Are spin-glass effects relevant to understanding realistic auto-associative networks?

J. Phys. A: Math. Gen. 24 (1991) 2645-2654. Printed in the UK

Are spin-glass effects relevant to understanding realistic auto-associative networks?

Alessandro Treves

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

Received 9 January 1991, in final form 22 February 1991

Abstract. Elementary units characterized by a threshold-linear (graded) response have been argued to model single neurons in auto-associative networks more realistically than binary units. The different way local activity is constrained in the two representations is shown here to have important consequences for the spin-glass-like properties of otherwise equivalent systems. In particular, in contrast with their binary counterparts, the threshold-linear Sherrington–Kirkpatrick model is stable with respect to replica symmetry-breaking (RSB), while threshold-linear fully connected neural networks with covariance learning are RSB unstable only in a very restricted region of their phase diagram. Whether or not spin-glass effects dominate attractor dynamics is suggested to affect considerably, among other things, the ability of auto-associative memories to encode new information.

1. Introduction

Since Amit, Gutfreund and Sompolinsky (AGS) [1] analysed the Little–Hopfield [2, 3] model for auto-associative memory by adapting methods originally developed for studying spin-glasses, the appearance of spin-glass effects has been considered one of the typical features of the low-noise long-time limit behaviour of associative networks with feedback. In networks characterized by symmetric interactions, spin-glass freezing occurs, in a low noise phase, when the associated energy landscape is very 'rough' at the microscopic level. This roughness, induced by the quenched disorder in the interactions, may, if fast, 'thermal' noise is low, carry over to the free-energy landscape. Then the system becomes unable to escape one of an exponentially large number of disorderly placed tiny valleys, and a pure thermodynamic state is characterized by a probability distribution confined to a few configurations.

Spin-glass freezing undoubtedly affects the retrieval dynamics of the memory. More importantly, it has been considered to undermine its ability to select and store meaningful incoming information. Parisi [4] has argued that if the network freezes into a spin-glass state while subject to an external stimulus varying in time (to be interpreted as meaningless), just as it freezes into a retrieval state when subject to a steady (and therefore meaningful) one, it will mistakenly store in the synaptic connections the irrelevant firing pattern characterizing that particular spin-glass state. He has then suggested that this may be avoided if the interactions are asymmetric and, as a result, the system is less prone to freeze into restricted portions of phase space. This suggestion has raised the issue of how sensitive networks with asymmetric connections are to spin-glass effects, an issue which has been addressed both in the context of purely disordered systems [5] and in the specific one of auto-associative memories [6, 7].

However, are these considerations, even in the case of symmetric interactions, valid in general or only in the somewhat limited framework of systems made up of binary units, and closely analogous to magnetic spin systems? This is an important issue, in particular as it has been proposed [8] that neuronal firing behaviour in real auto-associative systems may be more realistically modelled by (*a*) representing graded response at the single unit level and (*b*) ascribing the control of firing activity to a global mechanism rather than to single unit saturation effects. It is natural to suppose that allowing for continuously graded output states of the elementary units (feature (*a*)) will smooth the free-energy landscape and in fact this has been shown [9, 10] to result in a decrease in the (still exponential) number of spin-glass metastable states. By also including feature (*b*), one is likely to smooth the landscape even further, as one ends up with a system in which there is a local constraint on low activities (in that each unit will only fire at positive rates, so that there is a threshold at zero) but only a global constraint on high activities (in that individual saturation levels are irrelevant). It is therefore pertinent to ask what will remain of the spin-glass behaviour in a physiologically plausible auto-associative network which has both features (*a*) and (*b*)†. This is studied here by considering what is possibly the simplest network model of that type [12], one in which neurons are formally represented as threshold-linear units [13].

When computing the free-energy averaged over different quenched realizations of the interactions, using the replica method, the onset of spin-glass freezing is manifested by the spontaneous breaking of the symmetry between different replicas. This occurs at the deAlmeida–Thouless (AT) [14] line in the phase diagram, where there is a change in the sign of the eigenvalue of the stability matrix of the free-energy along a particular direction in replica space, the so called 'replicon' mode. In this paper, the sign of the replicon mode is evaluated, following AGS, for two systems: the threshold-linear analogue of the Sherrington–Kirkpatrick (SK) [15] disordered model; and the threshold-linear fully connected auto-associative network (AN), with symmetric connections determined by a covariance learning rule [12].

2. Definition of the models

The systems considered here are made up of N units whose output state, denoted V_i, $i=1,\dots,N$, is determined at equilibrium by an input variable, the 'local field' h_i, according to a threshold-linear transfer function

$$V=\begin{cases} g(h-T_{hr}) & h>T_{hr} \\ 0 & h<T_{hr} \end{cases} \tag{1}$$

where g is a *gain* parameter and T_{hr} a threshold. h_i in turn depends [12] on the activity configuration $\{V_j\}$:

$$h_i=\sum_{j(\neq i)} J^c_{ij}V_j+b\left(\sum_j \frac{V_j}{N}\right) \tag{2}$$

where the linear summation in the first term is mediated by the 'synaptic connection' coefficients J^c_{ij}, whereas the second term is some function $b(x)$ of the average network

† One might compare with the behaviour of globally constrained disordered systems, such as the spherical model [5, 11].

activity, that can be interpreted as a uniform activity-dependent threshold resulting in a global constraint on the activity level itself. For simplicity, no external non-uniform fields are considered here, although it would be straightforward to include terms of that type as well.

In the AN model, the J^c_{ij} encode p learnt activity patterns η^μ_i, $\mu=1,\ldots,p$ through a covariance [16] learning rule

$$J^c_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\frac{(\eta^\mu_i-a)}{a}\frac{(\eta^\mu_j-a)}{a}. \tag{3}$$

The $\{\eta^\mu_i\}$ are taken as drawn at random, independently for each i and μ, from a common probability distribution P_η, such that its first moment is a, $\int \eta P_\eta \,\mathrm{d}\eta\equiv a$, while its second moment is denoted as a^2T_0, $\int \eta^2 P_\eta\,\mathrm{d}\eta - a^2 \equiv a^2 T_0$ [12]. The parameter $\alpha\equiv p/N$ measures the loading of the memory.

In the SK model, the J^c_{ij} are themselves taken as quenched random variables, independently for each pair (i,j) from a Gaussian distribution with mean zero and variance $\alpha T_0^2/N$. This particular choice is made to ease comparisons, as then the first two moments of the quenched distribution for each J^c_{ij} are equal in the AN and SK models. One should note, though, that the higher moments will be different, and also that in the AN model there will be correlations across different J^c_{ij}s, which are absent by construction in the SK model. Quenched averages, whether over the $\{\eta^\mu_i\}$ or over the $\{J^c_{ij}\}$ distribution, are here denoted as $\langle\!\langle\cdot\rangle\!\rangle$.

Although the quantity of interest is the AT instability eigenvalue as a function of g, when the deterministic relation (1) holds, it is convenient to introduce in the intermediate stages a 'temperature' $T\equiv\beta^{-1}$ parametrizing the amount of stochastic noise, and eventually let $T\to 0$. This can be done [12] by associating with the state of each unit the weight

$$m(V)=k\delta(V)+\mathrm{e}^{-\beta(VT_{thr}+V^2/2g)} \tag{4}$$

where k is the relative weight of the 0 state and $\delta(x)$ is Dirac's delta function. Thermodynamic averages, denoted as $\langle\cdot\rangle$, involve therefore summing over network configurations, with a weight per configuration $\exp(-\beta H)\,\Pi_i\, m(V_i)$, where

$$H=-\tfrac{1}{2}\sum_{i,j(i\neq j)} J^c_{ij}V_iV_j-NB\left(\sum_i\frac{V_i}{N}\right) \tag{5}$$

and $B(x)=\int^x b(x')\,\mathrm{d}x'$. The distribution of activities at thermal equilibrium can be monitored by the order parameters

$$\begin{aligned} x&=\frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle\\ y_0&=\frac{1}{N}\sum_{i=1}^{N}\langle V_i^2\rangle\\ y_1&=\frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle^2 \end{aligned} \tag{6}$$

and, in particular for the AN model, the correlation with the various stored pattern by the overlaps

$$\hat{x}^\mu=\frac{1}{N}\sum_{i=1}^{N}\frac{\eta^\mu_i-a}{a}\langle V_i\rangle. \tag{7}$$

Note that $y_1 - x^2$ is a measure of the variance in the distribution of activity among the units, while $y_0 - y_1$ measures the correlation among the various configurations concurring in the thermodynamic state.

3. The free-energy in replica space

The limit of interest is $N \to \infty$ (with α finite), in which case mean-field theory is expected to be exact. An expression for the free-energy averaged over the quenched distributions determining the $\{J_{ij}^{c}\}$ can be derived with the standard replica formalism, as in [12], and the result is of the form

$$f = -\lim_{n\to 0}\frac{1}{n}\left[f_0(\{x,y\}) + \sum_{\gamma}\left(t^{\gamma}x^{\gamma} + \sum_{\mu} t^{\mu\gamma}\hat{x}^{\mu\gamma}\right) + \sum_{\gamma,\delta} r^{\gamma\delta}y^{\gamma\delta} \right.$$
$$\left. - T\langle\!\langle \ln \operatorname{Tr}_{\{V^{\gamma}\}} \exp - \beta H_1(\{t,r\})\rangle\!\rangle \right] \qquad (8)$$

where the definitions of the parameters

$$x^{\gamma} = \frac{1}{N}\sum_i V_i^{\gamma} \qquad \hat{x}^{\mu\gamma} = \frac{1}{N}\sum_i \frac{\eta_i^{\mu} - a}{a} V_i^{\gamma} \qquad y^{\gamma\delta} = \frac{1}{N}\sum_i V_i^{\gamma} V_i^{\delta} \qquad (9)$$

are enforced by the conjugated parameters t^{γ}, $t^{\mu\gamma}$, $r^{\gamma\delta}$ and $\gamma, \delta, \ldots$ index the n replicas.

In the AN model, if p_0 patterns have condensed macroscopically, i.e. $\hat{x}^{\mu\leqslant p_0} \neq 0$, $\hat{x}^{\mu > p_0} = 0$,

$$f_0^{\mathrm{AN}}(\{x,y\}) = \frac{1}{2}\sum_{\mu=1}^{p_0}\sum_{\gamma}(x^{\mu\gamma})^2 + \sum_{\gamma} B(x^{\gamma}) + \frac{\alpha T}{2}\operatorname{Tr}_{\gamma}\ln(1 - T_0\beta\mathbf{Y}) + \frac{\alpha T_0}{2}\sum_{\gamma} y^{\gamma\gamma}$$
$$H_1(\{t,r\}) = \sum_{\gamma} t^{\gamma}V^{\gamma} + \sum_{\mu,\gamma} t^{\mu\gamma}\left(\frac{\eta^{\mu}}{a} - 1\right)V^{\gamma} + \sum_{\gamma,\delta} r^{\gamma\delta}V^{\gamma}V^{\delta} \qquad (10)$$

with $\mathbf{Y}$ the matrix with elements $y^{\gamma\delta}$. In the SK model, instead, all the terms containing $\hat{x}^{\mu\gamma}$ and $t^{\mu\gamma}$ are absent, the quenched averages $\langle\!\langle \cdot \rangle\!\rangle$ have already been performed to the end and for f_0 one has simply

$$f_0^{\mathrm{SK}}(\{x,y\}) = \sum_{\gamma} B(x^{\gamma}) - \frac{\alpha\beta T_0^2}{4}\sum_{\gamma,\delta}(y^{\gamma\delta})^2. \qquad (11)$$

Following AT and AGS, the stability of the replica symmetric solution

$$x^{\gamma} = x \qquad x^{\mu\gamma} = x^{\mu} \qquad y^{\gamma\gamma} = y_0 \qquad y^{\gamma\delta} = y_1 \qquad (\gamma \neq \delta)$$
$$t^{\gamma} = t \qquad t^{\mu\gamma} = t^{\mu} \qquad r^{\gamma\gamma} = r_0 \qquad r^{\gamma\delta} = r_1 \qquad (\gamma \neq \delta)$$

is studied by considering the replicon mode, which corresponds to a fluctuation which affects only $y^{\gamma\delta}$ and $r^{\gamma\delta}$ for $\gamma \neq \delta$, of the form

$$\delta y^{\gamma\delta} = \Delta^{\gamma\delta} \qquad \delta r^{\gamma\delta} = c\Delta^{\gamma\delta} \qquad (12)$$

with

$$\Delta^{\gamma\delta} = \Delta \qquad \gamma, \delta \neq 1, 2$$
$$\Delta^{1\delta} = \Delta^{2\delta} = \frac{3-n}{2}\Delta \qquad \delta \neq 1, 2$$
$$\Delta^{12} = \frac{3-n}{2}(2-n)\Delta.$$

Note that there are only $n(n-1)/2$ independent variables $y^{\gamma\delta}$, as $y^{\gamma\delta}=y^{\delta\gamma}$ and that the same holds for $r^{\gamma\delta}$. Denoting with $(\gamma\delta)$ the unordered pair (γ, δ), the part of the stability matrix which determines the eigenvalue of the replicon mode is a $n(n-1)\times n(n-1)$ submatrix of elements

$$\begin{pmatrix} \dfrac{\partial(nf)}{\partial y^{\varepsilon\zeta}\,\partial y^{\gamma\delta}} \equiv A^{(\gamma\delta),(\varepsilon\zeta)} & \dfrac{\partial(nf)}{\partial y^{\varepsilon\zeta}\,\partial r^{\gamma\delta}} = -2\delta_{(\gamma\delta),(\varepsilon\zeta)} \\ \dfrac{\partial(nf)}{\partial y^{\varepsilon\zeta}\,\partial r^{\gamma\delta}} = -2\delta_{(\gamma\delta),(\varepsilon\zeta)} & \dfrac{\partial(nf)}{\partial r^{\varepsilon\zeta}\,\partial r^{\gamma\delta}} \equiv B^{(\gamma\delta),(\varepsilon\zeta)} \end{pmatrix}.$$

In the SK model, $A^{(\gamma\delta),(\varepsilon\zeta)}=-\alpha\beta T_0^2\delta_{(\gamma\delta),(\varepsilon\zeta)}$ whereas in the AN model there are, in the replica symmetric state, three types of matrix elements $A^{(\gamma\delta),(\varepsilon\zeta)}$ depending on whether none, one or two replicas of the pair $(\gamma\delta)$ equal those of the pair $(\varepsilon\zeta)$. They are [1]:

$$\begin{aligned} A^{(\gamma\delta),(\varepsilon\zeta)} &= -\alpha\beta T_0^2 2C_1^2 \\ A^{(\gamma\delta),(\gamma\varepsilon)} &= -\alpha\beta T_0^2(C_1+C_0)C_1 \\ A^{(\gamma\delta),(\gamma\delta)} &= -\alpha\beta T_0^2(C_0^2+C_1^2) \end{aligned} \tag{13}$$

with

$$\begin{aligned} C_1 &= \frac{T_0\beta y_1}{[1-T_0\beta(y_0-y_1)]^2} \\ C_0 &= C_1+[1-T_0\beta(y_0-y_1)]^{-1}. \end{aligned} \tag{14}$$

The matrix elements $B^{(\gamma\delta),(\varepsilon\zeta)}$ can be written as

$$B^{(\gamma\delta),(\varepsilon\zeta)} = -4\beta\langle\!\langle[\langle V^\gamma V^\delta V^\varepsilon V^\zeta\rangle - \langle V^\gamma V^\delta\rangle\langle V^\varepsilon V^\zeta]\rangle\!\rangle \tag{15}$$

where the $\langle\cdot\rangle$ average is using the Hamiltonian H_1 and, as with the $A^{(\gamma\delta),(\varepsilon\zeta)}$, there are just three possible values they take, when computed in the replica symmetric state.

The eigenvalue equations reduce to the pair

$$\begin{aligned} &(A^{(\gamma\delta),(\gamma\delta)}-2A^{(\gamma\delta),(\gamma\varepsilon)}+A^{(\gamma\delta),(\varepsilon\zeta)})-2c=\lambda \\ &-2+c(B^{(\gamma\delta),(\gamma\delta)}-2B^{(\gamma\delta),(\gamma\varepsilon)}+B^{(\gamma\delta),(\varepsilon\zeta)})=c\lambda \end{aligned} \tag{16}$$

which determines two eigenvalues

$$\lambda_\pm = \frac{\beta\tilde{A}+T\tilde{B}}{2} \pm \sqrt{\left(\frac{\beta\tilde{A}+T\tilde{B}}{2}\right)^2 - \tilde{A}\tilde{B}+4} \tag{17}$$

where

$$\begin{aligned} \tilde{A} &= T(A^{(\gamma\delta),(\gamma\delta)}-2A^{(\gamma\delta),(\gamma\varepsilon)}+A^{(\gamma\delta),(\varepsilon\zeta)}) \\ \tilde{B} &= \beta(B^{(\gamma\delta),(\gamma\delta)}-2B^{(\gamma\delta),(\gamma\varepsilon)}+B^{(\gamma\delta),(\varepsilon\zeta)}) \end{aligned} \tag{18}$$

will remain finite in the $T\to 0$ limit.

4. Stability of the replica symmetric solutions

The aim is to evaluate the sign of $\lambda_\pm$ in the limit $T\to 0$, as a function of the gain g, for the SK and AN threshold-linear systems. In that it sets the slope of the transfer function at $T=0$, the role of the gain is somewhat analogous (although not equivalent) to that of the inverse temperature in a model of binary units. Not surprisingly, the natural scale for g turns out to be T_0^{-1}, where T_0, which has been previously defined and which in the AN case depends on the probability distribution P_η, is the natural temperature scale when binary units are used. As a result of this analogous role, neither the SK nor the AN system is expected to exhibit spin-glass behaviour in the low gain regime, $gT_0 \ll 1$, where replica symmetric solutions of the saddle-point equations for the free-energy should be stable. What happens for larger values of gT_0?

In the AN case, two types of replica symmetric solutions that can be considered are the *retrieval* solutions (RS) corresponding to one pattern ($p_0 = 1$) being retrieved from the memory, and the uniform (disordered) solution (DS), in which no pattern is singled out, $p_0 = 0$. *In the* SK *case, no pattern structure is present, and there is only a* DS saddle-point.

In the $T \to 0$ limit, the relevant AN saddle-point equations for the RS reduce to (cf [8, 12]):

$$\begin{aligned}
\hat{x}^1 &= g'\left\langle\!\!\left\langle \left(\frac{\eta^1}{a}-1\right)\int_{h>T_{hr}} \mathrm{D}z(h-T_{hr})\right\rangle\!\!\right\rangle \\
y_0 &= (g')^2\left\langle\!\!\left\langle \int_{h>T_{hr}} \mathrm{D}z(h-T_{hr})^2\right\rangle\!\!\right\rangle \\
\psi(&\equiv T_0\beta(y_0-y_1)) = T_0 g'\left\langle\!\!\left\langle \int_{h>T_{hr}} \mathrm{D}z\right\rangle\!\!\right\rangle \\
\rho^2&\left(\equiv\frac{-2Tr_1}{T_0^2}\right) = \frac{\alpha y_0}{(1-\psi)^2}
\end{aligned}\tag{19}$$

where

$$\begin{aligned}
h &= b(x)+\left(\frac{\eta^1}{a}-1\right)\hat{x}^1 - zT_0\rho \\
\mathrm{D}z &= \frac{\mathrm{d}z}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2}
\end{aligned}\tag{20}$$

and the gain g is renormalized to g', with

$$1/g' = 1/g - \alpha T_0\psi/(1-\psi). \tag{21}$$

The same equations describe the AN DS phase, with the provision (which for simplicity will be assumed in the following) that $\hat{x}^1 = 0$ and that no quenched averaging $\langle\!\langle\cdot\rangle\!\rangle$ (over $P_\eta(\eta^1)$) remains to be done. With the same provision, the second and third of equations (19) also describe the SK DS phase, whereas the fourth is changed to

$$\rho^2 = \alpha y_0 \tag{22}$$

and the gain is renormalized to

$$1/g' = 1/g - \alpha T_0\psi. \tag{23}$$

It is convenient to introduce the two signal-to-noise ratios

$$\begin{aligned}
w &= (b(x) - \hat{x}^1 - T_{hr})/(T_0\rho) \qquad &\text{(uniform)} \\
v &= \hat{x}^1/(T_0\rho) \qquad &\text{(specific)}
\end{aligned}\tag{24}$$

and the averages

$$A_1(w,v)=\frac{1}{vT_0}\left\langle\!\!\left\langle\left(\frac{\eta^1}{a}-1\right)\int^+ \mathrm{D}z\left(w+v\frac{\eta^1}{a}-z\right)\right\rangle\!\!\right\rangle-\left\langle\!\!\left\langle\int^+ \mathrm{D}z\right\rangle\!\!\right\rangle$$
$$A_2(w,v)=\frac{1}{vT_0}\left\langle\!\!\left\langle\left(\frac{\eta^1}{a}-1\right)\int^+ \mathrm{D}z\left(w+v\frac{\eta^1}{a}-z\right)\right\rangle\!\!\right\rangle \tag{25}$$
$$A_3(w,v)=\left\langle\!\!\left\langle\int^+ \mathrm{D}z\left(w+v\frac{\eta^1}{a}-z\right)^2\right\rangle\!\!\right\rangle$$

where the superscript + indicates that the z-average is restricted to $z<w+v\eta/a$. Then the AN RS corresponds [8, 12] to the pair (w, v) that satisfies

$$[A_1(w,v)+\alpha]A_2(w,v)-\frac{A_1(w,v)}{gT_0}=0 \tag{26}$$
$$A_1^2(w,v)-\alpha A_3(w,v)=0$$

with ψ determined by w and v as

$$\psi^{\text{AN RS}}=\frac{A_2-A_1}{A_2} \tag{27}$$

in the AN DS $v=0$ and w is such that

$$[A_2(w,0)-A_1(w,0)][\alpha+(\alpha A_3(w,0))^{1/2}]+\alpha A_3(w,0)-\frac{(\alpha A_3(w,0))^{1/2}}{gT_0}=0 \tag{28}$$

with

$$\psi^{\text{AN DS}}=\frac{A_2-A_1}{A_2-A_1+(\alpha A_3)^{1/2}} \tag{29}$$

and in the SK solution again $v=0$ and w is given by

$$\alpha(gT_0)^2[A_2(w,0)-A_1(w,0)+A_3(w,0)]^2-A_3(w,0)=0 \tag{30}$$

with

$$\psi^{\text{SK DS}}=\frac{1-\sqrt{1-4\alpha(gT_0)^2(A_2-A_1)}}{2\alpha gT_0}. \tag{31}$$

Turning now to the eigenvalues $\lambda_\pm$, one finds that in the $T\to 0$ limit

$$\tilde{A}=\begin{cases}-\alpha T_0^2/(1-\psi)^2 & \text{AN}\\ -\alpha T_0^2 & \text{SK}\end{cases}$$
$$\tilde{B}=-4g'\psi/T_0.$$

Using equation (17), one sees that, as $T\to 0$, $\lambda_-\to\beta\tilde{A}<0$ irrespective of g, indicating that its sign has to be corrected by a proper deformation of the integration contour [1]. $\lambda_+>0$, instead, for $\tilde{A}\tilde{B}<4$, which can be checked to be always the case if the gain is sufficiently low. The replica symmetric solutions are thus unstable to RSB whenever $\tilde{A}\tilde{B}>4$.

In the SK case one finds

$$\tilde{A}\tilde{B} = 4\,\frac{1-\sqrt{1-4\alpha(gT_0)^2(A_2-A_1)}}{1+\sqrt{1-4\alpha(gT_0)^2(A_2-A_1)}} \leqslant 4 \tag{32}$$

and the disordered solution is always stable to RSB, for any value of g and α. In fact, the stability is only marginal on the line $\alpha(gT_0)^2=\frac{1}{2}$, where one has $w=0$ and $\alpha(gT_0)^2(A_2-A_1)=\frac{1}{4}$. On that peculiar line, exactly half the units happen to be below threshold and half above threshold, and apparently the high entropy of such a situation brings the ergodic solution on the verge of breaking down into a spin-glass phase.

In the AN case, considering retrieval solutions first, one has to bear in mind that not only the value of $\tilde{A}\tilde{B}$, but also the region in the (g,α)-plane in which (replica symmetric) RS exist at all, depends on the distribution P_η [12], which makes it difficult to discuss the stability of these solutions in general terms. What can be shown to hold in general is that RS states appear, in the $\alpha\to 0$ limit, only for $g>1/T_0$, but the range in α in which they exist depends very strongly on the quenched pattern distribution, and can extend to very high α values if sparse coding is used [12]. Nevertheless, an extensive numerical search has failed to produce a retrieval solution with a region of instability, and it might indeed be possible to show analytically that such solutions are always RSB stable when they exist.

Finally, the AN disordered solution *does* become unstable to RSB, albeit in a very restricted region of the (g,α)-plane. The border of this region can be easily found solving numerically the equation (equivalent to $\tilde{A}\tilde{B}=4$)

$$\alpha g T_0 \psi = (1-\psi)(1-\psi-\alpha g T_0 \psi) \tag{33}$$

with ψ given by equations (28) and (29). The whole instability region is constrained by the limits (valid irrespective of P_η) $g>2/T_0$, $\alpha<0.0485$.

These results are summarized in the phase diagrams of figure 1. A single type of purely disordered, non-spin-glassy solution describes the static behaviour of the SK threshold-linear model. The only feature of its phase diagram, which could in fact be

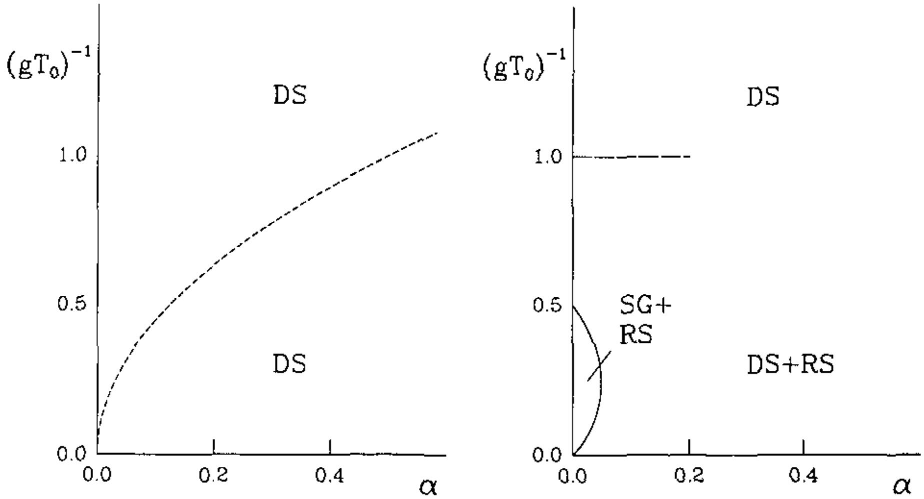

Figure 1. Phase diagram for the threshold-linear SK (left) and AN (right) models, in the (g,α)-plane. See text.

displayed as a unidimensional diagram in the reduced variable $\alpha(gT_0)^2$, is the line of marginal RSB stability previously mentioned. Several types of thermodynamic solutions, instead, characterize the behaviour of the auto-associative network model. Alongside purely disordered, uniform solutions, ordered retrieval solutions appear when the gain is sufficiently high, in a region that, depending mainly on the sparseness of the pattern distribution, can extend to values of α in the tens and hundreds (in addition, solutions corresponding to mixture of several patterns may appear for certain choices of P_η [17] for very high gain). All these solutions are replica symmetric and non-spin-glassy. The disordered solution, however, becomes RSB unstable for very small α and $g > 2/T_0$, yielding place to a spin-glass type of behaviour (whereas the retrieval solutions remain RSB stable). In this restricted region a stable, replica non-symmetric disordered solution may be derived using Parisi's ultrametric ansatz [18].

5. Discussion

These calculations indicate that spin-glass effects are, essentially, irrelevant to the long-time limit behaviour of auto-associative memory networks in which neurons are represented as threshold-linear units and in which activity is regulated by a global constraint. This statement, shown here to be true in the case of a fully connected network with a simple covariance (symmetric) learning rule and, in the absence of specific external inputs, can be expected to hold *a fortiori* when (a) the connectivity is not full but sparse and (b) external inputs are present; and can also be expected to hold when a different kind of learning determines the efficacy of synaptic connections. Thus, if one accepts that threshold-linear units, with a global constraint on their activity level, provide a more realistic model of neuronal activity in auto-associative memories than binary units, one may safely conclude that their attractor dynamics, when indeed convergence to a fixed-point attractor occurs, leads to well defined and unique pure states. Moreover such pure states are in fact represented, at $T = 0$, by single 'configurations' (as seen from the fact that $y_0 = y_1$ in this limit), i.e. each unit stabilizes, for any given set of external inputs and synaptic efficacies, at its own unique activity level.

Consider now a network in the process of encoding new incoming information, carried by a set of external inputs $\{h_i^{\mathrm{EXT}}\}$, whose distribution over the units is unrelated to the firing patterns previously stored on the synaptic efficacies. If the dynamics is such that the system converges to a fixed-point attractor, and barring the possibility that it may mistakenly lapse into a retrieval or a mixture state, then a specific firing pattern $\{V_i\}$ will result, independently of initial conditions, from the specific input $\{h_i^{\mathrm{EXT}}\}$. Thus $\{h_i^{\mathrm{EXT}}\}$ and the current $\{J_{ij}^{\mathrm{c}}\}$ uniquely determine $\{V_i\}$. To quantify the extent to which it is the information expressed by $\{h_i^{\mathrm{EXT}}\}$ rather than that contained in $\{J_{ij}^{\mathrm{c}}\}$ that determines the new firing pattern to be stored, one may introduce an appropriate information measure, which involves a quenched average over the synaptic efficacies. This is done in a related paper [19], and it is shown to yield important constraints on the relative strengths of external inputs (with respect to intrinsic connections) which are necessary in a real memory of this type to store reasonable amounts of new information. It is, however, the present result which validates the way these constraints are evaluated. If spin-glass effects *were* present, $\{h_i^{\mathrm{EXT}}\}$ and $\{J_{ij}^{\mathrm{c}}\}$ would not fix $\{V_i\}$, as the systems would, depending on initial conditions, end up in one of exponentially many possible configurations $\{V_i\}$. The uncertainty associated with this effective indeterminacy would be reflected, as indeed is the case with a system of binary units, in a drastic decrease in the amount of information that could be stored for each new pattern.

Acknowledgments

This work is part of a larger project carried out with Dr E T Rolls, and supported by grants from the MRC and from the EEC BRAIN initiative. A useful suggestion was given by Dr K Y M Wong. This paper was completed during a stay at the Racah Institute of Physics, in Jerusalem.

Note added. In an interesting paper [20], Hadeler and Kuhn have shown that another threshold-linear coupled system [13], in fact equivalent to the present one, has a unique solution if and only if the matrix $\mathbf{1}-g\mathbf{J}$ is positive definite. Therefore a calculation of its eigenvalues (possibly using the replica trick [21]) might yield results equivalent to those derived here.

References

[1] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30

[2] Little W A 1974 *Math. Biosci.* **19** 101
Little W A and Shaw G L 1975 *Behavioral Biology* **14** 115

[3] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554

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

[5] Crisanti A and Sompolinsky H 1987 *Phys. Rev.* A **36** 4922; 1988 *Phys. Rev.* A **37** 4865

[6] 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

[7] Treves A and Amit D J 1988 *J. Phys. A: Math. Gen.* **21** 3155

[8] Treves A and Rolls E T 1990 What determines the capacity of auto-associative memories in the brain? *Preprint* Department of Experimental Psychology, University of Oxford

[9] Waugh F R, Marcus C M and Westervelt R M 1990 *Phys. Rev. Lett.* **64** 1986

[10] Fukai T and Shiino M 1990 *Phys. Rev.* A **42** 7459

[11] Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801

[12] Treves A 1990 *Phys. Rev.* A **42** 2418

[13] Hartline H K and Ratliff F 1958 *J. Gen. Phys.* **41** 1049

[14] deAlmeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983

[15] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792

[16] Sejnowski T J 1977 *J. Math. Biol.* **4** 303

[17] Treves A 1990 *J. Phys. A: Math. Gen.* **23** 2631

[18] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115

[19] Treves A and Rolls E T 1991 Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network *Preprint* Department of Experimental Psychology, University of Oxford

[20] Hadeler K P and Kuhn D 1987 *Biol. Cybern.* **56** 411

[21] Edwards S F and Jones R C 1976 *J. Phys. A: Math. Gen.* **9** 1595