Metastable states in asymmetrically diluted Hopfield networks

J. Phys. A: Math. Gen. 21 (1988) 3155-3169. Printed in the UK

Metastable states in asymmetrically diluted Hopfield networks

A Treves[†] and Daniel J Amit[†]

Institute for Advanced Studies, Jerusalem, Israel

Received 1 February 1988, in final form 28 March 1988

Abstract. A count of the number of metastable states is employed to obtain indications on the retrieval and spin-glass properties of asymmetrically diluted neural networks. It is found that the main effect is on the retrieval states. Their position, distribution and number depend essentially on the normalised storage parameter $\tilde{\alpha}$, the ratio of the number of memories to the mean connectivity. The effect of asymmetrical dilution on metastable states uncorrelated with the memories depends on the dilution mode; the number of such states, however, still grows exponentially with system size, even for completely asymmetrical networks. To the extent that asymmetry destabilises this spin-glass phase it must be doing it by modifying the dynamics and not by eliminating metastable states.

It is also shown that there are no individual retrieval states with significant basins of attractions, for the symmetric as well as the asymmetric neural network.

1. Introduction

It was the imposition of symmetry on the synaptic connections (coupling constants) which led to a great clarification of the properties of neural networks (Hopfield 1982, 1984, Amit *et al* 1985a, b, 1987). But once an initial clarity was obtained, attention turned to the effects of asymmetry. Three different pressures have acted in this direction: biological plausibility, questions about the robustness of the results, given that no basic principle enforces symmetry, and a possible cognitive role for asymmetry. Parisi (1986) has argued that while asymmetry destabilises spin-glass states it opens up chaotic trajectories for the system, and those can be used to account for learning. In fact, trying to model learning within the standard Hopfield model, one faces the problem that near every input there is a spin-glass attractor. There is no way of distinguishing between an attractor that is a memory and one that is a spurious 'confused' state. If asymmetry can turn fixed points corresponding to confused states into chaotic trajectories, such that the correlated activity of pairs of neurons averages to zero, then the problem is solved without *ad hoc* mechanisms. Stimuli which are not within the basin of attraction of a memory will be ignored unless they are persistently imposed in the input. Following a different idea, Shinomoto (1987) has implemented the biological observation that each neuron has in most cases a unique function, either excitatory or inhibitory. This introduces an asymmetry and as a consequence provides a way of singling out memorised patterns from unmemorised ones.

To relax the symmetry constraint on the synaptic strengths J_{ij} is, however, a rather difficult problem, because one cannot define an energy function

$$H = -\tfrac{1}{2} \sum_{i \neq j} J_{ij} S_i S_j$$

[†] On leave from Racah Institute of Physics, Hebrew University, Jerusalem, Israel.

and one must resort to a full dynamical theory to describe the model. Attempts in this direction have focused on asymmetrically diluted models, where one starts from a symmetrical fully connected network of N neurons and introduces asymmetry by cutting off synapses until the connectivity of the network is lowered to a given value C. Asymmetry appears to act as an additional source of noise, weakening the spin-glass phase and improving retrieval (Hertz *et al* 1987, Feigelman and Ioffe 1986, 1987), as had been foreshadowed by Hopfield (1982).

In a remarkable contribution, Derrida *et al* (1987) have shown that the dynamics of an asymmetrically diluted network is exactly soluble in the limit of extreme dilution $C \ll \ln N$. The result is that, for a suitable choice of the parameters, the essential features of associative memory of the fully connected Hopfield model are preserved, and perhaps even improved, under extreme dilution. While this is an impressive statement, the DGZ theory cannot possibly be considered a closer approximation to biological reality. Keeping in mind the anatomical figure for the connectivity ($\approx 10^4$), if the number of neurons has to be bigger than the exponential of the connectivity, it will be a super-astronomical number.

It is therefore desirable to develop as many tools as possible which can provide insight in the intermediate regime, when the connectivity is high but not full and asymmetry is introduced in the process of dilution. One such approach is to try to count the average number of metastable states, i.e. states stable to all single spin flips, as a function of their overlap with the memorised patterns. Such a calculation was carried out for the standard (fully symmetric) Hopfield model (Gardner 1986), extending the counting of metastable states in the SK spin glass (Bray and Moore 1980, 1981, Tanaka and Edwards 1980, De Dominicis *et al* 1980). It was found that the structure of metastable states suggested by the calculation closely corresponds to the picture derived in the thermodynamic studies. At low loading levels, when the number of stored patterns p remains finite as $N \to \infty$, such counting reproduces *exactly* the number of metastable states predicted by the mean-field theory (Amit *et al* 1986). Near saturation, when $p/N = \alpha$ as $N \to \infty$, the correspondence is less direct. Yet the counting of states gives the main features of the thermodynamical analysis, and even goes beyond it in its sensitivity to metastable states with low barriers.

In fact, in the standard model, counting produces the following picture. Metastable states appear only in two distinct regions of phase space: in a wide band that extends continuously all the way from the states which are very weakly correlated with the stored patterns, and in a narrow band, disjoint from the first only below a certain critical value of α, where the metastable states are strongly correlated with a single stored pattern. This second group includes the retrieval states exposed by the thermodynamic analysis, and as $\alpha \to 0$ the mean correlation of states in this group with the stored pattern approaches unity exactly in the same way as the correlation of the retrieval states. The critical value of α, above which the gap between the two bands disappears, turns out to be 0.113, to be compared with $\alpha_c = 0.138$ (or 0.145) (Amit *et al* (1987); see also the theory with one replica symmetry breaking by Crisanti *et al* (1986)). It has been argued (Gardner 1986) that this critical value of α is not supposed to coincide with α_c, the storage capacity of the network, because the appearance of this cluster of strongly correlated metastable states is not immediately related to the thermodynamic behaviour. However, the two values *are* comparable. One concludes that the counting approach yields a rough approximation of the storage capacity, which is not drastically worse than the approximation involved in, say, assuming replica symmetry. Thus, one can extract from such calculations information about the retrieval properties of the system, which agrees qualitatively with that contained in the full solution of the model. This kind of approach is certainly less detailed and less transparent—for example, it does not give information on features like the stability to flipping clusters of spins, or the height of barriers between stable states. It has, however, the advantage that it does not require a symmetrical J_{ij}. This is just the situation we would like to apply it to.

2. The model

As in the studies mentioned above we introduce asymmetry by combining it with dilution. Specifically, we consider diluted 'Hebbian' synapses of the form $T_{ij} = J_{ij} w_{ij}$, where J_{ij} is the usual symmetric matrix

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$$

representing the storage of p random uncorrelated patterns, and w_{ij} equals 0 or 1 according to a given probability distribution. In general, w_{ij} may differ from w_{ji}, leading to an asymmetry of the synapses. If N is the number of neurons

$$C \equiv N \langle w_{ij} \rangle$$

is the mean resulting connectivity. Let the dilution parameter be

$$\gamma \equiv C/N \equiv \langle w_{ij} \rangle .$$

While previous dynamical studies have considered the cases $\gamma \to 0$ (Derrida *et al* 1987, Kree and Zippelius 1987) and $\gamma = \frac{1}{2}$ (Hertz *et al* 1987, Feigelman and Ioffe 1986, 1987) as $N \to \infty$, we shall allow γ to take arbitrary values.

One can consider any probability distribution for the w_{ij}. We shall focus on three particularly meaningful examples. The first:

$$P(w_{ij}, w_{ji}) = \gamma \delta(w_{ij} - 1) \delta(w_{ji} - 1) + (1 - \gamma) \delta(w_{ij}) \delta(w_{ji})$$

is a symmetric dilution (SD) considered by Sompolinsky (1986) and will serve for comparison. Next is a random dilution (RD), used by Derrida *et al* (1987) and Hertz *et al* (1987):

$$P(w_{ij}, w_{ji}) = P(w_{ij}) P(w_{ji})$$

with

$$P(w) = \gamma \delta(w - 1) + (1 - \gamma) \delta(w)$$

so that w_{ij} and w_{ji} are independent. The third one is an asymmetric dilution (AD)

$$P(w_{ij}, w_{ji}) = \begin{cases} (2\gamma - 1)\delta(w_{ij} - 1)\delta(w_{ji} - 1) + (1 - \gamma)\delta(w_{ij} - 1)\delta(w_{ji}) & \\ \quad + (1 - \gamma)\delta(w_{ij})\delta(w_{ji} - 1) & \gamma \geqslant \frac{1}{2} \\ \gamma\delta(w_{ij} - 1)\delta(w_{ji}) + \gamma\delta(w_{ij})\delta(w_{ji} - 1) & \\ \quad + (1 - 2\gamma)\delta(w_{ij})\delta(w_{ji}) & \gamma \leqslant \frac{1}{2}. \end{cases}$$

The case $\gamma = \frac{1}{2}$ of this dilution has been treated by Feigelman and Ioffe (1986, 1987) and by Kinzel (1987). In all three cases the mean connectivity is C.

Let us also define a parameter measuring the mean symmetry of the connections:

$$\lambda \equiv \langle T_{ij} T_{ji} \rangle / \langle T_{ij}^2 \rangle$$

and for our three cases we have

$$\lambda_{\mathrm{SD}}=1 \qquad \lambda_{\mathrm{RD}}=\gamma \qquad \lambda_{\mathrm{AD}}=\begin{cases}2-1/\gamma & \gamma\geqslant\frac{1}{2}\\ 0 & \gamma\leqslant\frac{1}{2}.\end{cases}$$

In fact, the calculations presented here can be carried out for an arbitrary distribution of the w_{ij}, and the result would depend on the corresponding parameters λ and γ only. Moreover, once one has chosen a given mode of dilution, λ is a function of γ, so in the following we shall omit λ as an explicit argument in functions that depend on both parameters.

3. The calculation

We denote by N_s the number of configurations stable to all single spin flips. In the absence of fast noise each spin in a stable network state is aligned with its local field, namely for $i=1,\ldots,N$

$$S_i=\operatorname{sgn}\left(\sum_j T_{ij}S_j\right).$$

The objective is to estimate N_s as a function of α, γ and the overlap

$$m=\frac{1}{N}\sum_i S_i\xi_i^{\mu_0}$$

between the state $\{S_i\}$ and one of the p stored patterns, $\{\xi_i^{\mu_0}\}$. In the $N\to\infty$ limit, for fixed T_{ij} this number behaves as $\exp[Nf(\alpha,m,\gamma)]$. What one would therefore have to compute is the quenched average over the T_{ij} of the function f. This implies introducing replicas and complicates the calculation considerably. So we will follow Gardner (1986) and limit ourselves to a computation of the quantity

$$\exp[NF(\alpha,m,\gamma)]\equiv\langle\exp[Nf(\alpha,m,\gamma)]\rangle_{T_{ij}}=\langle N_s\rangle_{T_{ij}}.$$

This quantity gives an upper bound for $\exp(N\langle f(\alpha,m,\gamma)\rangle_{T_{ij}})$, since the exponential function is convex. We shall see that even with this limitation we can extract interesting features.

Stability to single spin flips means that all spins are aligned with their local fields, i.e.

$$h_i\equiv\sum_{j\neq i}\left(\frac{1}{N}\sum_\mu\xi_i^\mu\xi_j^\mu\right)w_{ij}S_j=\lambda_iS_i \qquad \lambda_i\geqslant 0 \tag{1}$$

for all S_i. Then

$$\langle N_s\rangle_{T_{ij}}=\mathrm{Tr}_S\int_0^\infty\prod_i \mathrm{d}\lambda_i\left\langle\delta\left(\lambda_iS_i-\frac{1}{N}\sum_{i\neq j}\sum_\mu\xi_i^\mu\xi_j^\mu w_{ij}S_j\right)\right\rangle_{T_{ij}}. \tag{2}$$

The Tr_S is a sum over all 2^N possible configurations of the network and every configuration that satisfies (1) contributes 1 to this sum. For a given realisation of the random variables $\{\xi_i^\mu, w_{ij}\}$ the right-hand side gives, therefore, the total number of states satisfying (1) at every site.

In the thermodynamic limit one obtains (see the appendix for details)

$$\langle N_s\rangle=\left(\frac{N}{2\pi\det H}\right)^{1/2}\exp[NF(\alpha,m,\gamma)] \tag{3}$$

where $F(\alpha, m, \gamma)$ is the value of the function

$$F=\alpha\left(\frac{1}{2}\frac{(2-\lambda/\gamma+s)^2}{[2-(1+\lambda)/\gamma+r]}-s-1+\frac{\lambda}{2\gamma}+\tfrac{1}{2}\ln(1+r-1/\gamma)\right)$$

$$+\frac{1-m}{2}\ln\frac{2\phi(T^-)}{1-m}+\frac{1+m}{2}\ln\frac{2\phi(T^+)}{1+m} \qquad (4)$$

with

$$\phi(T)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{T}\mathrm{e}^{-t^2/2}\,\mathrm{d}t \qquad T^{\pm}=\frac{s\alpha\bullet m}{\sqrt{r\alpha}}$$

at the saddle point over the parameters r, s, and det H is the determinant of the second derivatives of F with respect to all saddle-point parameters. This is the expression for all three modes of dilution. They are distinguished by the different values of the symmetry parameter λ introduced above, as a function of the dilution parameter γ.

It is also possible to include in the calculation only those metastable states for which the magnitude of the local field is at all sites above a threshold: $S_i h_i \geqslant h_0$, or $\lambda_i \geqslant h_0$. This may be a way of selecting only states with sizable basins of attraction (Krauth and Mézard 1987, Gardner and Derrida 1988). The function F remains the same, and only the upper limits in the integrals of ϕ change, correspondingly, to $T^{\pm}=(s\alpha-h_0/\gamma\bullet m)/\sqrt{r\alpha}$. We shall examine the effect of non-zero threshold in a later section, while here we continue with $h_0=0$ (i.e. equation (1) as it stands).

In general the saddle-point equations in r and s have to be solved numerically. One has to keep in mind that $\langle N_s\rangle$ is an upper bound for exp $N\langle f\rangle$. If, for some values of α, m and γ, $F(\alpha, m, \gamma)<0$, this implies that there are typically no states stable to single spin flips, and hence no stable states at all, with those values of the parameters.

4. The limit $\alpha\to\infty$

The limit $\alpha\to\infty$ has been identified as the spin-glass limit in the symmetric fully connected models (Amit *et al* 1987). In this limit one can solve the equations analytically to find

$$F(\infty, m, \gamma)=-\frac{T^2}{2\lambda}+\ln 2\phi(T)-\tfrac{1}{2}\ln(1-m^2)-m\tanh^{-1}m\ldots \qquad (5)$$

where T is given by the saddle-point equation

$$T=\lambda\frac{\phi'(T)}{\phi(T)}.$$

This corresponds exactly to what one finds for asymmetrical SK spin glasses (Crisanti and Sompolinsky 1987). (For an SK spin glass the same calculation can be performed introducing the asymmetry either with dilution as described above or with couplings of the form $J_{ij}=J_{ij}^{\mathrm{S}}+J_{ij}^{\mathrm{A}}$, where the symmetrical and antisymmetrical parts have variances $J_0\cos\theta$ and $J_0\sin\theta$, respectively. In that case $\lambda=\cos 2\theta$.)

We see that for $\alpha\to\infty$ the dependence on m is strictly via the phase space factor, coming from the binomial distribution. In other words, when there are too many patterns, no individual one is meaningful anymore, and the overlap enters only as the magnetisation along an arbitrary diagonal of the N-dimensional hypercube. F is maximal for $m = 0$ and decreases with m, becoming negative at a value $m_0(\lambda)$. The value $F_{\max}$ at $m = 0$ corresponds to taking the saddle point also in m, and thus gives the total number of metastable states. Moreover, dilution as such has no effect and it is only asymmetry that affects the number of metastable states. As a function of the mean symmetry λ, F is a monotonically increasing function. $F_{\max}(\lambda = 1) = 0.1992$, as for the SK model (Tanaka and Edwards 1980). For $\lambda \to 0$, $F_{\max} \approx \lambda/\pi$. In particular, for a fully asymmetric network $\lambda = 0$, and $\langle N_s \rangle = 1$ (one can check that the prefactor of equation (3) yields 1 in this case). This is an average value, so for some realisations of the couplings the actual value will be higher than one, and for some zero (Crisanti and Sompolinsky 1987). We now turn to study finite values of α.

5. The critical value $\tilde{\alpha}^*$

A diluted network storing p patterns stores pN bits of information in CN synaptic strengths. It is therefore meaningful (Derrida *et al* 1987) to replace $\alpha \equiv p/N$ of the fully connected network with

$$\tilde{\alpha} \equiv p/C = \alpha/\gamma.$$

Intuitively, one expects the retrieval ability of the network to depend primarily on $\tilde{\alpha}$. In fact, this is what we find analysing the behaviour of F. The parameter α, instead, is still relevant for the description of the effects of the slow noise arising from random correlations between the stored patterns.

For small $\tilde{\alpha}$, $F(\tilde{\alpha}, m, \gamma)$ is positive in two distinct regions of the interval $0 \leqslant m \leqslant 1$, separated by a gap where $F < 0$: a narrow band close to $m = 1$, and a much broader region at $m = 0$, comprising the global maximum of F (see figure 1). This behaviour is the same as for the symmetric fully connected case (Gardner 1986), but here the

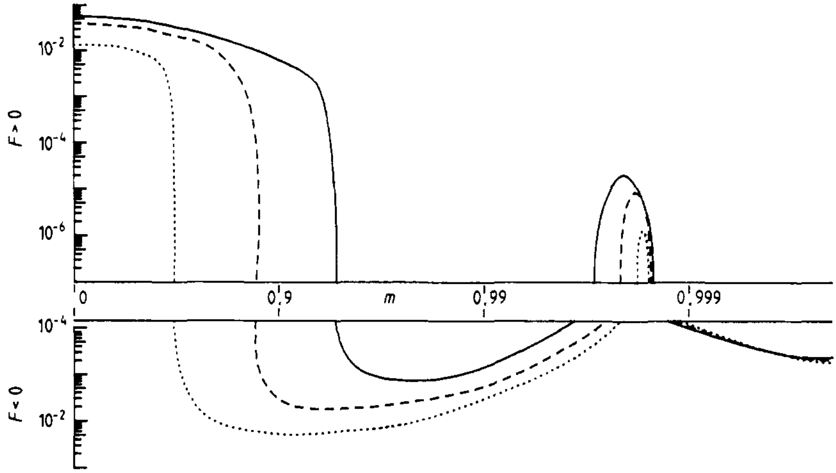

Figure 1. The behaviour of F with the overlap m for $\tilde{\alpha} = 0.1$ and RD. The curves are for different values of the parameter $k \equiv (1-\gamma)/\gamma$: full curve, $k = 0$ ($\alpha = 0.1$); broken curve, $k = 1$ ($\alpha = 0.05$); dotted curve, $k = 10$ ($\alpha = 0.009$). Increasing dilution does not change significantly the overlap of retrieval states (for $\tilde{\alpha} \to 0$ it does not change it at all, equation (6a)), while $F(m = 0)$ decreases according to equation (7).

parameter is $\tilde{\alpha}$. Thus, our bound allows for the existence of two groups of metastable states: one strongly correlated with an input pattern, comprising the retrieval states, and one centred around $m = 0$, which are all spurious metastable states, unrelated to retrieval.

For $\tilde{\alpha} > \tilde{\alpha}^*(\gamma)$ the gap disappears and the bound is not strong enough to prove whether the two groups remain distinct or coalesce. The fact that $\tilde{\alpha}^*(1)$ is close to α_c, the storage capacity found in the thermodynamic solution of the fully connected model (Amit *et al* 1986), suggests that, at least for values of γ close to 1, $\tilde{\alpha}^*(\gamma)$ gives a rough approximation of the storage capacity of asymmetric networks. Numerically one finds that $\tilde{\alpha}^*(\gamma)$ is a very mildly varying function of γ in the whole interval $0 < \gamma \leq 1$ (figure 2).

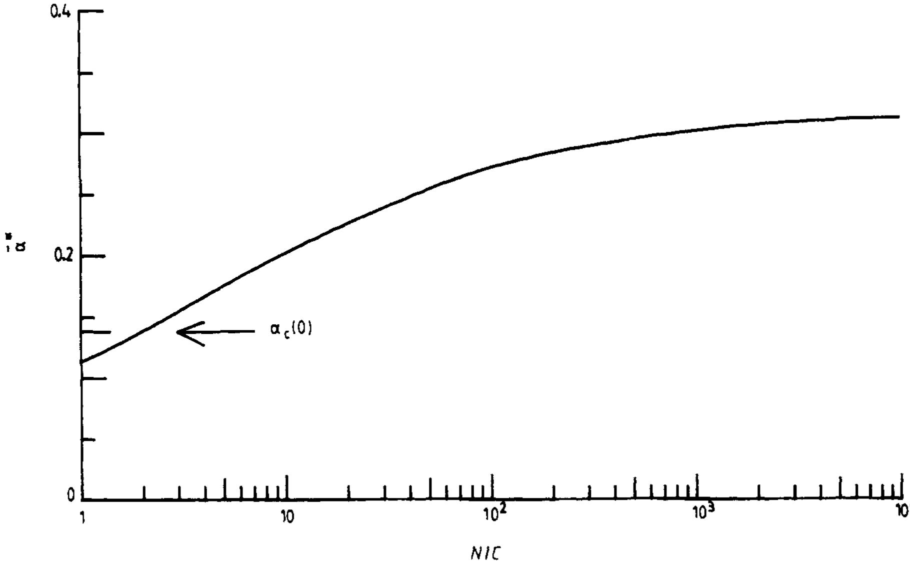

Figure 2. The critical value $\tilde{\alpha}$ as a function of the dilution, for the case of RD. The arrow shows the maximal capacity of the fully connected model.

Comparing $\tilde{\alpha}^*(\gamma)$ with $\alpha_c(\gamma)$ for the symmetrically diluted model (Sompolinsky 1986), one finds that the behaviour as a function of γ is very similar. Kinzel (1987) has determined $\tilde{\alpha}_c(\frac{1}{2})$ numerically for the AD model ($\lambda = 0$). His result, $\tilde{\alpha}_c(\frac{1}{2}) = 0.15$, agrees with a mild increase of $\tilde{\alpha}_c$ with dilution. Note that for the RD model at $\gamma \to 0$ (in fact, in the limit $C \ll \ln N$) one has (Derrida *et al* 1987) $\tilde{\alpha}_c = 2/\pi$. But here the transition is continuous and overlaps become very small as saturation is approached.

6. The retrieval states

One can expand $F(\tilde{\alpha}, m, \gamma)$ for $\tilde{\alpha}$ small and m close to 1 to find that F is positive in a narrow interval around m_0 where the Hamming distance per spin d_0 is

$$d_0 \equiv \tfrac{1}{2}(1 - m_0) \approx \left(\frac{\tilde{\alpha}}{2\pi}\right)^{1/2} \exp\left(\frac{-1}{2\tilde{\alpha}}\right). \qquad (6a)$$

The width of this little peak is

$$\Delta m \approx \lambda^{1/2} \frac{8}{\sqrt{\tilde{\alpha}}} d_0^{3/2} \tag{6b}$$

and the height of the peak is

$$F_0(\tilde{\alpha}, m_0, \gamma) \approx \lambda \frac{2}{\tilde{\alpha}} d_0^2. \tag{6c}$$

For $\lambda = 0$, the leading term in the expansion of Δm and F_0 goes to zero, so one has to consider the next order term, and this yields

$$\Delta m \approx \gamma^{1/2} \frac{16}{\sqrt{3\tilde{\alpha}}} d_0^2 \qquad \lambda = 0 \tag{6d}$$

$$F_0(\tilde{\alpha}, m_0, \gamma) \approx \gamma \frac{8}{3\tilde{\alpha}^2} d_0^3 \qquad \lambda = 0. \tag{6e}$$

If there are fixed points of the dynamics close to a stored pattern, they must be inside this band. Equation (6*a*) implies that their distance from the stored pattern is a function of $\tilde{\alpha}$ only, and not of the amount or type of dilution. Indeed, one has for all modes of dilution the same result as found for the specific case of symmetric dilution by Sompolinsky (1987). The equation, however, gives only the leading term, so this is exact in the $\tilde{\alpha} \to 0$ limit, and approximately true for $\tilde{\alpha}$ as high as 0.1, as shown in figure 1. Asymmetry, on the other hand, makes the width of the band shrink. Also the maximum of F goes down with asymmetry. Notice that it does not necessarily follow that if there are more attractors close to the same pattern asymmetry makes them coalesce: their mutual Hamming distance might be different from zero while their overlap with the pattern is the same.

For p finite as C, $N \to \infty$, $F \to 0$ and there is just a single stable state, coinciding at all sites with the stored pattern. As p increases with C, keeping $\tilde{\alpha}$ finite and fixed, many retrieval states appear around a given pattern. It is possible to obtain an intuitive understanding of the multiplicity of metastable states that arises at $\tilde{\alpha} \neq 0$ in simple physical terms. A small fraction of spins ($\Delta N/N$, of order $\exp(-1/2\tilde{\alpha})$) have local fields close to zero, and are therefore effectively decoupled from the rest of the spins that are frozen in their orientation by strong local fields. This system of weakly coupled spins can be viewed as a randomly diluted spin glass, which has an exponential number of metastable states,

$$\ln N_s \approx \Delta N \lambda$$

(cf equation (5)). Here $\lambda \approx \Delta N/N$, so one obtains

$$\ln N_s \approx N(\Delta N/N)^2 \approx N \exp(-1/\tilde{\alpha})$$

as given by equation (6*c*). The uncoupled spins do not contribute to the overlap with the stored pattern, which is therefore less than 1 by an amount of order $\exp(-1/2\tilde{\alpha})$, equation (6*a*). Also equation (6*b*) can be understood in similar terms, as the dispersion in m due to the random magnetisations of the spin-glass states.

Keeping in mind the value of λ for RD, equation (6*c*) implies that for this type of dilution the total number of strongly correlated metastable states, given by the saddle-point value F_0, scales as $\langle N_s^c \rangle \approx \exp(CG(\tilde{\alpha}))$, i.e. N is neatly substituted by C as the scaling factor in the exponent, with G a function of $\tilde{\alpha}$ alone. The size of the system is thus irrelevant, in this case, to both the number of retrieval states and their mean overlap with the stored patterns, and it only affects the fluctuations in this mean overlap, which go to zero as $N \to \infty$ (equation (6*b*), with $\lambda = C/N$).

7. The uncorrelated states

For all $\tilde{\alpha} \neq 0$, F has an absolute maximum at $m = 0$. The second derivative at the maximum goes as

$$\left.\frac{\mathrm{d}^2 F}{\mathrm{d}m^2}\right|_{m=0} \approx -\tfrac{1}{2}\alpha\pi \tag{7}$$

for α small; hence it is much smaller in absolute value than 1, indicating a distribution of metastable states much broader than the phase space distribution we have for $\alpha \to \infty$ (equation (5)). In other words, the bound gives configurations with macroscopic correlation with one of the stored patterns a better chance of being stable to single spin flips than random configurations. This is consistent with the fact that the remnant overlap in a Hopfield model with finite α is greater than in a spin glass, which itself is non-zero (Amit *et al* 1987, Kinzel 1985).

As for the height of the peak at $m = 0$, one can expand for α small

$$F(\alpha, 0, \gamma) \approx \tfrac{1}{2}\alpha\left(\ln\frac{2}{\alpha\pi} - 2 + \frac{\lambda}{\gamma}\right). \tag{8}$$

At fixed γ this value decreases with increasing asymmetry. However, it is constant with respect to dilution if one dilutes in the RD mode. AD, instead, lowers F, and SD enhances it, with decreasing γ. It is not clear whether these differences between different modes of dilution have immediate bearing upon the dynamical behaviour of the system in this uncorrelated portion of phase space. For $\alpha \neq 0$, $\exp[NF(\alpha, 0, \gamma)]$ is the value of $\langle N_s \rangle$ at the saddle point over the parameter m, and so it should also give the total number of metastable states, if for $m = 0$ fluctuations are negligible (Gardner 1986). For a fully asymmetric system equation (8) yields that this number is smaller by a factor $\exp(-\tfrac{1}{2}p)$ than for the symmetric fully connected one, in the small-α limit.

It is interesting to note at this point that, for all finite α, one has, even for fully asymmetric networks, $F(\alpha, 0, \gamma) > 0$, and only in the $\alpha \to \infty$ limit $F(\lambda = 0) \to 0$, recovering the corresponding result obtained for an SK spin glass (Crisanti and Sompolinsky 1987). In fact, in that limit

$$F(\alpha, 0, \gamma) \approx \frac{1}{\sqrt{\alpha}}\left(\frac{2}{\pi}\right)^{3/2} \gamma^{5/2}\,(1 - \tfrac{2}{3}\gamma^{1/2}) \qquad \lambda = 0.$$

Therefore $F(\lambda = 0)$ has a maximum for a finite value α_0; taking for example $\gamma = \tfrac{1}{2}$ one finds $\alpha_0 = 0.63$ and $F(\alpha_0, 0, \tfrac{1}{2}) = 0.0343$.

Since the uncorrelated states of the Hopfield model represent a spin-glass phase (Amit *et al* 1987), this might seem to contradict the finding that full asymmetry destroys, even at $T = 0$, the exponential growth in the number of stable states in a spin glass (Crisanti and Sompolinsky 1987). The apparent contradiction is readily explained, however, by observing that our definition of full asymmetry does not imply a total absence of correlations in the coupling matrix T_{ij}. In fact, higher-order correlations enter the game, and they do not vanish. For example, the third-order correlations

$$\langle T_{ij} T_{jk} T_{ki} \rangle = \gamma^3 \langle J_{ij} J_{jk} J_{ki} \rangle$$

are zero in an SK spin glass, due to the independence of the J_{ij}, while in the Hopfield model they are not zero and provide a feedback that tends to stabilise each spin in a frozen position. If the J_{ij} are given by the Hebb rule, one finds

$$\frac{\langle \Sigma_{j,k} J_{ij} J_{jk} J_{ki} \rangle}{(\langle \Sigma_j J_{ij}^2 \rangle)^{3/2}} = \frac{1}{\sqrt{\alpha}}$$

and, as $\alpha \rightarrow \infty$, the effects of third-order correlations (and, similarly, of higher-order ones) become irrelevant and one retrieves a pure SK spin glass.

8. Probing the robustness of metastable states

As mentioned above, one can choose to count only those metastable states where the local field is (in absolute value) above a threshold h_0: $h_i S_i \geqslant h_0$. These can be considered as the fixed points of a particular non-Hamiltonian dynamics where one includes in the local fields a self-coupling term with negative sign, $-h_0 S_i$. In general, however, one can simply focus on these states as a subset of metastable states, which are supposedly more robust to the destabilising effect of a finite temperature or an external noise, and are more likely to have sizable basins of attraction. To keep the discussion general, we shall refer to the maximum value of h_0 for which a state is still stable to all single spin flips as the stability parameter of such a state. If we restrict ourselves to these states, the factor F that gives their exponential growth is modified as a function of the threshold in the following ways.

For $\alpha \rightarrow \infty$, setting $h_0 = \tilde{h}_0 \sqrt{\alpha\gamma}$, we find

$$F(\infty, m, \gamma) = -\frac{(T+\tilde{h}_0)^2}{2\lambda} + \ln 2\phi(T) - \tfrac{1}{2}\ln(1-m^2) - m \tanh^{-1} m \tag{9}$$

where now T solves the saddle-point equation

$$T = \lambda \frac{\phi'(T)}{\phi(T)} - \tilde{h}_0.$$

The relevant stability parameters are of order $\lambda\sqrt{\alpha\gamma}$. The correspondence with the SK spin glass is achieved when the spin-glass interaction is normalised so that $\langle \Sigma_j J_{ij}^2 \rangle = \alpha\gamma$. $F_{\max}(\lambda = 1)$ decreases with increasing h_0 and becomes zero for $\tilde{h}_0 = 0.351$. This means that for all thresholds $h_0 < 0.351\sqrt{\alpha\gamma}$ there is still an exponentially large number of metastable states such that at every site the local field exceeds the threshold.

For $\alpha \rightarrow 0$, we find for the maximum corresponding to uncorrelated states

$$F(\alpha, 0, \gamma) \approx \tfrac{1}{2}\alpha \left(\ln \frac{2}{\alpha\pi} - 2 + \frac{\lambda}{\gamma} \right) - \frac{h_0}{\gamma} \tag{10}$$

which shows that the relevant stability parameters for these states are of order

$$h_0 \approx \gamma\alpha \ln \frac{1}{\alpha}.$$

Finally, for the retrieval states, one has, as $\tilde{\alpha} \rightarrow 0$,

$$F_0(\tilde{\alpha}, m_0, \gamma) \sim \frac{2}{\tilde{\alpha}} d_0(d_0\lambda - h_0/\gamma) \tag{11a}$$

with d_0 given by equation ($6a$). For $\lambda=0$ one has

$$F_0(\tilde{\alpha}, m_0, \gamma) \approx \frac{2}{\tilde{\alpha}} d_0 \left(\gamma \frac{4}{3\tilde{\alpha}} d_0^2 - h_0/\gamma \right) \qquad \lambda = 0. \qquad (11b)$$

From equation ($11a$) it follows that each retrieval fixed point has a very low stability parameter (h_0 of order $\exp(-1/2\tilde{\alpha})$), and it is thus expected to be easily destabilised by noise and to have a tiny basin of attraction. This does not imply, however, any retrieval difficulty. It is just an effect of the spin-glass multiplicity associated with a retrieval state at finite $\tilde{\alpha}$, as discussed above. Indeed, intuitive arguments similar to those used to account for the multiplicity of order $\ln N_s \approx N(\Delta N/N)^2$, combined with equation (9), lead to an estimate of the stability parameter of such spin-glass states of order $\Delta N/N \approx \exp(-1/2\tilde{\alpha})$, and to a multiplicity scaling with h_0 as predicted by equation (11).

Thus, very slight noise will destroy the spin-glass freezing of the uncoupled spins and cause much hopping around. This is consistent with the finding (Amit *et al* 1987) that the temperature of replica symmetry breaking is (for the fully connected network) of order $\exp(-1/2\tilde{\alpha})$. But all this will not affect the overlap with the stored pattern, which will remain fixed and large, as it is determined by the rest of the spins. In the same way, the basins of attraction of each individual metastable state can be tiny, but what is important macroscopically is the sum of all the basins of attraction corresponding to the same mean overlap m_0. As long as $\tilde{\alpha}$ is small, the distinction between a fixed point and a trajectory spanning the small subspace of phase space corresponding to the spins with weak local fields is irrelevant. As $\tilde{\alpha}$ grows so does the number of uncoupled spins, and eventually there is an abrupt opening up of the whole phase space at $\tilde{\alpha}_c$, and the network ceases to perform as a memory (Amit *et al* 1987). These considerations suggest the importance of looking at the relevant order parameter (m in this case) rather than at the stability of fixed points, when analysing a network subject to noise.

9. Conclusions

Estimating the number of states stable to single spin flips is an approach limited in its scope, whose results should be confirmed with other methods. Yet it has proven its plausibility in the standard Hopfield model, where a suggestive correspondence emerges with the features derived from the thermodynamic solution. We have applied the method to asymmetrically diluted Hopfield networks, for which a comprehensive treatment of the asymmetrical dilution with extensive connectivity is lacking. The results indicate that the retrieval properties persist, but features like the storage capacity, dispersion and retrieval quality now depend essentially on the value of $\tilde{\alpha}$, the ratio of the number of stored patterns to the connectivity. If we assume that the value $\tilde{\alpha}^*$, where the band of retrieval states merges into the wider band of uncorrelated metastable states, gives a fair approximation of the critical $\tilde{\alpha}_c$ up to which the model has retrieval properties, then the storage capacity does not change more than by a factor of three or four going from a fully connected to an extremely diluted system. The overlap of the retrieval fixed points with the stored pattern they are close to depends on $\tilde{\alpha}$ alone, while asymmetry might affect the structure of nearby stable states. This structure appears to correspond to a spin glass associated to spins with low local field. The typical number of uncorrelated states, on the other hand, depends essentially on α

and is otherwise a constant in randomly diluted models, while it decreases with increasing asymmetry if one forces an asymmetrical dilution. A fully asymmetric model, however, has still an exponential number of metastable uncorrelated states, due to the effect of higher-order correlations in the coupling matrix. The lesson is that, to the extent that asymmetry weakens the uncorrelated spin-glass behaviour, it does not do it by eliminating many metastable states, but rather by deforming the dynamics (see, e.g., Crisanti and Sompolinsky 1987). While confirming these predictions analytically requires rather cumbersome replica calculations, the essential features can be tested with the results of computer simulations.

Acknowledgments

This work is the outgrowth of a discussion with Dr G Grinstein at the IBM Europe Institute in Oberlech 1986. We have also benefited from discussions with Professor H Sompolinsky. This work was supported in part by a grant from the US-Israel Binational Science Foundation.

Appendix

We show how to obtain the expressions (3) and (4) for $\langle N_s \rangle$, starting from equation (2). Using the integral representation for the δ functions one has

$$\langle N_s \rangle = \mathrm{Tr}_S \int_0^\infty \prod_i \mathrm{d}\lambda_i \int_{-\infty}^{\infty} \prod_i \frac{\mathrm{d}\phi_i}{2\pi} \exp\Big(-\mathrm{i} \sum_i \phi_i S_i \lambda_i\Big) \times \Big\langle \exp\Big(\frac{\mathrm{i}}{N} \sum_{i \neq j} \sum_\mu \phi_i \xi_i^\mu \xi_j^\mu S_j w_{ij}\Big)\Big\rangle . \tag{A1}$$

We first carry out the quenched average over the w_{ij}. This is a product of terms

$$\prod_{i>j} b_{ij} \equiv \prod_{i>j} \langle \exp(a_{ij} w_{ij} + a_{ji} w_{ji}) \rangle$$

where

$$a_{ij} = \frac{\mathrm{i}}{N} \sum_\mu \phi_i \xi_i^\mu \xi_j^\mu S_i .$$

One finds

$$\ln b_{ij} = (a_{ij} + a_{ji})\gamma + \tfrac{1}{2}(a_{ij}^2 + a_{ji}^2)\gamma^2 k + a_{ij} a_{ji} \gamma^2 h + \ldots$$

where we have set

$$k = (1-\gamma)/\gamma$$

and the difference between the three modes of dilution is in the factor h:

$$h_{\mathrm{SD}} = k \qquad h_{\mathrm{RD}} = 0 \qquad h_{\mathrm{AD}} = \begin{cases} -k^2 & \gamma \geq \frac{1}{2} \\ -1 & \gamma \leq \frac{1}{2} . \end{cases}$$

The neglected terms, of higher order in a_{ij}, can be shown to give a vanishing contribution as $N \to \infty$.

Performing in the integrals (A1) the following transformation of variables:

$$\phi_i S_i \gamma \to \phi_i \qquad \lambda_i \to \lambda_i \gamma$$

one writes

$$\begin{aligned}\ln b_{ij} &= \mathrm{i}\frac{(\phi_i+\phi_j)}{N} S_i S_j \sum_{\mu} \xi_i^{\mu}\xi_j^{\mu} - \frac{k}{2}\frac{(\phi_i^2+\phi_j^2)}{N^2}\sum_{\mu,\nu} \xi_i^{\mu}\xi_j^{\mu}\xi_i^{\nu}\xi_j^{\nu} - h\frac{\phi_i\phi_j}{N^2}\sum_{\mu,\nu} \xi_i^{\mu}\xi_j^{\mu}\xi_i^{\nu}\xi_j^{\nu} \\ &= \mathrm{i}\frac{(\phi_i+\phi_j)}{N} S_i S_j \sum_{\mu} \xi_i^{\mu}\xi_j^{\mu} - \frac{k}{2}\frac{(\phi_i^2+\phi_j^2)}{N}\alpha - h\frac{\phi_i\phi_j}{N}\alpha + \ldots\end{aligned}$$

where only terms with $\mu = \nu$ have been kept. The terms with $\mu \neq \nu$ can also be neglected in the thermodynamic limit.

Then

$$\begin{aligned}\langle N_s\rangle &= \mathrm{Tr}_S \int_{-\infty}^{\infty}\prod_i \frac{\mathrm{d}\phi_i}{2}\int_0^{\infty}\prod_i\frac{\mathrm{d}\lambda_i}{\pi}\exp(-\mathrm{i}\Sigma_i\,\phi_i\lambda_i)\prod_{i>j}\exp\left(-\frac{k}{2}\frac{\phi_i^2+\phi_j^2}{N}\alpha - h\frac{\phi_i\phi_j}{N}\alpha\right) \\ &\quad\times\left\langle\prod_{i>j}\exp\left(\mathrm{i}\frac{(\phi_i+\phi_j)}{N}S_iS_j\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}\right)\right\rangle \\ &= 2^{-N}\,\mathrm{Tr}_S\int_{-\infty}^{\infty}\prod_i \mathrm{d}\phi_i D(\phi_i)\exp\left[-\alpha\left(\mathrm{i}\sum_i\phi_i + \frac{k}{2}\sum_i\phi_i^2 - \frac{h}{2}\sum_{i,j}\frac{\phi_i\phi_j}{N}\right)\right] \\ &\quad\times\left\langle\prod_{i,j}\exp\left(\mathrm{i}\phi_iS_iS_j\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}\right)\right\rangle \qquad \text{(A2)}\end{aligned}$$

where

$$D(\phi_i) = \int_0^{\infty}\frac{\mathrm{d}\lambda_i}{\pi}\exp(-\mathrm{i}\phi_i\lambda_i).$$

To perform the average over the ξ_i^{μ} we use a Gaussian transform

$$\begin{aligned}&\prod_{i,j}\exp\left(\mathrm{i}\phi_iS_iS_j\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}\right) \\ &\quad = N^p\int\prod_{\mu}\frac{\mathrm{d}m_{\mu}\mathrm{d}\tilde{m}_{\mu}}{2\pi}\exp\left(-N\sum_{\mu}m_{\mu}\tilde{m}_{\mu} + \sum_{i,\mu}(\mathrm{i}\phi_iS_i\xi_i^{\mu}m_{\mu} + S_i\xi_i^{\mu}\tilde{m}_{\mu})\right).\end{aligned}$$

If p_0 patterns condense, we perform the quenched average over the remaining $(p-p_0)$ ξ. Now one can take the trace, to find

$$\begin{aligned}\langle N_s\rangle &= N^p\int_{-\infty}^{\infty}\prod_i \mathrm{d}\phi_i D(\phi_i)\exp\left[-\alpha\left(\mathrm{i}\sum_i\phi_i + \frac{k}{2}\sum_i\phi_i^2 - \frac{h}{2}\sum_{i,j}\frac{\phi_i\phi_j}{N}\right)\right] \\ &\quad\times\int\prod_{\mu}\frac{\mathrm{d}m_{\mu}\mathrm{d}\tilde{m}_{\mu}}{2\pi}\langle \mathrm{e}^{N\tilde{G}}\rangle \qquad \text{(A3)}\end{aligned}$$

where

$$\begin{aligned}\tilde{G} &= -\sum_{\mu}m_{\mu}\tilde{m}_{\mu} + \frac{1}{N}\sum_i\ln\cosh\left(\sum_{\mu=1}^{p_0}(\mathrm{i}\phi_im_{\mu}+\tilde{m}_{\mu})\xi_i^{\mu}\right) \\ &\quad + \frac{1}{N}\sum_i\sum_{\mu>p_0}\ln\cosh(\mathrm{i}\phi_im_{\mu}\tilde{m}_{\mu})\end{aligned}$$

and the only average left is that over the discrete ξ corresponding to the condensed patterns. For finite p, $\alpha \to 0$, and one has the same equations one finds in the thermodynamics of the fully connected model with finite p.

We are interested in the finite-α limit. If one expands the ln cosh one has

$$\int \prod_{\mu > p_0} \frac{\mathrm{d}m_\mu\, \mathrm{d}\tilde{m}_\mu}{2\pi} \exp N\left(-\sum_{\mu > p_0} m_\mu \tilde{m}_\mu + \frac{1}{2N} \sum_i \sum_{\mu > p_0} (\mathrm{i}\phi_i m_\mu + \tilde{m}_\mu)^2 \right)$$
$$= N^{-(p-p_0)} \exp\left\{ -N\frac{\alpha}{2} \ln\left[\frac{1}{N} \sum_i \phi_i^2 + \left(1 - \frac{\mathrm{i}}{N} \sum_i \phi_i \right)^2 \right] \right\}.$$

Setting

$$u = \frac{1}{N} \sum_i \phi_i^2 \qquad v = \frac{\mathrm{i}}{N} \sum_i \phi_i$$

by introducing δ functions and representing them with integrals, one finally has

$$\langle N_S \rangle = N^{p_0+2} \int \frac{\mathrm{d}u\, \mathrm{d}v\, \mathrm{d}x\, \mathrm{d}y}{(2\pi)^2} \prod_\mu \frac{\mathrm{d}m_\mu\, \mathrm{d}\tilde{m}_\mu}{2\pi} \mathrm{e}^{NG} \qquad \mu = 1, \ldots, p_0 \tag{A4}$$

where

$$G = -\alpha(v + \tfrac{1}{2}ku - \tfrac{1}{2}hv^2) - m_\mu \tilde{m}_\mu + xu - yv - \tfrac{1}{2}\alpha \ln[u + (1-v)^2] + \ln \tilde{\Phi}$$

and

$$\tilde{\Phi} = \int \mathrm{d}\phi D(\phi) \exp(-x\phi^2 + \mathrm{i}y\phi) \langle \cosh[(\mathrm{i}\phi m_\mu + \tilde{m}_\mu)\xi^\mu] \rangle.$$

$\langle N_S \rangle$ may now be evaluated with the saddle-point method.

Considering only the case $p_0 = 1$ we have

$$\tilde{\Phi} = \frac{\mathrm{e}^{\tilde{m}}}{2} \left(1 + \operatorname{erf} \frac{y+m}{2\sqrt{x}} \right) + \frac{\mathrm{e}^{-\tilde{m}}}{2} \left(1 + \operatorname{erf} \frac{y-m}{2\sqrt{x}} \right).$$

Solving the saddle-point equations for u, v, $\tilde{m}$ to eliminate these variables, and substituting

$$2x/\alpha = r \qquad y/\alpha = s$$

one obtains formulae (3) and (4), where one should note, however, that det H is intended as the Hessian determinant with respect to the five variables $\tilde{m}$, x, y, u, v.

References

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev.* A **32** 1007

—— 1985b *Phys. Rev. Lett.* **55** 1530

—— 1986 unpublished

—— 1987 *Ann. Phys., NY* **173** 30

Bray A J and Moore M A 1980 *J. Phys. C: Solid State Phys.* **13** L469

—— 1981 *J. Phys. C: Solid State Phys.* **14** 1313

Crisanti A, Amit D and Gutfreund H 1986 *Europhys. Lett.* **2** 337

Crisanti A and Sompolinsky H 1987 *Phys. Rev.* A **36** 4922

De Dominicis C, Gabay M, Garel T and Orland H 1980 *J. Physique* **41** 923

Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

Feigelman M V and Ioffe L B 1986 *Europhys. Lett.* **1** 197

—— 1987 *Int. J. Mod. Phys.* B **1** 51

Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047

Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

Hertz J A, Grinstein G and Solla S A 1987 *Proc. Heidelberg Colloq. on Glassy Dynamics and Optimization, 1986* ed I Morgenstern and I L van Hemmen (Berlin: Springer) p 538

Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554

—— 1984 *Proc. Natl Acad. Sci., USA* **81** 3088

Kinzel W 1985 *Z. Phys.* B **60** 205

—— 1987 *Proc. Heidelberg Colloq. on Glassy Dynamics and Optimization, 1986* ed I Morgenstern and I L van Hemmen (Berlin: Springer) p 529

Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745

Kree R and Zippelius A 1987 *Phys. Rev.* A **36** 4421

Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L675

Shinomoto S 1987 *Biol. Cybern.* **57** 197

Sompolinsky H 1986 *Phys. Rev.* A **34** 2571

—— 1987 *Proc. Heidelberg Colloq. on Glassy Dynamics and Optimization, 1986* ed I Morgenstern and I L van Hemmen (Berlin: Springer) p 485

Tanaka F and Edwards S F 1980 *J. Phys. F: Met. Phys.* **10** 2769