Learning and retrieval in attractor neural networks above saturation

1991 J. Phys. A: Math. Gen. 24 715

(http://iopscience.iop.org/0305-4470/24/3/030)

J. Phys. A: Math. Gen. **24** (1991) 715–734. Printed in the UK

Learning and retrieval in attractor neural networks above saturation

M Griniasty and H Gutfreund

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

Received 26 July 1990

Abstract. Learning in the context of attractor neural networks means finding a synaptic matrix J_{ij} for which a certain set of configurations are fixed points of the network dynamics. This is achieved by a number of learning algorithms designed to satisfy certain constraints. This process can be formulated as gradient descent dynamics to the ground state of an energy function, corresponding to a specific algortihm. We investigate neural networks in the range of parameters when the ground-state energy is positive; namely, when a synaptic matrix which satisfies all the desired constraints cannot be found by the learning algorithm. In particular, we calculate the typical distribution functions of local stabilities obtained for a number of algorithms in this region. These functions are used to investigate the retrieval properties as reflected by the size of the basins of attraction. This is done analytically in sparsely connected networks, and numerically in fully connected networks. The main conclusion of this paper is that the retrieval behaviour of attractor neural networks can be improved by learning above saturation.

1. Introduction

The neural networks, which have attracted so much attention as models of associative memory, constist of N binary neutrons connected by synaptic couplings J_{ij}. The probability of neuron i to be, at time $t+1$, in one of the two states $S_i = +1$ (active) or $S_i = -1$ (idle) is given by

$$P(S_i) = [1 + \exp(-2\beta S_i h_i(t))]^{-1} \tag{1}$$

where h_i is the local field at the previous time step

$$h_i = \sum_{j \neq i} J_{ij} S_j \tag{2}$$

and β is a parameter related to the synaptic noise. In the absence of noise, when $\beta \to \infty$, the dynamics reduces to

$$S_i(t+1) = \operatorname{sgn}[h_i(t)]. \tag{3}$$

The basic idea is that any cognitive response is represented by a dynamical flow leading to an attractor of the network dynamics. In particular, the network serves as a model for associative memory, if a set of specified configurations, $\{\xi_i^\mu = \pm 1\}$ $(i = 1, \ldots, N;\ \mu = 1, \ldots, p)$, which are representations of the memorized concepts, are attractors of the network dynamics. This means that, starting from an initial configuration $\{S_i\}$, which has a sufficiently large overlap with one of the memorized patterns, the system will flow to that pattern as a fixed point of dynamics, or in the presence of noise the system will fluctuate in a close neighbourhood of that pattern in configuration space.

0305-4470/91/030715+20$03.50

Learning in the context of neural network models means finding a synaptic matrix J_{ij} which ensures a dynamical behaviour leading to specific attractors. The task is to organize the space of network states in basins of attraction around *a priori* known memory states.

In the history of attractor neural networks there have been two lines of approach to this problem. The first, following the work of Hopfield (1982, 1984), assumes a specific form of the J_{ij}'s on the stored patterns. The study of this line of approach focused on the performance of the network in the retrieval mode (Amit *et al* 1985, 1987). In the other approach, following the work of Gardner (1988), instead of making an ansatz about the form of the J_{ij} matrix, one considers the J_{ij}'s as dynamical variables, which are modified by a learning algorithm to satisfy certain constraints.

Learning as an error correction process. Gardner's (1988) approach to the problem of learning, leading to a statistical mechanics formulation in the space of J_{ij}, starts from the requirement that a set of p chosen patterns $\{\xi_i^\mu\}$, $i=1, N$; $\mu=1,\ldots,p$, are fixed points of the dynamics (3). This is clearly guaranteed when

$$\Delta_i^\mu \equiv \xi_i^\mu \sum_j \frac{J_{ij}}{\sqrt{N}} \xi_j^\mu > \kappa \tag{4a}$$

for every i and μ. We shall refer to the expressions on the left-hand side of these inequalities as the local stabilities. They are proportional to the local field on site i when the network is in state $\{\xi_i^\mu\}$. This state is a fixed point even for $\kappa=0$, but a finite κ is required to ensure significant basins of attraction. When $\kappa>0$, it is necessary to constraint the synaptic matrix in order to avoid the freedom in κ due to an overall scaling of the inequalities. A commonly used constraint is

$$\sum_j J_{ij}^2 = N. \tag{4b}$$

The search for a matrix J_{ij} which satisfies (3) and (4), proceeds by an error correction process. One defines the parameters

$$\varepsilon_i^\mu = \theta(\kappa - \Delta_i^\mu) \tag{5}$$

where $\theta(x)$ is the step function: $\theta(x)=1$ for $x>1$, and $\theta(x)=0$ otherwise. Starting from an initial J_{ij}^0, the current value of J_{ij} is modified, for each pattern and at each site, by

$$\Delta J_{ij} = \frac{1}{N} \varepsilon_i^\mu f(\Delta_i^\mu) \xi_i^\mu \xi_j^\mu. \tag{6}$$

Thus, J_{ij} is changed only when $\varepsilon_i^\mu = 1$, namely only in order to correct an 'error', when the corresponding inequaltiy is not satisfied. The range of possible choices of the functions $f(\Delta_i^\mu)$ has recently been discussed in a review paper by Abbott (1990). Let us mention here two cases.

(*a*) $f(\Delta_i^\mu)=\gamma$. This is the perceptron algorithm, with step-size γ, which has originally been formulated for the two-layer (perceptron) network (Rosenblatt 1962, Minsky and Pappert 1988). Gardner (1988) has shown that this algorithm can be applied to multiply connected networks and that the convergence theorem, which guarantees that a solution is found by this procedure in a finite number of steps if such a solution exists, can be extended to this case with a finite κ.

(*b*) $f(\Delta_i^\mu) = \gamma(\kappa - \Delta_i^\mu)$. This is the relaxation algorithm. It is a modification of an algorithm suggested recently by Abbott and Kepler (1989a). It converges (if a solution exists) for $0 < \gamma \leq 2$.

Gardner (1988) has shown that for a set of random uncorrelated patterns $\{\xi_i^\mu\}$, a matrix J_{ij}, which satisfies (4*a*) and (4*b*), exists with probability approaching 1 as $N \to \infty$, if the storage parameter $\alpha = P/N$ does not exceed the critical storage capacity $\alpha_c(\kappa)$. In particular, $\alpha_c = 2$ for $\kappa = 0$. For a given α, this defines a critical $\kappa_c(\alpha)$. For example, $\kappa_c(1.0) \approx 0.5$ and $\kappa_c(0.5) \approx 1.0$.

Any matrix which satisfies equations (4) will have the stored patterns as fixed points. This, however, does not guarantee the performance of the network as a useful model for associative memory. The latter depends on the behaviour of the network in the retrieval mode, where the relevant questions are the size of basins of attraction and the stability of attractors to synaptic noise. A useful measure, which characterizes this behaviour is the distribution function of the stability parameters Δ_i^μ. Its effect on the basins of attraction was investigated by Kepler and Abbott (1988), Gardner (1989) and by Krauth *et al* (1988a), and its effect on the stability to synaptic noise was studied by Amit *et al* (1990). It is clear that higher weight at larger values of Δ implies higher stability. One way to achieve this in a learning algorithm is to increase κ in (4) as much as possible. Krauth and Mezard (1987) suggested a modification of the perceptron algorithm, which finds a matrix J_{ij}, for which these inequalities are satisfied with the largest possible κ.

The question to be addressed in this paper is: is it of advantage to increase κ even further, violating some of these inequalities, but improving thereby the retrieval properties?

Learning as an optimization process. To investigate the problem of learning in the domain $\kappa > \kappa_c$, it is more convenient to consider the learning as an optimization process rather than as an error-correcting process. To this end one defines, in the space of matrices J_{ij}, a cost function $E(\{J\})$, which has a global minimum at the desired J_{ij}. A discretization of the gradient descent dynamics associated with the cost function,

$$\frac{\partial J_{ij}}{\partial t} = -\frac{\partial E}{\partial J_{ij}} \tag{7}$$

defines a learning algorithm which leads to the minimum E, provided that this function has a sufficiently smooth surface in the space of J_{ij}. To calculate this minimum, one assigns to each synaptic matrix a Boltzmann probability and defines the partition function, just like in the canonical formulation of statistical mechanics, as

$$Z = \int \mathrm{d}\Omega_J \prod_i \delta(J_{-i}^2 - N) \exp[-\eta E(J)] \tag{8}$$

where we have introduced explicitly the normalization condition (4*b*). Due to the obvious analogy, we shall refer to the cost function, occassionally, as an 'energy'. The minimum 'energy' is given, as usually, by

$$E_0 = -\lim_{\eta \to \infty} \frac{\mathrm{d}}{\mathrm{d}\eta} \langle \ln Z \rangle_\xi \tag{9}$$

where the average is over the possible realizations of the patterns $\{\xi_i^\mu\}$.

2. Specific choices of the cost function

The cost function E can usually be written as a sum of local terms for each pattern μ

$$E = \sum_{i,\mu} V(\Delta_i^\mu) \tag{10}$$

We shall investigate several cases of the function $V(\Delta)$.

2.1. The Gardner-Derrida cost function

$$V(\Delta) = \theta(\kappa - \Delta). \tag{11}$$

In this case the energy is just the number of errors in the learning process. It was used by Gardner and Derridda (1988) to estimate the lowest possible fraction of unsatisfied inequaltities (4) when $\alpha > \alpha_c(\kappa)$. Note, however, that this cost function does not define a gradient-descent algorithm leading to a solution corresponding to the lowest fraction of errors.

2.2. The perceptron cost function

$$V(\Delta) = (\kappa - \Delta)\theta(\kappa - \Delta). \tag{12}$$

This form takes into account the degree by which each unsatisfied inequality is violated. Discretization of the relaxation dynamics (7) defines a learning algorithm, in which the J_{ij} are modified by

$$\Delta J_{ij} = \frac{\gamma}{N} \sum_\mu \varepsilon_i^\mu \left(\xi_i^\mu \xi_j^\mu - \frac{\Delta_i^\mu J_{ij}}{\sqrt{N}} \right) \tag{13}$$

where ε_i^μ is defined in (5). Except for the second term, which is due to the normalization condition (4*b*), this is just a version of the perceptron algorithm in which the patterns are learned in parallel.

2.3. The adatron cost function

$$V(\Delta) = (\kappa - \Delta)^2 \theta(\kappa - \Delta). \tag{14}$$

This modification of the previous form leads to the algorithm

$$\Delta J_{ij} = \frac{\gamma}{N} \sum_\mu 2\varepsilon_i^\mu (\kappa - \Delta_i^\mu) \left(\xi_i^\mu \xi_j^\mu - \frac{\Delta_i^\mu J_{ij}}{\sqrt{N}} \right) \tag{15}$$

which is the relaxation algorithm, mentioned above, in which the patterns are learned in parallel.

A learning algorithm based on a modification of the cost function in (14), without imposing the normalization condition on the synaptic matrix and replacing κ by 1, has been recently investigated by Anlauf and Biehl (1989), who proposed the term 'adatron algorithm' as a combination of the perceptron and the adaline (see below) algorithms. They show that whenever a solution with zero 'energy' exists, this algorithm will find one with the highest minimal stability. This is achieved significantly faster than by the minover algorithm of Krauth and Mezard (1987).

In all the three cases, discussed above, $E_0 =$ at $\alpha < \alpha_c(\kappa)$, when all the constraints (4) are satisfied. At $\alpha > \alpha_c$, they have different ground-state energies.

2.4. The adaline cost function

The three forms of the cost function, listed above, correspond to the constraints (4). When the inequalities in these constraints are replaced by equalities, the appropriate cost function is

$$V = (\kappa - \Delta)^2. \tag{16}$$

The J_{ij}'s are modified, by the gradient-descent dyanmics, just as in (15), except that this is done at each step, regardless of whether a specific inequality is satisfied or not (the masking parameter ε is now missing).

A modification of this cost function, in which κ is set to one and the synaptic matrix is not normalized, leads to the adaline learning algorithm, introduced originally by Widrow and Hoff (1960). This algorithm has been studied extensively. It was shown (Diederich and Opper 1987) that it leads to the pseudo-inverse synaptic matrix (Personnaz *et al* 1985, Kanter and Sompolinsky 1987). The dynamics of this algorithm has been studied by Hertz *et al* (1989) and Kinzel and Opper (1990).

3. Theory—general

The calculation of the minimum of 'energy' (9), involves the average of ln Z over all the possible realizations of the patterns $\{\xi_i^\mu\}$. Such averages are frequently encountered in statistical mechanics and are treated by the replica method, represented by the identity

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\langle Z^n \rangle - 1}{n} \tag{17}$$

where Z^n is the partition function of n identical replicas of the system. The basic parameter, which appears in the calculation is the overlap between ground-state configurations in two replicas

$$q_{\alpha\beta} = \frac{1}{N} \sum_j J_{ij}^\alpha J_{ij}^\beta. \tag{18}$$

The treatment is particularly simple if one assumes that all the replicas have identical ground states. This assumption is known as the replica symmetric approximation, in which $q_{\alpha\beta}$ reduces to a parameter q. When the number of stored patterns increases, the number of ground-state configurations increases, and $q \to 1$, indicating the limit of a single such configuration.

The calculation, which by now has become standard, leads in the replica symmetric approximation and in the limit $q \to 1$, to the expression

$$\langle Z^n \rangle = \exp[nN\eta G(\alpha, x)] \tag{19}$$

where

$$G(\alpha, x) = -\frac{1}{2x} + \frac{\alpha}{\eta} \int \mathrm{D}t \log \int \mathrm{d}\lambda \exp[-\eta F(\lambda, x, t)]. \tag{20}$$

The function in the exponent is

$$F(\lambda, x, t) = V(\lambda) + \frac{(\lambda - t)^2}{2x} \tag{21}$$

and we have defined $x \equiv \eta(1-q)$. Here, and in what follows

$$\mathrm{D}t \equiv \frac{\mathrm{d}t}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}.$$

The function $G(\alpha, x)$ has to be calculated at the saddle point with respect to x. In the limit $\eta \to \infty$, the integrand is dominated by the minimum of $F(\lambda, t)$ obtained at some value $\lambda_0(t)$. Thus, in this limit

$$G(\alpha, x) = -\frac{1}{2x} + \alpha \int \mathrm{d}t\, F(\lambda_0(x, t), x, t). \tag{22}$$

In view of (9), (17) and (19), $G(\alpha, x)$ is indeed the minimum energy, provided that x assumes its saddle-point value. Equating the derivative of $G(\alpha, x)$ with respect to x to zero, one finds

$$\alpha^{-1} = \int \mathrm{D}t(\lambda_0(x, t) - t)^2. \tag{23}$$

For given α, this equation determines x. Inserting this relation into (22), one gets a compact expression for the ground-state energy per site per pattern

$$E_0 = \int \mathrm{D}t\, V(\lambda_0(x, t)). \tag{24}$$

The critical storage capacity α_c, is defined as the maximum storage for which the ground-state energy is zero. In view of (22) this corresponds to $x = \infty$. As $\alpha > \alpha_c$, x becomes finite and decreases.

Distribution of local stabilities. The distribution of local stabilities, in the state of minimum energy, is given by

$$\rho(\Delta) = \lim_{\eta \to \infty} \langle \delta(\Delta - \Delta_i^\mu) \rangle_{J,\xi} \tag{25}$$

where the average over J_{ij} is performed, again, with the Boltzman probability subject to the normalization condition, and the average on ξ is over the possible realizations of the stored patterns. The calculation proceeds along the lines of Kepler and Abott (1988). The result is

$$\rho(\Delta) = \lim_{\eta \to \infty} \int \mathrm{D}t \frac{\exp[-\eta F(\Delta, x, t)]}{\int \mathrm{d}\lambda \exp[-\eta F(\lambda, x, t)]} \equiv \int \mathrm{D}t\, \rho_t(\Delta) \tag{26}$$

where the values of x is determined from (23). The function $\rho_t(\Delta)$ is normalized to unity and sharply peaked around the minimum of F. Thus,

$$\rho(\Delta) = \int \mathrm{D}t\, \delta(\Delta - \lambda_0(x, t)). \tag{27}$$

Multiplying this equation by $V(\Delta)$ and integrating over Δ, one finds a natural expression for the ground-state energy

$$E_0 = \int \mathrm{d}\Delta\, V(\Delta)\rho(\Delta). \tag{28}$$

Stability to replica symmetry breaking. The calculation, leading to E_0, is based on the replica symmetric approximation. This approximation is valid as long as the fluctuation matrix in replica space, around the replica symmetry solution, is positive definite. The criterion for the stability of this solution can be cast in the same form as in Gardner and Derrida (1988)

$$\alpha\gamma_1\gamma_2 < 1. \tag{29}$$

In the limit $q \to 1$, one gets just as there, $\gamma_1 = (1-q)^2$.

The calculation of γ_2 proceeds along similar lines, and the result for a general cost function is given by

$$\gamma_2 = \frac{1}{(1-q)^2}\int \mathrm{D}t\left\{1 - \frac{1}{x(\partial^2 F/\partial\lambda^2)|_{\lambda_0(t)}}\right\}^2. \tag{30}$$

4. Theory—specific cost functions

Let us now apply the results of the last sections to study the specific energy functions, mentioned in the introduction.

4.1. The Gardner-Derrida cost function

This case has been studied previously and the results for the fraction of errors, stability to replica symmetry breaking (Gardner and Derrida 1988) and the distribution of stablities (Amit *et al* 1990) are known. We discuss it here only for completeness, in order to show how it fits into the general framework developed above.

The key parameters, $\lambda_0(x, t)$ and $F(\lambda_0, x, t)$, are

$$\begin{aligned} &\lambda_0 = t \qquad && F(\lambda_0, x, t) = 0 && \text{for } t > \kappa \\ &\lambda_0 = \kappa \qquad && F(\lambda_0, x, t) = (\kappa - t)^2/2x && \text{for } \kappa - \sqrt{2x} < t < \kappa \\ &\lambda_0 = t \qquad && F(\lambda_0, x, t) = 1 && \text{for } t < \kappa - \sqrt{2x}. \end{aligned} \tag{31}$$

From (22), one finds,

$$G(\alpha, x) = \alpha\int_{-\infty}^{\kappa-\sqrt{2x}} \mathrm{D}t + \alpha\int_{\kappa-\sqrt{2x}}^{x} \mathrm{D}t\frac{(\kappa-t)^2}{2x} - \frac{1}{2x}. \tag{32}$$

For a given α and κ, the value of x is determined, using (23), by

$$1 = \alpha\int_{\kappa-\sqrt{2x}}^{\kappa} \mathrm{D}t(\kappa - t)^2. \tag{33}$$

We wish to point out that our definition of the parameter x differs from that of Gardner and Derrida (1988). To compare with their results one has to replace $\sqrt{2x}$ by x.

The distribution function, $\rho(\Delta)$, is easily derived from (27), as

$$\rho(\Delta) = [H(\kappa - \sqrt{2x}) - H(\kappa))]\delta(\Delta - \kappa) + \frac{1}{2\pi}\mathrm{e}^{-\Delta^2/2}[\theta(\Delta - \kappa) + \theta(\kappa - \sqrt{2x} - \Delta)] \tag{34}$$

where

$$H(y) = \int_y^{\infty} \mathrm{D}t.$$

In addition to the δ-function and the truncated Gaussian at $\Delta > \kappa$, which also appear in $\rho(\Delta)$ at $\alpha = \alpha_c(\kappa)$ (Kepler and Abbott 1988, Gardner 1989, Krauth *et al* 1988b), we now get a contribution at negative values of κ, separated by a gap of $\sqrt{2x}$ from the $\Delta = \kappa$. It was pointed out by Amit *et al* (1990) that this part of $\rho(\Delta)$ has a destructive effect on the basins of attraction. The distribution function, $\rho(\Delta)$, is shown, for $\kappa = 1$ and $\alpha = 0.55$, in figure 1(*a*).

Since $\rho(\Delta)$ is normalized to unity, we can deduce the fraction of unsatisfied inequalities (4) from

$$f = \int_{-\infty}^{\kappa} \rho(\Delta)\, \mathrm{d}\Delta \tag{35}$$

which in our case gives

$$f = H(\sqrt{2x} - \kappa). \tag{36}$$

All these results are valid as long as the replica symmetric solution is stable. Based on (28)–(30), this is the case when

$$\alpha \int_{\kappa-\sqrt{2x}}^{\kappa} \mathrm{D}t < 1. \tag{37}$$

4.2. The perceptron cost function

Going through the same procedure as before, we find

$$\begin{array}{lll} \lambda_0 = t & F(\lambda_0, x, t) = 0 & \text{for } t > \kappa \\ \lambda_0 = \kappa & F(\lambda_0, x, t) = (\kappa - t)^2/2x & \text{for } \kappa - x < t < \kappa \\ \lambda_0 = t + x & F(\lambda_0, x, t) = \kappa - t - x/2 & \text{for } t < \kappa - x. \end{array} \tag{38}$$

The relation between α and x is now given by

$$1 = \alpha \int_{\kappa-x}^{\kappa} \mathrm{D}t (t-\kappa)^2 + \alpha x^2 H(x-\kappa) \tag{39}$$

and the distribution of stabilities is

$$\rho(\Delta) = [H(\kappa - x) - H(\kappa)]\delta(\Delta - \kappa) + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\Delta^2/2}\, \theta(\Delta - \kappa) + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\Delta - x)^2/2}\, \theta(\kappa - \Delta). \tag{40}$$

The main difference to the previous case is that there is no gap in $\rho(\Delta)$, and the stabilities, corresponding to the violated inequalities (4), are concentrated in a truncated Gaussian below κ. This is shown in figure 1(*b*), for the same parameters as in the previous model. The fraction of errors is in this case

$$f = H(x - \kappa). \tag{41}$$

The replica symmetric approximation is valid when

$$\alpha \int_{\kappa-x}^{\kappa} \mathrm{D}t > 1. \tag{42}$$

The boundary line for replica symmetry breaking is compared in figure 2 with the corresponding line for the previous case (37). The region of validity of the replica symmetric approximation, is in the present case, significantly larger.

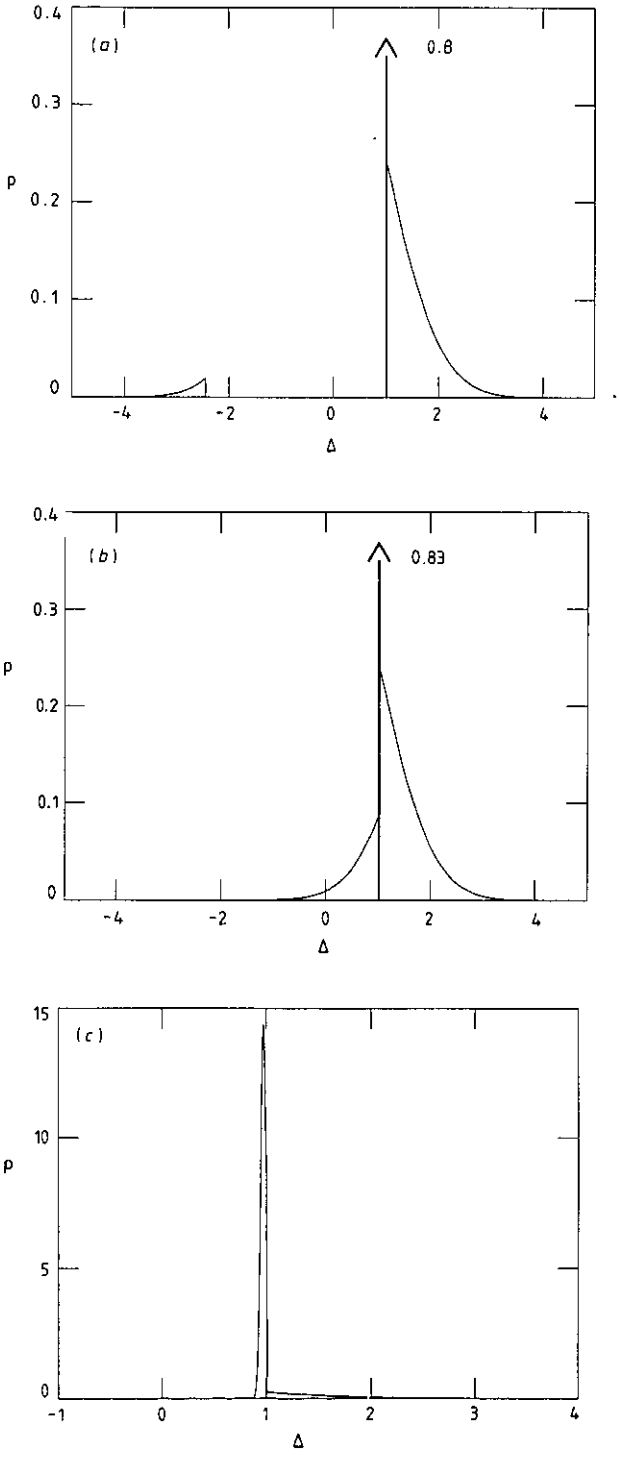

Figure 1. The distribution of local stabilities at $\alpha = 0.55$ and $\kappa = 1$ for the Gardner–Derrida (a), perceptron (b) and adatron (c) cost functions. The numbers in (a) and (b) indicate the height of the δ-function.

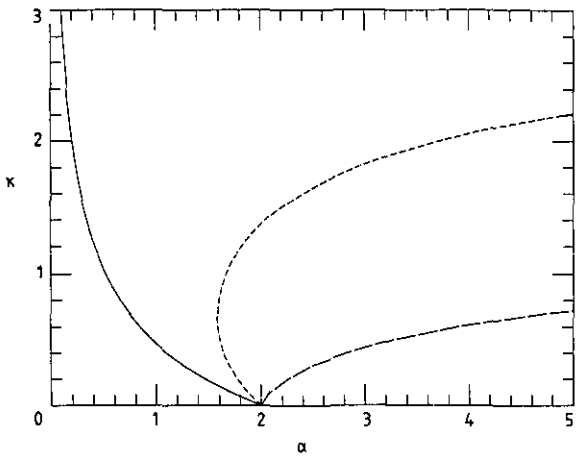

Figure 2. The critical line for replica symmetry breaking for the Gardner-Derrida (short-broken curve) and the perceptron (long-broken curve) cost function. The full curve is the $\alpha_c(\kappa)$ line.

4.3. The adatron cost function

In this case

$$\lambda_0 = t \qquad F(\lambda_0, x, t) = 0 \qquad \text{for } t > \kappa$$
$$\lambda_0 = (2x\kappa + t)/(2x+1) \qquad F(\lambda_0, x, t) = \{(\kappa - t)^2\}/(2x+1) \qquad \text{for } t < \kappa. \tag{43}$$

The equation for x is

$$\alpha \int_{-\infty}^{\kappa} \mathrm{D}t(t-\kappa)^2 = \left(\frac{2x+1}{2x}\right)^2 \tag{44}$$

and the distribution function, $\rho(\Delta)$, is given by

$$\rho(\Delta) = \theta(\Delta - \kappa)\frac{1}{\sqrt{2\pi}} e^{-\Delta^2/2} + \theta(\kappa - \Delta)\frac{2x+1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}[(2x+1)\Delta - 2x\kappa]^2\}. \tag{45}$$

Note, that there is no δ-function contribution at $\Delta = \kappa$, for any finite x. This function, evaluated at $\kappa = 1$ and $\alpha = 0.55$, is shown in figure 1(*c*). When $x \to \infty$, namely at $\alpha = \alpha_c$, the second term shrinks into a δ-function, and one gets the known result on the critical line $\alpha_c(\kappa)$. The fraction of errors is obtained by integrating the second term,

$$f = H(-\kappa). \tag{46}$$

We find the surprising result that the fraction of errors jumps discontinuously at $\alpha = \alpha_c(\kappa)$ and is then indepenent on α. This feature has also been pointed out by Anlauf and Biehl (1990).

The fraction of errors as function of κ, in the last three cases, is shown in figure 3 for $\alpha = 0.5$.

To check the criterion for stability to replica symmetry breaking, we find from (29) and 30,

$$\gamma_1\gamma_2 = H(-\kappa)\left(\frac{2x}{2x+1}\right)^2. \tag{47}$$

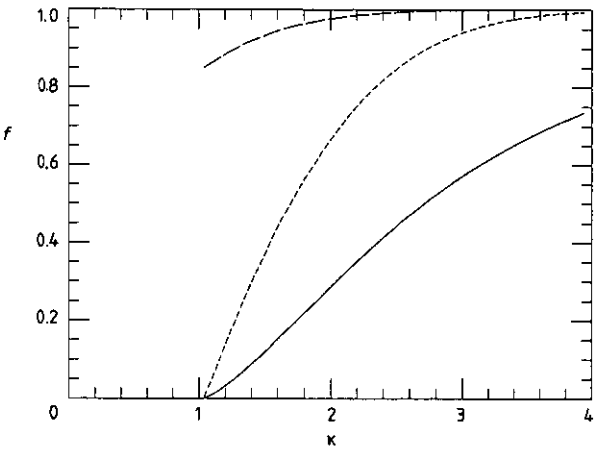

Figure 3. The fraction of errors in learning as function of κ for $\alpha = 0.5$, for the Gardner–Derrida (full curve) the perceptron (short-broken curve) and the adatron (long-broken curve) cost functions.

Using (44), one gets

$$\alpha\gamma_1\gamma_2 = H(-\kappa)\left(\int_{-\kappa}^{\infty} \mathrm{D}t(t+\kappa)^2\right)^{-1} \leqslant 1 \tag{48}$$

where the equality sign holds only for $\kappa = 0$. Thus, there is no replica symmetry breaking for any finite κ, and replica symmetry stability is marginal when $\kappa = 0$.

The distribution of local fields on the critical line $\alpha_c(\kappa)$ is the same in the three cases considered above, since, according to the classification of Abbott and Kepler (1989b) they belong the same universality class. The result is, however, very different away from this line.

4.4. The adaline cost function

In this case

$$\lambda_0 = (2x\kappa + t)/(2x+1) \qquad F(\lambda_0, x, t) = (t-\kappa)^2/(2x+1) \qquad \text{for all } t \tag{49}$$

and the relation between x and α, derived from (23), is

$$\alpha = \left(\frac{2x+1}{2x}\right)^2 \frac{1}{1+\kappa^2}. \tag{50}$$

At $x \to \infty$, we get the critical storage capacity $\alpha_c = (1+\kappa^2)^{-1}$. The distribution function of local stabilities, obtained from (26), is

$$\rho(\Delta) = \frac{2x+1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}[(2x+1)\Delta - 2x\kappa]^2\}. \tag{51}$$

On the critical line, when $x \to \infty$, this reduces to

$$\rho(\Delta) = \delta(\Delta - \kappa). \tag{52}$$

A straightforward calculation of (30), together with (29) and (50), gives

$$\alpha\gamma_1\gamma_2 = \frac{1}{1+\kappa^2} \tag{53}$$

so that the replica symmetric approximation is valid for $\kappa > 0$ and is marginally stable at $\kappa = 0$.

4.5. *The 'Hopfield' cost function*

Finally, as a pedagogical example, let us consider the cost function

$$V = -\Delta. \tag{54}$$

The minimum of (21) is, in this case, obtained for all t at $\lambda_0 = t + x$. Inserting this into the equation for G, and taking the saddle-point value with respect to x, we find the relation between α and x, to be $x = \sqrt{\alpha}$, and (27) gives immediately,

$$\rho(\Delta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\Delta - \frac{1}{\sqrt{\alpha}} \right)^2 \right]. \tag{55}$$

This is exactly the distribution of stabilities in the Hopfield model (Abbott and Kepler 1989b).

5. Basins of attraction

The time evolution of the system can be characterized by the overlaps of the network states, at time t, with the learned patterns

$$m_\mu(t) = \frac{1}{N} \sum_i \xi_i^\mu S_i(t). \tag{56}$$

An external stimulus is recognized as the concept μ if it imposes an initial state, which after some time drives the network to the attractor corresponding to this concept, namely, if asymptotically $m_\mu \approx 1$ and all the other overlaps are small. The smallest value of $m_\mu(t=0)$, which still leads to the corresponding attractor is a measure of the radius of the basin of attraction.

The overlap at time $t+1$, after one synchronous time step of the network dynamics, is related to $m(t)$ by

$$m(t+1) = \int \mathrm{d}\Delta\, \rho(\Delta)\, \mathrm{erf}\left(\frac{m(t)\Delta}{\sqrt{2(1-m^2(t))}} \right) \tag{57}$$

where erf is the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{x}} \int_0^x \mathrm{d}y \exp(-y^2). \tag{58}$$

In networks with asymmetric random sparse connectivity (Derrida *et al* 1987) in which each neuron is connected, on the average to C other neurons, so that $C \approx \log N$, (57) determines completely the dynamics of overlaps and can be iterated to its fixed point. This allows an analytical study of the static retrieval properties of such networks (Gardner 1989, Amit *et al* 1990).

Sparsely connected networks. We have calculated the fixed point overlaps of the dynamical equation (56), for the stability distributions associated with the cost functions discussed above. There are three possible scenarios.

(*a*) A single stable fixed point at zero. This corresponds to the lack of any retrieval capability.

(*b*) An unstable fixed point at zero and a stable one at $m^* > 0$. This represents a region of retrieval (the closer m^* to unity, the better the quality of retrieval) in which m^* is reached starting from any configuration for which the initial overlap with the target pattern is $m(t=0) > 0$. We shall refer to this case as retrieval with a full basin of attraction.

(*c*) Two stable fixed points at zero and $m^* > 0$, and one unstable fixed point at $0 < m_0 < m^*$. In this case m^* is reached only when starting from a configuration with $m(t=0) > m_0$. We shall refer to this case as retrieval with a partial basin of attraction. The size of the basin is characterized by m_0.

Figure 4(*a*) shows the retrieval phase diagram in the $\alpha - \kappa$ plane for the perceptron cost function. Retrieval above the critical line $\alpha_c(\kappa)$ is possible only at storage levels below unity. The replica symmetric approximation is valid (see figure 2) in the entire region where $m^* > 0$. In figure 4(*b*) we examine the dependence of m^* and m_0 as

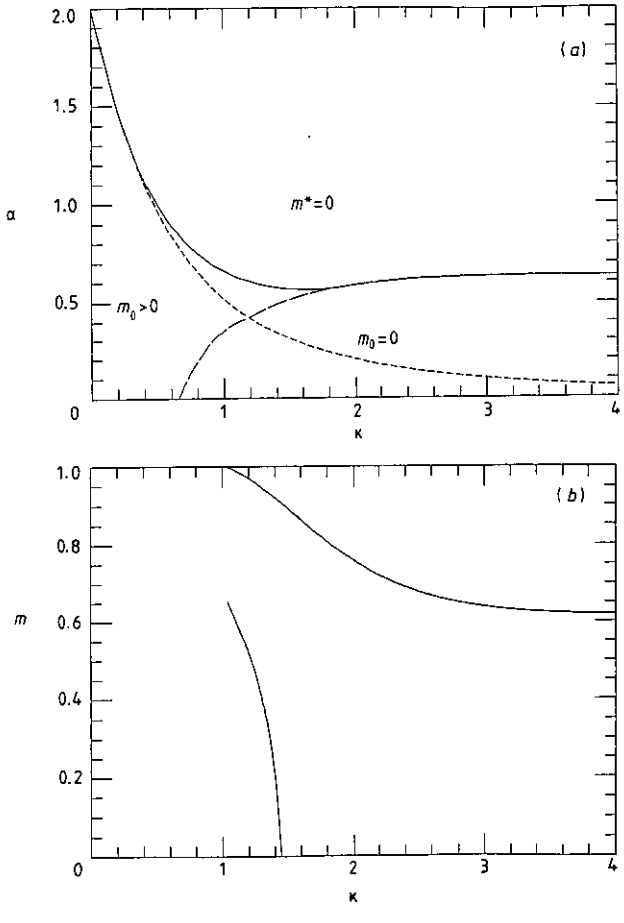

Figure 4. (*a*) Retrieval phase diagram for sparsely connected networks in case of the perceptron cost function. The full curve separates regions of retrieval and no retrieval. The long-broken curve separates between regions of full and partial basins of attraction. The short-broken curve is the curve $\alpha_c(\kappa)$. (*b*) Values of the fixed point overlap m^* (upper curve) and the overlap m_0 (lower curve), at the boundary of the basin of attraction, for $\alpha = 0.5$.

function of κ for $\alpha = 0.5$. Since we are interested in the behaviour above saturation, the curves start at $\kappa \approx 1$, which is the critical value of κ at this α. One observes that beyond the critical line the size of the basin of attraction increases but the quality of retrieval decreases.

The same calculation, for the adatron cost function, is presented in figures 5(a), (b). One finds that in this case the region of retrieval above the critical line is larger than in the previous case. Moreover, the retrieval properties (again for $\alpha = 0.5$) that the basin of attraction increases when κ is increased beyond κ_c and m^* is still very close to one. This continues to be the case even above $\kappa \approx 1.9$ which is the region of a full basin of attraction.

The results for the adaline cost function are shown in figures 6(a), (b). Unlike the previous two cases, there is now a significant region of retrieval beyond saturation also at high storage level. It is even possible, at low κ, to increase slightly the maximum storage capacity. The retrieval properties beyond κ_c, at $\alpha = 0.5$, are similar to the previous case.

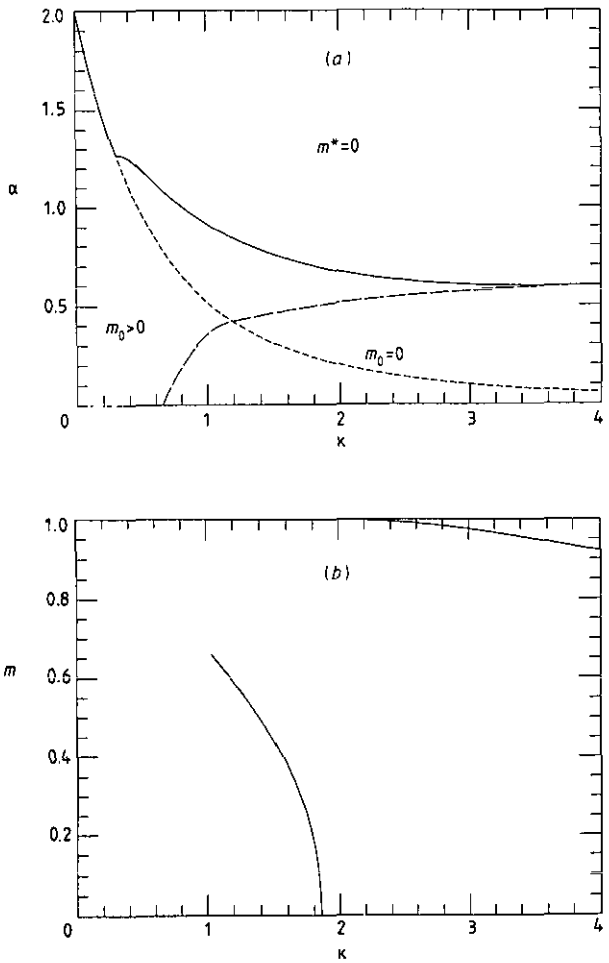

Figure 5. Same as figure 4, for the adatron cost function.

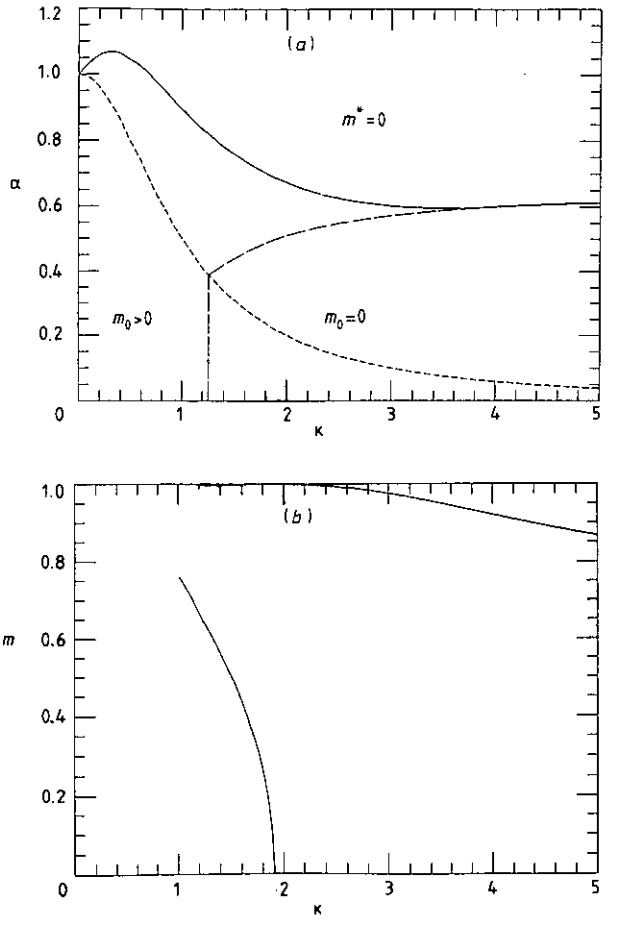

Figure 6. Same as figure 4, for the adaline cost function.

For the adaline case we also plot m_0 and m^* for a fixed $\kappa = 2$ as a function of α (figure 7). One gets perfect retrieval with a full basin of attraction well above the critical storage ($\alpha \approx 0.2$) at this κ. Only at $\alpha \approx 0.52$ does m_0 begin to increase (the basin of attraction decreases) until there occurs, at $\alpha \approx 0.67$, a sharp transition to $m^* = 0$.

Fully connected networks. In fully connected networks the relation (56) is valid only for a single time step. Kepler and Abbott (1988) suggested that it may, nevertheless, be used to estimate qualitatively the basins of attraction. They found, from numerical simulations, that an initial configuration with an overlap $m_\mu(0)$ with pattern μ will ultimately flow to this pattern, if the system goes in one time step more than half way to the desired attractor, namely, if

$$\frac{\mu_\mu(1) - m_\mu(0)}{1 - m_\mu(0)} \geqslant \frac{1}{2}. \tag{59}$$

Rather than using this relation for a qualitative comparison of the retrieval properties below and above saturation, we shall demonstrate the effect by numerical simulations.

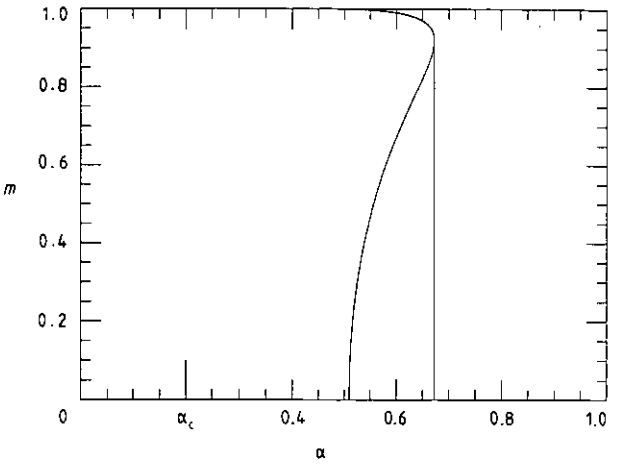

Figure 7. Values of m^* (upper curve) and m_0 (lower curve), for the adaline cost function, at fixed $\kappa = 2$ as a function of α.

We present results for the perceptron, adatron and adaline cost functions, obtained on networks of size $N = 200$ with 100 random uncorrelated patterns ($\alpha = 0.5$). In each case, a J_{ij} matrix is found by the corresponding algorithm. The latter is then used to calculate the distribution of stabilities and the size of the basins of attraction

Figure 8(a) shows the distribution of stabilities at $\kappa = 0.8$ and $\kappa = 1.2$, obtained with the perceptron algorithm, after averaging over several runs. The lower value of κ is slightly below saturation. In finite systems it is hard to find zero-energy solutions significantly closer to the theoretical critical value ($\kappa_c \approx 1.0$). The higher value of κ is well above saturation. The learning process, in this case, continues until the energy, computed at each step, no longer decreases. The final energy is 0.075, which should be compared with the theoretical value (for these α and κ) of 0.072. The theoretical fraction of errors in learning, which is the integrated weight below κ, is in this case, $f = 0.141$. In simulations we expect that about one half of the weight in the δ-function (which in this case is 0.744) is shifted below κ. Thus, we expect a fraction of errors of $f = 0.51$. The empirical value is 0.53. In figure 8(b) we show the overlaps of the fixed-point configurations, reached by the network dynamics (3), with one of the stored patterns, as a function of the initial overlap with that pattern. Each point is a result of averaging over 200 cases. One observes that, in this case, the stored patterns themselves are not fixed points of the dynamics and they flow to very close configurations with $m^* \approx 0.97$. This is due to the long tail in the distribution of stabilities, which extends to negative (though very small) values. We wish to point out, in passing, that the shape of $\rho(\Delta)$ for this algorithm is different from the theoretical form, which is obtained by averaging over all the possible solutions to the learning problem, and predicts a Gaussian tail above κ. In the present case, the maximum is above the threshold due to the fact that in the perceptron algorithm we keep modifying the interaction matrix by the same increment until the last memory is stabilized. In this process we reach local stabilities, which are significantly larger than necessary, if the goal is merely to satisfy the constraints (4).

Figure 9(a), (b) shows the same results for the relaxation algorithm, corresponding to the adatron cost function. The energy reached at $\kappa = 1.2$ is 0.02 while the theoretical

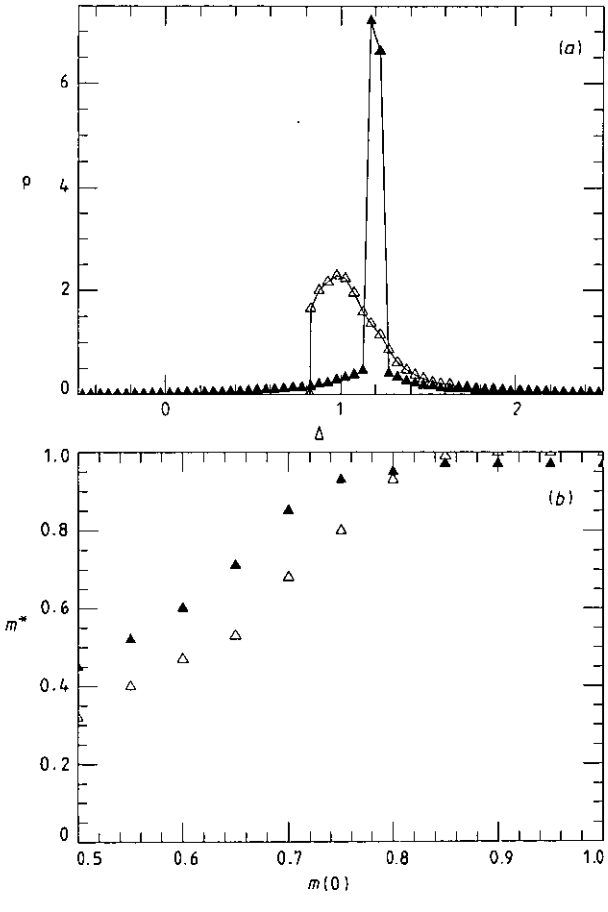

Figure 8. (*a*) The distribution of local stabilities at $\kappa = 0.8$ (open triangles) and $\kappa = 1.2$ (full triangles), obtained with the perceptron algorithm on networks of $N = 200$ with 100 stored uncorrelated patterns. (*b*) Fixed-point overlaps with the stored patterns as a function of the initial overlaps, averaged over 200 cases, for values of κ corresponding to (*a*).

value is 0.018. The empirical fraction of errors is 0.87, which should be compared with the theoretical $f = 0.885$. The tail of the distribution function below κ is much shorter than in the previous case, and is confined to positive values. The stored patterns, and configurations in significant neighbourhood around them, are therefore stable. Note that, unlike the previous case, with this algorithm the distribution of stabilities has the theoretical shape.

In figures 10(*a*), (*b*) we present similar results for the adaline algorithm at $\kappa = 1$ and $\kappa = 1.8$. The lower value correspond to the critical κ for $\alpha = 0.5$, and the theoretical distribution of stabilities should be a δ-function. At higher values of κ it broadens into a Gaussian distribution. For $\alpha = 0.5$ and $\kappa = 1.8$, we expect the average of the Gaussian to be at $\kappa \approx 1.2$ (from (50), (51)). This is indeed the case. The expected ground-state energy is 0.416. We find a value of 0.42. One observes again, even more than in the previous cases, that the basins of attraction are significantly increased in networks above saturation.

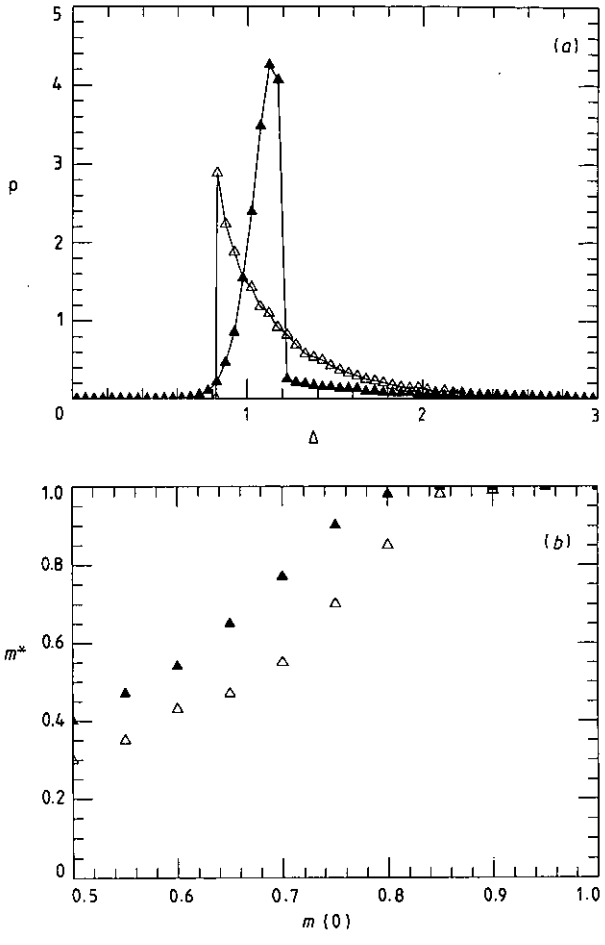

Figure 9. Same as figures 8(*a*), (*b*), for the relaxation algorithm, corresponding to the adatron cost function.

6. Conclusion

We have investigated several learning algorithms, formulated as gradient-descent dynamics of corresponding cost functions. These algorithms are well known and have been studied previously. These studies have, generally, been confined to the range of parameters below saturation, when the algorithms converge to solutions with zero 'energy', namely, when all the constraints of the learning problem are satisfied. The basic contribution of the present work is the emphasis on the region in the (α, κ)-plane, where this cannot be achieved.

We have explored the retrieval properties, at zero noise, in sparsely connected networks above saturation. For the algorithms associated with the perceptron and adatron cost functions, we find that retrieval is possible only at storage levels significantly below the critical storage capacity $\alpha_c = 2$. At these storage levels, it is of advantage to increase κ above its critical value and to improve thereby the retrieval behaviour. However, it is not possible to increase the storage capacity by allowing

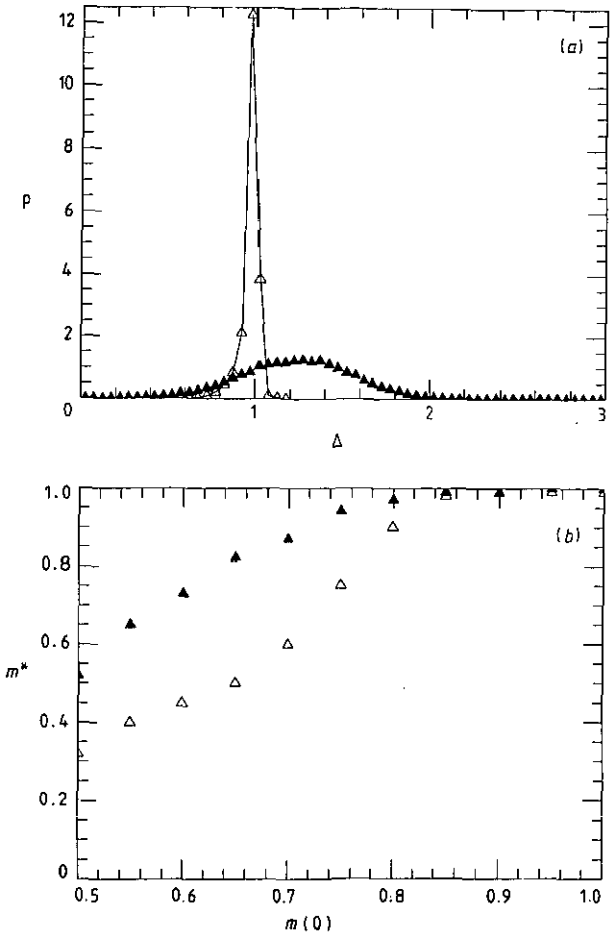

Figure 10. Same as figures 8(*a*), (*b*) for the adaline algorithm, except that now the open triangles correspond to $\kappa = 1$ and the full ones to $\kappa = 1.8$.

errors in learning. The same conclusion was reached by Amit *et al* (1990), who studied the Gardner-Derrida cost function. On the other hand, for the adaline algorithm, retrieval is possible near and slightly above the critical storage capacity, which in this case is $\alpha_c = 1$.

The properties of fully connected networks are very different from those of their sparsely connected counterparts. This was learned long ago by comparing the behaviour of the Hopfield network with its dilute version (Derrida *et al* 1987). We have not performed an extensive study, based on the empirical relation (59) or on numerical simulations, of the retrieval phase diagram for the fully connected networks. Instead, we have demonstrated the advantage of learning beyond saturation by specific examples of results of numerical simulations. A more systematic study of this problem will be discussed elsewhere.

We have concentrated, in this work, on attractor networks, but the methods and results are also relevant for feedforward networks, in particular, for a simple perceptron network. For example, in autoassociation problems the single-step dynamics of overlaps

(57) defines the mapping from the input to the output layer. A question of interest, in this context, is what algorithm guarantees the highest overlap in one time step. This problem will be discussed in a separate publication.

Acknowledgment

We are grateful to our colleagues Daniel Amit, Martin Evans and David Hansel for discussions of the issues raised in this paper. One of us (MG) acknowledges the support of the Levi Eshkol foundation of the Israeli Ministry of Science and Technology.

References

Abbott L F 1990 *Network* **1** 105
Abbot L F and Kepler T B 1989a *J. Phys. A: Math. Gen.* **22** L711
—— 1989b *J. Phys. A: Math. Gen.* **22** L745
Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
—— 1987 *Ann. Phys.* **173** 30
Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
—— 1990 *Parallel Processing in Neural Systems and Computers* ed R Eckmiller, G Hartmann and G Hanske (Amsterdam: Elsevier) p 153
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
—— 1989 *J. Phys. A: Math. Gen.* **22** 1969
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Hertz J A, Krogh A and Thobergsson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
—— 1984 *Proc. Natl. Acad. Sci. USA* **81** 3088
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Kinzel W and Opper M 1990 *Physics of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
Krauth W and Mezard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Krauth W, Nadal J-P and Mezard M 1988a *J. Phys. A: Math. Gen.* **21** 2995
Krauth W, Mezard M and Nadal J-P 1988b *Complex Systems* **2** 387
Minsky M and Pappert S 1988 *Perceptrons* (Cambridge, MA: MIT Press)
Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
Rosenblatt F 1962 *Principles of Neurodynamics* (Washington, DC: Spartan)
Widrow B and Hoff M E 1960 *WESCON Convention Report IV* (San Francisco: Western Periodicals Company)