Retrieval properties of diluted attractor neural networks

1996 J. Phys. A: Math. Gen. 29 3041

(http://iopscience.iop.org/0305-4470/29/12/012)

J. Phys. A: Math. Gen. **29** (1996) 3041–3051. Printed in the UK

Retrieval properties of diluted attractor neural networks

C Rodrigues Neto and J F Fontanari

Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos SP, Brazil

Received 11 October 1995, in final form 21 February 1996

Abstract. We investigate the dependence of the retrieval properties of the pseudo-inverse and optimal attractor neural networks on the fraction of stored patterns α, the temperature T and the margin parameter κ. Phase diagrams in the full space of parameters are presented in the regime of extreme dilution, i.e. when the connectivity C satisfies the condition $C \ll \ln N$, where N is the number of neurons. Furthermore, we study analytically the neighbourhood of a stored pattern for both models by calculating the average fraction of unstable sites ϵ in a pattern that differs by d sites from a given stored pattern. This analysis may shed light on the properties of the basins of attraction of the stored patterns.

1. Introduction

The usefulness of attractor neural networks as models of associative memory was first pointed out to the statistical physics community by Hopfield (1982). The basic idea is to specify the synaptic weights J_{ij} between the N neurons that compose the network so that a given set of $P = \alpha N$ binary patterns $\boldsymbol{\xi}^l = (\xi_1^l, \ldots, \xi_N^l)$, $l = 1, \ldots, P$ become the equilibrium states of the following stochastic dynamics: given the state of the network at time t then the state of neuron i at time $t+1$ will take on the value $\sigma = \pm 1$ with a probability

$$\frac{1}{1 + \exp[-2\sigma h_i(t)/T]} \tag{1.1}$$

where $h_i(t) = \sum_{j \neq i} J_{ij} S_j(t)$ and T is by definition the temperature. The parameter that measures the proximity of state $\boldsymbol{S}(t)$ to pattern $\boldsymbol{\xi}^l$ is the retrieval overlap

$$m_t^l = \frac{1}{N} \sum_i S_i(t) \xi_i^l \,. \tag{1.2}$$

It is usually assumed that the components ξ_i^l are statistically independent random variables drawn from the distribution

$$p(\xi_i^l) = \tfrac{1}{2}\delta(\xi_i^l - 1) + \tfrac{1}{2}\delta(\xi_i^l + 1) \tag{1.3}$$

where $\delta(x)$ is the Dirac delta function.

The equilibrium properties of the Hopfield model, where the weights are given by the Hebb rule, have been thoroughly studied by Amit *et al* (1985, 1987). Moreover, in the limit of extreme random dilution, i.e. when the connectivity of the network C satisfies the condition $C \ll \ln N$, Derrida *et al* (1987) have solved exactly the neural dynamics (1.1). In

0305-4470/96/123041+11$19.50

this paper we study the retrieval properties of two classical associative memory models—the pseudo-inverse and optimal attractor neural networks—which we describe in what follows.

The pseudo-inverse attractor neural network model (Kohonen 1984, Personnaz *et al* 1986) can store perfectly a set of linearly independent patterns. The weights J_{ij} are obtained by picking the minimal norm solution of the set of P linear equations

$$\Delta_i^l \equiv \frac{1}{\sqrt{C}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l = 1 \qquad l = 1, \ldots, P \tag{1.4}$$

for each $i = 1, \ldots, N$. The quantity Δ_i^l is termed the stability of the component ξ_i^l. The equilibrium properties of the pseudo-inverse model have been studied by Kanter and Sompolinsky (1987) within the replica-symmetric framework in the regime of non-zero temperature. In particular, these authors have found that the pseudo-inverse storage capacity for random patterns is $\alpha_c = 1$.

The pseudo-inverse model, however, is not optimal in the sense that its storage capacity is not maximal. In a remarkable paper, Gardner has shown how the ensemble of weights of the optimal attractor neural network could be characterized within the statistical mechanics framework (Gardner 1988). In fact, the weights of the optimal attractor neural network must satisfy the inequalities

$$\Delta_i^l \geqslant \kappa \tag{1.5}$$

for all i and l. The margin parameter $\kappa \geqslant 0$ is introduced in order to ensure that the stable patterns $\boldsymbol{\xi}^l$ possess a finite basin of attraction, although it is not at all obvious that the size of the basins of attraction must vanish for $\kappa = 0$. It has been shown that the storage capacity of the optimal attractor neural network for random patterns decreases with increasing κ and, in particular, that $\alpha_c = 2$ for $\kappa = 0$ (Gardner 1988).

In another remarkable contribution, Gardner has shown how the dynamics of the diluted optimal attractor neural network could be solved exactly in the zero-temperature regime (Gardner 1989). At non-zero temperature the retrieval overlap with a given stored pattern obeys the equation (Amit *et al* 1990)

$$m_{t+1} = \int_{-\infty}^{\infty} \mathrm{d}\Delta\, \mathcal{P}(\Delta) \int_{-\infty}^{\infty} \mathrm{D}y\, \tanh\left[\frac{1}{T}\left(m_t \Delta + y Q^{1/2} \sqrt{1 - m_t^2}\right)\right] \tag{1.6}$$

where $\mathrm{D}y = \mathrm{d}y/\sqrt{2\pi}\mathrm{e}^{-y^2/2}$ is the Gaussian measure and $Q = Q_i = 1/N \sum_j J_{ij}^2$ is the squared norm of the synaptic weights. Here $\mathcal{P}(\Delta)$ is the distribution of probability of the stabilities defined as (Keppler and Abbot 1988)

$$\mathcal{P}(\Delta_i^l) = \left\langle \delta\left(\Delta_i^l - \frac{1}{\sqrt{C}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l\right)\right\rangle. \tag{1.7}$$

The notation $\langle\ \rangle$ stands for the averages over the patterns ξ_i^l and over the ensemble of weights that satisfy equations (1.4) or inequalities (1.5) depending on whether we are considering the pseudo-inverse or the optimal attractor model. We note that $\mathcal{P}(\Delta_i^l) = \mathcal{P}(\Delta)$ because of the statistical independence of the random variables ξ_i^l.

What is remarkable about equation (1.6) is that it applies to any neural network model; the dependence on the specific model enters through $\mathcal{P}(\Delta)$ only. It must be emphasized, however, that it applies only to the case where there is only one condensed pattern at $t = 0$. The diluted version of the pseudo-inverse model has been studied by Opper *et al* (1989) in the regime of zero temperature, while Amit *et al* (1990) have carried out a thorough analysis of the diluted version of the optimal attractor neural network in the saturation regime, i.e.

at $\alpha = \alpha_c(\kappa)$. Actually, the zero-temperature analysis of Gardner (1989) was also restricted to this regime.

The goal of this paper is twofold. First, we consider the diluted versions of the pseudo-inverse and optimal attractor neural networks. We extend the analysis of Opper *et al* (1989) for the pseudo-inverse model to the non-zero temperature regime, thus obtaining the phase diagram of the model in the space (α, T). Furthermore, we obtain the phase diagram of the optimal attractor neural network in the full space of parameters (α, κ, T). As mentioned above, this phase diagram was known at the saturation regime $\alpha = \alpha_c(\kappa)$ only. Due to the acknowledged importance of these two associative memory models, we think the presentation of their complete phase diagrams is a worthwhile endeavour. Second, we investigate the nature of the neighbourhood of a stored pattern in the pseudo-inverse and optimal attractor neural network models. To achieve this we evaluate the fraction ϵ of unstable sites in a test pattern $\boldsymbol{\eta}^l$ that is at a fixed Hamming distance d to the stored pattern $\boldsymbol{\xi}^l$. Clearly, the dependence of ϵ on d contains much information about the neighbourhood of the stored pattern $\boldsymbol{\xi}^l$. On the one hand, if $\epsilon = d$, so that every site we flip from $\boldsymbol{\xi}^l$ becomes an unstable site, then the neighbourhood of this pattern is expected to be quite smooth. On the other hand, if a very small deviation from the fixed point $\boldsymbol{\xi}^l$ leads to an abrupt increase in the number of unstable sites then its neighbourhood resembles a golf course. In this case we can conclude that the basin of attraction of $\boldsymbol{\xi}^l$ is vanishingly small.

The remainder of this paper is organized as follows. In section 2 we consider the pseudo-inverse model. The phase diagram in the space (α, T) is presented for the diluted version of the model and the dependence of ϵ on d is calculated. Section 3 is devoted to the analysis of the optimal attractor neural network. The phase diagram in the full space of parameters (α, κ, T) is presented and the neighbourhood of a stored pattern is investigated. Finally, in section 4 we summarize our results and present some concluding remarks.

2. The pseudo-inverse model

As mentioned above, the weights of the pseudo-inverse network are given by the minimal norm solution of equation (1.4). Although this solution can be expressed in terms of the correlation matrix $C_{kl} = 1/N \sum_i \xi_i^k \xi_i^l$ (Kohonen 1984),

$$J_{ij} = \frac{1}{N} \sum_{kl} \xi_i^k \xi_j^l (C^{-1})_{kl} \tag{2.1}$$

this explicit formulation is not necessary to the analysis of the diluted network for which the knowledge of the probability distribution of the stabilities Δ_i^l suffices. In order to evaluate $\mathcal{P}(\Delta_i^l)$ we introduce the energy function

$$E_i(h) = \sum_l (\Delta_i^l - 1)^2 + h\delta\left(\Delta_i^l - \frac{1}{\sqrt{C}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l\right) \tag{2.2}$$

so that

$$\mathcal{P}(\Delta_i^l) = -\lim_{\beta \to \infty} \frac{1}{\beta} \frac{\partial \overline{\ln Z_i}}{\partial h}\bigg|_{h=0} \tag{2.3}$$

where Z_i is the partition function

$$Z_i(h) = \prod_j \int_{-\infty}^{\infty} \mathrm{d}J_{ij} \, \exp(-\beta E_i(h)) \tag{2.4}$$

and the bar indicates an average over the random variables ξ_i^l. Here β and h are simply auxiliary variables whose physical significance is irrelevant to our analysis. In the following we will omit the indices i and l since the statistical independence of ξ_i^l makes all sites and all patterns equivalent. Although the procedure to calculate $\mathcal{P}(\Delta)$ is standard (Gardner and Derrida 1988), we must be careful with the choice of the normalization Q: for $\alpha \leqslant 1$, it must be chosen as the smallest value for which $E_i(h=0)=0$, while for $\alpha > 1$ it must be chosen so as to minimize $E_i(h=0)$ (Fontanari 1993). For $\alpha \leqslant 1$ we find $Q = \alpha/(1-\alpha)$ and

$$\mathcal{P}(\Delta) = \delta(\Delta - 1) \tag{2.5}$$

while for $\alpha > 1$ we find $Q = 1/(\alpha - 1)$ and

$$\mathcal{P}(\Delta) = \sqrt{\frac{\alpha^2}{2\pi(\alpha-1)}}\exp\left[-\frac{\alpha^2}{2(\alpha-1)}\left(\Delta - \frac{1}{\alpha}\right)^2\right]. \tag{2.6}$$

Finally, we note that replacing the 1 on the right-hand side of (1.4) by a positive parameter, say κ, has no effect in our analysis since this parameter can be eliminated altogether simply by redefining the normalization of the weights.

2.1. Diluted version

Once $\mathcal{P}(\Delta)$ is known, the analysis of the dynamics of the diluted neural network becomes straightforward. The different phases in the space (α, T) are determined by the fixed points $m_{t+1} = m_t = m^*$ of equation (1.6). In order to study these fixed points for a general neural network model it is convenient to define the function

$$g(m) = m - \int_{-\infty}^{\infty} \mathrm{d}\Delta\, \mathcal{P}(\Delta) \int_{-\infty}^{\infty} \mathrm{D}y\, \tanh\left[\frac{1}{T}\left(m\Delta + yQ^{1/2}\sqrt{1-m^2}\right)\right] \tag{2.7}$$

so that the fixed points are the roots of

$$g(m) = 0\,. \tag{2.8}$$

We note that since $g(-m) = -g(m)$ the paramagnetic fixed point $m^* = 0$ is always a root. Moreover, since $-m^*$ is also a root we consider in the following analysis only the non-negative roots of equation (2.8). Expansion of this equation in powers of m yields

$$m^* = 6^{1/2}\sqrt{\frac{g'(m=0)}{g'''(m=0)}} \tag{2.9}$$

for the non-zero solution. Thus, the equation

$$g'(m=0) = 0 \tag{2.10}$$

determines the continuous transition line between the retrieval ($m^* > 0$) and the paramagnetic ($m^* = 0$) phases, provided that $g'''(m=0) \neq 0$. As usual in this sort of mean-field analysis, the simultaneous vanishing of $g'''(m=0)$ and $g'(m=0)$ determines the tricritical point (TCP). Furthermore, equation (2.10) also gives the limit of stability of the paramagnetic fixed point $m^* = 0$: if $g'(m=0) > 0$ then it is an attractive fixed point while if $g'(m=0) \leqslant 0$ it is a repulsive fixed point. For the pseudo-inverse and the optimal attractor models a numerical analysis of equation (2.8) shows that if $m^* = 0$ is stable then either it is the only root or there are two additional positive roots. If, however, $m^* = 0$ is unstable then there is only one additional positive root. Finally, to determine the discontinuous transition line we must solve $g(m) = 0$ and $g'(m) = 0$ simultaneously, since

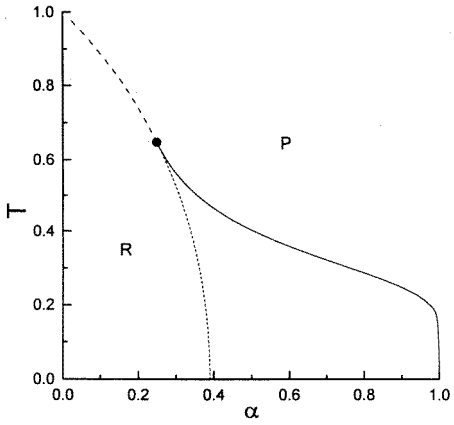

Figure 1. Phase diagram of the pseudo-inverse attractor neural network showing the retrieval (R) and the paramagnetic (P) phases. The full curve is the discontinuous transition, the long-broken curve the continuous transition, and the short-broken curve the limit of stability of the paramagnetic phase. The tricritical point occurs at $\alpha = 0.249$ and $T = 0.648$.

at that transition, which occurs in a region where $m^* = 0$ is stable, the two positive roots coalesce into a double root before disappearing altogether.

Reverting to the pseudo-inverse model, we have found that for $\alpha > 1$ the only root of equation (2.8) is the paramagnetic one, $m^* = 0$. A similar result has been found by Amit *et al* (1990) in their analysis of the optimal attractor network for $\alpha > \alpha_c(\kappa)$, i.e. in the regime where the patterns cannot be perfectly stored. The phase diagram for $\alpha \leqslant 1$ is presented in figure 1, where the full curve represents the discontinuous transition, the long-broken curve the continuous transition, and the short-broken curve the limit of stability of the paramagnetic fixed point. The TCP is located at $\alpha = 0.249$ and $T = 0.648$. The short-broken curve intersects the $T = 0$ line at $\alpha = 1/(1 + \pi/2) \approx 0.389$, in agreement with the result of Opper *et al* (1989). Between the full and the short-broken curves both phases coexist. We note that the phase diagram of the fully connected model presents a discontinuous transition only (Kanter and Sompolinsky 1987).

2.2. Neighbourhood of a stored pattern

As pointed out in the introduction, the neighbourhood of a stored pattern can be analysed by calculating the distribution of stabilities of a test pattern $\boldsymbol{\eta}^l$ that possesses a fixed overlap with the stored pattern $\boldsymbol{\xi}^l$. More specifically, the components of the test pattern are generated according to the conditional probability distribution

$$p(\eta_i^l|\xi_i^l) = \frac{1+b}{2}\delta(\eta_i^l - \xi_i^l) + \frac{1-b}{2}\delta(\eta_i^l + \xi_i^l) \tag{2.11}$$

where $0 \leqslant b \leqslant 1$ measures the overlap between $\boldsymbol{\eta}^l$ and $\boldsymbol{\xi}^l$. This parameter is related to the Hamming distance d between these two patterns: $d = (1-b)/2$. We focus on the fraction of unstable sites of $\boldsymbol{\eta}^l$, i.e. the ratio ϵ between the number of sites for which the stabilities

$$\Lambda_i^l = \frac{1}{\sqrt{N}}\eta_i^l \sum_{j\neq i} J_{ij}\eta_j^l \tag{2.12}$$

are negative and the total number of sites N. The statistical independence of the variables η_i^l and ξ_i^l for different sites allows us to write this ratio as

$$\epsilon = \int_{-\infty}^{0} \mathrm{d}\Lambda^l \, \mathcal{W}(\Lambda^l) \tag{2.13}$$

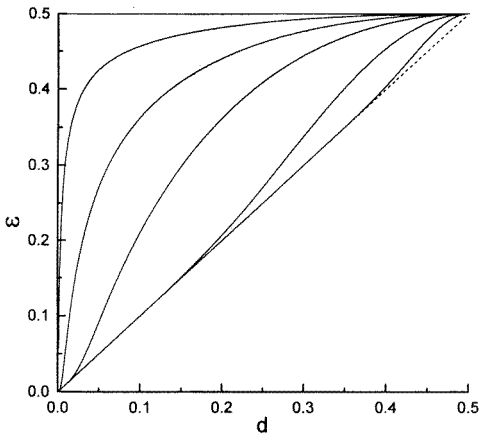

Figure 2. Fraction of unstable sites of pattern $\boldsymbol{\eta}$ as a function of its Hamming distance to stored pattern $\boldsymbol{\xi}$ for the pseudo-inverse attractor neural network. The parameters are (from top to bottom) $\alpha = 1$, 0.99, 0.9, 0.6, 0.1 and 0.01. The broken curve is $\epsilon = d$ which coincides with the curve of ϵ for $\alpha = 0$.

where $\mathcal{W}(\Lambda^l)$ is the probability distribution of the stabilities of test pattern $\boldsymbol{\eta}^l$,

$$\mathcal{W}(\Lambda^l) = \left\langle\!\!\left\langle \delta\left(\Lambda^l - \frac{1}{\sqrt{N}} \eta_i^l \sum_{j \neq i} J_{ij} \eta_j^l \right) \right\rangle\!\!\right\rangle. \tag{2.14}$$

Here the notation $\langle\!\langle\ \rangle\!\rangle$ stands for the averages over the patterns $\boldsymbol{\eta}^l$ and $\boldsymbol{\xi}^l$, as well as over the ensemble of weights that satisfy equations (1.4) or inequalities (1.5).

For the pseudo-inverse model, the distribution defined in equation (2.14) can be calculated by defining an appropriate energy function, analogously to equation (2.2). The final result is

$$\mathcal{W}(\Lambda) = \sqrt{\frac{1}{2\pi Q(1-b^2)}} \exp\left[-\frac{(\Lambda - b\eta_i\xi_i)^2}{2Q(1-b^2)} \right] \tag{2.15}$$

where $Q = \alpha/(1-\alpha)$ and we have dropped the pattern index for simplicity. Using the distributions (2.11) and (1.3) to carry out the averages over η_i and ξ_i, respectively, yields the following expression for the fraction of unstable sites in $\boldsymbol{\eta}$:

$$\epsilon = \frac{1-b}{2} + bH\left[\frac{b}{\sqrt{Q(1-b^2)}} \right] \tag{2.16}$$

where $H(x) = \int_x^\infty \mathrm{D}t$. This quantity is shown in figure 2 as a function of the Hamming distance $d = (1-b)/2$ for several values of α. For $\alpha = 0$ we find $\epsilon_0 = d$. As α increases towards $\alpha_c = 1$ the neighbourhood of the stored pattern starts to take a golf-course shape: for small d the number of errors is a very sensitive function of the Hamming distance, while for large d it becomes practically independent of the distance to the stored pattern. In particular, at $\alpha = \alpha_c$ we find $\epsilon_c = 0$ if $d = 0$ and $\epsilon_c = \frac{1}{2}$ otherwise.

3. The optimal attractor model

In this case the distribution of probability of the stabilities is given by (Kepler and Abbot 1988, Gardner 1989)

$$\mathcal{P}(\Delta) = \sqrt{\frac{1}{2\pi(1-q)}}\,\Theta(\Delta - \kappa) \int_{-\infty}^{\infty} \mathrm{D}y\, \frac{\exp[-\Xi^2(\Delta, y)/2]}{H[\Xi(\kappa, y)]} \tag{3.1}$$

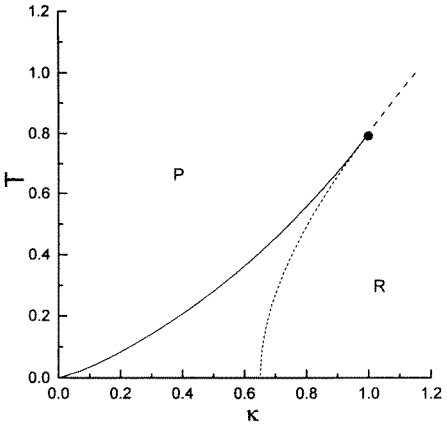

Figure 3. Phase diagram of the optimal attractor neural network in the plane $\alpha = 0$. The tricritical point occurs at $\kappa = 1$ and $T = 0.799$. The convention is the same as for figure 1.

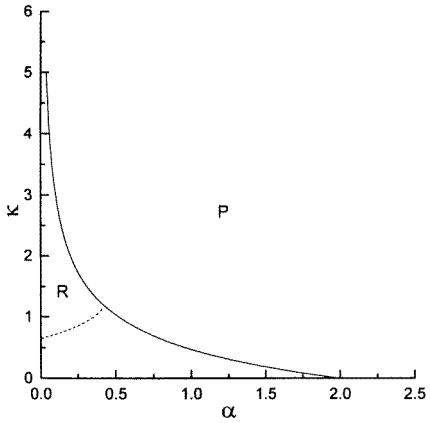

Figure 4. Phase diagram of the optimal attractor neural network in the plane $T = 0$. The convention is the same as for figure 1.

where

$$\Xi(x, y) = \frac{x - y\sqrt{q}}{\sqrt{1-q}} \,. \tag{3.2}$$

Here $\Theta(x) = 1$ if $x > 0$ and 0 otherwise. The parameter q measures the overlap between two distinct sets of optimal weights. It is given by the solution of the equation

$$q = \alpha\sqrt{\frac{1-q}{2\pi}} \int_{-\infty}^{\infty} \mathrm{D}y \, (\kappa - y/\sqrt{q}) \frac{\exp[-\Xi^2(\kappa, y)/2]}{H[\Xi(\kappa, y)]} \,. \tag{3.3}$$

In the saturation regime $\alpha = \alpha_c(\kappa)$ one has $q \to 1$, so that equation (3.1) simplifies considerably. Since for Boolean neural networks the choice of the weights normalization is irrelevant, we have set $Q = 1$ as usual.

3.1. Diluted version

The procedure to obtain all the transition lines of the phase diagram was already presented in section 2.1 so, in this section, we present the final results only. We consider first the simplest case, namely, the phase diagram in the plane $\alpha = 0$ which is shown in figure 3. At $T = 0$, the paramagnetic fixed point becomes unstable for $\kappa > 0.651$. The TCP is located at $\kappa = 1$ and $T = 0.799$. For large κ the continuous transition is very well approximated by the linear relation $T = \kappa$.

The zero-temperature limit is also simple since the retrieval fixed point is $m^* = 1$ in this limit. The phase diagram in the plane $T = 0$, shown in figure 4, does not present the continuous transition between the retrieval and the paramagnetic phases. The short-broken curve, which delimits the region of stability of the fixed point $m^* = 0$, intersects the $\alpha = 0$ axis at $\kappa = 0.651$. Moreover, it intersects the discontinuous transition line, given by $\alpha = \alpha_c(\kappa)$, at $\alpha = 0.42$ and $\kappa = 1.2$, in agreement with the results of Gardner (1989) and Amit *et al* (1990). Although we use the same convention to represent the discontinuous transitions in figures 3 and 4 they are qualitatively different: the transition in figure 3 is due to the disappearance of the retrieval fixed point of equation (2.8), while the one depicted in figure 4 simply delimits the region where a set of weights that satisfies the inequalities

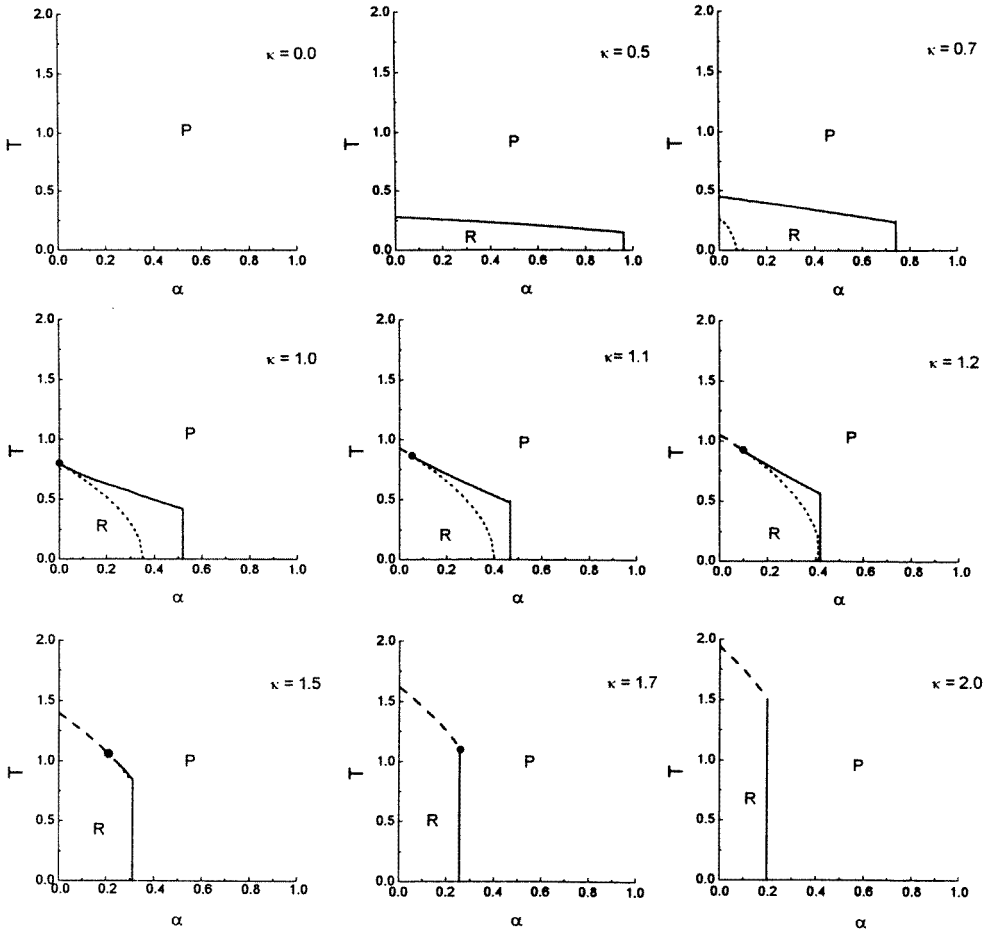

Figure 5. Phase diagram of the optimal attractor neural network in the planes $\kappa = 0$, 0.5, 0.7, 1.0, 1.1, 1.2, 1.5, 1.7 and 2.0. The convention is the same as for figure 1.

(1.5) exists, and therefore has nothing to do with equation (2.8). We will refer to the line $\alpha = \alpha_c(\kappa)$ as the termination line. For $\alpha > \alpha_c(\kappa)$ we could consider the ensemble of weights that minimize the number of violations of the stability criterion, in a similar way as we have done for the pseudo-inverse model in the region $\alpha > 1$. However, as mentioned before, the paramagnetic fixed point is the only asymptotic solution of the resulting dynamic equation in this case (Amit *et al* 1990).

Finally, we turn to generic values of the control parameters α, κ and T. In this case, the calculation of $g(m)$, equation (2.7), involves the evaluation of a triple integral that, however, can be reduced to a double integral through an appropriate change of the integration variables that allows for the analytical integration over Δ. The phase diagram in the planes $\kappa = 0$, 0.5, 0.7, 1.0, 1.1, 1.2, 1.5, 1.7 and 2.0 is presented in figure 5. For $\kappa = 0$ the retrieval fixed point is unstable, although any non-zero value of κ can stabilize it. This can be easily seen by verifying the stability condition $g'(m = 1) > 0$ at $T = 0$, since if this fixed point is unstable at $T = 0$, it is very likely to be unstable for non-zero temperatures too. In particular, in this limit we find $g'(m^* = 1) \to -\infty$ for $\kappa = 0$, and $g'(m^* = 1) = 1$ for $\kappa > 0$. As κ

increases from zero, the region of stability of the retrieval fixed point also increases. The discontinuous transition ends abruptly at the termination line $\alpha = \alpha_c(\kappa)$. For $\kappa \geqslant 0.651$ there appears a region near the origin $T = 0$ and $\alpha = 0$ where the paramagnetic fixed point is unstable. Following our convention, the boundary of this region is represented by the short-broken curve. This curve first intersects the discontinuous transition line at $\alpha = 0$ and $\kappa = 1$; it first intersects the termination line at $T = 0$ and $\kappa = 1.2$. The tricritical point (TCP) generated at $\kappa = 1$ reaches the termination line at $\kappa = 1.7$. As a result, the discontinuous transition and the TCP disappear for $\kappa > 1.7$. Besides, the retrieval fixed point becomes the only stable fixed point below the continuous transition line. As expected, increasing the margin parameter κ increases the robustness of the network to noise and decreases its storage capacity.

Comparing the phase diagram of the pseudo-inverse model, presented in figure 1, with the phase diagram of the optimal attractor model, presented in figure 5, we can conclude that for the same storage capacity $\alpha_c = 1$ (achieved by $\kappa \approx 0.470$) the pseudo-inverse performs better. In fact, besides being more robust to noise, the pseudo-inverse network presents a regime where the retrieval fixed point is the only stable fixed point, while for the optimal attractor network this regime is present for $\kappa > 0.651$ only.

3.2. Neighbourhood of a stored pattern

In this case the distribution of probability of the stabilities of the test pattern $\boldsymbol{\eta}$ is given by

$$\mathcal{W}(\Lambda) = \frac{\mathrm{e}^{-\Lambda^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{D}y \int_{-\infty}^{\infty} \mathrm{D}x \, \frac{H[\Xi_1]}{H[\Xi_2]} \tag{3.4}$$

where

$$\Xi_1 = \frac{\kappa - b\Lambda\eta_i\xi_i - \sqrt{q(1-b^2)}x}{\sqrt{(1-q)(1-b^2)}} \tag{3.5}$$

and

$$\Xi_2 = \frac{\kappa - b\eta_i\xi_i\left(\Lambda q + \sqrt{q(1-q)}y\right) - \sqrt{q(1-b^2)}x}{\sqrt{1-q}} \,. \tag{3.6}$$

As before, the triple integral that appears in the evaluation of ϵ can be reduced to a double integral through an appropriate change of the integration variables. The dependence of ϵ on d for $\kappa = 0$ is presented in figure 6. In particular, for $\alpha = 0$ we find

$$\epsilon_0 = \frac{1-b}{2} + \frac{b}{\pi} \arccos b \tag{3.7}$$

while for $\alpha = \alpha_c = 2$ we find $\epsilon_c = 0$ for $b = 1$ and $\epsilon_c = 1/4 + \epsilon_0/2$ otherwise. It is interesting to compare these results with those for the pseudo-inverse network. While the dependence of ϵ_0 on d is very simple for the pseudo-inverse model, namely $\epsilon_0 = d$, it is quite complex for the optimal attractor: any small deviation from the stored pattern leads to an abrupt increase in the number of unstable sites, since all order derivatives of ϵ_0 diverge at $d = 0$. This result seems to corroborate the conclusion drawn from our analysis of the diluted network that the basins of attraction of the stored patterns vanish for $\kappa = 0$.

In order to better compare the pseudo-inverse and the optimal attractor models, we choose $\kappa \approx 0.470$ so that the storage capacity of both models become the same ($\alpha_c = 1$). Figure 7 shows the dependence of ϵ on d for this case. For small d we find $\epsilon \approx d$, that contrasts with the non-analytic behaviour of the case $\kappa = 0$. It is interesting that a non-zero value of κ guarantees a smooth neighbourhood, i.e. $\epsilon \approx d$ even at the saturation regime

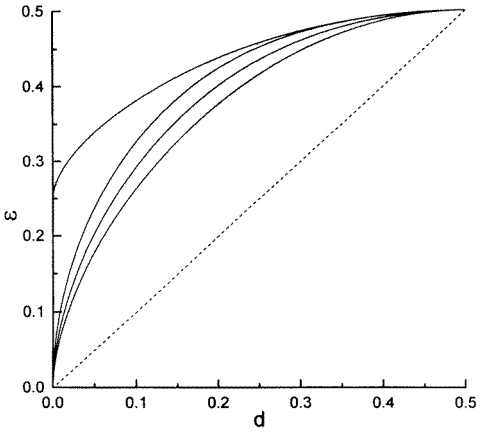

Figure 6. Fraction of unstable sites of pattern η as a function of its Hamming distance to stored pattern ξ for the optimal attractor neural network. The parameters are $\kappa = 0$ and (from top to bottom) $\alpha = 2$, 1, 0.5 and 0. The broken curve is $\epsilon = d$.

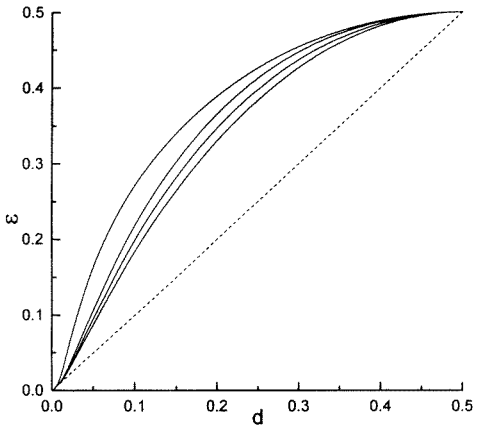

Figure 7. Same as figure 6 but for $\kappa = 0.470$ and (from top to bottom) $\alpha = 1$, 0.5, 0.25 and 0.

$\alpha = \alpha_c$. Larger values of κ increase the range of d for which $\epsilon \approx d$. In particular, for $\kappa \to \infty$ we find $\epsilon \to d$ for $\alpha \leqslant \alpha_c \to 0$. This result leads us to conjecture that the smoother the neighbourhood of a stored pattern, the larger its basin of attraction.

4. Conclusion

In this paper we have presented a comparison between the retrieval properties of two well known associative memory models: the pseudo-inverse and optimal attractor neural networks. The first part of our analysis deals with extremely diluted neural networks, for which the one-step dynamics (Kepler and Abbot 1989) becomes exact for all times (Derrida *et al* 1987, Gardner 1989). The phase diagram of the pseudo-inverse model showing the regions of existence of the paramagnetic and retrieval phases is presented in the space (α, T), thus generalizing the zero-temperature analysis of Opper *et al* (1989). Moreover, the phase diagram of the optimal attractor neural network is presented in the full space of parameters (α, κ, T). We note that this model had been studied at the saturation regime $\alpha = \alpha_c(\kappa)$ only (Gardner 1989, Amit *et al* 1990). Although our analysis of the diluted networks is quite straightforward, in the sense that the dynamic equation (2.7) is well known (Amit *et al* 1990), it is justified since the fixed points of this equation have not been investigated for arbitrary values of the parameters α, κ and T.

In the second part of our analysis we present an original analytical technique to study the nature of the neighbourhood of a stored pattern ξ^l. More specifically, we calculate analytically the fraction of sites ϵ that become unstable when d sites of a stored pattern are flipped. We think this is an important piece of information to characterize an associative memory model. In particular, we can easily think of an artificial dynamics that guarantees the retrieval of the stored patterns in the case that $\epsilon \approx d$: the dynamics must be such that only site flips that decrease the number of unstable sites are allowed. Thus, the d unstable sites will eventually be flipped and the stored pattern recovered. We note that the local structure of the neighbourhood of a stored pattern, characterized by $\epsilon(d)$, is independent of the degree of dilution of the network, so the results presented in figures 2, 6 and 7 are valid

both for the diluted and for the fully connected neural networks.

The existence of fixed points other than the stored patterns is an important issue in the analysis of the pseudo-inverse and the optimal attractor neural networks. For the pseudo-inverse model, this issue was addressed by Kuhlmann and Anlauf (1994) who have calculated an upper bound to the total number of metastable states as a function of the Hamming distance to a stored pattern. In particular, they have found that there are areas around the patterns in which nearly no metastable states exist. The size of these areas decreases as α increases towards $\alpha_c = 1$. Such analysis is complementary to ours, which presents the fraction of unstable sites of a *typical* (not necessarily metastable) state as a function of its Hamming distance to a stored pattern. In fact, the finding that $\epsilon \approx d$ for α not too near α_c seems to corroborate the results of Kuhlmann and Anlauf (1994). To the best of our knowledge, however, no similar analysis has been carried out for the optimal attractor neural network.

The results presented in this paper allow for a better comparison between the retrieval performances of the pseudo-inverse and the optimal attractor models. On the one hand, the analysis of the diluted neural networks shows that for the storage level $\alpha_c = 1$ the pseudo-inverse model is more robust to noise and its attractive fixed points possess larger basins of attraction. On the other hand, the analysis of the neighbourhood of a stored pattern indicates that for small α the pseudo-inverse attractive fixed points possess a smoother neighbourhood than the optimal attractor's, while for α near the saturation regime this situation is reversed. As mentioned before, we believe that the smoother the neighbourhood, the larger the basin of attraction of the stored pattern. Moreover, both analyses indicate that for $\kappa = 0$ the optimal attractor neural network is useless as an associative memory model, since the basins of attraction of the stored patterns vanish for all values of α and T.

Acknowledgments

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). JFF thanks Dr W K Theumann for discussions on the calculations of section 3.2.

References

Amit D J, Gutfreund A and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007
——1987 *Ann. Phys., NY* **173** 30
Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **2** 337
Fontanari J F 1993 *J. Phys. A: Math. Gen.* **26** 6147
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
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kepler T B and Abbot L F 1988 *J. Physique* **49** 1657
Kohonen T 1984 *Self-Organisation and Associative Memory* (Berlin: Springer)
Kuhlmann P and Anlauf J K 1994 *J. Phys. A: Math. Gen.* **27** 5857
Opper M, Kleinz J, Köhler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407
Personnaz L, Guyon I and Dreyfus G 1986 *Phys. Rev.* A **34** 4217