Dynamics of a neural network with hierarchically stored patterns

J. Phys. A: Math. Gen. 23 (1990) 1801-1810. Printed in the UK

Dynamics of a neural network with hierarchically stored patterns

S Bacci†, G Mato‡ and N Parga

Centro Atómico Bariloche and Instituto Balseiro, 8400 S C de Bariloche, Argentina

Received 31 May 1989, in final form 5 December 1989

Abstract. We study the dynamics of a strongly diluted and a non-diluted hierarchical model. The first one is solved exactly. The phase diagram and the size of the attraction basins of the fixed points of the dynamics are calculated and their behaviour is compared.

1. Introduction

Neural networks have been proposed as models of content-addressable or associative memories (Little 1974, Hopfield 1982) in an attempt to explain properties of the nervous systems in terms of the behaviour of two-state neurons.

The thermodynamical properties of these models were extensively studied (Pereto 1984, Amit *et al* 1985a, 1987, Crisanti *et al* 1986) and they can be considered fully understood: the system can store and retrieve a quantity of patterns proportional to the number of neurons with a small number of errors (the maximum value of the coefficient of proportionality is $\alpha = 0.14$). However, these models have a number of drawbacks as candidates for modelling real nervous systems. The most important unrealistic features are the symmetry of the synaptic connections, the full connectivity and the quasi-orthogonality of the stored patterns.

One of the central features of the human brain is its ability to store and to access a huge and diverse amount of information, which consists not of isolated facts but rather in patterns correlated in a complex, general way with each other. Then it is desirable to study models which can provide insight in an intermediate regime when there is a special type of correlation, where the patterns are organised according to a hierarchical structure.

If the synaptic connections are not symmetric then the thermodynamical formalism cannot be applied because there is no Hamiltonian that allows the use of the equilibrium approach. So we are forced to develop a dynamical theory to describe the behaviour of the system, as well as to have some insight into the nature of its attractors. The solution of this problem is greatly simplified when there is not only asymmetry but also strong dilution (Derrida *et al* 1987). In this model the number of synapses per neuron is lower than the logarithm of the number of neurons. This is not a realistic assumption because in the human brain there are about 10^{10} neurons and each of them is connected to 10^4 other neurons. The number of stored patterns is proportional to

† Fellowship granted by Comisión Nacional de Energia Atómica.

‡ Fellowship holder of the Consejo Nacional de Investigaciones Cientificas y Técnicas.

the average coordination number $\tilde{C}$ of a neuron and the maximum coefficient is $\alpha = 2/\pi$. This model has been generalised (Kree *et al* 1987) for a connectivity per neuron tending to zero in the thermodynamic limit without the restriction $\tilde{C} \ll \ln(N)$.

Exact solutions of the dynamical problem were also found for different models with a finite number of stored patterns (Riedel *et al* 1988, Coolen *et al* 1988). All these solutions consist of a set of coupled recurrence equations for the overlaps between the state of the spins and some chosen memorised patterns.

Storing correlated patterns forces us to modify the learning rule because Hebb's rule is not useful in this case. Rules that can store an extensive number of patterns with arbitrary correlation have been proposed (Personnaz *et al* 1985) but they are non-local. On the other hand, many studies have been done on simpler rules that store hierarchically organised patterns (Parga *et al* 1986, Dotsenko 1986, Feigelman *et al* 1987, Gutfreund 1988, Bacci *et al* 1989a, b) through different techniques such as numerical simulations and thermodynamical analysis.

The scheme developed by Feigelman *et al* for generating p hierarchically correlated patterns is the following: first for each category we choose N_a random numbers $\xi_i^\alpha = 1, -1$ ($\alpha = 1, \ldots, N_a$ with equal probability); second we select $p = RN_a$ random numbers $\beta_i^{\alpha\gamma} = 1, 0$ ($\alpha = 1, \ldots, N_a$; $\gamma = 1, \ldots, R$) with probability c or $1-c$ ($0 < c < 1/2$). The value of the pattern (α, γ) is given by

$$\xi_i^{\alpha\gamma} = \xi_i^\alpha (1 - 2\beta_i^{\alpha\gamma}). \qquad (1)$$

This process is repeated independently for each site i. The orthogonal patterns ξ_i^α are called classes. It is easy to see that two patterns in the same class have an overlap $q = (1-2c)^2$ and that two patterns belonging to different classes have a null overlap.

The learning rule is given by

$$J_{ij} = \frac{1}{\tilde{C}} \sum_{\alpha=1}^{N_a} \xi_i^\alpha \xi^\alpha \left[1 + \frac{1}{\Delta} \sum_{\gamma=1}^{R} (\beta_i^{\alpha\gamma} - c)(\beta_j^{\alpha\gamma} - c) \right] \qquad (2)$$

where Δ is a parameter that controls the relative weight of the classes and the memorised patterns in the learning rule. For $\Delta = \Delta_0 = c(1-c)$ this learning rule becomes the one proposed by Parga *et al* (1986). The storage capacity of this system is of the same order of magnitude as the storage capacity in the Hopfield model.

In this work we study the dynamical behaviour of neural networks with hierarchically organised patterns. In section 2 this will be done for a learning rule that is a strongly diluted version of (2). In section 3 the non-diluted case is examined and an exact solution is found for a finite number of stored patterns. An extension for an extensive number of memorised patterns is considered. Section 4 contains a comparison of the results and the conclusions.

2. Dynamics of the strongly diluted model

In this model the synaptic interactions are:

$$T_{ij} = J_{ij} c_{ij} \qquad (3)$$

where the J_{ij} are given by (2) and c_{ij} are random variables with distribution

$$P(c_{ij}) = (1 - \tilde{C}/N)\delta(c_{ij}) + (\tilde{C}/N)\delta(c_{ij} - 1) \qquad (4)$$

where $\tilde{C} \ll \log(N)$.

The field $h_i(t)$ is evaluated at each site i

$$h_i(t) = \sum_{j \neq i} T_{ij} s_j(t) \tag{5}$$

and all the spins are updated simultaneously according to

$$s_i(t+1) = \begin{cases} +1 & \text{with probability } 1/[1+\exp(-2h_i(t)/T)] \\ -1 & \text{with probability } 1/[1+\exp(2h_i(t)/T)] \end{cases} \tag{6}$$

where T is introduced as a generalised temperature. Using the master equation for the spin distribution we can calculate the thermal average of spin s at time $t+1$, i.e. $\langle s_i(t+1)\rangle_T = \tanh(h_i(t)/T)$.

We consider the evolution of a configuration that has a macroscopic overlap with one class and two memorised patterns in this class. The relative quantities are

$$\begin{aligned} m_\alpha(t) &= 1/N \sum_{i=1}^{N} \xi_i^\alpha \langle s_i(t)\rangle_T \\ m_{\alpha\gamma}(t) &= 1/N \sum_{i=1}^{N} \xi_i^\alpha (1 - 2\beta_i^{\alpha\gamma}) \langle s_i(t)\rangle_T \end{aligned} \tag{7}$$

but instead of studying the temporal evolution of the overlap with the patterns it is more convenient to consider the overlaps with the fluctuation, which are defined by

$$U_{\alpha\gamma}(t) = \frac{1}{N\Delta} \sum_{i=1}^{N} \xi_i^\alpha (c - \beta_i^{\alpha\gamma}) \langle s_i(t)\rangle_T \tag{8}$$

and are related to the overlaps with the patterns through

$$m_{\alpha\gamma} = (1-2c) m_\alpha + 2\Delta U_{\alpha\gamma}. \tag{9}$$

To solve the dynamics we split the internal field in two parts: one corresponding to the condensed patterns and the second to all the others:

$$\begin{aligned} \xi_1^1 h_i(t) = {} & m_1 + (c - \beta_i^{11}) U_{11} + (c - \beta_i^{12}) U_{12} \\ & + \left\langle \sum_{\alpha>1} \sum_{y_r} \xi_1^1 \xi_i^\alpha \xi_{y_r}^\alpha S_{y_r}(t) + \frac{1}{\Delta} \sum_{(\alpha,\gamma)} \sum_{y_r} \xi_i^1 \xi_i^\alpha (\beta_i^{\alpha\gamma} - c) \xi_{y_r}^\alpha (\beta_{y_r}^{\alpha\gamma} - c) S_{y_r}(t) \right\rangle \end{aligned}$$

with $(\alpha, \gamma) \neq (1, 1), (1, 2)$. Here $\langle\ \rangle$ denotes average over the disorder.

Due to the condition of strong dilution (Derrida *et al* 1987) spins $s_i(t)$ depend on different sites at $t = 0$ and so they are uncorrelated. Moreover, the values $s_i(t)$ and $s_i(t+n)$ are also uncorrelated for $n \geq 1$ because all the sites in the tree of ancestors are different with probability one. The correlations being absent, we can replace the second term of the field by a Gaussian variable, the dispersion of which can be calculated from (10) giving $\sqrt{\alpha}$. For diluted models α is defined as $(p-1)/\tilde{C}$. It is important to remark that it is not enough that $\tilde{C}/N \to 0$ in the thermodynamic limit; a logarithmic dilution to neglect correlations is necessary (Kree *et al* 1987).

After averaging over the remaining classes and memorised patterns we can write the recurrence equations for m_1, U_{11} and U_{12}:

$$\begin{aligned} m_1(t+1) &= c^2 W + c(1-c)(X+Y) + (1-c)^2 Z \\ U_{11}(t+1) &= \frac{\Delta_0}{\Delta} \{-cW + (c-1)X + cY + (1-c)Z\} \\ U_{12}(t+1) &= \frac{\Delta_0}{\Delta} \{-cW + (c-1)Y + cX + (1-c)Z\} \end{aligned} \tag{11}$$

where

$$W=\int_{-\infty}^{\infty}\mathrm{d}y\frac{\mathrm{e}^{-y^2}}{\sqrt{\pi}}\tanh((m_1+(c-1)U_{11}+(c-1)U_{12}-\sqrt{2\alpha'})/T)$$

$$X=\int_{-\infty}^{\infty}\mathrm{d}y\frac{\mathrm{e}^{-y^2}}{\sqrt{\pi}}\tanh((m_1+(c-1)U_{11}+cU_{12}-\sqrt{2\alpha'})/T)$$

$$Y=\int_{-\infty}^{\infty}\mathrm{d}y\frac{\mathrm{e}^{-y^2}}{\sqrt{\pi}}\tanh((m_1+cU_{11}+(c-1)U_{12}-\sqrt{2\alpha'})/T)$$

$$Z=\int_{-\infty}^{\infty}\mathrm{d}y\frac{\mathrm{e}^{-y^2}}{\sqrt{\pi}}\tanh((m_1+cU_{11}+cU_{12}-\sqrt{2\alpha'})/T)$$

and $\alpha'=\alpha(\Delta_0/\Delta)^2$.

Starting from particular initial values we iterate these equations until a fixed point is reached. Let us summarise the picture that emerges.

Figure 1 shows the behaviour at $T=0$. Above the curve α_c^A $m_1=m_{11}=m_{12}=0$. This means that the system is not able to remember anything. Between α_i^A and α_c^R $m_1\neq 0$ and $m_{11}=m_{12}=0$. This fixed point corresponds to the class but not the stored patterns. Between α_c^R and α_c^S $m_1\neq 0$ and $m_{11}\neq 0, m_{12}=0$ or $m_{11}=0, m_{12}\neq 0$. The system remembers pattern 1 or 2 but not both. Finally below α_c^S $m_1\neq 0$ and $m_{11}=m_{12}\neq 0$. The system cannot distinguish between the two patterns. This behaviour is qualitatively different to that obtained by Derrida *et al* because in that case the symmetric solutions appear for a value of α greater than the one for which both patterns can be distinguished ($m_1\neq 0$ and $0\neq m_{11}\neq m_{12}\neq 0$), moreover for this solution the system can distinguish between the two patterns although none of their overlaps are zero. Perhaps this is due to the fact that in Derrida *et al* (1987), correlated patterns are being stored using Hebb's rule.

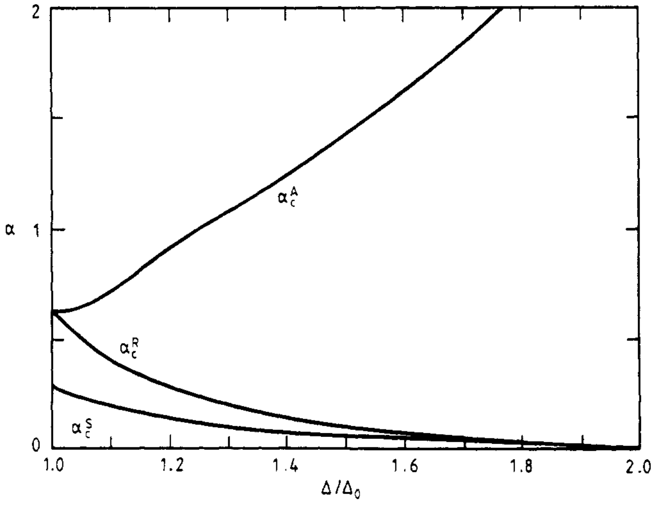

Figure 1. Storage capacity as a function of the parameter Δ/Δ_0 at $T=0$ for the strongly diluted model with $q=0.5$. Above α_c^A all the overlaps are null. Between α_c^A and α_c^R only the overlap with the class is non-null. Between α_c^R and α_c^S there is also a non-null overlap with one pattern. Below α_c^S all the overlaps are non-zero; furthermore the overlaps with the two stored patterns are equal.

In figure 2 we show the retrieval diagram for $\alpha = 0$. We can see the same four regions.

In figures 3 and 4 the diagrams $\alpha - T$ are shown for two values of Δ/Δ_0 (1 and 1.5 respectively). In figure 3 α_c^A and α_c^R coincide at $\Delta/\Delta_0 = 1$, as is also predicted by the thermodynamical analysis of the non-diluted model (Feigelman *et al* 1987).

Some analytical properties of the phase diagrams are possible to obtain. The value α_c^A at $T=0$ is given by $f'(0) = 1$ where

$$f(m) = \mathrm{erf}(m/\sqrt{2\alpha'}) \qquad (12)$$

because the other overlaps are zero. We can solve it to obtain $\alpha_c^A = (2/\pi)(\Delta/\Delta_0)^2$.

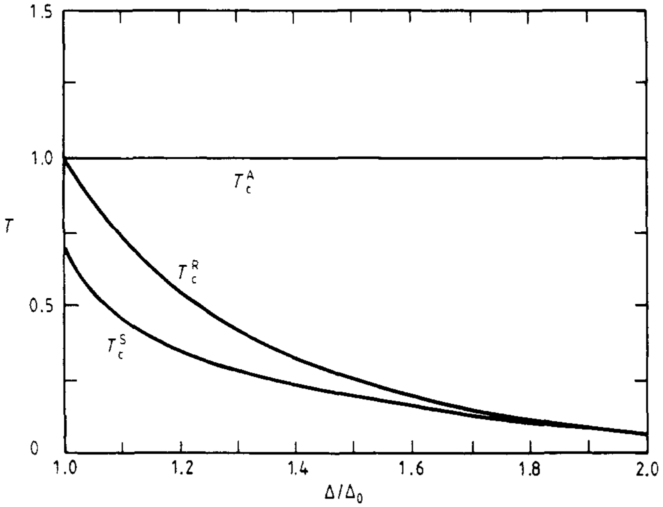

Figure 2. Critical temperature as a function of the parameter Δ/Δ_0 for the strongly diluted model with $q = 0.5$.

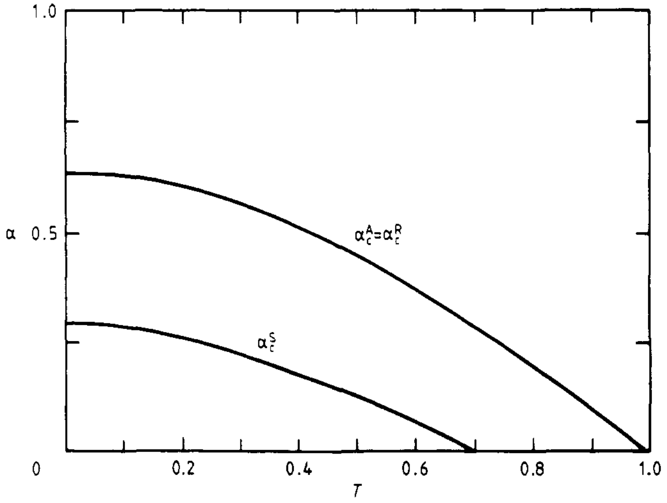

Figure 3. Storage capacity as a function of temperature at $\Delta/\Delta_0 = 1$ for the strongly diluted model with $q = 0.5$.

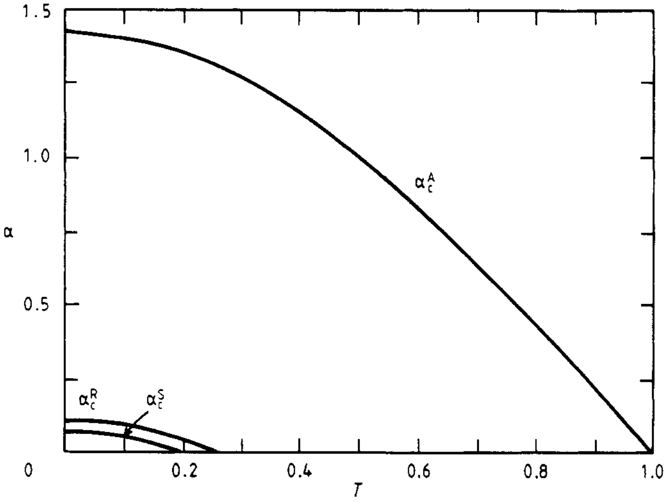

Figure 4. Storage capacity as a function of temperature at $\Delta/\Delta_0 = 1.5$ for the strongly diluted model with $q = 0.5$.

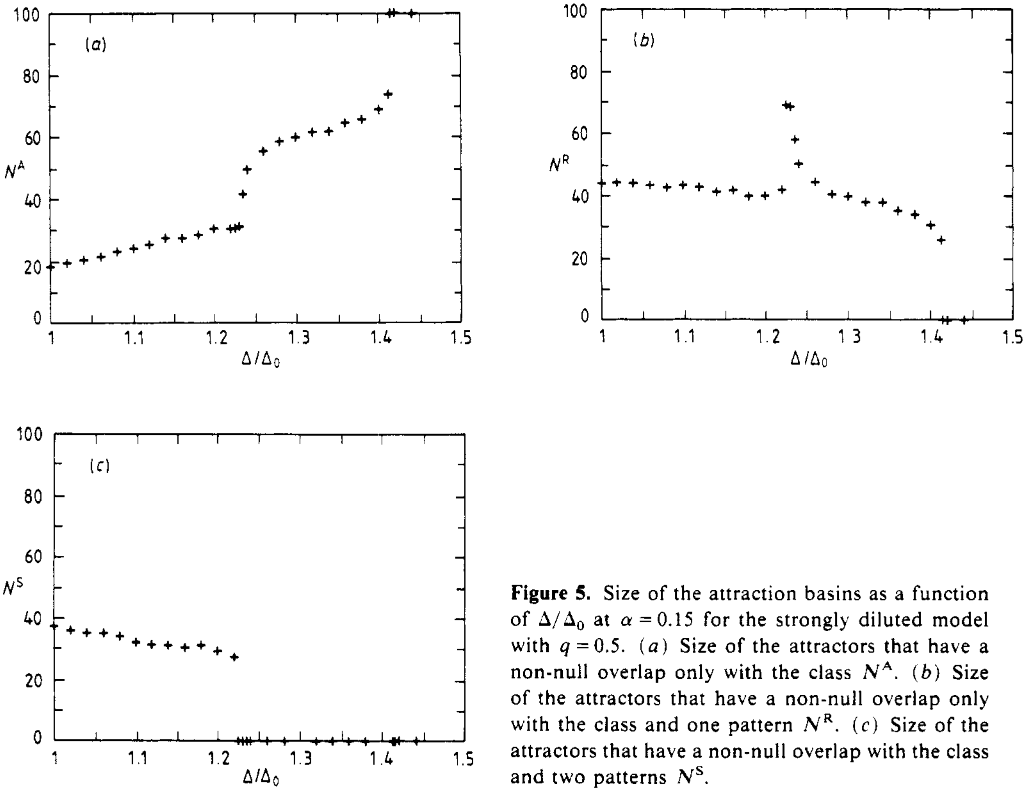

Figure 5. Size of the attraction basins as a function of Δ/Δ_0 at $\alpha = 0.15$ for the strongly diluted model with $q = 0.5$. (a) Size of the attractors that have a non-null overlap only with the class N^A. (b) Size of the attractors that have a non-null overlap only with the class and one pattern N^R. (c) Size of the attractors that have a non-null overlap with the class and two patterns N^S.

Similarly the value T_c^A at $\alpha = 0$ is given by $g'(0) = 1$ where

$$g(m) = \tanh(m/T) \tag{13}$$

from where the value $T_c^A = 1$, independently of Δ/Δ_0, is derived.

We can obtain this information because this transition is second order justifying the expansion in powers. The method is not applicable to the other transitions because they are first order as we shall see below.

Although the configurations follow a heat bath dynamics the dynamical process in (11) is deterministic and the configuration space is divided into separate basins of attraction of the fixed points of (11). We define the size of each attractor as the relative number of initial configurations which flow to that attractor. We choose randomly three values of the overlaps and iterate (11) until a fixed point is reached. This procedure is repeated 10 000 times to obtain the size with an error about 1%. In figures 5(*a*, *b*, *c*), we can see the size of the basin of the states that have overlap only with the class (N^A), with only one stored pattern (N^R) and with the two stored patterns respectively (N^S) for $\alpha = 0.15$, $T = 0$ as a function of Δ/Δ_0. The transitions at α_c^R and α_c^S have a first-order character because the size of the corresponding attraction basins changes suddenly. When one attractor reaches the border between two basins its basin of attraction suddenly shrinks to zero and is incorporated in the other. A similar behaviour is obtained keeping Δ/Δ_0 fixed and varying α.

In figures 6(*a*, *b*) we show schematically the shape of the attraction basins when we condense only one pattern and the corresponding category for $\Delta/\Delta_0 = 1$ and $\Delta/\Delta_0 = 1.5$. The calculation was done iterating (11) with $U_{12} = 0$. The arrows show the flow of the fixed points as a function of α for $0 < \alpha < \alpha_c^R$. When $\Delta/\Delta_0 > 1$ the

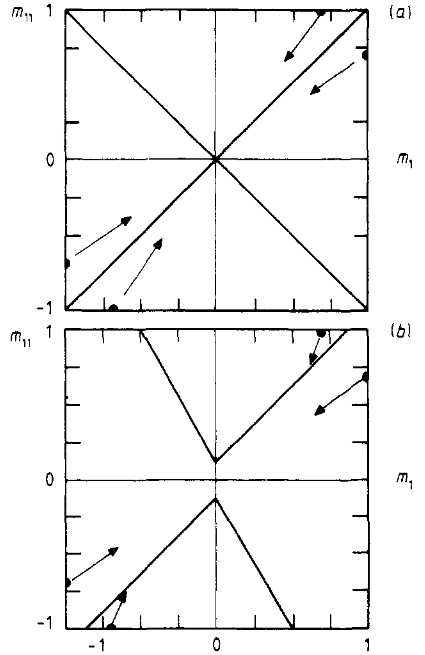

Figure 6. Schematic shapes of the attraction basins for $\alpha = 0.01$, $q = 0.5$ and (*a*) $\Delta/\Delta_0 = 1$; (*b*) $\Delta/\Delta_0 = 1.5$. The arrows show the flow of the attractors as a function of α. For $\alpha > 2/\pi$ the only fixed point is zero.

transition is first order because a macroscopic fraction of the phase space changes from one fixed point to another. On the other hand for $\Delta/\Delta_0 = 1$ both fixed points collapse simultaneously to zero which becomes the only stable fixed point.

3. Dynamics of the non-diluted model

The synaptic interactions are given by (2). Following van Hemmen *et al* (1986) the solution can be achieved introducing sublattices $I(\bar{x}) = \{i/(\xi_i^\alpha, \beta_i^{\alpha\gamma}) = \bar{x}\}$ where $\bar{x}$ is a vector with $N_a + RN_a$ components. Their first N_a components are 1 or -1 with equal probability and the last RN_a components are 1 or 0 with probability c or $1-c$. If the total number of patterns p verifies $2^p \gg N$ (which corresponds to $\alpha = 0$) then each sublattice contains a macroscopic number of sites and it is possible to find recurrence equations for the magnetisation of each sublattice and for the overlaps.

If we suppose that only overlaps with one class and two memorised patterns are relevant the recurrence equations are the same as for the strongly diluted model (11) with $\alpha = 0$.

The problem of solving the dynamics with an extensive number of patterns has not yet been solved. According to a proposal by Riedel *et al* (1988) an effective Hamiltonian can be introduced. This includes a Gaussian noise term with dispersion $\sqrt{\alpha r}$ where r is a known function of q (Amit *et al* 1985b). With this we can obtain the same fixed point equations as with the thermodynamical analysis without breaking the replica symmetry.

Introducing this noise term the recurrence equations at $T = 0$ are

$$\begin{aligned} m_1(t+1) &= c^2 \operatorname{erf}(w) + c(1-c)(\operatorname{erf}(x) + \operatorname{erf}(y)) + (1-c)^2 \operatorname{erf}(z) \\ U_{11}(t+1) &= \frac{\Delta_0}{\Delta}\{-c \operatorname{erf}(w) + (c-1)\operatorname{erf}(x) + c \operatorname{erf}(y) + (1-c)\operatorname{erf}(z)\} \\ U_{12}(t+1) &= \frac{\Delta_0}{\Delta}\{-c \operatorname{erf}(w) + (c-1)\operatorname{erf}(y) + c \operatorname{erf}(x) + (1-c)\operatorname{erf}(z)\} \end{aligned} \tag{14}$$

with

$$r = (1-M)^{-2}$$

and

$$M = (2/\sqrt{\pi\alpha' r})\{c^2 \exp(-w^2) + c(1-c)(\exp(x^2) + \exp(-y^2)) + (1-c)^2 \exp(-z^2)\}$$

where

$$\begin{aligned} w &= (m_1 + (c-1)U_{11} + (c-1)U_{12})/\sqrt{2\alpha' r} \\ x &= (m_1 + (c-1)U_{11} + cU_{12})/\sqrt{2\alpha' r} \\ y &= \frac{m_1 + cU_{11} + (c-1)U_{12}}{\sqrt{2\alpha' r}} \\ z &= (m_1 + cU_{11} + cU_{12})/\sqrt{2\alpha' r}. \end{aligned}$$

These equations were iterated until a fixed point was reached. The retrieval diagram obtained is shown in figure 7. In this case the transition in which all the overlaps go to zero is first order, according to the thermodynamical result obtained by Feigelman *et al*, and the curve is not quadratic because Δ/Δ_0 does not appear only as a factor of α' but also in the equation that defines r.

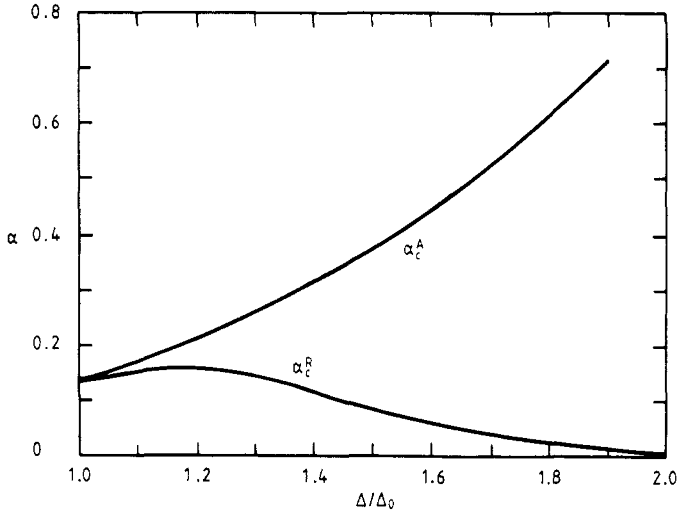

Figure 7. Storage capacity as a function of Δ/Δ_0 at $T=0$ for the non-diluted model.

The transition in which the overlaps with the stored patterns go to zero shows the same behaviour to that obtained by the thermodynamical analysis, having a maximum value for $\Delta/\Delta_0 > 1$.

4. Conclusions

Comparing results of sections 2 and 3 we can see that the behaviour of the systems with strong dilution and with no dilution is similar. The only important differences are as follows.

(*a*) The value of α_c: in the strongly diluted model it is $2/\pi = 0.636$, while in the non-diluted model it is 0.138 which, in fact, seems to have no importance since the content of information stored per synapse is similar in both cases (Canning *et al* 1988).

(*b*) The order of the transition in which all overlaps go to zero: it is a first-order transition for the non-diluted model and second order for the strongly diluted one.

(*c*) The shape of the curve α_c^R (see figure 1 and 7). In the non-diluted model the maximum value is not at $\Delta/\Delta_0 = 1$ but at about $\Delta/\Delta_0 = 1.25$.

We suppose that these effects are due to temporal and spatial correlations that are absent in the strongly diluted model but not in the non-diluted one at $\alpha \neq 0$ and affect the equations through the parameter *r*.

From figures 3 and 4 we can see that the system behaves in a similar fashion to that the Hopfield model, i.e. for low values of α and T there are mixed states that disappear before the network makes the transition to the paramagnetic state. In Derrida *et al* (1987) the behaviour is very different. When correlated patterns are stored the mixed states appear for higher values of T and α, after that the retrieval states have been destroyed. The reason for the difference is that in the article by Derrida *et al* the correlated patterns are being stored using Hebb's rule. It is easy to see that using this rule it is possible to store only one category, otherwise the internal field correspondent to the non-condensed patterns would not have zero mean value.

It is interesting to remark that with the technique used in this work remanence effects, that were analysed numerically in the Hopfield model (Amit *et al* 1985b) and in the hierarchical model (Bacci *et al* 1989a, b), cannot be found because one of the hypotheses was that all the patterns not included in the recurrence equations had an overlap order $1/\sqrt{N}$. The same assumption is made to find the thermodynamical solution.

References

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev. Lett.* **55** 1530

——1985b *Phys. Rev.* A **32** 1007

——1987 *Ann. Phys., NY* **173** 30

Bacci S, Alfaro J, Wiecko C and Parga N 1989a *J. Physique* **50** 757

Bacci S, Mato G and Parga N 1989b *Int. J. Neural Systems* **1** 69

Canning A and Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 3275

Coolen A C C and Ruijgrok Th W 1988 *Phys. Rev.* A **38** 4253

Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

Dotsenko V S 1986 *Physica* **140A** 410

Feigelman M V and Ioffe L B 1987 *Int. J. Mod. Phys.* B **1** 51

Gutfreund H 1988 *Phys. Rev.* A **55** 1530

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

Kree R and Zippelius A 1987 *Phys. Rev.* A **36** 4421

Little W A 1974 *Math. Biosci.* **19** 101

Parga N and Virasoro M A 1986 *J. Physique* **47** 1857

Peretto P 1984 *Biol. Cybern.* **50** 51

Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359

Riedel U, Kuhn R and van Hemmen J L 1988 *Phys. Rev.* A **38** 1105

Van Hemmen J L and Kuhn R 1986 *Phys. Rev. Lett.* **57** 913