A feature retrieving attractor neural network

1993 J. Phys. A: Math. Gen. 26 2333

(http://iopscience.iop.org/0305-4470/26/10/008)

J. Phys. A: Math. Gen. **26** (1993) 2333–2342. Printed in the UK

A feature retrieving attractor neural network

D O'Kane and D Sherrington
Theoretical Physics, University of Oxford, 1 Keble Rd, Oxford OX1 3NP, UK

Received 3 August 1992

Abstract. A mean-field analysis is presented of the retrieval behaviour of a modular network of binary neurons capable of storing and retrieving patterns made up of associated features.

1. Introduction

Ever since Hopfield's watershed paper in which the analogy between memory storage/retrieval and the groundstates of complex spin-glass type systems was first proposed, and the following work by Amit *et al* (AGS) [4] in which the theoretical pattern storage limit of Hopfield's model was determined, much work has been done in extending this approach to models of a more practical and biological nature. One such extension, introduced by O'Kane and Treves [8], considered an architecture involving a system of internally strongly coupled neural networks, or modules, randomly but dilutely interconnected. The original study was stimulated by considerations of the anatomy of neocortex [1] and of the concept of patterns constituted from particular associations of features, the latter being stored in individual modules. The original paper of O'Kane and Treves was formulated in terms of graded response neurones and was confined to a study of a noiseless dynamics. This present paper extends the analysis to binary McCulloch–Pitts neurons with noisy dynamics. As well as confirming qualitative aspects of the earlier study, such as the existence of stable memory-glass states in which only local, but no global retrieval occurs, it permits an interesting interpolation between two extremes of statistically homogeneous networks discussed earlier: that of full connectivity where the phase transition between retrieval and non-retrieval is first order; and that of dilute random connectivity where the transition is second order.

2. The model

The overall architecture of the model is one consisting of M distinct modules of N neuronal units each. A unit can take two states $S = \pm 1$ and is connected to all $N-1$ other units within the same module and to L units distributed at random throughout the remaining $M-1$ modules. The total number of connections to a neuron is therefore $C = L + N - 1$ and we denote by $\gamma = L/C$ the fraction of long-range connections. All connections are symmetric and governed by Hebbian learning rules. P patterns are stored on both the short- and long-range connections. However, each of these (global network) patterns is made up of M *features*, one per module, each drawn from a repertoire of D features stored in that module—see figure 1. We assume that the distribution of ξ representing each feature d within a module m (which would result from some unspecified storage process)

0305-4470/93/102333+10$07.50

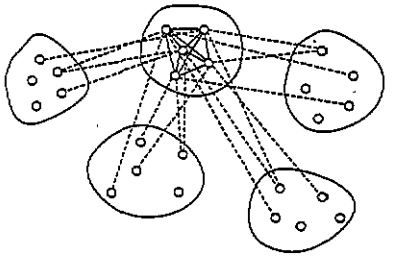
is given independently for each unit i (and d, m) and $\xi = \pm 1$ with equal probability. Note that $i = 1, \ldots, N$; $d = 1, \ldots, D$ and $m = 1, \ldots, M$. A global pattern, labelled p (with $p = 1, \ldots, P$), is then a random combination $\{d_1^p, \ldots, d_m^p, \ldots, d_M^p\}$. For simplicity, we shall assume that $P/D \equiv \mu$ is integral, and that features are assigned to patterns by randomly partitioning the P patterns in each module into D groups of μ elements (there are $P!/(\mu!)^D$ possible such assignments). The system is taken to obey random sequential dynamics

$$\text{Prob}(S_{i_m}(t+1) = \pm 1) = \tfrac{1}{2}\left\{1 \pm \tanh\beta\left(\sum_{j_m} J_{i_m j_m} S_{j_m} + \sum_{m'(\neq m)} \sum_{j_{m'}} J_{i_m j_{m'}} S_{j_{m'}}\right)\right\} \tag{1}$$

where β is the usual measure of synaptic gain/noise and the synaptic efficacies take the form

$$J_{i_m j_m}^{\text{short}} = \frac{\mu}{C} \sum_{d=1}^{D} \xi_{i_m}^d \xi_{j_m}^d \qquad J_{i_m j_n}^{\text{long}} = \frac{c_{i_m j_n}}{C} \sum_{p=1}^{P} \xi_{i_m}^p \xi_{j_n}^p \qquad m \neq n. \tag{2}$$

Variable $c_{i_m j_n} = 1, 0$ depending upon whether or not there is a link between neurone i_m in module m and j_n in module n. We further take the specific $c_{i_m j_n}$ to be chosen randomly with a probability c $(= L/N(M-1))$, $(1-c)$ that $c_{i_m j_n} = 1, 0$ and concentrate on the limit of small c and large M.



module

pattern	1	2	3	4	5
1	J	P	C	M	X
2	I	Q	C	M	Z
3	K	P	B	L	Y
4	J	R	A	L	Z
5	K	R	B	N	X
6	I	Q	A	N	Y

Figure 1. Schematic representation of the architecture of the network (top, only the connections relative to one module are drawn for clarity) and of the corresponding memory organization (middle). Full curves represent short-range connections while the broken curves denote long-range connections. The table at the bottom gives an example of how features could combine into patterns when $\mu = 2$. Boldface letters denote one particular pattern being retrieved.

3. The free energy

The attractor macrostates correspond to the asymptotic distribution of microstates generated by (1). They are given by the appropriate ergodicity-breaking thermodynamic equilibrium

states of the Hamiltonian

$$H = -\sum_m \sum_{(i_m,j_m)} J^{\text{short}}_{i_m j_m} S_{i_m} S_{j_m} - \sum_{(m,n)} \sum_{i_m,j_n} J^{\text{long}}_{i_m j_n} S_{i_m} S_{j_n} \tag{3}$$

and are obtainable from a minimization of the corresponding free energy. In the standard manner [3], we take a statistically relevant average over the specific choices of patterns and connectivities, using a replica procedure to perform the appropriate quenched averages [9]. There are now three relevant macroscopic order parameters characterizing the thermodynamic states

(i) the overlap with feature m in module d

$$m^{d\gamma}_m = \frac{1}{N} \sum_{i_m} \langle \xi^d_{i_m} \langle S^\gamma_{i_m} \rangle \rangle_\xi \tag{4}$$

(ii) the overlap of the whole system with pattern p

$$m^{p\gamma} = \frac{1}{MN} \sum_m \sum_{i_m} \langle \xi^p_{i_m} \langle S^\gamma_{i_m} \rangle \rangle_\xi \tag{5}$$

(iii) the Edwards–Anderson 'spin-glass' order parameter

$$q^{\gamma\delta} = \frac{1}{MN} \sum_m \sum_{i_m} \langle \langle S^\gamma_{i_m} \rangle \langle S^\delta_{i_m} \rangle \rangle_\xi \tag{6}$$

where $\langle . \rangle$ denotes a thermodynamic average and $\langle . \rangle_\xi$ an average over the pattern choice and the connectivity. Assuming a single retrieved pattern, integrating out the order parameters pertaining to unretrieved patterns, dropping terms of higher order in c, and taking the replica-symmetric ansatz, we obtain as the free energy per neuron in the thermodynamic limit ($N \to \infty$ $L \to \infty$, γ fixed)

$$\begin{aligned} f = {} & \frac{\alpha}{2} + \frac{\mu(1-\gamma)}{2M} \sum_m (m^{d'}_m)^2 + \frac{1}{2}\gamma (m^{p'})^2 - \frac{1}{\beta M} \sum_m \left\langle \int \mathrm{D}z \, \ln(2\cosh \beta h_m) \right\rangle_\xi \\ & + \frac{\alpha}{2\beta\mu(1-\gamma)} \left(\ln(1 - \beta\mu(1-\gamma)(1-q)) - \frac{\beta\mu(1-\gamma)q}{(1-\beta\mu(1-\gamma)(1-q))} \right) \\ & + \frac{\alpha\beta(r-\gamma)(1-q)}{2} + \frac{\alpha\beta\gamma(1-q^2)}{4} \end{aligned} \tag{7}$$

where

$$h_m = (\mu(1-\gamma) m^{d'}_m + \gamma m^{p'})\xi + z\sqrt{\alpha r} \qquad \int \mathrm{D}z = \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-z^2/2) \tag{8}$$

and

$$r = q \left(\gamma + \frac{\mu(1-\gamma)}{(1-\beta\mu(1-\gamma)(1-q))^2} \right). \tag{9}$$

d' labels the retrieved feature in each module and p' the retrieved global pattern and $\alpha = P/C$. Equations characterizing the properties of the attractor states as a function of the order parameters are given by extremizing f.

3.1. The saddle-point equations

The macroscopic properties of the model are therefore described by the following coupled saddle-point equations.

$$m_m^{d'} = \int \mathrm{D}z \tanh \beta(\gamma m^{p'} + \mu(1-\gamma)m_m^{d'} + z\sqrt{\alpha r})$$

$$q = \int \mathrm{D}z \tanh^2 \beta(\gamma m^{p'} + \mu(1-\gamma)m_m^{d'} + z\sqrt{\alpha r})$$

$$m^{p'} = \frac{1}{M}\sum_m m_m^{d'}. \tag{10}$$

These describe the four distinct phases which the system may assume. One is the 'paramagnetic' non-retrieval phase with $m_m^{d'} = m^{p'} = q = 0$. The others are described below.

3.2. Retrieval phase

If there is local and global retrieval, $m_m^{d'} = m^{p'} = m$, and the equations take the form:

$$g_m = m - \int \mathrm{D}z \tanh \beta((\gamma + \mu(1-\gamma))m + z\sqrt{\alpha r}) = 0$$

$$g_q = q - \int \mathrm{D}z \tanh^2 \beta((\gamma + \mu(1-\gamma))m + z\sqrt{\alpha r}) = 0. \tag{11}$$

As these equations are only analytically soluble in a few limited cases, a numerical method was used. This involved using a simplex descent algorithm to find the maximum value of α having a zero-valued, but finite m minimum to the function

$$\hat{O}(m, q, \alpha, \mu, \gamma, \beta) = g_m^2 + g_q^2. \tag{12}$$

This method allowed us to solve for a range of values of μ and γ, and the resulting $\alpha - T$ phase diagrams are given in figures 2(*a*) and 2(*b*). Two interesting limits exist. First, when $\gamma = 0$, the equations become analogous to the fully connected Hopfield model under the mapping $\alpha_{\mathrm{Hopf}} = \alpha/\mu$ and $\beta_{\mathrm{Hopf}} = \beta\mu$, and so the retrieval transition is first order. Second, when $\gamma = 1$, equation (11) maps directly onto the SK model whose analogous ferromagnetic-to-paramagnetic phase transition is second order. At some intermediate stage the system must pass through a point at which the transition changes its order.

This point can be determined as follows. To begin with, note that at α_c the condition for a second-order transition is

$$\left.\frac{\partial^2 f}{\partial m^2}\right|_{m=0} = 0 \tag{13}$$

which leads to

$$T(\gamma + \mu(1-\gamma))^{-1} = \int \mathrm{D}z \operatorname{sech}^2 \beta(z\sqrt{\alpha r}) = 1 - q_{(m=0)}. \tag{14}$$

Consequently a plot of q against T on the retrieval/non-retrieval boundary—see figure 3—should be linear in the second-order transition regime. To determine the tricritical point at which a second-order transition becomes first order also requires that at α_c

$$\left.\frac{\partial^4 f}{\partial m^4}\right|_{m=0} = 0. \tag{15}$$

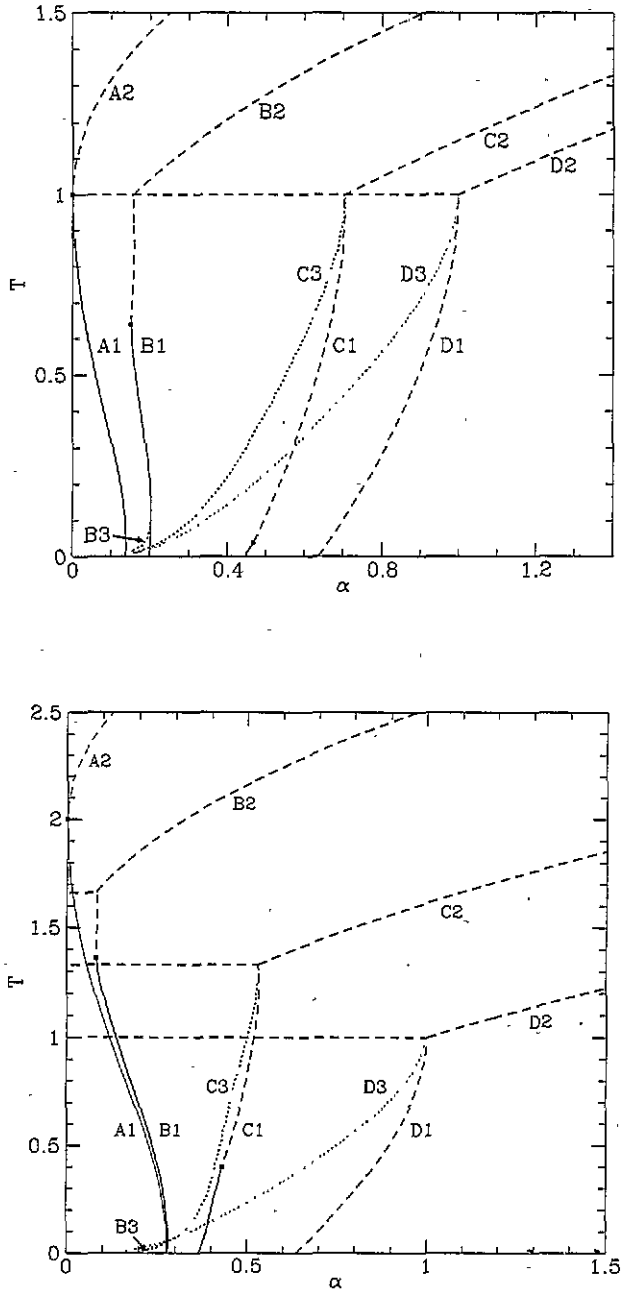


Figure 2. $\alpha - T$ phase diagrams for (*a*) $\mu = 1$ and (*b*) $\mu = 2$. The lines are labelled corresponding to the following key: A, B, C, D denote $\gamma = 0, \frac{1}{3}, \frac{2}{3}, 1$ respectively; while 1, 2, 3 denote the retrieval/no-retrieval phase boundary, the paramagnetic/spin-glass phase boundary and the de Almeida–Thouless line, respectively. A solid line and a dashed line denote the first- and second-order phase transition respectively. A square represents the tricritical point.

Combining (13) and (15), we arrive at

$$\frac{2}{3} = \beta(\gamma + \mu(1-\gamma)) \int \mathrm{D}z \, \mathrm{sech}^4 \beta(z\sqrt{\alpha r}) \tag{16}$$

which describes a point along the phase boundaries marked by squares in figures 2 and 3.

The value of α at the triple point was determined by expanding in m and q about T_c to give

$$\alpha^* = \frac{\gamma^2}{\gamma^3 + \mu(1-\gamma)(\gamma + \mu(1-\gamma))^2}. \tag{17}$$

In the large μ regime, certain terms may be neglected and the equations become equivalent to those of the AGS model under the mappings

$$\alpha_c = \mu(1-\gamma)\alpha_{AGS}(T) \qquad \text{and} \qquad \beta = \beta_{AGS}/\mu(1-\gamma) \tag{18}$$

showing that the critical capacity increases with μ, but only with a constant of proportionality aymptotically equal to the fraction of short-range connections. This signifies the fact that when many features are present, it is the local connections which dominate.

When $T \to 0$, $q \to 1$ and we may easily calculate the full μ dependence of α_c since the coupled equations simplify into a single-dimensional self-consistent equation in y

$$\frac{(\mathrm{erf}\, y)^2}{2\alpha y^2} = \frac{\gamma}{(\gamma + \mu(1-\gamma))^2} + \frac{\pi\mu(1-\gamma)(\mathrm{erf}\, y)^2}{((\gamma + \mu(1-\gamma))\sqrt{\pi}\, \mathrm{erf}\, y - 2\mu(1-\gamma) y \exp(-y^2))^2} \tag{19}$$

where y is defined as follows

$$y = \frac{m(\gamma + \mu(1-\gamma))}{\sqrt{2\alpha r}}. \tag{20}$$

The critical value α_c is the highest value of α for which there is a non-zero y solution to this equation. The μ dependency of α for $T = 0$ is plotted in figure 4 for different values of γ. As in the finite-temperature case, we also find that the transition switches from being second order to first order with increasing μ, provided $\gamma \neq 1$. As in the other figures the tricritical points are marked by squares. In the limit of $\gamma = 1$ the capacity equation becomes equivalent to that for the model proposed by Derrida *et al* (DGZ) [6]. However, note that the DGZ model has asymmetric connections so that despite the equivalence of the capacity equations, the models are still fundamentally different—a point which has already been discussed by Watkin and Sherrington [10].

3.3. Spin-glass phase

In the spin-glass phase, $m^d = m^p = 0$, but $q \neq 0$. Investigation of the form of the function g_q showed that the transition from a spin-glass state to a retrieval state is always second order. Hence, one may perform an expansion in q to give the quartic in T_g

$$T_g^4 - [2\mu(1-\gamma)]T_g^3 - [\alpha\gamma + \alpha\mu(1-\gamma) - \mu^2(1-\gamma)^2]T_g^2 + [2\mu\gamma\alpha(1-\gamma)]T_g - \alpha\mu^2(1-\gamma)^2\gamma = 0. \tag{21}$$

It is the largest root which corresponds to the relevant solution. It interpolates between the $\gamma = 0$ solution, $T_g = \mu + \sqrt{\alpha\mu}$, which has been shown in section 3.2 to map directly onto the fully connected Hopfield model, and the $\gamma = 1$ case which has a solution $T_g = \sqrt{\alpha}$.

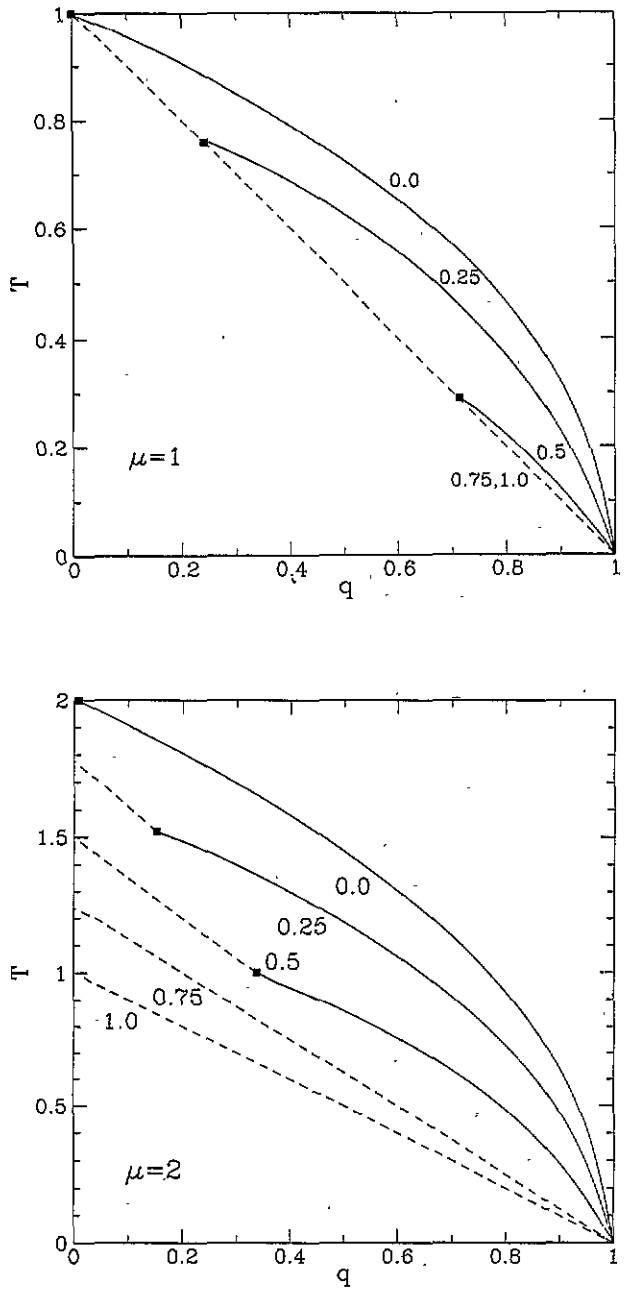


Figure 3. Plot of spin-glass order parameter against temperature for (*a*) $\mu = 1$ and (*b*) $\mu = 2$ for the different values of γ labelling each line. A solid line and a dashed line denote the first- and second-order phase transition respectively. The tricritical point as predicted by (16) is denoted by a square.

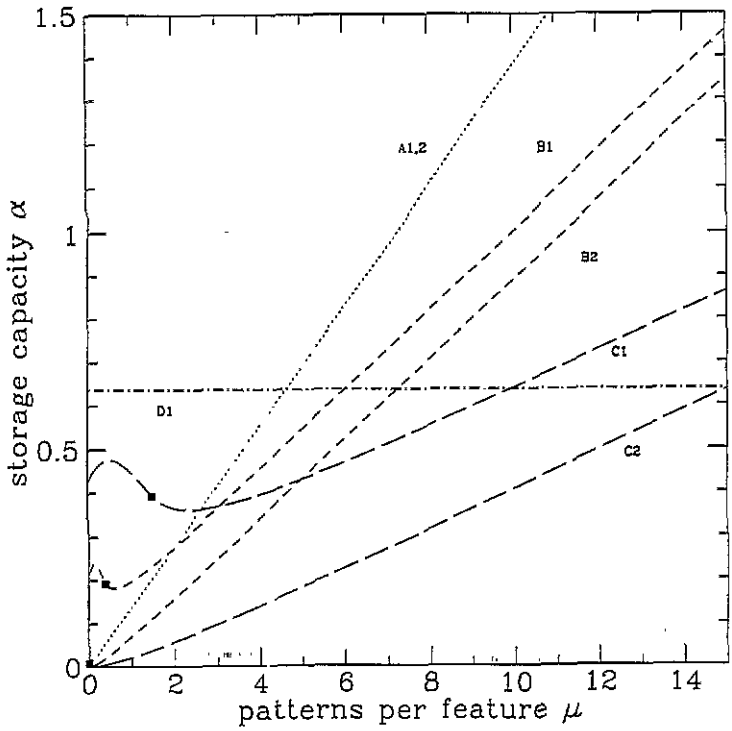


Figure 4. $\alpha - \mu$ phase diagrams for zero temperature. The labelling is: A, B, C, D denote $\gamma = 0, \frac{1}{3}, \frac{2}{3}, 1$ respectively; while 1, 2 denote the retrieval/no-retrieval phase boundary and the memory-glass/no memory phases boundary respectively. A square marks the tricritical point.

3.4. Memory-glass phase

A memory-glass phase is one in which there is only local and no global retrieval, i.e. $m = m_m^{d'} \sim O(1)$ but $m^{p'} = 0$; and its phase boundaries can be found by solving

$$\begin{aligned} m &= \int Dz \tanh \beta(\mu(1-\gamma)m + z\sqrt{\alpha r}) \\ q &= \int Dz \tanh^2 \beta(\mu(1-\gamma)m + z\sqrt{\alpha r}) \end{aligned} \tag{22}$$

to give figure 5. As in 3.3, when $T = 0$ the equations simplify to a one-dimensional equation, this time of the form

$$(\operatorname{erf} y)^2 = 2\alpha y^2 \left(\frac{\gamma}{(\mu(1-\gamma))^2} + \frac{\pi(\operatorname{erf} y)^2}{\mu(1-\gamma)(\sqrt{\pi}\operatorname{erf} y - 2y\exp(-y^2))^2} \right). \tag{23}$$

As μ tends to infinity, the memory-glass capacity equation becomes equivalent to the retrieval-state equation with $\mu \gg 1$, which is to be expected since the capacity of a network with dominant local retrieval should approach one which has local but no global retrieval.

3.5. De Almeida–Thouless line

So far, in this paper we have assumed replica symmetry in the q and r order parameters. However, it has been noted [2] that this may result in certain regions wherein the free energy

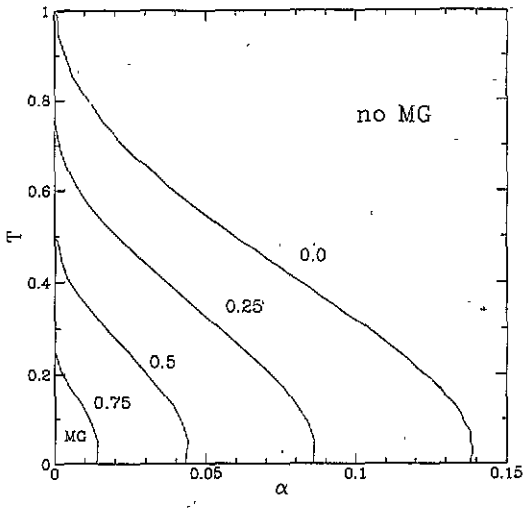


Figure 5. $\alpha - T$ phase diagram for the memory-glass (MG) phase. Each line is labelled with the corresponding value of γ.

is unstable under small fluctuations in q and r about the replica-symmetric saddle-point. Specifically, the stability test determining this is the condition for the eigenvalues of the Hessian matrix to become negative. The line across which this occurs is given by

$$\frac{\alpha r \beta^2}{q} \int \mathrm{Dz}\ \mathrm{sech}^4\, \beta((\gamma + \mu(1-\gamma))m + z\sqrt{\alpha r}) = 1 \tag{24}$$

This de Almeida–Thouless line [2] has been calculated for different values of μ and γ and is also displayed in figure 2. In the region of T, α small these lines all begin at the origin and move away from the zero-temperature axis exponentially slowly. They then begin to rise until they intersect the retrieval boundary. Below the line, replica symmetry must be broken and the retrieval boundary is therefore incorrect.

4. Discussion

In terms of its performance as a feature storing/retrieving system, this model displays some interesting properties. To begin with, it is found that as the number of shared features—μ—becomes large, the number of global pattern configurations which can be retrieved grows asymptotically proportionally to μN. Although good, there is a possible drawback since the memory-glass state becomes asymptotically as persistent as the retrieval state as μ increases, although we anticipate that the basins of attraction of these states are smaller than those of the true retrieval states. Unfortunately, the requirement that M be large, as well as $N \to \infty$, means that the system is too large to attempt to carry out reliable simulations to test this. One way to avoid falling into the memory-glass state would be to increase the noise in the dynamics, i.e. the temperature T, since full retrieval states are stable until $T_c = \gamma + \mu(1-\gamma)$, but memory-glass states have a lower critical temperature $\mu(1-\gamma)$. For γ greater than ~ 0.4, this can be done without any decrease in the pattern storage capacity.

This model is closely related to that studied by Krey and Poeppel [7], who investigated a network segmented into units each storing separate patterns ('letters'). They too impose a

tendency for the letters to become organized into preferred 'words'. Despite this, the system possesses stable states consisting of non-preferred words—equivalent to our memory-glass states. However, they do not investigate the effects of dilution.

It is interesting to compare the behaviour of our model when $\mu = 1$, with that of the arbitrarily dilute AGS (Amit–Gutfreund–Sompolinsky) model [5] whose low-temperature behaviour is similar but whose critical capacity falls to zero at T_c. The difference can be understood from signal-to-noise considerations. In the diluted AGS model, the signal comes from as many units as the 'fully connected' noise. However, in the modular model, the signal comes from $N + L$ units whilst the 'fully connected' noise comes from only N units. The signal-to-noise ratio is, therefore, greater in this model, and so it better withstands the increase in noise to the system resulting from an increase in temperature.

An interesting, but unknown feature of this model, is how its replica-symmetry-breaking scheme interpolates between the known $\gamma = 1$ limit where the full Parisi Ansatz is the correct solution, to the $\gamma = 0$ AGS model whose replica-symmetry-broken solution is difficult to obtain. It has been conjectured by Canning and Naef [5] that the AT line always intersects the retrieval phase boundary at the point where the latter's gradient becomes infinite—which it certainly seems to do in this model—and that, furthermore, the full replica-symmetry-broken retrieval phase line may be found by simply drawing a line vertically from this intersection to the zero-temperature axis.

Acknowledgment

DO'K would like to acknowledge a Research Studentship from the Department of Education for Northern Ireland.

References

[1] Abeles M 1991 *Corticonics* (Cambridge: Cambridge University Press)
[2] De Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[3] Amit D J 1989 *Modeling Brain Function* (New York: Cambridge University Press)
[4] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30–67
[5] Canning A and Naef J P 1992 *J. Physique* I **2** 1792
[6] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167–173
[7] Krey U and Poppel G 1989 *Z. Phys.* B **76** 513–520
[8] O'Kane D and Treves A 1992 *J. Phys. A: Math. Gen.* **25** 5055–5069
[9] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[10] Watkin T L H and Sherrington D 1991 *Europhys. Lett.* **14** 791–796