Random dilution in a neural network for biased patterns

1989 J. Phys. A: Math. Gen. 22 2103

(http://iopscience.iop.org/0305-4470/22/12/014)

J. Phys. A: Math. Gen. **22** (1989) 2103-2118. Printed in the UK

Random dilution in a neural network for biased patterns

M R Evans

Department of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK

Received 26 January 1989, in final form 20 February 1989

Abstract. A neural network introduced by Tsodyks and Feigel'man suitable for the storage of biased patterns is studied in a randomly diluted form. Coupled evolution equations are derived for the two order parameters needed to describe a configuration near to a stored pattern. These equations are studied numerically and are found to exhibit non-linear effects such as spiralling trajectories and limit cycles. The bifurcations by which the transition to no memory occurs are illustrated. It is seen that a nominated pattern may not be within the basin of attraction of the memory fixed point correlated with it. The structure of a memory fixed point is investigated and found to be more complicated than a single configuration. Finally the situation where two patterns are highly correlated is examined and phase boundaries separating the regimes of no memory, undistinguishing memory and distinguishing memory are constructed. Multiple storage of a pattern does not improve its recall without appropriate modification of the thresholding parameter.

1. Introduction

The modelling of McCulloch-Pitts (1943) neural networks by spin systems has become an area of considerable interest within the physics community. In general the aim of such models is to choose a matrix of interactions (synaptic connection strengths) between the N neurons so that $l = \alpha N$ nominated configurations (patterns) are fixed points of some prescribed dynamics. The initial motivation in studying such models was to ascertain their dynamic behaviour because the conditions under which an input or starting configuration evolves to a pattern define how the network acts as an associative memory (Hebb 1949, Little 1974).

Although recent studies have been made concerning such basins of attraction (Forrest 1988, Kepler and Abbott 1988, Krauth *et al* 1988a, b). It is the static properties of models that have been more extensively studied (Hopfield 1982, Amit *et al* 1985a, b, 1987, Gardner 1986, Bruce *et al* 1987). A mean-field theory was developed initially for the Hopfield-Little model and yielded a maximum storage capacity α_c (Amit *et al* 1985b). In this model the neurons are two-state variables S_i taking on values ± 1 (firing and non-firing), the interactions are given by the Hebb (1949) rule and patterns to be stored are unbiased in that they contain equal numbers of firing and non-firing neurons. Some attention has now turned to the situation where the patterns to be stored are biased (Amit *et al* 1987). In this case the models that yield the highest α_c use variables V_i for the neurons, where $V_i = 1, 0$ (Willshaw and Longuet-Higgins 1970, Tsodyks and Feigel'man 1988, Tsodyks 1988, Buhmann *et al* 1988, Horner 1988). In the limit of maximum bias, where nearly all neurons in a pattern are not firing and take the value zero, their expressions for α_c are of the same form as Gardner's estimate for a network with unspecified but optimal interaction strengths (Gardner 1987).

0305-4470/89/122103+16$02.50

The model introduced by Tsodyks and Feigel'man (1988) and independently by Buhmann *et al* (1988) uses a modified Hebb conection rule defined by

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{l} (\eta_i^\mu - p)(\eta_j^\mu - p) \tag{1}$$

where η_i^μ denotes the value of the ith neuron in the μth nominated pattern. The bias of the patterns is p so that the quenched random variables η follow a distribution

$$P(\eta) = p\delta(\eta - 1) + (1-p)\delta(\eta). \tag{2}$$

We shall be concerned with the case of large bias where $p \ll 1$. The energy function is defined as

$$H = -\frac{1}{2N} \sum_{i \neq j} J_{ij} V_i V_j + \theta \sum_i V_i. \tag{3}$$

The term θ is an external field (known as the thresholding) which is essential if the storage capacity is to be increased over the unbiased case. Mean-field theory (Tsodyks and Feigel'man 1988, Buhmann *et al* 1988) shows that at the optimum value of θ and zero temperature a first-order phase transition to no memory occurs at

$$\alpha_c \sim \frac{1}{2p|\log p|}.$$

The role of θ in restricting the number of ones in a configuration becomes more apparent when one considers the effective field at a site i

$$h_i = \sum_{j \neq i} J_{ij} V_j - \theta \tag{4}$$

and the updating rule of the dynamics

$$\begin{aligned} V_i(t+\Delta t) &= 1 \qquad \text{with probability} \qquad \left[1 + \exp\left(-\frac{2h_i(t)}{T}\right)\right]^{-1} \\ V_i(t+\Delta t) &= 0 \qquad \text{with probability} \qquad \left[1 + \exp\left(\frac{2h_i(t)}{T}\right)\right]^{-1} \end{aligned} \tag{5}$$

where T is the temperature. Mean-field theory, however, gives no information on the dynamical behaviour of the model, which is the aspect we examine in this paper. Specifically we consider parallel dynamics whereby at each time step all sites are updated synchronously by (5) and we examine flows in phase space of a version of the model in which the dynamics can be solved exactly. The points to be investigated are as follows.

(i) To what extent are the flows restricted to the region of phase space that has the same bias as the patterns?

(ii) What are the qualitative features of the basins of attraction of the patterns and how do they change as the system becomes saturated at α_c?

(iii) How does varying the thresholding change the flows?

In order to carry out this work we need to derive evolution equations for a network configuration. In a fully connected network the temporal evolution of configurations is a difficult problem to tackle due to the correlations between sites (Gardner *et al* 1987). Only for a finite number of patterns can self-averaging flow equations be derived (Coolen and Ruijgrok 1988). However, it has been shown that if one randomly dilutes

a network by cutting most of the connections then the dynamics can be solved exactly (Derrida *et al* 1987). This approach forsakes an Hamiltonian, because the connections become asymmetric, thus the model is defined solely by its dynamics. Derrida *et al* (1987) found that randomly diluting the Hopfield–Little model changes its characteristics considerably. In particular the transition at α_c becomes second rather than first order. However, we will show that when we randomly dilute the connection rule (1) the transition remains first order and the fixed-point structure is unchanged.

The latter result has been noted by Tsodyks (1988) where the connection rule (1) was also studied in a randomly diluted form. Tsodyks (1988) considered the model defined by continuous-time dynamics, as opposed to the discrete-time dynamics considered in the present work, and used the dynamic functional method (Kree and Zippelius 1987) to obtain the fixed points. However, information was not obtained on the basins of attraction of the fixed points which is the main problem addressed in the present work by deriving evolution equations for the order parameters.

The randomly diluted connection rule is given by

$$J_{ij} = C_{ij} \sum_{\mu=1}^{l} (\eta_i^\mu - p)(\eta_j^\mu - p) \tag{6}$$

where the distribution of C_{ij} is given by

$$P(C_{ij}) = \frac{C}{N}\delta(C_{ij}-1) + \left(1-\frac{C}{N}\right)\delta(C_{ij}) \tag{7}$$

where C is the mean number of other neurons to which a particular neuron is connected. The connection rule is now asymmetric because C_{ij} is independent of C_{ji}. When $C \ll \log N$ the dynamics become exactly soluble due to the elimination of correlations between sites (Derrida and Pomeau 1986, Derrida *et al* 1987).

To express the overlap of the configuration of the system with a single pattern (taken to be pattern 1) one requires two order parameters

$$m_1^{(1)} = \frac{1}{Np}\sum_{i=1}^{N} \langle \eta_i^{(1)} V_i \rangle \tag{8}$$

$$m_2^{(1)} = \frac{1}{Np}\sum_{i=1}^{N} \langle (1-\eta_i^{(1)}) V_i \rangle. \tag{9}$$

It is also useful to consider

$$r = m_1^{(1)} + m_2^{(1)}. \tag{10}$$

These parameters have straightforward physical interpretations: m_1 measures the number of 'correct ones' (sites that are 1 in the configuration and pattern); m_2 measures the number of 'incorrect ones' (sites that are 1 in the configuration and 0 in the pattern); r measures the 'activity' (the number of ones in the configuration). The choice of m_1 and m_2 is convenient because they correspond to averages over 1-sites and 0-sites of the pattern respectively. The asymmetry between the fields at the two types of site necessitates this separation.

2. Derivation of evolution equations

We now proceed to derive equations for the evolution of the order parameters. The calculation outlined here is an extension of that performed by Derrida *et al* (1987).

Consider the situation where $m_1^{(1)}$ is finite but $m_1^{(\mu)} \sim O(1/Np)$, for $\mu > 1$. This requires $r \sim O(1)$. In this situation the configuration is near to pattern 1 and we need only consider the two order parameters associated with pattern 1. With this in mind we will drop the pattern 1 superscript from the order parameters. By calling a site that takes value one in pattern 1 a '1-site' we can express the order parameters as

$$m_1(t+\Delta t) = \langle\langle V_i(t+\Delta t)\rangle\rangle_{1\text{-sites}} \tag{11}$$

$$= \left\langle \left[1+\exp\left(-\frac{2h_i(t)}{T} \right) \right]^{-1} \right\rangle_{1\text{-sites}} \tag{12}$$

$$m_2(t+\Delta t) = \langle\langle V_i(t+\Delta t)\rangle\rangle_{0\text{-sites}} \tag{13}$$

$$= \left\langle \left[1+\exp\left(-\frac{2h_i(t)}{T} \right) \right]^{-1} \right\rangle_{0\text{-sites}} \tag{14}$$

where the single angular bracket indicates an average over sites and the double angular brackets an additional thermal average. To perform the site averages we must construct expressions for the field distributions. As the configurations are near to pattern 1 we can split the field into a signal term from pattern 1 and a noise term from the other patterns:

$$h_i(t) = (\eta_i^1 - p) \sum_{k=1}^{K} (\eta_k^1 - p) V_k(t) + \sum_{\mu=2}^{l} \sum_{k=1}^{K} (\eta_i^\mu - p)(\eta_k^\mu - p) V_k(t) - \theta \tag{15}$$

where the k index labels the K sites that are connected to site $i (C_{ik} = 1)$. For a site k that is a 1-site

$$V_k(t) = 1 \text{ with probability } m_1(t) \tag{16}$$

and for a site k that is a 0-site

$$V_k(t) = 1 \text{ with probability } \frac{p}{(1-p)} m_2(t). \tag{17}$$

We can now write down the probability that

$$h_i(t) = (\eta_i^1 - p)(S_1(1-p) - S_2 p) + N_1(1-p)^2 - N_2 p(1-p) + N_3 p^2 - \theta \tag{18}$$

as

$$P(S_1, S_2, N_1, N_2, N_3)$$

$$= \sum_{K=0}^{N} \frac{C^K e^{-C}}{K!} \frac{K!}{S_1! S_2! S_3!} (pm_1(t))^{S_1} (pm_2(t))^{S_2} (1-pr(t))^{S_3} \delta_{S_1+S_2+S_3, K}$$

$$\times \frac{(K(l-1))!}{N_1! N_2! N_3! N_4!} (p^3 r(t))^{N_1} (2(1-p)p^2 r(t))^{N_2}$$

$$\times ((1-p)^2 pr(t))^{N_3} (1-pr(t))^{N_4} \delta_{N_1+N_2+N_3+N_4, K(l-1)}.$$

The S terms are the signal, coming from the first term on the RHS of (15), and the N terms are noise. For example one of the K signal terms will be an S_1 term if $\eta_k^1 = 1$ and $V_k = 1$, the probability of which is pm_1. As $N \to \infty$ with $C \ll \log N$ the multinomial probabilities reduce to independent binomial probabilities for S_1, S_2, N_1, N_2, N_3. One finds that the N_1 distribution has the largest variance by at least a factor $1/p$. Thus

in the limit $p \ll 1$ an average over the field is equivalent to a Gaussian average over the N_1 noise with the other terms in (18) set to their means. On performing such an average for m_1 and m_2 and discarding terms of order p one obtains

$$m_1(t+\Delta t) = \int_{-\infty}^{\infty} \frac{dy \exp[-y^2]}{\pi^{1/2}} \left[1+\exp\left(\frac{-2}{T_0}(m_1(t)+(2\alpha pr(t))^{1/2}y-\theta_0)\right)\right]^{-1}$$
$$m_2(t+\Delta t) = \frac{1}{p}\int_{-\infty}^{\infty} \frac{dy \exp[-y^2]}{\pi^{1/2}} \left[1+\exp\left(\frac{-2}{T_0}((2\alpha pr(t))^{1/2}y-\theta_0)\right)\right]^{-1} \tag{19}$$

where

$$T_0 = T/Cp \qquad \alpha = l/C \qquad \theta_0 = \theta/Cp.$$

At $T=0$ the integrals appearing in (19) can be written as Gauss error functions to give

$$m_1(t+\Delta t) = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{m_1(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right]$$
$$m_2(t+\Delta t) = \frac{1}{2p}\left[1-\operatorname{erf}\left(\frac{\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right] \tag{20}$$

where

$$\operatorname{erf}(x) = \frac{2}{\pi^{1/2}}\int_0^x e^{-t^2}\,dt. \tag{21}$$

One could also consider the model defined by random sequential dynamics where at each time step a site is chosen randomly and updated by rule (5). This model would have, in place of the map (20), the flow

$$\frac{dm_1}{dt} = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{m_1-\theta_0)}{(2\alpha pr)^{1/2}}\right)\right] - m_1$$
$$\frac{dm_2}{dt} = \frac{1}{2p}\left[1-\operatorname{erf}\left(\frac{\theta_0}{(2\alpha pr)^{1/2}}\right)\right] - m_2 \tag{22}$$

which is similar in form to the Wilson and Cowan (1972) evolution equations for populations of interacting excitatory and inhibitory neurons. The fixed points of (20) and (22) are identical and one would expect that the trajectories in the m_1-m_2 plane near to these fixed points are qualitatively the same. For simplicity only the model using parallel dynamics is considered further.

2.1. Discussion of evolution equations and their fixed-point structure

In studying the map (19) we are primarily interested in fixed points that are highly correlated with the patterns. These fixed points will be referred to as *memories* to distinguish them from the nominated patterns. Remarkably the fixed point equations of (19) are equivalent to the saddle point equations derived for the fully connected model (Tsodyks and Feigel'man 1988). This suggests that the Gaussian treatment of the noise from uncondensed patterns used here is correct even for the fully connected model in the limit $p \ll 1$. This can be explained by realising that in a V-model network it is only the 1s that interact and for the case of patterns with high bias comparatively few 1s are shared by patterns. Thus there is little correlation within the noise.

We can use the analysis of Tsodyks and Feigel'man (1988) to write down approximate forms for critical quantities. Firstly at zero temperature a memory exists up to a value of α given by

$$\alpha_c(p) \simeq \frac{\theta_0^2}{2p|\log p|} \tag{23}$$

for $1-\theta \gg |\log p|^{-1/2}$ and

$$\alpha_c(p) \simeq \frac{(1-\theta_0)^2}{p|\log(1-\theta_0)|} \tag{24}$$

for higher values of θ. As $\alpha \to 0$ the transition temperature is

$$T_c \simeq \frac{(1-\theta_0)}{|\log(1-\theta_0)|}. \tag{25}$$

In these equations it can be seen that θ_0 controls the performance of the network. Figure 1 illustrates how θ_0 controls the position of the memory at $T=0$. At the lower θ_0 values the activity is greatest but that means there are incorrect 1s present; at higher values the activity is lowered but we have more incorrect 0s.

Bearing in mind that θ is present in order to control the level of activity then a qualitative look at how r comes into the map (19) is appropriate. The width of the

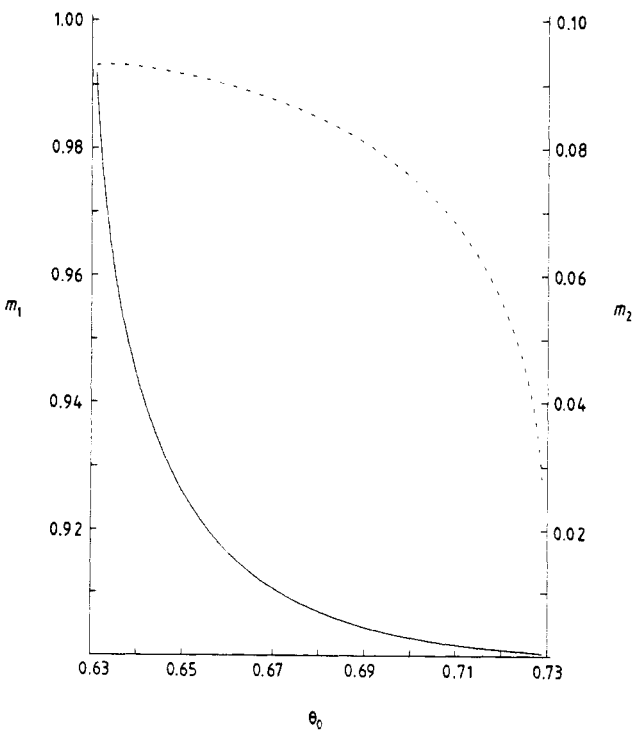

Figure 1. A plot of the m_1 (broken curve) and m_2 (full curve) values of the memory fixed point against θ_0 for $p = 10^{-4}$ and $\alpha = 200$.

noise is $2\alpha r$. An increase in m_1 or m_2 will increase the noise which in general will decrease m_1 but increase m_2. This constitutes a feedback loop in the activity which may become delicately balanced. The outcome of this is some curious non-linear effects to be discussed in the next section. The fact the activity is not rigidly constrained by the dynamics is reflected by the existence of two fixed points that are not correlated with the patterns. These are the 'all zeros' fixed point $m_1 = m_2 = 0$ and the high activity fixed point with $r \sim O(1/p)$. The second of these is outwith the region of validity of (20) because for such a large activity the configuration will have finite m_1 overlaps with more than one pattern. Although this fixed point's position cannot be determined by (20) it does exist in the model and its effect on trajectories near to a pattern is evident in the next section.

2.2. Numerical study of evolution equations

One of the key points to investigate in a neural network is how it acts as an associative memory. How does a given input or starting configuration evolve and under what conditions will the system iterate to a memory, thus making an association between the input and that memory? Armed with the map (19) we are in a position to examine the basins of attraction of the memories, in particular to see how they behave as α increases.

Figures 2 and 3 show trajectories in the m_1-m_2 plane at zero temperature, $p = 10^{-4}$ and selected values of θ and α. They show sequences in which α increases past some critical value. In figure 2 at $\alpha = 260$ the memory is an attractive node and there is a saddle point to its lower left which limits the memory's basin of attraction. There is also a saddle point out of the frame which acts as a watershed for trajectories reaching the high activity fixed point. At $\alpha = 261$ the qualitative structure remains the same. However a trajectory starting at the pattern (marked by a cross) is no longer within the memory's basin of attraction. When α reaches 262 the memory and lower saddle point have annihilated. This sequence of memory loss can be classified as a saddle-node bifurcation.

In figure 3 the memory starts as a spiral. The two saddle points mentioned above are again present. At $\alpha = 258$ the pattern is no longer within the memory's basin of

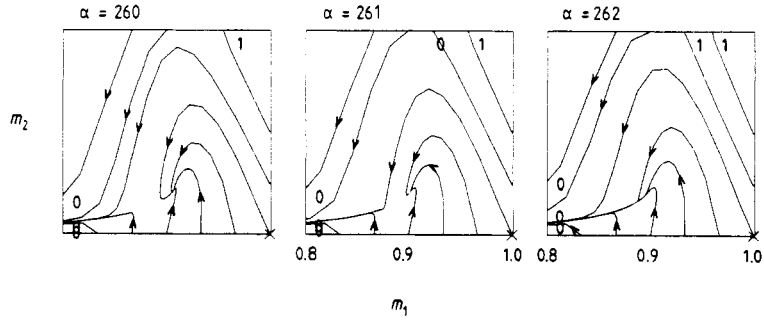

Figure 2. A sequence of frames showing trajectories of (20) near to the pattern which is marked by a cross. In the sequence α passes through its critical values. When a trajectory leaves the frame, a 1 indicates that it continues to the high activity fixed point and a 0 indicates that it continues to the all-zeros fixed point. The fixed parameter values are $p = 10^{-4}$ and $\theta_0 = 0.699$.

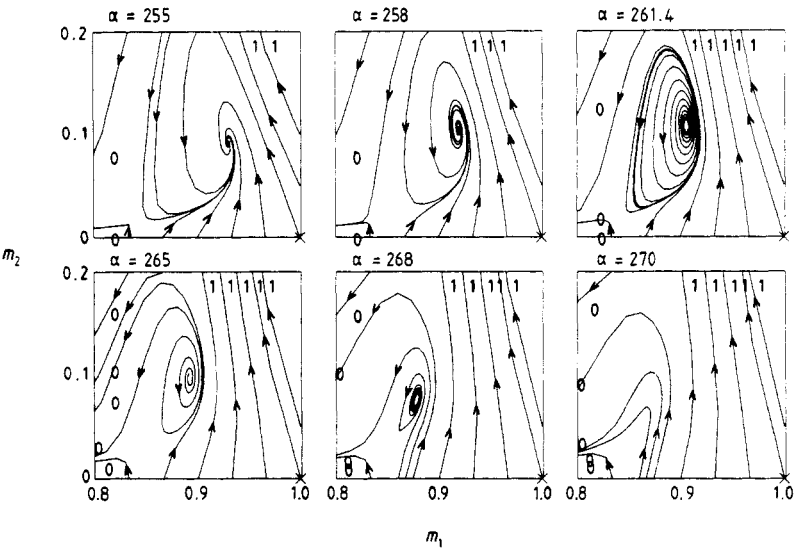

Figure 3. A sequence of trajectories as in figure 2. The fixed parameter values are $p = 10^{-4}$ and $\theta_0 = 0.6917$.

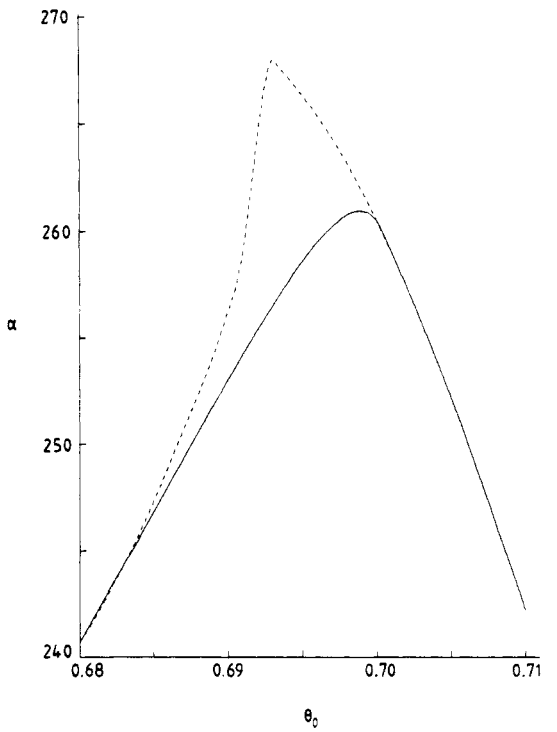

Figure 4. A plot of α_1 (full curve) and α_2 (broken curve) against θ_0 for $p = 10^{-4}$. The range of θ_0 is that within which both α values are maximised.

attraction. The memory destabilises at $\alpha = 261.4$ and trajectories starting from it spiral out to an attractive limit cycle. At $\alpha = 265$ this limit cycle is no longer present and all trajectories leave the frame. However, at $\alpha = 268$ the spiral restabilises and so a memory is again present. When α reaches 270 the lower saddle and memory have annihilated and there are no fixed points present in the frame. The destabilisation and restabilisation of the memory, of which this sequence is an example, will be referred to as *intermittency*.

These two figures show that the transition to no memory is first order and illustrate the two types of memory loss that are present in the model—saddle attractor bifurcations and spiral destabilisations. At lower θ_0 values the same mechanisms are present but spirals destabilise to the high activity fixed point and the higher saddle point is involved in the bifurcations.

One can interpret the various memory losses as the basins of attraction of the all zeros and high activity fixed points becoming large enough to encroach on the memory. The value of θ then determines which of these basins of attraction reaches the memory first. Figure 3 is in the intermediate-θ regime where both non-memory fixed points are strongly affecting the memory. This gives rise to spiralling, limit cycles and intermittency.

The figures also illustrate the need for a more careful definition of the critical value of α. When considering neural networks as associative memories one would like the pattern to be within the basin of attraction of the memory fixed point so that there is a clear association between them. The value of α at which this is no longer so will be called α_1. However, α_1 is not a parameter directly relevant to stability analysis of the fixed points, so unless one can find formulae for the basins of attraction it is only from numerical studies of flows as performed in this work that α_1 can be determined.

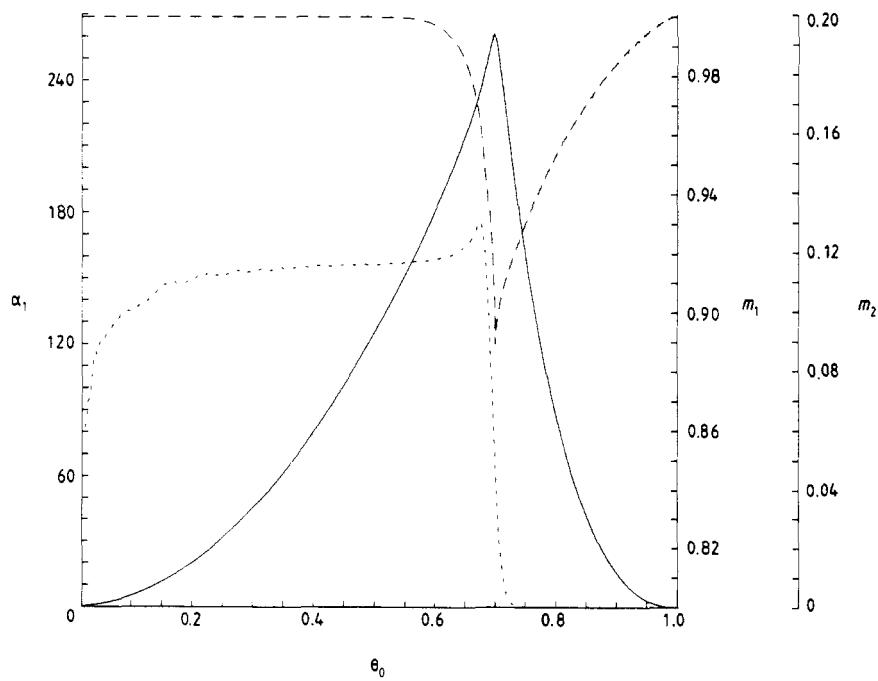

Figure 5. A plot of α_1 against θ_0 at $p = 10^{-4}$ with the values of the order parameter memory fixed points at α_1 superimposed. Full curve, α_1; broken curve, m_1; dotted curve, m_2.

A second α value of importance is α_2 which we define to be the lowest value at which there is no memory. By using α_2 we avoid dealing with any intermittancy that may be present.

Figure 4 shows that α_1 and α_2 differ only in the region of θ values that give the highest storage. The slight kink on the left of α_2 curve is where the attractor becomes a spiral. Figure 5 shows α_1 has the θ dependence given by (23) and (24), and how the position of the memory at α_c moves sharply in the transition region between the two forms.

3. Structure of the memory

Derrida *et al* (1987) showed, by following the evolution of the overlap between two different configurations that were within the basin of attraction of the same memory, that a memory was not a single configuration. The method will be followed in the present work to investigate whether the same result is obtained and if so whether the asymmetry between 1s and 0s is reflected in the structure of the memory. We take two configurations $\{V_i\}$ and $\{\tilde{V}_i\}$ which have finite overlaps m_1 and $\tilde{m}_1$ with the same pattern, but infinitesimal overlaps with the other patterns. We define an overlap parameter between the two configurations

$$q = q_1 + q_2 \tag{26}$$

where

$$\begin{aligned} q_1 &= \frac{1}{Np}\sum_{i=1}^{N}\langle \eta_i (V_i - \tilde{V}_i)^2\rangle \\ q_2 &= \frac{1}{Np}\sum_{i=1}^{N}\langle (1-\eta_i)(V_i - \tilde{V}_i)^2\rangle. \end{aligned} \tag{27}$$

Here q_1 measures the number of 1-sites that differ in the two configurations; q_2 measures the number of 0-sites that differ. If $q \to 0$ as $t \to \infty$ then the two configurations eventually become identical. Using the same averaging as for the evolution of the m_1 and m_2 order parameters

$$\begin{aligned} q(t+\Delta t) &= \tfrac{1}{2}\left\langle 1 - \tanh\left(\frac{h_i(t)}{T_0}\right)\tanh\left(\frac{\tilde{h}_i(t)}{T_0}\right)\right\rangle_{1\text{-sites}} \\ q_2(t+\Delta t) &= \frac{(1-p)}{2p}\left\langle 1 - \tanh\left(\frac{h_i(t)}{T_0}\right)\tanh\left(\frac{\tilde{h}_i(t)}{T_0}\right)\right\rangle_{0\text{-sites}}. \end{aligned} \tag{28}$$

The fields h_i and $\tilde{h}_i$ are correlated, because of the two configurations' overlap, and the probability that

$$\begin{aligned} h_i = (\eta_i^1 - p)[(1-p)(S_1+S_2) - p(S_5+S_6)] + (1-p)^2(N_1^1+N_2^1+N_5^1+N_6^1) \\ - p(1-p)(N_1^2+N_2^2+N_5^2+N_6^2) + p^2(N_1^3+N_2^3+N_5^3+N_6^3) - \theta \end{aligned} \tag{29}$$

$$\begin{aligned} \tilde{h}_i = (\eta_i^1 - p)[(1-p)(S_1+S_3) - p(S_5+S_7)] + (1-p)^2(N_1^1+N_3^1+N_5^1+N_7^1) \\ - p(1-p)(N_1^2+N_3^2+N_5^2+N_7^2) + p^2(N_1^3+N_3^3+N_5^3+N_7^3)) - \theta \end{aligned} \tag{30}$$

is given by

$$P(S_1, S_2 \ldots S_8, N_1^1, N_1^2, N_1^3, N_2^1, \ldots, N_8^3)$$

$$= \sum_{K=0}^{N} \frac{C^K \, \mathrm{e}^{-C}}{K!} \frac{K!}{S_1! \ldots S_8!} \left(\frac{p(m_1 + \tilde{m}_1 - q_1)}{2}\right)^{S_1} \left(\frac{p(m_1 - \tilde{m}_1 + q_1)}{2}\right)^{S_2}$$

$$\times \left(\frac{p(-m_1 + \tilde{m}_1 + q_1)}{2}\right)^{S_3} \left(p(1 - \tfrac{1}{2}(m_1 + \tilde{m}_1 + q_1))\right)^{S_4}$$

$$\times \left(\frac{p(m_2 + \tilde{m}_2 - q_2)}{2}\right)^{S_5} \left(\frac{p(m_2 - \tilde{m}_2 + q_2)}{2}\right)^{S_6} \left(\frac{p(-m_2 + \tilde{m}_1 + q_1)}{2}\right)^{S_7}$$

$$\times \left(1 - p\left(1 + \frac{1}{2}(m_2 + \tilde{m}_2 + q_2)\right)\right)^{S_8} \delta_{S_1 + S_2 + \ldots + S_8, K}$$

$$\times \frac{(S_1(l-1))!}{N_1^1! N_1^2! N_1^3!} p^{2N_1^1} (2p(1-p))^{N_1^2} (1-p)^{2N_1^3} \delta_{N_1^1 + N_1^2 + N_1^3, S_1(l-1)}$$

$$\vdots$$

$$\times \frac{(S_8(l-1))!}{N_8^1! N_8^2! N_8^3!} p^{2N_8^1} (2p(1-p))^{N_8^2} (1-p)^{2N_8^3} \delta_{N_8^1 + N_8^2 + N_8^3, S_8(l-1)}. \tag{31}$$

The S and N refer to the signal and noise terms of equation (15). The subscripts on these letters correspond to the eight possible situations at a site k:

	η_k	V_k	$\tilde{V}_k$
(1)	1	1	1
(2)	1	1	0
(3)	1	0	1
(4)	1	0	0
(5)	0	1	1
(6)	0	1	0
(7)	0	0	1
(8)	0	0	0

The superscripts on the noise terms indicate the contribution to the field. For example N_1^1 is the number of $(1-p)^2$ contributions to the field from sites of type (1) in the above table. As $N \to \infty$ followed by C and with $p \ll 1$ an average over the field reduces to a triple Gaussian average over

$$N_1^1 + N_5^1 \qquad N_2^1 + N_6^1 \qquad N_3^1 + N_7^1.$$

At $T = 0$ we obtain

$$q_1(t + \Delta t) = \frac{1}{2}\left[1 - \int_{\infty}^{\infty} \frac{\mathrm{d}y \, \mathrm{e}^{-y^2}}{\pi^{1/2}} \operatorname{erf}\left(\frac{\sqrt{\alpha p(m_1 + m_2 + \tilde{m}_1 + \tilde{m}_2 - q)}\, y + m_1 - \theta_0}{\sqrt{\alpha p(m_1 + m_2 - \tilde{m}_1 - \tilde{m}_2 + q)}}\right)\right.$$

$$\left. \times \operatorname{erf}\left(\frac{\sqrt{\alpha p(m_1 + m_2 + \tilde{m}_1 + \tilde{m}_2 - q)}\, y + \tilde{m}_1 - \theta_0}{\sqrt{\alpha p(-m_1 - m_2 - \tilde{m}_1 - \tilde{m}_2 + q)}}\right)\right] \tag{32}$$

$$q_2(t+\Delta t)=\frac{(1-p)}{2p}\left[1-\int_{\infty}^{\infty}\frac{dy\,e^{-y^2}}{\pi^{1/2}}\,\mathrm{erf}\left(\frac{\sqrt{\alpha p(m_1+m_2+\tilde{m}_1+\tilde{m}_2-q)}\,y-\theta_0}{\sqrt{\alpha p(m_1+m_2-\tilde{m}_1-\tilde{m}_2+q)}}\right)\right.$$
$$\left.\times\,\mathrm{erf}\left(\frac{\sqrt{\alpha p(m_1+m_2+\tilde{m}_1+\tilde{m}_2-q)}\,y-\theta_0}{\sqrt{\alpha p(-m_1-m_2+\tilde{m}_1+\tilde{m}_2+q)}}\right)\right] \tag{33}$$

where on the RHS of (33) and (34) m_1, m_2, $\tilde{m}_1$, $\tilde{m}_2$, q are the values at time t. As $t\to\infty$ m_1, $\tilde{m}_1$, m_2, $\tilde{m}_2$ attain the values given by the fixed points of (20) and the fixed points q_1^* and q_2^* are given by

$$q_1^*=\frac{1}{2}\left[1-\int_{-\infty}^{\infty}\frac{dy\,e^{-y^2}}{\pi^{1/2}}\,\mathrm{erf}^2\left(\frac{\sqrt{\alpha p(2m_1^*+2m_2^*-q^*)}\,y+m_1^*-\theta_0}{\sqrt{\alpha pq^*}}\right)\right] \tag{34}$$

$$q_2^*=\frac{(1-p)}{2p}\left[1-\int_{-\infty}^{\infty}\frac{dy\,e^{-y^2}}{\pi^{1/2}}\,\mathrm{erf}^2\left(\frac{\sqrt{\alpha p(2m_1^*+2m_2^*-q^*)}\,y-\theta_0}{\sqrt{\alpha pq^*}}\right)\right]. \tag{35}$$

As $q\neq 0$ the memory is not a unique configuration. Given this one might suppose that a memory is any corrupted version of the pattern with the appropriate number of incorrect 1s and 0s. This would mean $q_2^*=2m_2^*$, $q_1^*=2(1-m_1^*)$. However on inspecting figure 6 one sees that these equations are not obeyed. One therefore concludes that a memory configuration is a noisy version of the pattern with the incorrect sites not chosen randomly. For the parameter values in figure 6 the numbers of incorrect 1s and 0s are approximately equal but as $q_2^*\simeq m_2^*$ and $q_1^*<\frac{1}{2}(1-m_1^*)$ the incorrect 0s vary more than the incorrect 1s between memory configurations.

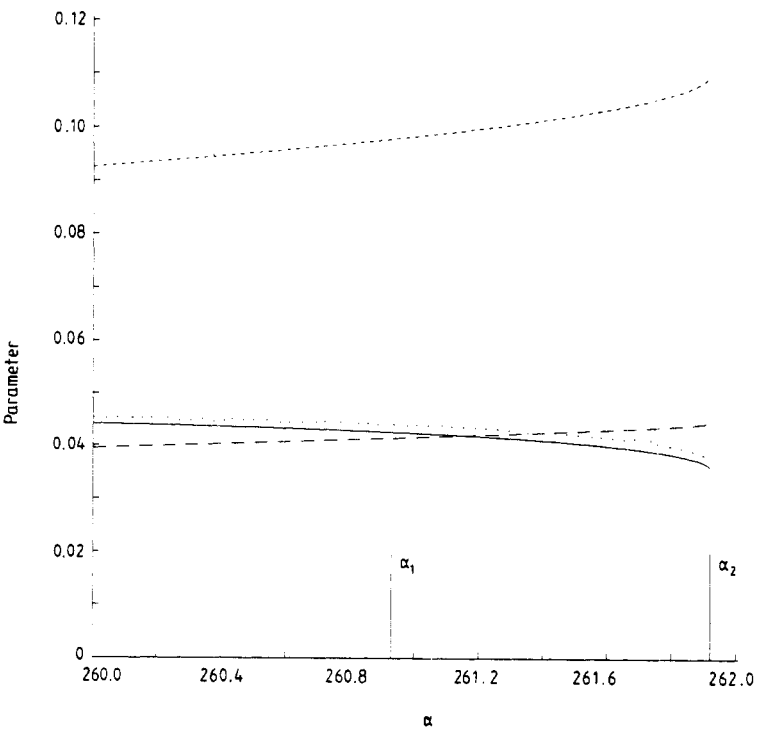

Figure 6. A plot of the fixed points of the overlap parameters (q_1, q_2) and the order parameters (m_1, m_2) against α for $\theta_0=0.699$ and $p=10^{-4}$. Full curve, m_2; short broken curve, $1-m_2$; long broken curve, q_1; dotted curve, q_2.

4. Higher than random overlap between patterns

In the Hopfield model it has been found that multiple storage of a pattern improves its recall (Fontanari and Köberle 1988a). Storing a pattern twice is the extreme case of storing two patterns that have higher than random overlaps with each other. The latter situation has been studied in both the random diluted (Derrida *et al* 1987) and fully connected (Fontanari and Köberle 1988b) Hopfield model. In both cases it gives rise to three distinct phases: at low α there is a separate memory for each pattern; as α is increased there is a single memory equally correlated with both patterns and at higher α there is no memory. In Derrida *et al* (1987) the range of α values for the undistinguishing memory includes values that are greater than α_c for the recall of the random patterns. This also applies at the higher values of the correlation between the two patterns in Fontanari and Köberle (1988b). In this section we shall see that this improvement in recall is *not* exhibited in the model studied in the present paper.

We define an overlap parameter between two patterns

$$Q=\frac{1}{Np}\sum_{i=1}^{N}\eta_i^1\eta_i^2 \tag{36}$$

and use order parameters as before with the pattern index reinstated:

$$m_1^{(1)}=\frac{1}{Np}\sum_{i=1}^{N}\eta_i^1 V_i$$

$$m_1^{(2)}=\frac{1}{Np}\sum_{i=1}^{N}\eta_i^2 V_i$$

$$m_2^{(1)}=\frac{1}{Np}\sum_{i=1}^{N}(1-\eta_i^1)V_i.$$

We let patterns 1 and 2 have overlap $Q\sim O(1)$ and all other pattern pairs have overlap $Q\sim O(p)$. In deriving evolution equations for the three order parameters required we use the same techniques outlined in the previous sections and so here we will just write down the zero temperature result:

$$m_1^{(1)}(t+\Delta t)=\frac{Q}{2}\left[1+\operatorname{erf}\left(\frac{m_1^{(1)}(t)+m_1^{(2)}(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right]+\frac{1-Q}{2}\left[1+\operatorname{erf}\left(\frac{m_1^{(1)}(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right]$$

$$m_1^{(2)}(t+\Delta t)=\frac{Q}{2}\left[1+\operatorname{erf}\left(\frac{m_1^{(1)}(t)+m_1^{(2)}(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right]+\frac{1-Q}{2}\left[1+\operatorname{erf}\left(\frac{m_1^{(2)}(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right] \tag{37}$$

$$m_2^{(1)}(t+\Delta t)=\frac{1}{2p}\left[1-\operatorname{erf}\left(\frac{\varepsilon_0}{(2\alpha pr(t))^{1/2}}\right)\right]+\frac{1-Q}{2}\left[1+\operatorname{erf}\left(\frac{m_1^{(2)}(t)-\theta_0}{(2\alpha pr(t))^{1/2}}\right)\right].$$

These equations exhibit the three fixed-point regimes mentioned above. However, to construct a phase diagram we must pick a criterion for the phase boundaries. In figure 7 the criterion used is that starting from one of the patterns the fixed point to which the map (38) iterates classifies the phase to which the α, Q values belong. This is equivalent to using an α_1 definition for the critical value of α. Figure 7 shows that only at comparatively high Q values is the undistinguishing memory phase present. In fact there is only a small range of Q values where all three phases are present. Except at very low Q values the storage capacity is *reduced* compared with § 2.

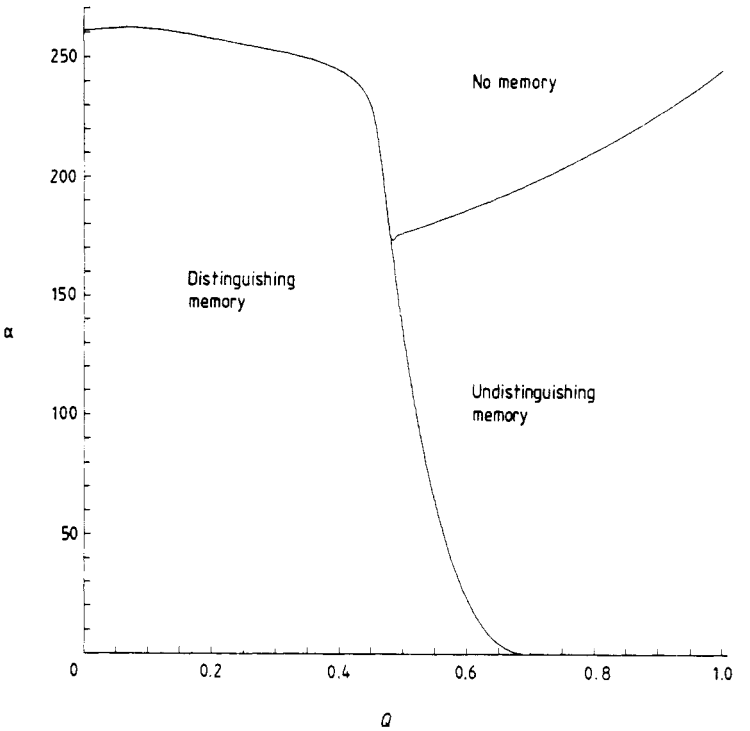

Figure 7. A phase diagram representing the three phases discussed in the text. The thresholding is $\theta_0 = 0.699$.

These observations reflect the importance of the thresholding parameter and activity of a configuration. A memory that does not distinguish between the patterns ($m_1 = Q$) will have too low an activity to be a fixed point at low Q. Although a Q overlap between patterns will strengthen the connections between shared 1s, this will lead to higher activities in trajectories near to the patterns. This may destabilise the memories that distinguish between the patterns at lower α than before. Consider the case of storing a pattern twice ($Q = 1$). Equations (38) reduce to (20) if

$$\theta_0 = 2\theta_0'$$

$$\alpha = 4\alpha'$$

where the primed values are those for storing the pattern once. Hence an increase in the maximum α for recall of the pattern is possible only if θ_0 is suitably increased.

5. Discussion

We have shown that the method of random dilution gives considerable insight into the flows and therefore recall properties of a V-model network. Several key features of V-model networks have emerged in this work. The mathematical structure of the memory fixed point changes with the thresholding. The importance of the thresholding lies in its control of the activity of a configuration, which is not constant in flows due to the distinct active and passive roles of the 1s and 0s. It might be fruitful to consider more general thresholdings by introducing site or time dependence. The first scheme

might allow the storing of patterns with higher than random overlaps to improve performance. The second scheme might be able to exploit the dependence of basin of attraction on thresholding to allow wayward trajectories to be captured by the memories.

Several surprising features of the model have also been illustrated. Firstly the fixed-point equations of the map (20) are equivalent to the mean-field equations of the fully connected model, implying that random dilution has not changed the features of the model to a great extent. Tsodyks (1988) also noted this 'universality' for the continuous-time model. Secondly in considering the storage capacity there is a need for careful definition of the critical value of α, due to the possibilities of intermittancy and the pattern being outwith the basin of attraction of the memory. Finally the breakdown of a memory does not occur through its basin of attraction vanishing continuously; instead we see more complicated mechanisms. These lead to the possibility of the memory taking the form of a limit cycle. It would be interesting to use the methods of bifurcation theory to analyse the flow (22) to give more information on this topic.

It is worthwhile reiterating a comment of Buhmann *et al* (1988) that fixed points uncorrelated with the patterns are clearly distinguishable from memories due to the differing levels of activity. Thus one can tell, without knowing the patterns, whether an association has been made between an input and a stored pattern.

Acknowledgments

The author wishes to thank Alastair Bruce for helpful discussions and acknowledges invaluable guidance from the late Elizabeth Gardner. Financial support was from the SERC.

References

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev. Lett.* **55** 1530
—— 1985b *Phys. Rev.* A **32** 1007
—— 1987 *Phys. Rev.* A **35** 2293
Bruce A D, Gardner E and Wallace D J 1987 *J. Phys. A: Math. Gen.* **20** 2909
Buhmann J, Divko R and Schulten K 1988 *Preprint* Technische Universität München
Coolen A C C and Ruijgrok Th W 1988 *Phys. Rev.* A **38** 4253
Derrida B and Pomeau Y 1986 *Europhys. Lett.* 1 45
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Fontanari J F and Köberle R 1988a *J. Phys. A: Math. Gen.* **21** 2477
—— 1988b *J. Phys. A: Math. Gen.* **21** 1253
Forrest B F 1988 *J. Phys. A: Math. Gen.* **21** 245
Gardner E 1987 *Europhys. Lett.* **4** 481
Gardner E, Derrida B and Mottishaw P 1987 *J. Physique* **48** 741
Hebb D O 1949 *The Organisation of Behaviour* (New York: Wiley)
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2544
Horner H 1988 *Z. Phys.* B to be published
Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Krauth W, Nadal J-P and Mézard M 1988a *J. Phys. A: Math. Gen.* **21** 2995
—— 1988b *Complex Systems* **2** 387
Kree R and Zippelius A 1987 *Phys. Rev.* A **36** 4421
Little W A 1974 *Math. Biosci.* **19** 101

McCulloch W S and Pitts W A 1943 *Bull. Math. Biophys.* **5** 115
Tsodyks M V 1988 *Europhys. Lett.* **7** 203
Tsodyks M V and Feigel'man M V 1988 *Europhys. Lett.* **6** 101
Willshaw D J and Longuet-Higgins H C 1970 *Machine Intelligence* (Edinburgh: Edinburgh University Press) ch 5, p 351
Wilson H R and Cowan J D 1972 *Biophys. J.* **12** 1