Robustness against random dilution in attractor neural networks

1991 J. Phys. A: Math. Gen. 24 L743

(http://iopscience.iop.org/0305-4470/24/13/008)

J. Phys. A: Math. Gen. 24 (1991) L743–L749. Printed in the UK

LETTER TO THE EDITOR

Robustness against random dilution in attractor neural networks

A Komoda†, R Serneels†, K Y M Wong‡ and M Bouten†

† Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

‡ Department of Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK

Received 8 March 1991

Abstract. We study the robustness of attractor neural networks against random disruption of a fraction of the synaptic couplings. For the maximally stable network (MSN), we determine the effect of different degrees of dilution on the overall storage capacity, the size of the basins of attraction and the attractor overlap. Comparison with corresponding results for the Hopfield model indicates that, although MSN is more robust at low degrees of dilution, the Hopfield network becomes more robust at high dilution.

Memory retrieval in attractor neural network models is a collective phenomenon. Information of one of the many stored patterns is encoded distributively among the synapses, so that the stored pattern is associated with a dynamically stable configuration of the neuronal states. In common with many other collective phenomena in physics, it is a *global* behaviour which is very robust against *local* uncorrelated disruptions. In the presence of thermal noise or dilution of synapses, for example, macroscopic retrieval states can still survive (although their overlap with the associated patterns may be affected) [1,2].

In a previous paper [3], we have investigated the robustness of the stored memories against random cutting of a fraction of the connections in a *feed-forward* neural network. We have seen that perfect output cannot be sustained even for perfect input, once a small fraction of the connections is destroyed. As a result, the storage capacity of *perfect retrieval* immediately drops from a finite value to zero on random dilution. Nevertheless, this does not genuinely reflect the robustness of the system, for the output can still have a high probability of being correct if the patterns have been stored during the learning process with a large aligning field.

Robustness against random dilution may be more clearly demonstrated in *attractor* neural networks, in which the outputs of the neurons are fed back to the input to maintain an iterative dynamics. For a feedback architecture, the network state can still drift towards an attractor which is highly correlated with the stored pattern, although it cannot retrieve the pattern perfectly on random dilution. Thus it is still meaningful to consider the notion of robustness in terms of the attractor overlap, basin size and storage capacity of *attraction*.

In the present letter, we generalize the notion of robustness against random dilution to attractor neural networks. We shall concentrate on the effects of the disruption of synapses on the attractor overlap, basin size and storage capacity. As

0305-4470/91/130743+07$03.50

we shall see, the networks deteriorate on increasing dilution using all of the three performance criteria, and the disruptive effects are especially marked for networks with small aligning fields.

It has been pointed out that the effects of random dilution are equivalent to those of static synaptic noise [2,4]. Both introduce fluctuations to the local fields of each neuron, proportional to the magnitude of disruption. Here we notice that in dilute networks, the effects of random dilution are, furthermore, equivalent to those of thermal noise [5]. This is because configuration correlations beyond one time step are negligible in dilute networks, rendering static and dynamic synaptic noises indistinguishable. Thus the dilution fraction can be represented by an effective noise temperature, and there is a direct correspondence between the phase diagrams of the two cases. Consequently, a network robust against random dilution should also be robust against thermal noise, and vice versa.

The determination of the dynamical features of the system requires solution of the dynamical equations over many time steps. For a network with high connectivity, this is a very difficult problem. Most analytical calculations have therefore been done for the highly diluted structure of Derrida *et al* [6]. In this model, each neuron is connected to only a small number C of the total number of N neurons with $\ln C \ll \ln N$ [7]. The random and high dilution basically reduces the dynamical problem to the solution of a single time step.

In this letter, we consider such a highly diluted network in which p patterns are stored in the non-vanishing synaptic coefficients J_{ij}. We are particularly interested in the behaviour of the so-called maximally stable network (MSN) [8], because it has the highest storage capacity in the absence of disruption [5], and our previous study has revealed its strong robustness against random dilution [3]. Networks with other synaptic presciptions, such as the Hebbian network, will be discussed afterwards.

In the MSN the coupling coefficients fulfil the following two requirements:

(i) the normalization conditions

$$\sum_j J_{ij}^2 = C \qquad (i = 1, 2, \cdots, N) \tag{1}$$

(ii) the stability conditions for the aligning fields

$$\Lambda_i^\mu \equiv \frac{1}{\sqrt{C}} \sum_j \xi_i^\mu J_{ij} \xi_j^\mu > K \qquad (\mu = 1, \cdots, p \quad i = 1, \cdots, N). \tag{2}$$

The maximum allowed value for the stability parameter K is determined, for each value of the storage ratio $\alpha = p/C$, by Gardner's equation [9]

$$\frac{1}{\alpha} = \int_{-\infty}^{K(\alpha)} \frac{\mathrm{d}t}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2)(K(\alpha) - t)^2. \tag{3}$$

We consider parallel dynamics in which the state of every neuron $S_i(t)$ is updated at each time step

$$S_i(t+1) = \mathrm{sgn}\left(\frac{1}{\sqrt{C}} \sum_j J_{ij} S_j(t)\right) \qquad (i = 1, \cdots, N). \tag{4}$$

Knowing the overlap $m(t)$ of the network state with one of the stored patterns at time t, it is easy to calculate the overlap $m(t+1)$ with the same pattern at the subsequent time $t+1$ [10, 11]:

$$m(t+1) = \int \mathrm{d}\Lambda\, P(\Lambda) \operatorname{erf}\left(\frac{m(t)\Lambda}{\sqrt{2(1-m^2(t))}}\right). \tag{5}$$

This expression depends solely on the distribution $P(\Lambda)$ of the aligning field Λ_i^μ. For the highly diluted network, one is allowed to iterate (5) over many time steps. This makes it possible to determine the attractor overlap and basins of attraction by studying the fixed points of the map (5). Such a calculation has been done for MSN by Gardner [10]. For MSN, the $P(\Lambda)$ distribution is given by a normalized Gaussian truncated below K plus a delta function at K:

$$P(\Lambda) = \delta(\Lambda - K)\tfrac{1}{2}[1 + \operatorname{erf}(K/\sqrt{2})] + \theta(\Lambda - K)(1/\sqrt{2\pi}) \exp(-\tfrac{1}{2}\Lambda^2). \tag{6}$$

To study the robustness of MSN against random cutting of synapses, we execute in the highly diluted network a further random cutting of a fraction of the remaining synapses, after the synaptic coefficients have been prescribed according to (1) to (3). Each of the non-vanishing synapses acquires an independent probability $(1-f)$ of being disrupted so that, after our destructive action is over, every neuron remains connected on average to only fC other neurons. The connections that survive the cutting keep their previously assigned value J_{ij}. The synaptic coefficient in the final network can therefore be written as $c_{ij}J_{ij}$ where $c_{ij} = 1$ or 0 with probability f and $(1-f)$, respectively. The new aligning fields are

$$\gamma_i^\mu = \frac{1}{\sqrt{fC}} \sum_j \xi_i^\mu c_{ij} J_{ij} \xi_j^\mu \tag{7}$$

where we have adapted the normalizing pre-factor to take account of the decreased number of terms in the sum. The distribution $P(\gamma)$ of the γ_i^μ has been calculated in [3] and could be used, after correcting for the altered normalization factor, directly in the dynamical equation (5). An alternative and simpler expression can be obtained by noting that the γ_i^μ, for fixed J_{ij} and ξ_i^μ and thus fixed Λ_i^μ, are Gaussian variables with mean $\sqrt{f}\Lambda_i^\mu$ and variance $1-f$. Retracing the derivation of (5) and averaging now over the c_{ij} as well yields

$$m(t+1) = \int \mathrm{d}\Lambda\, P(\Lambda) \operatorname{erf}\left(\frac{\sqrt{f}m(t)\Lambda}{\sqrt{2(1-fm^2(t))}}\right) \tag{8}$$

where $P(\Lambda)$ is again the distribution (6) of the aligning fields Λ_i^μ in MSN. Putting $f = 1$ in (8) reproduces (5) as it should.

The fixed points of the iterative map (8) for different values of K (or α) and f can only be obtained numerically. Figure 1 shows the results for four different values of f. Full curves represent stable fixed points, while broken curves show unstable fixed points. The latter define the basins of attraction. For $f = 1$, we recover the results of Gardner [10] with a stable fixed point $m = 1$ for all $\alpha \leqslant \alpha_c = 2$. This retrieval attractor has a wide basin of attraction only when α is smaller than $\alpha_B = 0.42$ and a

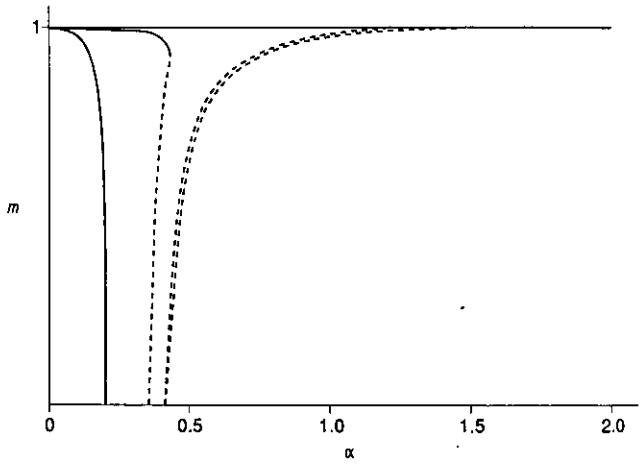

Figure 1. Fixed point overlaps of the randomly diluted MSN network for four different degrees of dilution (From right to left: $f = 1$, 0.99, 0.80, 0.40). The stable fixed points are shown as full curves and the unstable ones as broken curves.

narrow basin for $\alpha > 0.42$. It disappears discontinuously at $\alpha_c = 2$. For $f = 0.99$, the picture remains qualitatively the same. Below the critical storage ratio $\alpha_c = 1.18$, a stable retrieval fixed point exists with an overlap m very close to 1. This attractor has a wide basin of attraction only when α is smaller than 0.416 and a narrow basin for larger values of α. Above the critical storage $\alpha_c = 1.18$, the only attractor is $m = 0$. For these large values of α, a random disruption of 1 per cent of all connections in MSN has a devastating effect on the stored memories. This confirms our previous finding [3] that a network with small stability parameter K is extremely vulnerable to random cutting of even a very small fraction of the synapses. For increased damage, $f = 0.8$, we can still recognize the same behaviour but the difference between the critical value $\alpha_c = 0.43$ and the boundary for wide retrieval $\alpha_B = 0.354$ has become much smaller. When the damage becomes very large like for $f = 0.4$, we observe a different behaviour. The interval of narrow retrieval has disappeared altogether. For values of α larger than $\alpha_c = \alpha_B = 0.205$, no retrieval is possible. For lower values of α, we always have wide retrieval. The retrieval overlap, however, is no longer close to 1 but decreases continuously to zero at α_c.

The transition point between the two regimes observed in figure 1 can be determined analytically by studying equation (8) for very small values of m [5]. Denoting the right hand member of (8) by $f_K(m)$ and expanding in powers of m, one obtains the approximate fixed-point equation

$$m = f_K'(0)m + \frac{1}{3!}f_K'''(0)m^3 \qquad (m \ll 1). \tag{9}$$

The fixed point $m = 0$ changes its stability on iteration when $f_K'(0) = 1$, or

$$\frac{\pi}{2f} = \left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{K^2}{2}\right) + \frac{K}{2}\left[1 + \operatorname{erf}\left(\frac{K}{\sqrt{2}}\right)\right]\right\}^2. \tag{10}$$

For each value of f, this equation determines the value $K_B(f)$ or $\alpha_B(f)$ which confines the interval of wide retrieval. The sign of $f_K'''(0)$ in (9) determines the curvature of

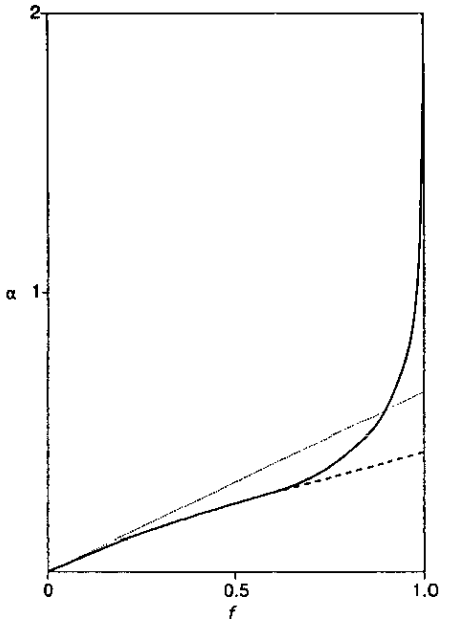

Figure 2. Phase diagram for the randomly diluted MSN network. The full curve shows α_c, the boundary between retrieval and no retrieval. The broken curve shows $\alpha_B(f)$, the transition between wide and narrow retrieval. The dotted line shows the retrieval boundary for the dilute Hopfield network.

the lines of fixed points at $m = 0$ in figure 1. This curvature differentiates the two regimes in the following way. For $f_K'''(0) < 0$, $m = 0$ is the only fixed point when the system lies on the phase boundary defined by (10), which therefore separates the regimes of wide retrieval and non-retrieval. On the other hand, for $f_K'''(0) > 0$, an other stable fixed point is present when the system lies on the phase boundary (10), which therefore separates the regimes of wide and narrow retrieval. This results in a tricritical point, on the phase boundary (10), which satisfies $f_K'(0) = 1$ and $f_K'''(0) = 0$ simultaneously. The condition $f_K'''(0) = 0$ yields

$$(1-K^2)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{K^2}{2}\right)+\frac{K}{2}\left[1+\operatorname{erf}\left(\frac{K}{\sqrt{2}}\right)\right](3-K^2)=0. \qquad (11)$$

Solving this equation numerically yields $K^* = 1.70$. Putting this value in (3) and in (10) yields $\alpha^* = 0.26$ and $f^* = 0.53$. This latter values fix the boundary curve between the two regimes in figure 1.

The results are summarized in the (α,f) phase diagram in figure 2. The full curve shows $\alpha_c(f)$ while the broken curve represents $\alpha_B(f)$. The curve $\alpha_c(f)$ marks the boundary between retrieval and no retrieval. The curve $\alpha_B(f)$ divides the domain of retrieval into wide and narrow retrieval. Both curves merge with the same slope [5] for values of f smaller than $f^* = 0.53$ which defines the tricritical point. For values of f larger than 0.53, the retrieval overlap is in general close to 1 for all values of α below $\alpha_c(f)$ where it jumps discontinuously to 0. For values of f smaller than 0.53, the retrieval overlap decreases continuously from the value 1 at $\alpha = 0$ tot 0 at $\alpha_c(f)$. This different behaviour corresponds to the two regimes in figure 1.

The dotted line in figure 2 shows the storage capacity for the highly diluted Hopfield model, i.e. the Hebbian network [6]. Since in this case $J_{ij} \sim \sum_\mu \xi_i^\mu \xi_j^\mu$ for all synaptic coefficients, the effects of dilution before and after the learning stage become indistinguishable. Random cutting of the remaining synapses after learning merely causes a further reduction of the very low connectivity. The critical storage capacity α_c therefore is given by the straight line $2\pi/f$. Alternately, the same result is obtained from the dynamical equation (8) by noting that the aligning field distribution $P(\Lambda)$ for the Hopfield model is a normalized Gaussian of mean $1/\sqrt{\alpha}$. The resultant equation is equivalent to that of the undiluted case if we substitute α/f for α. No difference exists between α_c and α_B because the attractor always has a wide attraction basin.

Comparing the MSN and the diluted Hopfield model, we note that the Hopfield model has a higher storage capacity than the MSN for f smaller than 0.89. Only for larger values of f (i.e. little damage) does the MSN perform better than the randomly diluted Hopfield model. This can be considered as a manifestation of the principle of specialization, in the sense that the MSN performs better in the presence of low disruption, whereas the Hopfield model performs better in a highly disruptive environment [8].

In the limit of very high dilution ($f \to 0$), the phase lines of the two networks become tangents to each other. This is because in the limit of very low storage capacity, the aligning field distribution $P(\Lambda)$ for the two cases are nearly identical. For the MSN it is essentially a delta function peaked at $K \sim 1/\sqrt{\alpha} \gg 1$, whereas for the Hopfield model it is a Gaussian peaked at $1/\sqrt{\alpha} \gg 1$. It is interesting to note a similar feature for the effects of thermal noise on the two networks [5]. In the limit of high noise temperature, the two networks have the same storage capacity asymptotically.

It is natural to compare the effects of random dilution and thermal noise. To this end, we can rewrite the dynamical equation (8) as

$$m(t+1) = \int d\Lambda\, P(\Lambda) \operatorname{erf}\left(\frac{m(t)\Lambda}{\sqrt{1 - m(t)^2 + T_{\text{eff}}^2}} \right) \tag{12}$$

where $T_{\text{eff}}^2 = (1-f)/f$ can be considered as the effective noise temperature due to random dilution. This dynamical equation is identical to the case of introducing a Gaussian noise of magnitude T_{eff} to the local fields during updating [8]. We note in passing that the equivalent *static* synaptic noise due to random dilution has again the same expression [2, 4]. In practice, thermal noise is dynamical, i.e. it fluctuates from one time step to the other, and therefore has a nature different from dilution noise, which is static. For the case of thermal noise studied in [5], we have a probabilistic dynamics in a perfect structure of the network, whereas here we are concerned with a deterministic dynamics in a disrupted network with static, structural noise. However, since configurational correlations beyond one time step are negligible in the highly diluted network, the effects of the two kinds of noise become indistinguishable. The phase diagram in figure 2 is therefore identical to that for the thermal noise [5] after a rescaling of the axes.

Our work has demonstrated that the MSN has a strong robustness specialized to a low degree of dilution, whereas the Hopfield model is specialized to an opposite environment. It is therefore natural to consider the more general issue of finding the

synaptic prescription with optimal robustness at a fixed degree of random dilution. The corresponding issue in the case of thermal noise has recently been addressed using the principle of adaptation [8]. There an appropriate performance function corresponding to a training overlap is defined, and the synaptic prescription optimizing the performance function is chosen from the space of all synaptic coefficients. The optimally adapted network is obtained by setting the training overlap to be the same as the attractor overlap, to be determined self-consistently. An interesting result is that the Hopfield model has a higher storage capacity than *any* other synaptic prescription for a noise temperature higher than 0.38. Employing the notion of effective noise temperature in (12), this implies that the Hopfield model has the highest storage capacity for a fraction f less than 0.87. The case of optimal adaptation in the presence of random dilution will be considered in detail in a separate publication [12].

We thank D Sherrington and E Domany for very meaningful discussions. This work is partially supported by a grant from the Science and Engineering Research Council of the United Kingdom and by the Programme on Inter-University Attraction Poles of the Belgian Government.

References

[1] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[2] Sompolinsky H 1987 *Heidelberg Colloqium on Glassy Dynamics (Springer Lecture Notes in Physics 275)* ed J L van Hemmen and I Morgenstern (Berlin: Springer)
[3] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643
[4] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[5] Amit D J , Evans M R , Horner M and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
[6] Derrida B, Gardner E, Zippelius A 1987 *Europhys. Lett.* **4** 167
[7] Kree R and Zippelius A *Preprint*
[8] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659
[9] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[10] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[11] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[12] Wong K Y M and Bouten M 1991 Optimal robustness against random dilution in neural networks, in preparation