N} > 0$. Storage of noisy patterns can be achieved by demanding the stability Δ^μ of pattern μ to be larger than or equal to a value $\kappa \geqslant 0$

$$\Delta^\mu > \kappa. \tag{2}$$

Learning here means finding synaptic couplings J_i that satisfy equation (2) for as many patterns as possible. Below a critical value α_c of the ratio $\alpha = p/N$, all patterns can be learnt perfectly and the space of networks which solve a given learning task has non-zero volume. As one approaches the storage capacity, this freedom in choice shrinks and finally leaves only a single network that solves the problem at hand. A further increase in the number of patterns to be stored leads to errors in learning. It seems reasonable to assume that the freedom in choice of network should increase with the number of errors and finally, in the limit of infinite storage, any network should solve the learning problem equally well, taking a random guess at the output.

We want to consider learning as an optimization process. This means that we define a cost function which has as absolute minima networks with the desired properties (2). A learning rule corresponding to such a cost function would then be any dynamic process that minimizes the cost function. We will not address the question of this dynamic process and therefore use both words, learning rule and cost function, synonymously. Following [4], we consider three different cost functions of the form

$$E = \sum_{\mu} V(\Delta^\mu).$$

(i) Gardner–Derrida cost function

$$V(\Delta^\mu) = \theta(\kappa - \Delta^\mu).$$

This cost function simply counts the number of errors.

(ii) Perceptron cost function

$$V(\Delta^\mu) = (\kappa - \Delta^\mu)\theta(\kappa - \Delta^\mu).$$

Here errors are linearly weighted with their distance to the stability threshold κ.

(iii) Adatron cost function

$$V(\Delta^\mu) = (\kappa - \Delta^\mu)^2\theta(\kappa - \Delta^\mu).$$

Errors are weighted quadratically with their distance to the target stability κ.

Since the different cost functions vanish for $\alpha < \alpha_c$, they give, on average, the same solutions. They differ only for $\alpha > \alpha_c$, so that it may be advantageous to choose different cost functions for different applications. In particular, it is clear from the very definition of the cost functions that the Gardner–Derrida cost function produces as few errors as possible, irrespective of their stability, while the perceptron cost function leads to more errors with smaller deviations from the threshold stability κ, and the adatron cost function leads to yet more errors (in fact, as many as for a random guess) with, again, less deviation from κ.

3. Free energy

In order to study the performance of these learning rules one calculates their ground-state energy by considering the free energy in the limit $\beta \to \infty$.

$$f(\xi^\mu) = -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{N\beta}\log Z$$

$$= -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{N\beta}\log\int \mathrm{D}J\,\mathrm{e}^{-\beta E}$$

$$\mathrm{D}J = \prod_{i=1}^{N}\mathrm{d}J_i\,\delta((J)^2 - N).$$

The spherical constraint $J^2 = N$ has been chosen to get rid of the invariance $(J, \kappa) \to (\lambda J, \lambda\kappa)$. One calculates the free energy rather than the partition function itself because the former is assumed to be self-averaging, i.e.

$$F(\xi_\mu) = \langle F\rangle_{\xi^\mu}$$

for almost any realization of the disorder ξ^μ. To perform the average over the patterns ξ^μ with the distribution $P(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1)$, we apply the replica trick $\langle\log Z\rangle = \lim_{n\to 0}(\langle Z^n\rangle - 1)/n$. Using standard techniques [3,4] one finds

$$\langle Z^n\rangle = \int\prod_{a<b}\mathrm{d}q^{ab}\int\prod_{a<b}\mathrm{d}\phi^{ab}\int\prod_a \mathrm{d}E^a$$

$$\times\exp\left\{N\left(\alpha G_1(q^{ab}) + G_2(\phi^{ab}, E^a) + \mathrm{i}\sum_{a<b}\phi^{ab}q^{ab}\right)\right\}$$

where

$$\exp\{G_1(q^{ab})\} = \int \prod_a d\lambda^a \int \prod_a dt^a \times \exp\left\{ -\beta \sum_a V(\lambda^a) - i \sum_a t^a \lambda^a - \tfrac{1}{2} \sum_{a \neq b} q^{ab} t^a t^b - \tfrac{1}{2} \sum_a (t^a)^2 \right\}$$

$$\exp\{G_2(E^a, \phi^{ab})\} = \int \prod_a dJ^a \exp\left\{ -i \sum_a E^a ((J^a)^2 - 1) + i \sum_{a<b} \phi^{ab} J^a J^b \right\}.$$

In order to solve $\langle Z^n \rangle$ by saddle-point integration for $n \to 0$, one needs an ansatz for the order parameters. The simplest is the ansatz of replica symmetry,

$$q^{ab} = q \qquad (a \neq b) \qquad \phi^{ab} = \phi \qquad (a \neq b) \qquad E^a = E\,.$$

Note that for $\alpha < \alpha_c$ there are several solutions of (1), so that we have $q < 1$ for $\beta \to \infty$. For $\alpha > \alpha_c$, the solution with minimal fraction of errors becomes unique, implying $q \to 1$ for $\beta \to \infty$ with $\beta(1-q) = x = O(1)$. With this ansatz one finds [3, 4]

$$\langle f \rangle = \max_x \left\{ -\frac{1}{2x} + \alpha \int Dz \left[V(\lambda_0) + \frac{(\lambda_0 - z)^2}{2x} \right] \right\} \tag{3}$$

where λ_0 is the minimum of the the square bracket for given z and $Dz = dz/\sqrt{2\pi} \exp\{-z^2/2\}$.

In order to test the correctness of the RS ansatz, one can calculate its local stability [3, 4], which is necessary but not sufficient. Here we test the global stability of the RS solution by using a one-step RSB ansatz and comparing the free energy of this more general ansatz with that of RS. Since f^{RS} is maximized with respect to q, the RS solution must be rejected irrespective of its possible local stability if $f^{RSB} > f^{RS}$.

In one-step RSB, the ansatz for the saddle point is of the form

$$q^{ab} = \begin{pmatrix} Q_1 & Q_0 & \cdots & Q_0 \\ Q_0 & Q_1 & \cdots & Q_0 \\ \cdots & \cdots & \cdots & \cdots \\ Q_0 & Q_0 & \cdots & Q_1 \end{pmatrix}$$

where Q_0 is an $m \times m$ matrix with elements q_0, and Q_1 is an $m \times m$ matrix with elements q_1 on the off diagonals and 0 on the diagonal. For ϕ^{ab} the ansatz is equivalent to that of q^{ab} and

$$E^a = E.$$

Using standard techniques we arrive at the following results. For $\alpha < \alpha_c$, the solution space is connected, RS is correct, and we find no RSB. For $\alpha > \alpha_c$, the volume of solution space vanishes. Performing the limits $q_1 \to 1$ and $\beta \to \infty$, such that $x = \beta(1 - q_1)$ remains finite, brings us into the error regime of the network. We also make the ansatz that m scales with $1/\beta$, such that $w = m\beta$ is finite, and find a self-consistent solution.

The free energy is given by

$$\langle f \rangle = \min_{x, q_0, w} \left\{ \frac{q_0}{2x(1 + w\Delta q)} + \frac{\log(1 + w\Delta q)}{2wx} + \frac{\alpha}{wx} \int Dz_0 \log \int Dz_1 \exp\left\{ -wx \left[V(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x} \right] \right\} \right\} \tag{4}$$

with λ_0 minimizing the square bracket for given values of z_0 and z_1 and $\Delta q = 1 - q_0$. In the limit $q_0 \to 1$, this reduces to the RS result of [4].

We solved numerically the three-dimensional minimization with the results discussed in section 5.

4. Distribution of pattern stabilities

An interesting quantity to look at is the distribution of pattern stabilities, Δ^μ. It provides information on the deviation of errors from the threshold stability κ and thereby permits the calculation of the rate of errors and determines the dynamics of related attractor neural networks. The density $\rho_\xi(\Delta)$ of the distribution of stabilities is the relative volume of solution space with stability Δ:

$$\rho_\xi(\Delta) = \lim_{\beta\to\infty} \frac{1}{Z} \int \mathrm{D}J \exp\Big[-\beta \sum_\mu V(\Delta^\mu)\Big] \delta(\Delta - \Delta^\nu).$$

Since the patterns are independently identically distributed random variables, we can set $\nu = 1$. If we assume that $\rho_\xi(\Delta)$ is self-averaging, we can find its distribution by calculating its average value $\rho(\Delta)$ using the replica trick.

$$\rho(\Delta) = \lim_{\beta\to\infty} \lim_{n\to 0} \left\langle Z^{n-1} \int \mathrm{D}J \exp\left[-\beta \sum_\mu V(\Delta^\mu)\right] \delta(\Delta - \Delta^1) \right\rangle$$

where the brackets denote the pattern average. The calculation is similar to that of the free energy except for the average over pattern 1 [9, 10].

We find for the distribution of local stabilities

$$\rho(\Delta) = \int \mathrm{D}z_0 \frac{\int \mathrm{D}z_1 \exp\left\{-wx\left[V(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right\} \delta(\Delta - \lambda_0)}{\int \mathrm{D}z_1 \exp\left\{-wx\left[V(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right\}}. \tag{5}$$

The learning error, e, is the fraction of unsatisfied inequalities (1), i.e. the number of patterns with stability $\Delta < \kappa$. We find them by integrating the distribution of stabilities:

$$e = \int_{-\infty}^{\kappa} \mathrm{d}\Delta\, \rho(\Delta).$$

Information on the 'badness' of errors, as contained in the distribution of stabilities, is thereby lost.

The distribution of local stabilities has important implications for the performance of the network during retrieval. Let us consider an attractor neural network of binary neurons, which are updated according to equation (1). If we initialize the network in a state $S_i(t_0)$ with an overlap $m(t_0) = \frac{1}{N}\sum_i \xi_i^1 S_i(t_0)$ with the first pattern, then in the next time step this overlap will have evolved according to

$$m(t + \Delta t) = \int \mathrm{d}\Delta\, \rho(\Delta) \operatorname{erf}\left(\frac{m(t)\Delta}{\sqrt{2[1 - (m(t))^2]}}\right)$$

with $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}y \exp(-y^2)$. For strongly diluted networks [11] this equation can be iterated to determine the fixed points of the dynamics. A stable fixed point, $m^* \neq 0$, characterizes the retrieval quality; an unstable one characterizes the minimal overlap of the initial state which is required for retrieval. Hence $\rho(\Delta)$ determines the basin of attraction in strongly diluted nets [9].

5. Discussion

In this paragraph we will evaluate the one-step RSB expressions of the last two sections for specific cost functions. We thereby test the global stability of the RS approximation in the regime where it is locally stable and find results also in the regime of local instability of the RS ansatz. For the three cost functions to be considered the lines of instability of the RS ansatz are shown in figure 1. The leftmost line is the line of storage capacity.

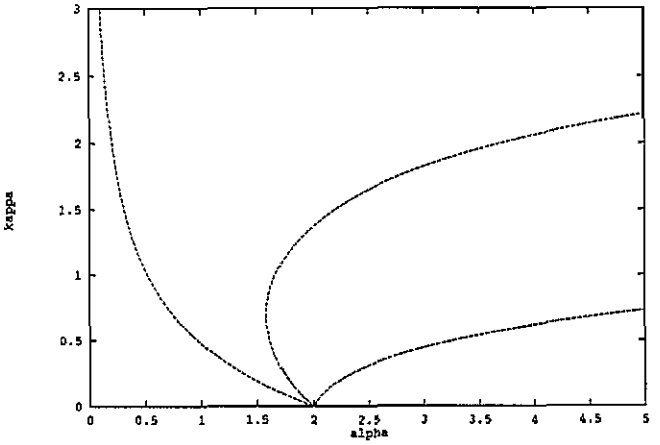

Figure 1. Storage capacity and AT lines. The leftmost line is the critical line of storage capacity. Next, to the right, is the line of local instability of the RS ansatz for the Gardner–Derrida cost function; the rightmost line is that for the Perceptron cost function. The line of instability for the Adatron cost function is the $\kappa = 0$ axis.

5.1. Gardner–Derrida cost function

Evaluating the free energy,

$$\langle f \rangle = \min_{x,q_0,w} \left\{ \frac{q_0}{2x(1+w\Delta q)} + \frac{\log(1+w\Delta q)}{2wx} + \frac{\alpha}{wx} \int \mathrm{D}z_0 \log \left[\int_{\frac{A-\sqrt{z_0}}{\sqrt{\Delta q}}}^{\frac{A}{\sqrt{\Delta q}}} \mathrm{D}z_1 \exp\left\{ -\frac{w}{2}(A - z_1\sqrt{\Delta q})^2 \right\} + \exp\{-wx\} H\left(\frac{\sqrt{2x} - A}{\sqrt{\Delta q}} \right) + H\left(\frac{A}{\sqrt{\Delta q}} \right) \right] \right\} \tag{6}$$

where $A = \kappa - z_0\sqrt{q_0}$ and $H(x) = \int_x^\infty \mathrm{D}y$, and comparing RS and RSB results, as done in figure 2, one finds that RS is *globally unstable for* $\alpha > \alpha_c$, independent of its local stability. This implies that, above saturation, the solution space is disconnected.

Note that for the Gardner–Derrida learning rule the free energy is equal to the rate of errors, since the cost function simply counts the number of errors.

The transition to RSB at $\alpha = \alpha_c$ can also be seen by looking at the order parameter q_0. It measures the average overlap of solutions from different regions of solution space. For small storage rates, it has the behaviour shown in figure 3. For $\alpha > \alpha_c$, q_0 branches off continuously from the replica-symmetric value $q_0 = q_1 = 1$. In the limit of infinite storage, $\alpha \to \infty$, we expect q_0 to approach 0 since, in this case, any network, i.e. any vector of

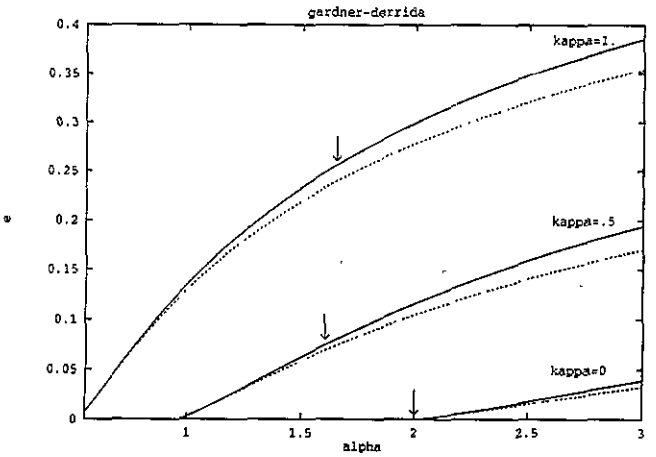

Figure 2. Rate of errors for the Gardner–Derrida learning rule. Lower curves correspond to RS and upper curves to RSB. The arrows indicate the points of local instability of the RS solution.

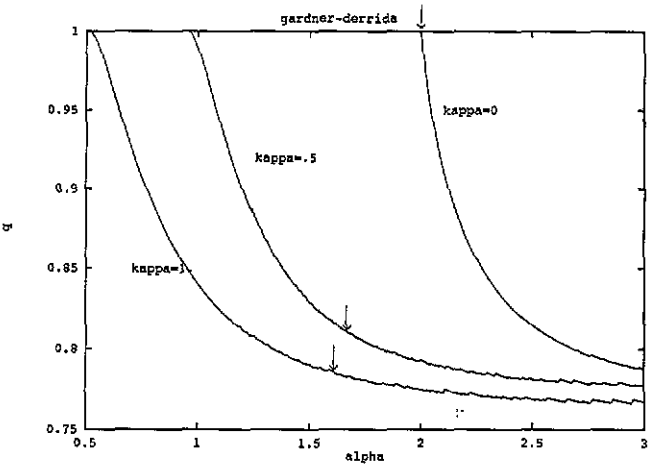

Figure 3. The order parameter q_0 shows the degree of RSB.

couplings, should do equally well on the given problem, just as well as by taking a random guess at the output. However, this is not what we find. We solved self-consistently the saddle-point equations in the limit $\alpha \to \infty$ with the ansatz

$$x \to 0 \qquad wx \to 0 \qquad w(1-q_0) \to \infty \qquad \text{for} \quad \alpha \to \infty$$

and find the following behaviour of the order parameters:

$$1-q_0 \approx \frac{9c_1}{\log\alpha} \qquad x \approx 9^{3/2}/c_0^2 \frac{1}{\alpha(\log\alpha)^{3/2}} \qquad wx \approx 1/9^{3/4} \frac{(\log\alpha)^{3/4}}{\sqrt{\alpha}}$$

where $c_0 = \frac{2}{3\sqrt{\pi}} \exp\left\{-\kappa^2/2\right\}$ and $c_1 = \frac{1}{2\pi\sqrt{2}} \exp\left\{-\kappa^2/2\right\}$. Since the smallest overlap scale q_0 of a correct RSB solution should tend to zero as $\alpha \to \infty$, we conclude that one-step RSB is incorrect at high storage levels. We have not studied at what point a transition to higher-order RSB occurs.

Evaluating the distribution of stabilities, the numerator in (4) becomes

$$\delta(\Delta-\kappa)\int_{\frac{\Delta-\sqrt{2x}}{\sqrt{\Delta q}}}^{\frac{A}{\sqrt{\Delta q}}} Dz_1 \exp\left[-\frac{w}{2}(A-z_1\sqrt{\Delta q})^2\right]$$
$$+\frac{\exp(-wx)}{\sqrt{2\pi\Delta q}}\exp\left[-\frac{(\Delta-z_0\sqrt{q_0})^2}{2\Delta q}\right]\Theta(\kappa-\sqrt{2x}-\Delta)$$
$$+\frac{1}{\sqrt{2\pi\Delta q}}\exp\left[-\frac{(\Delta-z_0\sqrt{q_0})^2}{2\Delta q}\right]\Theta(\Delta-\kappa)$$

so that three terms contribute to the distribution, leading to a gap of width $\sqrt{2x}$. As shown in figure 4 in one-step RSB, the δ-peak is decreased and the gap is narrower than in RS. Formally, this is because the order parameter x takes on smaller values in RSB.

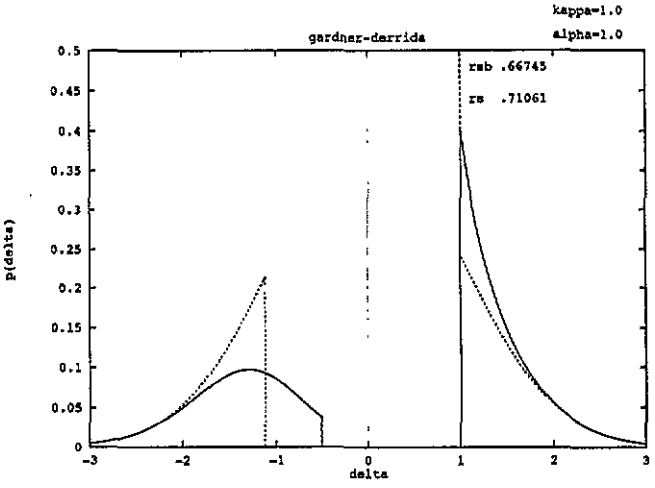

Figure 4. The distribution of stabilities for the Gardner–Derrida cost function at $\alpha = 1.0, \kappa = 1$, where RS is locally stable. The RSB curve is the full line, and the RS is the dotted line. The δ-peak has a height of 0.710 in RS and 0.667 in RSB.

We have used the distribution of stabilities to calculate the stable and unstable fixed points of the diluted dynamics. The effects of the RSB appear to be rather small. In figure 5, we show the fixed points along the line $\alpha = 0.35$ and we find a significant change due to RSB only for small values of the retrieval overlap m.

5.2. Perceptron cost function

The analytic expression for the free energy is

$$\langle f\rangle = \min_{x,q_0,w}\left\{\frac{q_0}{2x(1+w\Delta q)}+\frac{\log(1+w\Delta q)}{2wx}\right.$$
$$+\frac{\alpha}{wx}\int Dz_0 \log\left[\int_{\frac{A-\kappa}{\sqrt{\Delta q}}}^{\frac{A}{\sqrt{\Delta q}}} Dz_1 \exp\left\{-\frac{w}{2}(A-z_1\sqrt{\Delta q})^2\right\}\right.$$
$$\left.\left.+\exp\left\{\frac{wx}{2}(x+wx\Delta q-2A)\right\}H\left(\frac{x(1+w\Delta q)-A}{\sqrt{\Delta q}}\right)+H\left(\frac{A}{\sqrt{\Delta q}}\right)\right]\right\} \tag{7}$$

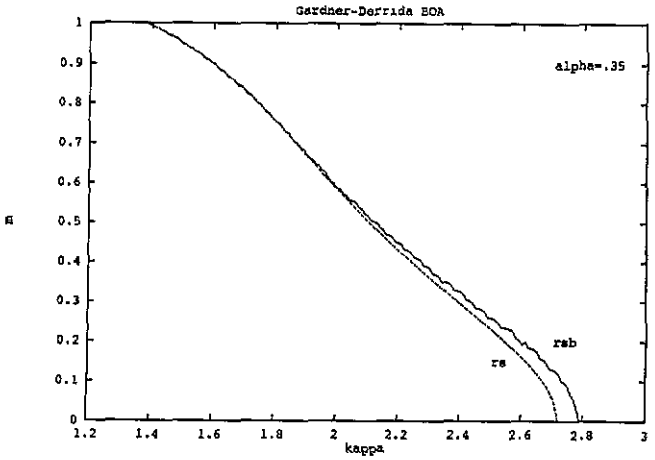

Figure 5. Fixed points of the diluted dynamics for the Gardner–Derrida learning rule at $\alpha = 0.35$. The lower and upper curve show the stable fixed point in RS and in one-step RSB, respectively. The unstable fixed point is at $m = 0$.

where, again, $A = \kappa - z_0\sqrt{q_0}$. In the region where RS is locally stable, we do not find numerically an RSB solution different from the RS solution. Thus the RS solution is *globally stable below the AT line*. Our RSB yields firstly results in the regime above the AT line. But we have not tested either local or global stability of the one-step RSB solution.

Since the difference in free energy of the RS and RSB solutions is of the order of 10^{-4}, we show only the behaviour of the order parameter q_0 in figure 6 which expresses the degree of symmetry breaking.

The numerator of the distribution of local stabilities (4) takes the form

$$\delta(\Delta - \kappa) \int_{\frac{A-x}{\sqrt{\Delta q}}}^{\frac{A}{\sqrt{\Delta q}}} \mathrm{D}z_1 \exp\left[-\frac{w}{2}(A - z_1\sqrt{\Delta q})^2\right]$$

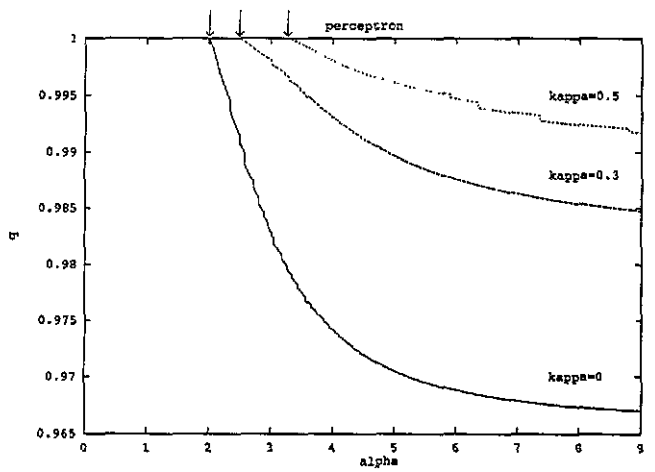

Figure 6. Order parameter q_0 shows RSB for the perceptron cost function. Arrows again indicate the points of local instability of the RS approximation.

$$+\frac{\exp\left\{-wx(\kappa-\Delta+\frac{1}{2}x)\right\}}{\sqrt{2\pi\Delta q}}\exp\left[\frac{(\Delta-z_0\sqrt{q_0})^2}{2\Delta q}\right]\Theta(\kappa-\Delta)$$

$$+\frac{1}{\sqrt{2\pi\Delta q}}\exp\left[\frac{(\Delta-z_0\sqrt{q_0})^2}{2\Delta q}\right]\Theta(\Delta-\kappa)$$

showing that for this cost function there is no gap in the distribution. The distribution of local stabilities is not changed significantly by RSB. The δ-peak decreases slightly in comparison to RS, as shown in figure 7.

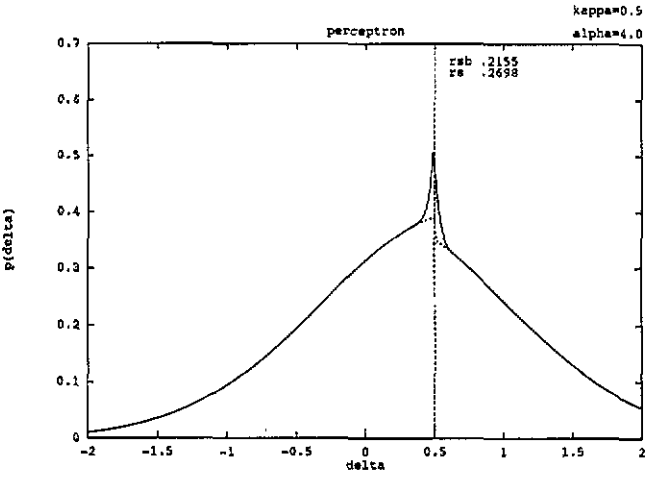

Figure 7. Distribution of stabilities for the perceptron cost function at $\alpha = 4$ and $\kappa = 0.5$. The RSB curve is shown by the full line, and the RS curve is dotted. The δ-peak has weights 0.270 and 0.215 in RS and RSB, respectively.

Figure 8 shows that the differences between the RS and RSB rate of errors are much greater than those in the free energy, which is sensible since the free energy can be written as

$$f = \int_{-\infty}^{\kappa} \mathrm{d}\Delta\, \rho(\Delta)\Delta\,.$$

We expect that the effects of one-step RSB on the fixed points of the diluted dynamics of attractor neural networks are even less than for the Gardner–Derrida learning rule, since the changes in the distribution of stabilities are less significant.

5.3. Adatron cost function

Here the free energy takes the form

$$\langle f\rangle = \min_{x,q_0,w}\left\{\frac{q_0}{2x(1+w\Delta q)}+\frac{\log(1+w\Delta q)}{2wx}+\frac{\alpha}{wx}\int \mathrm{D}z_0 \right.$$

$$\left.\times\log\left\{\int_{-\infty}^{\frac{A}{\sqrt{\Delta q}}}\mathrm{D}z_1\exp\left[-\frac{wx}{2x+1}(A-z_1\sqrt{\Delta q})^2\right]+H\left(\frac{A}{\sqrt{\Delta q}}\right)\right\}\right\}. \qquad (8)$$

Similarly to the perceptron cost function, we cannot find an RSB solution of the adatron learning rule below the AT line where the RS results are stable. Since for this learning rule the AT line lies along the $\kappa = 0$ axis, we do not find RSB at all.

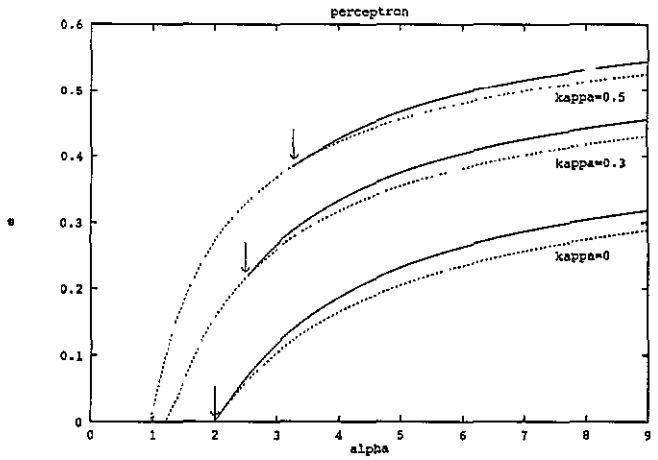

Figure 8. The rate of errors for the perceptron learning rule. The arrows indicate the AT line.

6. Conclusions

In the regime where perfect storage becomes impossible, we have studied the perceptron using three different learning rules formulated as the minimization of the cost function. We have tested the global stability of the replica-symmetric ansatz by considering a one-step replica symmetry breaking ansatz. This amounts to testing the connectedness of the solution space. We have found that for the Gardner–Derrida cost function, the replica symmetry is *globally unstable* above the critical line of storage $\alpha_c(\kappa)$, even where it is locally stable. The minimal number of errors and the distribution of local stabilities are markedly modified by replica symmetry breaking (see figures 2 and 4), while the retrieval properties of the corresponding attractor neural network are only slightly affected (figure 5). For the perceptron and adatron cost functions, on the other hand, local stability of the replica symmetric ansatz implies global stability. For the latter this should not come as a surprise since the cost function is a convex function of the couplings, as can be verified easily by calculating the matrix of second derivatives. This implies that local minima (with respect to the J's) are also global minima.† For the perceptron cost function, our calculation in one-step replica symmetry breaking provides approximate results at values of the storage ratio above the AT line (figures 7 and 8).

We suppose that the qualitatively different behaviour of the learning rules comes from the discreteness of the Gardner–Derrida cost function which takes on values $0, \ldots, p$ only, while the other cost functions are continuous. A similar situation occurs in the theory of annealed dilution in neural networks with continuous weights [12, 13]. If all couplings are present, replica symmetry holds for all $\alpha < \alpha_c$. If, however, a certain degree of dilution is assumed, discrete variables, $c_{ij} = 0, 1$, describing the missing bonds enter and replica symmetry breaks down. It is intuitive that, although replica symmetry breaking can occur with both discrete and continuous degrees of freedom, it is more natural in the former case, since the associated breaking of the solution space into different disconnected pieces is much more likely.

† We thank the referees for pointing this out to us.

References

[1] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[2] Cover T 1965 *IEEE Trans. Electron. Comput.* **14** 257
[3] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[4] Griniasty M and Gutfreund H 1990 *J. Phys. A: Math. Gen.* **24** 715
[5] Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A. Math. Gen.* **23** 3361
[6] Fontenari J F and Meir R 1991 *Phys. Rev.* A **45** 8874
[7] Biehl M and Opper M 1991 *Phys. Rev.* A **44** 6888
[8] Erichsen R, Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** L61
[9] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[10] Kepler T B and Abbot L T 1988 *J. Physique* **49** 1657
[11] Gardner E, Derrida B and Zippelius A 1987 *Europhys. Lett.* **4** 167
[12] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643
[13] Garcés R, Kuhlmann P and Eissfeller H 1992 *J. Phys. A: Math. Gen.* **25** L1335