Basin of attraction of the optimal perceptron with biased patterns

1991 J. Phys. A: Math. Gen. 24 3701

(http://iopscience.iop.org/0305-4470/24/15/035)

J. Phys. A: Math. Gen. 24 (1991) 3701-3706. Printed in the UK

Basin of attraction of the optimal perceptron with biased patterns

A Engel† and Th Schnelle

Institut für Theoretische Physik der Humboldt-Universität, Invalidenstrasse 42, O-1040 Berlin, Germany

Received 24 January 1991, in final form 26 March 1991

Abstract. For an extremely diluted neural network model designed by the optimal Gardner rule that stores patterns with low level of activity the basin of attraction is studied analytically. The recursion relation for the overlap with a stored pattern is two-dimensional and yields a richer bifurcation scenario than in the case of patterns with symmetric statistics. The bias in the patterns gives rise to ferromagnetic attractors which compete with the patterns.

There are several interesting parmeters characterizing the performance of attractor neural networks studied by statistical physicists as models for associative memories (for an introduction see Amit (1989) and Geszti (1990)). These include the storage capacity α_c usually defined as ratio of the number of stored patterns p and the number of neurons N, the typical basin of attraction describing the error-correcting abilities of the system, the robustness, the retrieval time, the learning time and the ability to generalize. Many studies have been devoted to the determination of the storage capacity α_c for a variety of pattern statistics and learning rules, in part because this quantity is accessible to analytical calculations within the replica approach (Amit *et al* 1985, Gardner 1988). On the other hand it is clear that the typical size of the basin of attraction of a pattern is at least as important to characterize an associative memory as the storage capacity. Since the critical storage capacity is reached when the basins of attraction shrink to zero there is a certain complementarity between storage capacity and attraction basin as also found in numerical simulations (Forrest 1988) and analytical studies of special cases (Gardner 1989, Opper *et al* 1989).

In the present paper we determine the basin of attraction realized by the learning rule giving maximal storage capacity for patterns $\{\xi_i^\mu\}$ with low level of activity. The pattern statistics is given by the distribution

$$P(\xi_i^\mu) = \frac{1+a}{2}\delta(\xi_i^\mu - 1) + \frac{1-a}{2}\delta(\xi_i^\mu + 1). \tag{1}$$

To be able to study the problem analytically we consider an extremely diluted model where the average connectivity C of the neurons scales as $\log N$. As is well known (Derrida *et al* 1987) the dynamics of the system can then be described by a set of recursion relations for macroscopic order parameters. The problem has been solved

† Present address: Institut für Theoretische Physik der Universität Göttingen, Bunsenstrasse 9, W-3400 Göttingen, Germany

0305-4470/91/153701+06$03.50

already by Gardner for the special case $\langle\langle \xi_i^\mu \rangle\rangle = a = 0$ (Gardner 1989). In this case the maximal storage capacity is $\alpha_c = 2$; however, only for $\alpha \leqslant \alpha_B = 0.42$ the attraction basins are large. For $a \neq 0$ the correlations between the patterns allow much larger values of α_c (Gardner 1988) and it is hence interesting to see how the attraction basins are modified. In the following we will choose $a < 0$ so the '+' sites are the 'signal' and the '−' sites act as 'background'.

As in the case $a = 0$ the recursion relation for the retrieval overlap is determined by the probability distribution of the stabilities

$$\Delta_i = \xi_i^1 C^{-1/2} \sum_j J_{ij} \xi_j^1. \tag{2}$$

However, now this probability distribution is different at signal and background sites, i.e. it depends on ξ_i^1.

Mathematically this gives rise to a two-dimensional recursion relation of the form

$$m^+(t+1) = f^+(m^+(t), m^-(t)) \qquad m^-(t+1) = f^-(m^+(t), m^-(t)) \tag{3}$$

for the order parameters

$$m^\pm(t) = \frac{2}{(1 \pm a)N} \sum_i^\pm \xi_i^1 S_i(t) \tag{4}$$

where $\Sigma^\pm$ means the sum over all sites with $\xi_i^1 = \pm 1$ and $\frac{1}{2}(1 \pm a)/N$ is the number of signal and background sites respectively. Physically it means that the error-correcting abilities of the network depend on whether the initial condition $\{S_i^0\}$ lies within the subspace spanned by the patterns or perpendicular to it (Amit *et al* 1987, Evans 1989).

Let us consider the ensemble of initial conditions $\{S_i^0\}$ giving rise to the initial overlaps m_0^+ and m_0^- defined by (4) for $t = 0$. The corresponding overlaps $m^+(t=1)$ and $m^-(t=1)$ are determined by the distribution of the aligned fields

$$\lambda_i = \xi_i^1 C^{-1/2} \sum_j J_{ij} S_j^0 = C^{-1/2} \left(\sum_j^+ \xi_i^1 J_{ij} S_j^0 + \sum_j^- \xi_i^1 J_{ij} S_j^0 \right) \tag{5}$$

via

$$m^\pm(t=1) = \frac{2}{(1 \pm a)N} \sum_i^\pm \operatorname{sign} \lambda_i. \tag{6}$$

The two parts of λ_i defined by (5) are both Gaussian random variables by the central limit theorem and it is easy to determine their first two moments by splitting the stabilities (2) into the two parts

$$\Delta_i^\pm = \xi_i^1 C^{-1/2} \sum_j^\pm J_{ij} \xi_j^1. \tag{7}$$

In this way we find

$$P(\lambda_i) = \left(2\pi \left[(1 - m_0^{+2}) \frac{1+a}{2} + (1 - m_0^{-2}) \frac{1-a}{2} \right] \right)^{-1/2} \times \exp\left\{ - \frac{[\lambda_i - m_0^+ \Delta_i^+ - m_0^- \Delta_i^-]^2}{2\left[(1 - m_0^{+2}) \frac{1+a}{2} + (1 - m_0^{-2}) \frac{1-a}{2} \right]} \right\} \tag{8}$$

and from (6)

$$m^{\pm}(t=1)=\frac{2}{(1\pm a)N}\sum_i^{\pm}\operatorname{erf}\left(\frac{m_0^+\Delta_i^+ + m_i^-\Delta_i^-}{\{(1+a)[(1-m_0^{+2})]+(1-a)[(1-m_0^{-2})]\}^{1/2}}\right) \tag{9}$$

where $\operatorname{erf}(x)=2/\sqrt{\pi}\int_0^x \exp(-t^2)\,dt$ denotes the standard error function. Assuming the overlaps to be self-averaging we can replace the site average by the ensemble average over the patterns and get

$$m^{\pm}(t=1)=\int_{-\infty}^{\infty} d\Delta^+\,d\Delta^-\,P^{\pm}(\Delta^+,\Delta^-)\operatorname{erf}\left(\frac{m_0^+\Delta^+ + m_0^-\Delta^-}{\{(1+a)(1-m_0^{+2})+(1-a)(1-m_0^{-2})\}^{1/2}}\right) \tag{10}$$

where

$$P^{\pm}(\Delta^+,\Delta^-)=\left\langle\!\!\left\langle \delta\left(\Delta^+ - \xi_i^1 C^{-1/2}\sum_j^+ J_{ij}\xi_j^1\right)\delta\left(\Delta^- - \xi_i^1 C^{-1/2}\sum_j^- J_{ij}\xi_j^1\right)\right\rangle\!\!\right\rangle_{\xi}. \tag{11}$$

Here $P^+(\Delta^+,\Delta^-)$ refers to the signal sites ($\xi_i^1=1$) and $P^-(\Delta^+,\Delta^-)$ to the background sites ($\xi_i^1=-1$). $P^{\pm}(\Delta^+,\Delta^-)$ can be determined using the projector operator formalism introduced by Elizabeth Gardner (Gardner 1988, Gardner 1989). The result becomes particularly simple near saturation ($q\to 1$) where it reads

$$P^{\pm}(\Delta^+,\Delta^-)=\delta(\Delta^+-\Delta^-\mp M)\left[\delta(\Delta-\kappa)\tfrac{1}{2}\operatorname{erfc}\left(\frac{\pm aM-\kappa}{[2(1-a^2)]^{1/2}}\right)\right.$$
$$\left.+\,\theta(\Delta-\kappa)[2\pi(1-a^2)]^{-1/2}\exp\left\{-\frac{(\Delta\mp aM)^2}{2(1-a^2)}\right\}\right]. \tag{12}$$

Here $\operatorname{erfc}(x)=1-\operatorname{erf}(x)$ is the complementary error function. The parameter M gives the ferromagnetic bias in the couplings J_{ij} and is defined by

$$M=C^{-1/2}\sum_j J_{ij}.$$

κ is the usual stability parameter. After learning all stabilities Δ_i as defined by (2) are larger than or equal to κ. Large values of κ should give rise to large basins of attraction (Forrest 1988). M and κ are related to the storage capacity α by the equations (Gardner 1988)

$$0=\left\langle\!\!\left\langle \int_{-\kappa+aM\xi}^{\infty}\frac{dt}{(2\pi)^{1/2}}e^{-t^2/2}\left(\frac{\kappa-aM\xi}{(1-a^2)^{1/2}}+t\right)\right\rangle\!\!\right\rangle_{\xi} \tag{13}$$

$$\alpha_c^{-1}=\left\langle\!\!\left\langle \int_{-\kappa+aM\xi}^{\infty}\frac{dt}{(2\pi)^{1/2}}e^{-t^2/2}\left(\frac{\kappa-aM\xi}{(1-a^2)^{1/2}}+t\right)^2\right\rangle\!\!\right\rangle_{\xi}. \tag{14}$$

Here the ξ-average is to be performed with the distribution (1).

Note that the difference $(\Delta^+-\Delta^-)$ is a non-random quantity so that $P^{\pm}(\Delta^+,\Delta^-)$ depends on the total stability $\Delta=\Delta^++\Delta^-$ only. The distribution for this total stability consists of a δ-peak at $\Delta=\kappa$ and a Gaussian tail for $\Delta>\kappa$ as in the case $a=0$.

Equations (10) and (12) combine to give $m(t=1)$ in terms of m_0. For the extremely diluted model this recursion relation is valid for all times due to the absence of correlation loops (Derrida *et al* 1987, Gardner 1989). Therefore we get the following

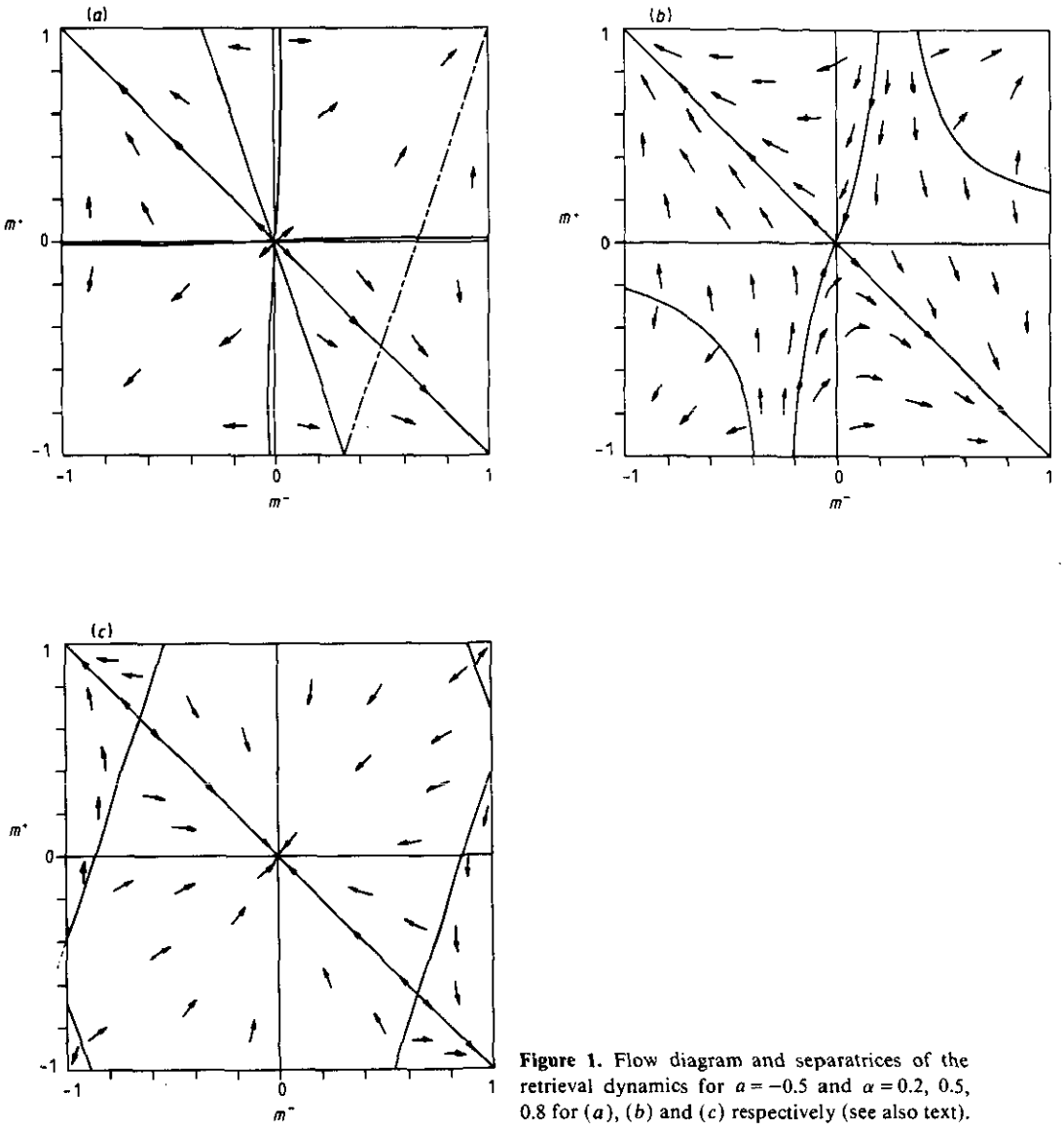

Figure 1. Flow diagram and separatrices of the retrieval dynamics for $a=-0.5$ and $\alpha=0.2$, 0.5, 0.8 for (a), (b) and (c) respectively (see also text).

explicit expression for the order parameter dynamics (cf (3)):

$$m^{\pm}(t+1)=\tfrac{1}{2}\operatorname{erfc}\left(\frac{\pm aM-\kappa}{[2(1-a^2)]^{1/2}}\right)$$
$$\times\operatorname{erf}\left(\frac{\kappa[m^+(t)+m^-(t)]\pm M[m^+(t)-m^-(t)]}{2[(1+a)(1-[m^+(t)]^2)+(1-a)(1-[m^-(t)]^2)]^{1/2}}\right)$$
$$+[2\pi(1-a^2)]^{-1/2}\int_{\kappa}^{\infty}\mathrm{d}\Delta\exp\left\{-\frac{(\Delta\mp aM)^2}{2(1-a^2)}\right\}$$
$$\times\operatorname{erf}\left(\frac{\Delta[m^+(t)+m^-(t)]\pm M[m^+(t)-m^-(t)]}{2[(1+a)(1-[m^+(t)]^2)+(1-a)(1-[m^-(t)]^2)]^{1/2}}\right). \tag{15}$$

The numerical analysis of this recursion relation yields the flow diagrams shown in

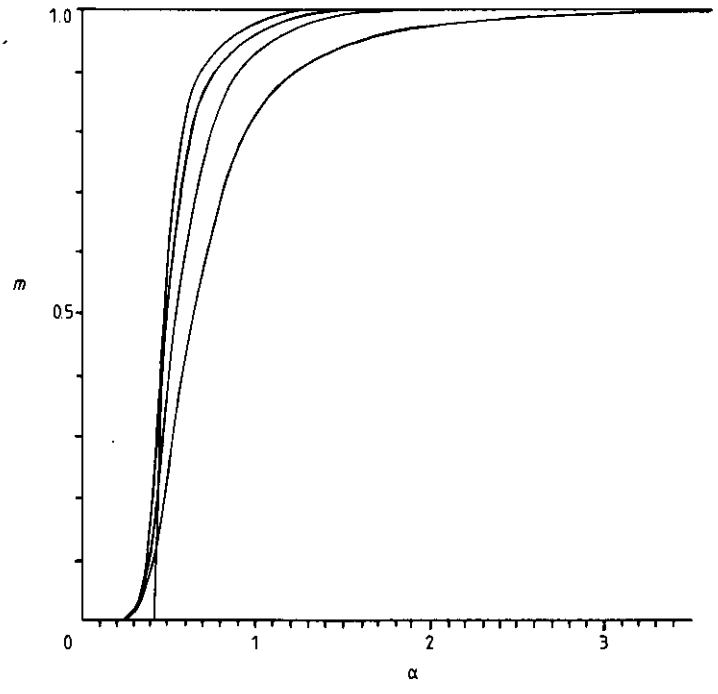

Figure 2. Fixpoints of the recursion relation for uniform noise in the initial condition ($m_0^+ = m_0^-$) as a function of α for $a = 0$, -0.5, -0.7, -0.9 (top from left to right). The corresponding values of α_c are 2, 2.4, 3.1 and 6.1 respectively.

figure 1. For $\kappa \geqslant 0$ there are always five fixpoints. Two are given by $m^+ = m^- = \pm 1$ and correspond to full retrieval of the pattern or its negative. The third one is given by $m^+ = m^- = 0$ and means no retrieval. The remaining two fixpoints are $m^+ = -m^- = \pm 1$ and correspond to ferromagnetic states (black and white screen respectively). Their existence is a direct consequence of the ferromagnetic bias M in the synaptic couplings.

For small α (large κ) the zero overlap fixpoint is unstable and the attraction basins of the patterns are large (figure 1(a)). Nevertheless not all initial conditions with $m = \frac{1}{2}(1+a)m^+ + \frac{1}{2}(1-a)m^- > 0$ are attracted by the retrieval fixpoint as in the case $a = 0$ (Gardner 1989). The dotted line in figure 1(a) corresponds to $m = 0$. As can be clearly seen a large number of the points to the right of this line evolve towards the ferromagnetic attractors. If, however, the noise is the same for signal and background sites, i.e. $m_0^+ = m_0^-$, the system will safely retrieve the pattern. It was proposed (Amit *et al* 1987) to confine the dynamics to configurations within the subspace spanned by the patterns, i.e. to those with $1/N \sum_i S_i = a$, in order to improve the retrieval quality. Such configurations are represented by the dashed-dotted line in figure 1(a) and give a basin of attraction significantly smaller than for $m_0^+ = m_0^-$. This is probably due to a smaller number of spurious attractors in the extremely diluted model as compared with the fully connected case studied by Amit *et al.*

For larger α (smaller κ) the zero overlap fixpoint becomes a saddle with the unstable directions pointing to the ferromagnetic attractors (figure 1(b)). Consequently these now attract very many configurations and the basins of attraction of the patterns become smaller. The latter are quantitatively given by the separatrices and show the expected anisotropy depending on the relation between m_0^+ and m_0^-.

For even larger values of α ($\kappa \to 0$) the zero overlap fixpoint becomes stable (figure 1(*c*)) as in the case $a = 0$ and the basins of attraction of the patterns shrink to zero.

In figure 2 we have plotted the fixpoint structure $m(\alpha)$ along the diagonal $m^+ = m^-$ of figure 1 giving the optimal attraction basin. For comparison we have included the $a = 0$ result of Gardner (1989). With increasing $|a|$ the region where all initial conditions with $m > 0$ are attracted by the pattern become smaller. Of course, for larger values of α the unstable fixpoint can be smaller than in the case $a = 0$ allowing for $\alpha_c > 2$. Normalizing all curves to $\alpha_c = 2$ shows, however, that no *relative* improvement of the basins of attraction occurs.

In conclusion we have shown that the attraction basins of the optimal perceptron storing patterns with low level of activity vary in a non-trivial manner with the mean activity a and the storage capacity α. Although for $a \to 1$ large values of α_c can be obtained not only does the amount of stored information decrease (Gardner 1988) but also the typical basins of attraction become extremely small.

Let us finally note that patterns with 'magnetization' as discussed here are only the simplest kind of correlated patterns. An obvious generalization is pattern hierarchies as introduced by Parga and Virasoro (1986). The storage capacity of the optimal perceptron storing hierarchically correlated patterns has been determined recently (Engel 1990). We expect a similar behaviour of the attraction basins as found in the present paper with the role of the ferromagnetic attractors played by the ancestor patterns.

References

Amit D J 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)
Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
—— 1987 *Ann. Phys., NY* **173** 30
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Engel A 1990 *J. Phys. A: Math. Gen.* **23** 2587
Evans M R 1989 *J. Phys. A: Math. Gen.* **22** 2103
Geszti T 1990 *Physical Models of Neural Networks* (Singapore: World Scientific)
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
—— 1989 *J. Phys. A: Math. Gen.* **22** 1969
Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Opper M, Klein J, Kohler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407
Parga N and Virasoro M A 1986 *J. Physique* **47** 1857