Basins of attraction near the critical storage capacity for neural networks with constant stabilities

1989 J. Phys. A: Math. Gen. 22 L407

(http://iopscience.iop.org/0305-4470/22/9/010)

J. Phys. A: Math. Gen. 22 (1989) L407-L411. Printed in the UK

LETTER TO THE EDITOR

Basins of attraction near the critical storage capacity for neural networks with constant stabilities

M Opper, J Kleinz, H Köhler and W Kinzel

Institut für Theoretische Physik III, Justus-Liebig-Universität Giessen, Heinrich-Buff-Ring 16, D-6300 Giessen, Federal Republic of Germany

Received 12 October 1988, in final form 14 February 1989

Abstract. The dynamics of neural networks with constant stabilities is studied analytically for states in the vicinity of a stored pattern. Only two parameters of the synaptic matrix determine the dynamics of the neurons in the cases considered. The resulting equations are able to predict the basin of attraction near the critical storage capacity. We compare our analytical results with simulations of two network models.

Neural networks have recently been investigated using models and methods of the statistical mechanics of disordered materials (Mezard *et al* 1987). Such models may be considered as content addressable memory. They allow one to store a set of P patterns $S^\nu = (S_1^\nu, \ldots, S_N^\nu)$, $\nu = 1, \ldots, P$, as attractors in a system of N interconnected two-state neurons $S_i \in \{-1, +1\}$ and to retrieve them exactly or almost perfectly from a distorted input version. Many static features of simple spin-glass-type networks are now well understood. For the special case of networks with symmetric coupling matrices stationary states can be calculated by means of equilibrium statistical mechanics (Amit *et al* 1987). A more difficult and still unsolved problem is the determination of the basin of attraction for a network with $P = \alpha N$ extensively many random patterns, i.e. the maximum amount of initial distortion under which the network will be able to retrieve a pattern. The quantity of interest is the overlap between a state $S_i(t)$, $i = 1, \ldots, N$, and a pattern S^1 at time t

$$m(t) = N^{-1} \sum_i S_i^1 S_i(t)$$

evolving from a noisy initial state

$$S_i(0) = \begin{cases} S_i^1 & \text{with probability } \frac{1}{2}(1 + m(0)) \\ -S_i^1 & \text{with probability } \frac{1}{2}(1 - m(0)). \end{cases} \tag{1}$$

If $m(\infty) = 1$, the network can retrieve the pattern perfectly. One is interested in the critical initial overlap $m(0)$, above which retrieval is possible with a high probability. We assume that the dynamics of the network is given by a parallel update of the neurons

$$S_i(t+1) = \operatorname{sgn}(h_i(t)) \tag{2}$$

where h_i is the internal field

$$h_i(t) = \sum_j J_{ij} S_j(t)$$

and J_{ij} denote the synaptical couplings. We shall use $J_{ii} = 0$ throughout the letter.

Exact results for the dynamics of the overlap could be given only in the case of strongly diluted networks (Derrida *et al* 1987) and for one-layer (Krauth *et al* 1988a)

and multilayered feed-forward networks (Meir and Domany 1987, 1988). Apart from various numerical simulations only a few approximate analytical approaches exist for totally connected single-layer feedback systems.

Kinzel (1985) has derived a mean-field approximation for the Hopfield model which neglects correlations completely; this approximation becomes exact in the extremely diluted network (Derrida *et al* 1987). Horner (1988) has matched a result valid for short times onto the asymptotic (fixed point) region to find an approximation to the dynamics of the Hopfield model. Krauth *et al* (1988b) studied the dynamics of an auxiliary model, the one-pattern model which approximates an actual system by incorporating only the stability and asymmetry of the actual network as parameters. The dynamics of the model itself must be solved by simulations apart from the first few parallel time steps.

In this letter we show that for a special class of such networks, namely those with constant stabilities of the stored patterns, the dynamics of the overlap can be solved in the vicinity of the attractor $m(\infty)=1$. Thus one is able to study very easily the retrieval properties of the networks near the critical storage capacity, where the basin of attraction is a small region near $m=1$.

The special condition to be fulfilled by the networks is that for all patterns ν and neurons i the so-called stabilities

$$\lambda_i^\nu = S_i^\nu \sum_j J_{ij} S_j^\nu \tag{3}$$

are *equal* to a positive constant λ.

Such networks are of great interest because learning algorithms, which enable the systems to store a prescribed set of patterns, are naturally defined by any procedure which minimises the quadratic form

$$H = \sum_{\nu i} \left(\lambda S_i^\nu - \sum_j J_{ij} S_j^\nu \right)^2$$

with respect to the couplings J_{ij}. For continuous J_{ij} a corresponding learning algorithm which leads to the pseudo-inverse coupling matrix (Kohonen 1988, Personnaz *et al* 1986) has been recently discussed (Diederich and Opper 1987). Even for a model with binary synapses $J_{ij}=\pm 1/\sqrt{N}$, a case which may be of technical importance, a simple descent of H will lead to a network with only a small variation $\Delta\lambda_i^\nu$ of stabilities (Köhler 1989).

Our analysis of the network's dynamics is based on the exact expression for the overlap at time $t+1$ averaged over the initial conditions (1)

$$m(t+1) = N^{-1} \sum_i S_i^1 \langle S_i(t+1) \rangle = N^{-1} \sum_i \int \mathrm{d}h \, P_t(h, i) \, \mathrm{sgn}(h) \tag{4}$$

where $P_t(h, i)$ denotes the probability distribution of the internal field at site (neuron) i and time t and $\langle \cdot \rangle$ denotes the average. For the pattern $\boldsymbol{S}^1$ which has to be recognised by the network we have chosen $S_i^1=1$, $i=1,\ldots,N$, for simplicity.

If $m(t)=1$, i.e. $\boldsymbol{S}(t)=\boldsymbol{S}^1$, the constant stabilities (3) lead to $P_t(h, i)=\delta(h-\lambda)$. Thus for $m(t)$ in the vicinity of 1, the field distribution can to lowest order be taken as Gaussian with a variance

$$\Delta_i^2(t) = \langle h_i^2(t) \rangle - \langle h_i(t) \rangle^2$$

converging to zero for $m(t) \to 1$.

Due to the independence of the $S_i(0)$ the assumption of Gaussian distributed internal fields is exact for $t=0$ and all values of $m(0)$. In this case one has

$$\langle h_i(0)\rangle = \lambda m(0) \qquad \Delta_i^2(0) = (1-m^2(0))J^2 \tag{5}$$

Where $J^2 = \Sigma_j J_{ij}^2$ is self-averaging with respect to the random patterns for $N \to \infty$. For the following time step $t=1$, the mean and variance of the field distributions can be written as

$$\begin{aligned} &\langle h_i(1)\rangle = \lambda m(1) \\ &\Delta_i^2(1) = (1-m^2(1))J^2 + \sum_{\substack{k,l \\ k\neq l}} J_{ik}J_{il}\langle[\mathrm{sgn}(h_k(0)) - m(1)][\mathrm{sgn}(h_l(0)) - m(1)]\rangle \end{aligned} \tag{6}$$

where we have used the fact that the distributions for $t=0$ are independent of site i.

The second part of the variance, containing correlations at different sites, can be evaluated using the joint density of $h_k(0)$ and $h_l(0)$.

However, since the covariance

$$\langle \Delta h_k(0)\, \Delta h_l(0)\rangle = \sum_j J_{kj}J_{lj}(1-m^2(0))$$

is of order $1/\sqrt{N}$, an expansion up to first order in these fluctuations is sufficient. Thus we find

$$\Delta_i^2(1) = (1-m^2(1))J^2 + \frac{2}{\pi J^2}\exp\left(-\frac{\lambda^2 m^2(0)}{J^2(1-m^2(0))}\right)\sum_{\substack{j,k,l \\ k\neq l}} J_{ik}J_{il}J_{kj}J_{lj}. \tag{7}$$

For $m(0) \simeq 1$ the second term can be neglected in comparison with the first. In this limit the dynamics is equivalent to that for strongly diluted networks, where site-site correlations vanish.

For $t>1$ the above arguments can be successively repeated, so that we can establish the one-step recursion

$$\langle h_i(t)\rangle \simeq \lambda m(t) \qquad \Delta_i^2(t) \simeq (1-m^2(t))J^2 \qquad m(t+1) \simeq \mathrm{erf}\left(\frac{\lambda m(t)}{(2\Delta_i^2(t))^{1/2}}\right) \tag{8}$$

which is asymptotically valid for $m(t) \simeq 1$.

The basin of attraction near $m(t)=1$ is limited by the unstable fixed points of (8).

We have compared our analytical result with numerical simulations for two network models. Figure 1 displays the case of continuous (pseudo-inverse) couplings, where for random patterns one has $\lambda = 1-\alpha$ and $J^2 = \alpha(1-\alpha)$. The simulational data were taken from Kanter and Sompolinsky (1987). Basins of attraction for a binary coupling model are shown in figure 2. For the latter system we had to account for small but non-zero fluctuations of the stabilities λ_i^ν. To lowest order this yields an additional contribution to the variance, which is given by

$$\Delta_i^2(t) = (1-m^2(t))J^2 + (\Delta\lambda)^2 m^2(t) \tag{9}$$

where $(\Delta\lambda)^2$ is the variance of the stabilities. In figure 2 we take a binary coupling matrix $J_{ij} = \pm 1/\sqrt{N}$ which was obtained from a descent algorithm by minimising the quadratic deviation from equation (3) with $\lambda = 1$ (Köhler 1989). We compare the basins of attraction obtained from direct simulation of the corresponding network with that given by the approximation, (8). Since the stabilities are distributed, we take the parameters $\langle\lambda_i^1\rangle$ and $(\Delta\lambda)^2$ from the numerical data. (Note that λ in (8) has to be replaced by $\langle\lambda_i^1\rangle$.) For this model the matrix of couplings is in general non-symmetric.

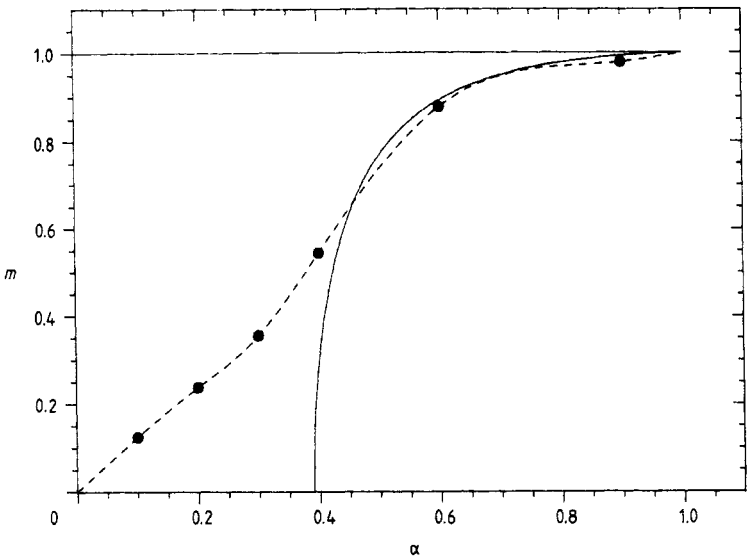

Figure 1. Overlap m as a function of storage capacity α for networks with pseudo-inverse couplings. $m = 1$ is the attractor, and the basin of attraction extends to the lower curve. The points are numerical data from Kanter and Sompolinsky (1987).

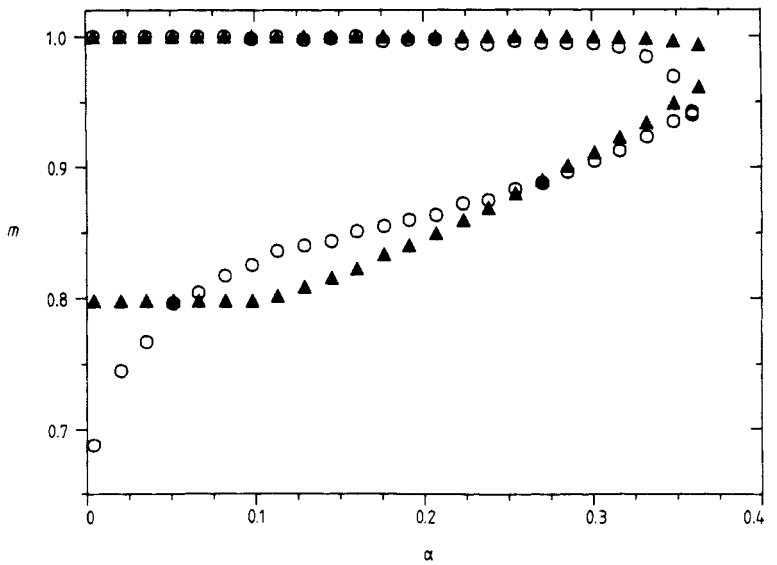

Figure 2. Same as figure 1 for networks with binary couplings. The open circles are data from numerical simulations (Köhler 1989) with $N = 256$. The full triangles are given by the present theory where the parameters of the coupling matrix are taken from simulations.

The present results show that the asymptotic dynamics (8) is in fact able to predict the shape of the basins of attraction near the critical storage capacity for networks with constant stabilities. In order to cover the whole region of capacities at least approximately one must, however, account for the correlations between different neurons as well as for the non-Gaussian shape of the field distribution. Such a programme is currently being investigated.

This work has been supported by the Deutsche Forschungsgemeinschaft, and is part of the Diploma thesis of JK.

References

Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **2** 337
Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
Horner H 1988 *Phys. Blätter* **44** 29
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kinzel W 1985 *Z. Phys.* B **60** 205
Köhler H 1989 *Diploma thesis* Justus-Liebig-Universität, Giessen
Kohonen T 1988 *Self Organisation and Associative Memory* (Berlin: Springer)
Krauth W, Mézard M and Nadal J P 1988a *Complex Systems* **2** 387
Krauth W, Nadal J P and Mézard M 1988b *J. Phys. A: Math. Gen.* **21** 2995
Meir R and Domany E 1987 *Phys. Rev. Lett.* **59** 359
—— 1988 *Phys. Rev.* A **37** 608
Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
Personnaz L, Guyon I and Dreyfus G 1986 *Phys. Rev.* A **34** 4217