Competitive attraction in neural networks with sign-constrained weights

1992 J. Phys. A: Math. Gen. 25 2227

(http://iopscience.iop.org/0305-4470/25/8/033)

J. Phys. A: Math. Gen. 25 (1992) 2227–2242. Printed in the UK

Competitive attraction in neural networks with sign-constrained weights

K Y M Wong†‡ and C Campbell§

† Department of Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK
§ Department of Engineering Mathematics, Queen's Building, Bristol University, Bristol BS8 1TR, UK

Received 7 June 1991, in final form 6 December 1991

Abstract. The presence of many attractors in neural networks give rise to interesting competitive phenomena. In this paper we consider dilute recurrent (or attractor) neural networks with sign-constrained weights and storing uncorrelated patterns with maximal stability. The dynamics of these networks is governed by the competitive effects of retrieval, non-retrieval and uniform (i.e. ferromagnetic) attractors, which result in basin encroaching, shrinking, splitting and wedging. We have found the parameter regions in which each of these attractors exist. The basins of attraction of the uniform attractors enlarge at the expense of the other attractors even when the weight signs are slightly imbalanced, but can be compensated by the introduction of a dynamical threshold.

1. Introduction

The purpose of this paper is twofold. First, we study the dynamical properties of attractor neural networks with sign-constrained weights. Secondly, using the sign-constrained network as a particular example, we demonstrate that the presence of many attractors in neural networks can give rise to interesting competitive attractor and transient behaviours.

Neural networks with sign-constrained weights are of interest for biological, technological and cognitive reasons. Biological synapses are either excitatory or inhibitory, and the nature of a synapse is believed to be unchanging [1, 2]. These features can be modelled with sign-constrained weights.

They are also of interest in the construction of neural hardware. The existence of both positive and negative weights in the electronic and optical implementation of neural networks increases the complexity of the circuit. In both cases a solution is to use sign-constrained neural networks (either with all positive weights and a threshold [3] or all negative weights [4, 5]).

Furthermore, the enforcement of weight sign constraints can give rise to an exploitable cognitive feature—namely an ability to distinguish between the recognition or non-recognition of an input pattern [6, 7]. If excitatory synapses are more numerous than inhibitory ones, an initial input distant from a stored pattern will drift

‡ Present address: Department of Physics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

0305-4470/92/082227+16$04.50

towards a uniform state, in which all the neurons are firing (or non-firing) simultaneously. Moreover, if the weight sign imbalance is sufficiently large, only retrieval and uniform attractors are stable. The uniform attractor states can therefore be regarded as the attractors for unrecognized inputs, and can be readily distinguished from retrieval attractors.

Following the above motivations, we have formulated a learning algorithm for perceptrons with sign-constrained weights [8]. This algorithm is guaranteed to converge provided that a solution exists. Furthermore, for uncorrelated patterns, we have found that the storage capacity of this perceptron is always half that of the corresponding unconstrained perceptron, irrespective of the distribution of weight signs [9]. This independence from the weight sign distribution is an example of gauge invariance. It is interesting to explore to what extent the dynamics of recurrent (or attractor) networks is also gauge invariant.

In this paper we will extend this study by considering the dynamics of attractor neural networks with sign-constrained weights. In general, the calculation of the dynamics of neural networks is an intractable problem, because correlations between different time steps cannot be neglected. As the system evolves, the number of correlation parameters grows substantially. Here we shall focus on the solvable case of dilute networks, in which each neuron is fed by C other neurons with $1 \ll \ln C \ll \ln N$, so that correlations beyond one time step are negligible [10]. We will consider the conditions for the existence of attractors, the size of the basins of attraction, the transient behaviour which determines the approach to the fixed points, and the extent of gauge invariance in the dynamics of these models. Preliminary results have already been presented in [11].

It turns out that three types of attractors are possible: retrieval state (i.e. the attractor configuration is a stored pattern), non-retrieval state (i.e. no correlation with the stored patterns), or uniform state (i.e. all 1's or all -1's simultaneously). On varying the storage level and the weight sign bias, the attracting power of these attractors changes, giving rise to complex competitive effects. There exist parameter regions which favour one attractor over the others. The favoured attractor then has a wide basin of attraction at the expense of the others. In some parameter regions, they may even encroach on their neighbouring basins. The transient approach to these attractors is also affected by the competition between attractors. As a result, a basin of attraction may be further divided into valleys under the influence of a strong neighbouring attractor, accompanied by bifurcations of saddle points and repellers along their adjoining basin boundaries.

In fact, this picture of competitive attraction is present in any dynamical system possessing many attractors. Attractor neural networks with weight sign constraints are a particular example exhibiting these effects, but they should also be present in other attractor neural networks. For example, similar features have been observed in a related sign-constrained model [12], the two pattern network [13] and a thresholded asymmetric model [14].

As we shall see, a very small weight sign bias, of the order $O(C^{-1/2})$ (C being the connectivity of a neuron), is sufficient to enlarge the basin of the uniform attractor disproportionately. Weight sign biases of the order $O(1)$ cause the retrieval attractors to narrow their basins drastically, which is an undesirable feature in most neural network implementations. This also shows that the gauge invariance for the static storage of patterns is no longer upheld in the network dynamics.

One way of compensating the effect of the uniform attractors is to introduce a

dynamical threshold to each neuron. The dynamical threshold restrains the average firing rate so that the state does not move towards the uniform states. Interestingly, gauge invariance can be restored to the network dynamics by the use of dynamical thresholds.

We present the dynamical equations in section 2. The conditions governing the stability of the fixed points and saddle points, and hence the existence of the attractors, are discussed in section 3. In section 4 we will describe the geometry of the attractor basins and their accompanying transient behaviours, and the phase diagram of the dynamical behaviour in the space of the storage level and weight sign bias. In section 5 we discuss the issue of gauge symmetry in network dynamics, and the effects of introducing a dynamical threshold. In section 6 we summarize and discuss the implications of our study.

2. Dynamical equations

Consider a randomly diluted asymmetric neural network, in which the neuronal state at node i ($i = 1, \ldots, N$) takes the possible values ± 1, and is fed by C other nodes $j = i_1, \ldots, i_c$ through the synapses J_{ij}. The weight sign constraints can be enforced by requiring $J_{ij} g_{ij} \geqslant 0$ where $g_{ij} = \pm 1$. The subcase $g_{ij} = g_j$ corresponds to enforcing the same sign on all the synapses emanating from each individual neuron. The dynamics of an attractor neural network can be either synchronous or asynchronous. For synchronous dynamics, all the neuronal states are updated according to

$$S_i^{t+1} = \mathrm{sgn}(h_i^t) \qquad h_i^t = \frac{1}{\sqrt{C}} \sum_j J_{ij} S_j^t \tag{2.1}$$

while for asynchronous dynamics, a neuron is randomly chosen at each time step and its state updated acccording to (2.1).

We will consider the network storing p random and unbiased patterns $\{\xi_i^\mu\}$ ($\mu = 1, \ldots, p$). We are interested in the dynamical evolution of a state configuration not only having a macroscopic overlap m^t with a stored pattern (say pattern 1) but also having an overlap a^t with a uniform state ($S_i = +1$ for all i, say). The two relevant dynamical variables are therefore the overlaps with $\{\xi_i^1\}$ and with the positive uniform state (i.e. the activity), respectively given by

$$m^t = N^{-1} \sum_i \xi_i^1 S_i^t \qquad a^t = N^{-1} \sum_i S_i^t. \tag{2.2}$$

Equivalently, the dynamics can be described by two alternative parameters

$$m_\pm^t = N_\pm^{-1} \sum_{\{i|\xi_i^1 = \pm 1\}} \xi_i^1 S_i^t \tag{2.3}$$

i.e. $m_\pm^t$ are the overlaps on those neurons having $\xi_i^1 = \pm 1$ respectively ($N_\pm$ are the number of neurons for which $\xi_i^1 = \pm 1$). Since we have assumed that the stored patterns have independent components and a random distribution, m^t and a^t are related to $m_\pm^t$ by

$$m^t = \tfrac{1}{2}(m_+^t + m_-^t) \qquad a^t = \tfrac{1}{2}(m_+^t - m_-^t). \tag{2.4}$$

The dynamical equations can be obtained by generalizing the single variable iterative equation for dilute networks [15, 16, 17]. For state configurations with a macroscopic overlap m^t with pattern 1, it has been shown that the dynamical equation is completely determined by the distribution of the aligning fields Λ_i^1. Analogously, in the present problem with two dynamical variables $m_\pm^t$, the dynamics is determined by the joint distribution of the two aligning fields

$$\Lambda_{i\pm}^1 = \frac{\xi_i^1}{\sqrt{C_{i\pm}^1}} \sum_{\{j|\xi_j^1=\pm 1\}} J_{ij}\xi_j^1 \tag{2.5}$$

where $C_{i\pm}^1$ is the number of neurons feeding neuron i for which $\xi_j^1 = \pm 1$, and is equal to $C/2$ in the limit $C \gg 1$. They satisfy the equations

$$\frac{1}{\sqrt{2}}\Lambda_{i+}^1 + \frac{1}{\sqrt{2}}\Lambda_{i-}^1 = \Lambda_i^1; \quad \frac{1}{\sqrt{2}}\Lambda_{i+}^1 - \frac{1}{\sqrt{2}}\Lambda_{i-}^1 = \xi_i^1 M_i \tag{2.6}$$

where $M_i \equiv C^{-1/2}\sum_j J_{ij}$ can be considered as the aligning field of the uniform state. For input overlaps $m_\pm^t$, the local field at neuron i is a sum of two Gaussian variables, of means $m_\pm^t \sum_{\xi_j^1=\pm 1} J_{ij}\xi_j^1/\sqrt{C}$ and variance $[1-(m_\pm^t)^2]C_{i\pm}^1/C$. Hence the dynamical equations are

$$\begin{aligned} &\text{synchronous dynamics :} && m_\pm^{t+1} = f_\pm(m_+^t, m_-^t) \\ &\textit{asynchronous dynamics :} && \mathrm{d}m_\pm^t/\mathrm{d}t = f_\pm(m_+^t, m_-^t) - m_\pm \end{aligned} \tag{2.7}$$

where $f_\pm$ are the retrieval functions

$$f_\pm(m_+, m_-) = \int \mathrm{d}\Lambda_+ \mathrm{d}\Lambda_- \rho_\pm(\Lambda_+, \Lambda_-)\mathrm{erf}\left(\frac{m_+\Lambda_+/\sqrt{2} + m_-\Lambda_-/\sqrt{2}}{\sqrt{2(1-m_+^2/2 - m_-^2/2)}}\right) \tag{2.8}$$

and the double field distribution is defined by

$$\rho_\pm(\Lambda_+, \Lambda_-) = \overline{\langle \delta(\Lambda_+ - \Lambda_{i+}^1)\delta(\Lambda_- - \Lambda_{i-}^1)\rangle}_{\xi_i^1=\pm 1} \tag{2.9}$$

where the overbar represents averaging over the fraction of weight space which stores the patterns and complies with the sign constraints, and the angular brackets represent averaging over the stored patterns. Alternatively, the retrieval functions can be expressed in terms of a^t and m^t

$$f_\pm(a, m) = \int \mathrm{d}\Lambda \mathrm{d}M \rho(\Lambda, M)\mathrm{erf}\left(\frac{m\Lambda \pm aM}{\sqrt{2(1-m^2-a^2)}}\right) \tag{2.10}$$

where the joint distribution of Λ and M is defined by

$$\rho(\Lambda, M) = \overline{\langle \delta(\Lambda - \Lambda_i^1)\delta(M - M_i)\rangle}. \tag{2.11}$$

We note immediately that the dynamical equations are invariant under the transformations $a \to -a$ and $m \to -m$. This implies that the dynamical behaviour is symmetric with respect to the a and m axes (i.e. $m = 0$ and $a = 0$ respectively).

For networks with no bias in the weight sign distribution (i.e. the number of positive and negative synapses feeding a neuron are the same), $M = 0$. In this case, the dynamical equation (2.7) is identical to the single parameter equation in [15,16,17] for $a^1 = 0$.

However, the most interesting case is for networks with a bias in the sign distribution given by†

$$B = \frac{1}{\sqrt{C}} \sum_j g_{ij} \qquad \text{for all } i \tag{2.12}$$

i.e. the fractional weight imbalance is of the order $O(C^{-1/2})$. This implies $M \sim O(1)$, and the attracting power of the retrieval and uniform states become comparable.

The above dynamical formulation is completely general, and applies to any synaptic prescription. Here we are interested in the maximally stable network [8], for which all aligning fields are bounded below by a positive stability parameter κ, which is related to the storage level $\alpha \equiv p/C$ by [9]

$$\frac{1}{2\alpha} = \int_{-\infty}^{\kappa} \frac{dt}{\sqrt{2\pi}} e^{-t^2/2} (\kappa - t)^2 \tag{2.13}$$

provided that α is less than the storage capacity $\alpha_c = 1$. For this network, the distribution of M is a delta function peaked at $\sqrt{C} \int dJ P(J) J$, where the weight distribution $P(J)$ for the sign-constrained network was derived in [9], and is given by a truncated Gaussian of width 2 plus a delta function‡

$$P(J) = \Theta(\pm J) \frac{\exp(-J^2/4)}{\sqrt{4\pi}} + \frac{1}{2}\delta(J) \tag{2.14}$$

for excitatory and inhibitory synapses respectively. This gives

$$M = \frac{B}{\sqrt{\pi}} \tag{2.15}$$

and the joint weight distribution $\rho(\Lambda, M)$ is now

$$\rho(\Lambda, M) = \rho(\Lambda)\delta\left(M - \frac{B}{\sqrt{\pi}}\right) \tag{2.16}$$

where the aligning field distribution $\rho(\Lambda)$ is given by a truncated Gaussian plus a delta function peak [15, 17]

$$\rho(\Lambda) = \theta(\Lambda - \kappa) \frac{\exp(-\Lambda^2/2)}{\sqrt{2\pi}} + \frac{1}{2}[1 + \operatorname{erf}(\kappa/\sqrt{2})]\delta(\Lambda - \kappa). \tag{2.17}$$

Substituting (2.16) in (2.10), the retrieval functions become,

$$f_{\pm}(a, m) = \int d\Lambda \rho(\Lambda) \operatorname{erf}\left(\frac{m\Lambda \pm aM}{\sqrt{2(1 - m^2 - a^2)}}\right). \tag{2.18}$$

This is our central result. Its consequences for the attractor structure will be studied in the following sections.

† We have mistakenly written down the wrong scaling in [11].

‡ We have mistakenly missed out the delta function part in [9].

3. Attractor and transient

3.1. Attractors and their stability

For $\alpha < 1$ these dynamical equations possess three types of attractor: (i) *retrieval attractors* $(a,m) = (0,\pm 1)$; (ii) *uniform attractors* $(a,m) = (\pm 1,0)$; (iii) *non-retrieval attractors* $(a,m) = (0,0)$. The stability of these attractors are determined by the eigenvalues of the matrix $\partial g_\pm / \partial m_\pm$ where $g_\pm(m_+,m_-) \equiv f_\pm(m_+,m_-) - m_\pm = 0$. For a stable attractor the two eigenvalues must both be negative. We find that the retrieval attractors are always stable below the storage capacity and the uniform attractors are stable for positive non-zero B. On the other hand the non-retrieval attractor is a stable fixed point for $\alpha \geqslant \alpha^* = 0.21$, and for B positive, $M \leqslant \sqrt{\pi/2}$ or $B \leqslant B^* = 2.22$. α^* is half the corresponding value for unconstrained networks [9], the difference arising from the weight sign constraints.

3.2. Saddle points and their stability

It is also interesting to consider the stability of saddle points which determine the character of transient behaviour. Particularly informative are the a and m saddle points lying on the a and m axes. The a saddle point is located at $(a,m) = (a^*,0)$, where a^* is the unstable fixed point of the equation

$$a^* = \operatorname{erf}\left(\frac{a^* M}{\sqrt{2(1-a^{*2})}}\right). \tag{3.0}$$

It merges with the non-retrieval attractor point when a^* becomes zero for $B \geqslant B^*$. It turns into an unstable fixed point (or a repeller) at low storage levels when

$$2\frac{\exp[-a^{*2}M^2/2(1-a^{*2})]}{\sqrt{2\pi(1-a^{*2})}} \int \mathrm{d}\Lambda \rho(\Lambda)\Lambda > 1. \tag{3.1}$$

The m saddle point is located at $(a,m) = (0,m^*)$, where m^* is the unstable fixed point of the equation

$$m^* = \int \mathrm{d}\Lambda \rho(\Lambda) \operatorname{erf}\left(\frac{m^*\Lambda}{\sqrt{2(1-m^{*2})}}\right). \tag{3.2}$$

It merges with the non-retrieval attractor point when m^* becomes zero for $\alpha \leqslant \alpha^*$. It turns into a repeller at high weight sign bias when

$$2M \int \mathrm{d}\Lambda \rho(\Lambda) \frac{\exp[-m^{*2}\Lambda^2/2(1-m^{*2})]}{\sqrt{2\pi(1-m^{*2})}} > 1. \tag{3.3}$$

The regions of stability of the fixed points and saddle points are shown in figure 1.

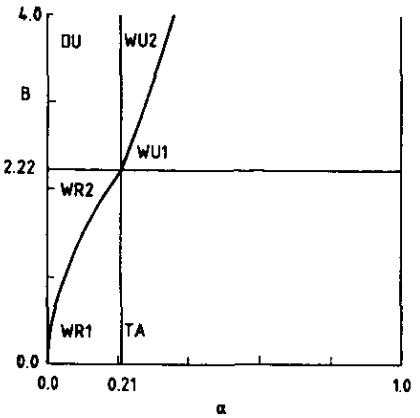

Figure 1. Regions of stability of fixed points and saddle points, and phase diagram in the space of storage level α and weight sign bias B. See section 4 for the names of the phases. The retrieval and uniform attractors are stable in all regions, while the non-retrieval attractor is only stable in region TA. The a saddle point exists in regions WR1 and TA, but turns into a repeller in region WR2. The m saddle point exists in regions WU1 and TA, but turns into a repeller in region WU2.

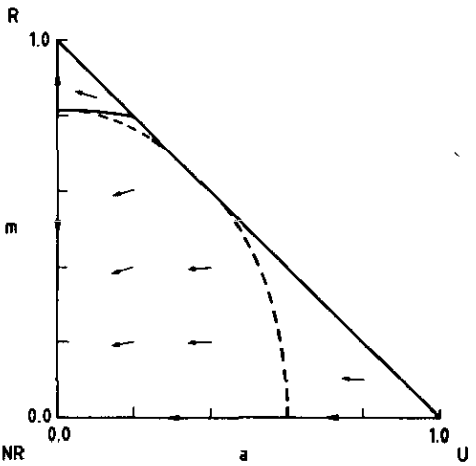

Figure 2. The basin boundary and transient boundaries for $\alpha = 0.3$ and $B = 0$. In figures 2–4, we have restricted the figures to the first quadrant (since the other quadrants are symmetrically similar). The labels R, U and NR represent the retrieval, uniform and non-retrieval fixed points respectively. The arrows on the diagrams show the direction of movement in the parameter space. Basin boundaries are plotted in solid lines, attractor lines in thin lines, transient boundaries $\mathrm{d}m/\mathrm{d}t = 0$ in dashed lines, and $\mathrm{d}a/\mathrm{d}t = 0$ in dotted lines.

4. Basins of attraction and transient behaviour

The shape and location of the basin boundary lines may depend on whether we are considering attractors for synchronous or asynchronous dynamics. However, we have observed no significant difference between the two types of dynamics in the cases of positive B we studied. In figures 2 and 3 we plot the basin boundaries for asynchronous dynamics. Moreover, we have plotted the (unique) *attractor lines* which emanate from a saddle point and converge to an attractor. All flow lines in the basin

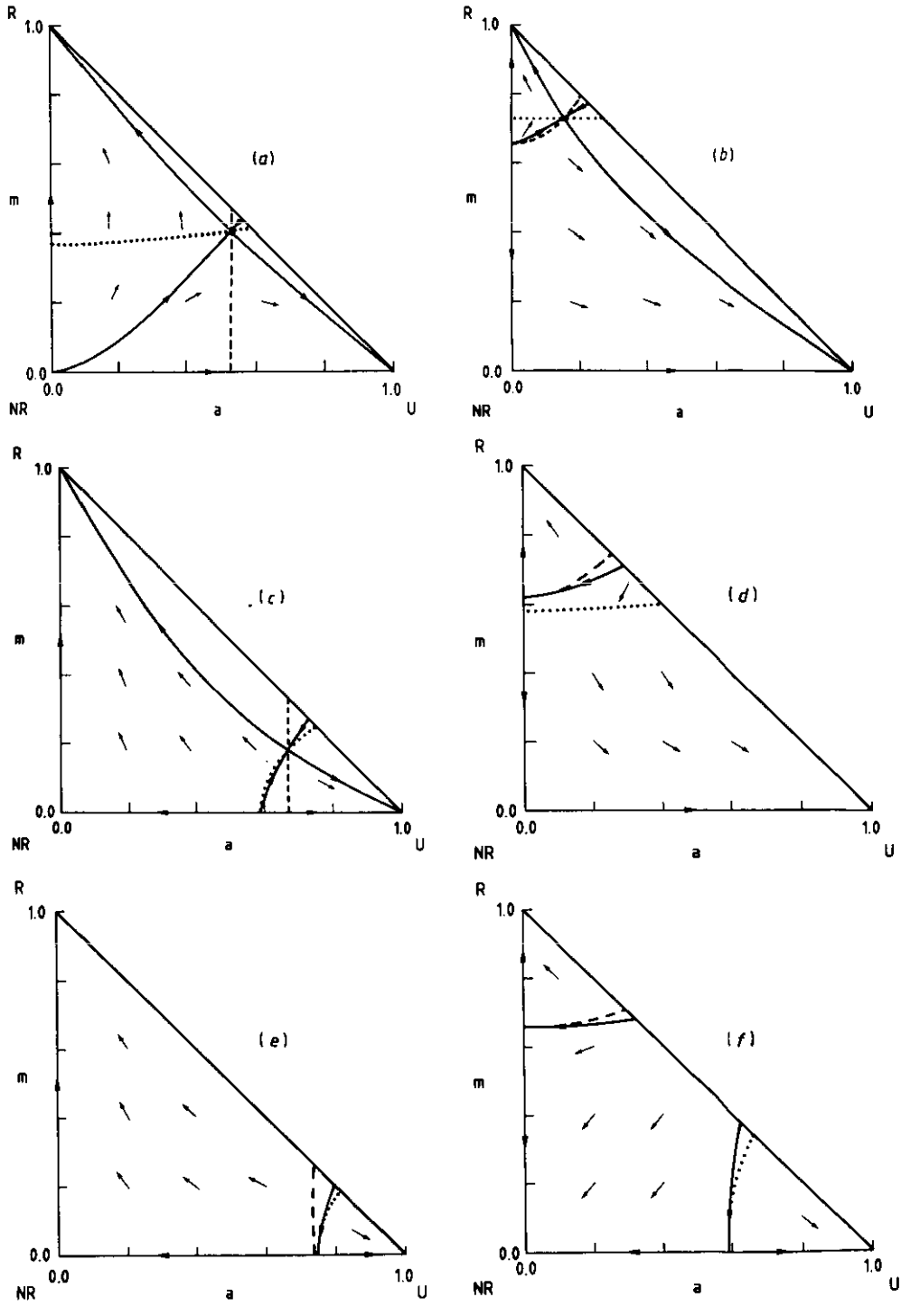

Figure 3. The basin boundary, attractor lines and transient boundaries for $(\alpha, B) = (a)$ (0.15, 2.5); (b) (0.25, 3.0); (c) (0.15, 2.0); (d) (0.25, 2.5); (e) (0.15, 1.8); (f) (0.25, 2.0). These are examples of regions DU, WU2, WR2, WU1, WR1 and TA in figure 1 respectively. (The ordering in layout corresponds to the relative positions in figure 1.)

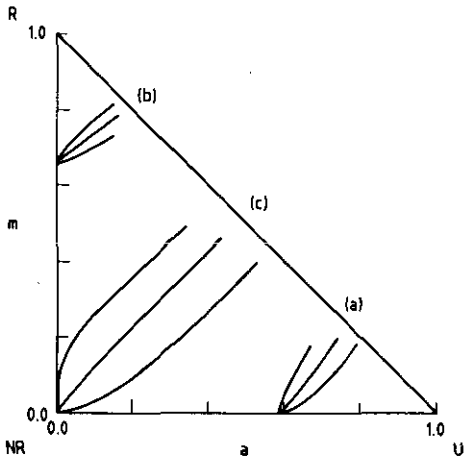

Figure 4. Effects of changing α and B on the basin boundary between the retrieval and uniform attractors in the regions of non-axial attraction. The segments of the basin boundaries plotted here start from the axial-repeller and end at the non-axial saddle point. Starting from the left, $(a)B = 2.0$ and $\alpha = 0.15, 0.12, 0.09$; $(b)B = 2.5$ and $\alpha = 0.20, 0.18, 0.15$; $(c)\alpha = 0.25$ and $B = 4.0, 3.5, 3.0$.

of attraction tend to approach one of these attractor lines on their flow towards the attractor point.

We have also plotted boundaries of transient behaviour. The dashed and dotted lines represent the curves $\mathrm{d}m/\mathrm{d}t = 0$ and $\mathrm{d}a/\mathrm{d}t = 0$ respectively. In both synchronous and asynchronous dynamics, they demarcate four types of transient states: (i) *retrieval transients*—m^t increases and a^t decreases; (ii) *non-retrieval transients*—both m^t and a^t decrease; (iii) *uniform transients*—m^t decreases and a^t increases; (iv) *active transients*—both m^t and a^t increase. Normally, the behaviour near the retrieval, non-retrieval and uniform attractor states (where they exist) are retrieval, non-retrieval and uniform transients respectively. However, in the vicinity of their basin boundaries, other transients are usually present.

There exist points of intersection of the basin boundaries and transient boundaries. Since basin boundaries meet the a and m axes at saddle points or repellers, the conditions $\mathrm{d}a/\mathrm{d}t = 0$ and $\mathrm{d}m/\mathrm{d}t = 0$ must be satisfied at the intersections. Thus these meeting points also lie on the transient boundaries.

Furthermore, there are regions of the parameter space (namely regions DU, WU2 and WR2 in figure 1, see figures 3(*a*)–(*c*)) where the two boundaries of transience $\mathrm{d}a/\mathrm{d}t = 0$ and $\mathrm{d}m/\mathrm{d}t = 0$ meet. In these cases, their meeting point is another saddle point lying off the axes of symmetry, and we have a confluence of the transient and basin boundaries.

4.1. Zero weight bias

When $B = 0$ and $\alpha < \alpha^*$ any state with a small (positive) macroscopic overlap with the stored pattern will inevitably move towards the retrieval fixed point since this is the only attractor present (the only transients present are retrieval transients).

For $\alpha > \alpha^*$ a non-retrieval attractor appears, as illustrated in figure 2. In the neighbourhood of the retrieval and non-retrieval states respectively, retrieval and non-retrieval transients are present. In addition, there exist regions of retrieval transients outside the basin of the retrieval attractor, i.e. there is a small transient increase in m^t

before moving towards the non-retrieval state. The region of retrieval transients may be disconnected in two for higher values of α (figure 2), one in the neighbourhood of the uniform fixed point, and the other in the vicinity of the retrieval attractor. Alternatively, it may be connected for lower α.

4.2. Basin encroaching and shrinking

For positive B the different regions of attractor and transient behaviour are illustrated in figure 1, and the dynamical evolution of each region are shown in figures 3(a)–(f). The transition between these different behaviours can be interpreted as the result of competing attraction of the retrieval and uniform attractors.

In general, the attracting power of an attractor depends on two factors: the storage level α and weight bias B. Low α favours the retrieval attractor, since the aligning fields of the stored pattern is strong, and high B favours the uniform attractor. When an attractor is favoured, its basin of attraction widens at the expense of the others, resulting in the shrinking, or even the disappearing, of the weaker attractors.

For high α and low B, both the retrieval and uniform attractors are not exceedingly strong, allowing for the presence of the non-retrieval attractor. This corresponds to the *triple attractor* region (TA). For the particular example in figure 3(f), there are regions of non-retrieval transients lying within the basin boundaries of the retrieval and uniform attractors. States in these regions undergo an initial small movement away from the attractor before entering the retrieval or uniform transient regions surrounding the attractors. This shows that the non-retrieval attractor is sufficiently strong, influencing not only the flow within its own basin, but also that in its neighbouring basins.

For low α and low B, the retrieval attractor is favoured and encroaches on the non-retrieval attractor. This corresponds to the *wide retrieval* region (WR1 and WR2), where both the retrieval and uniform attractors are stable, but the retrieval attractor has a wider basin of attraction, which includes the whole positive m axis and part of the a axis (Figure 3(c, e)). Similarly, for high α and high B, the uniform attractor is favoured in the *wide uniform* region (WU1 and WU2), as shown in figure 3(b, d).

For low α and high B, both the retrieval and uniform attractors are strong, corresponding to the *duopoly region* (DU). As shown in figure 3(a), both the retrieval and uniform attractors have wide basins of attraction along their respective axes.

4.3. Basin splitting

In regions WR1, WU1 and TA, no repellers are present on the basin boundaries. The axes of symmetry, which connect the saddle and attractor points, are therefore attractor lines. The uniform and retrieval basins can then be considered as a single valley. The active transient is absent, though the other three types of transients occur. These regions can be described as regions of *axial attraction*.

In regions WR2, WU2 and DU, repellers are present on the basin boundaries. Saddle points, and hence the attractor lines, are located off the axes of symmetry. The repeller lies on an axis of symmetry, which can therefore can be considered as a *valley boundary* separating two valleys. When all quadrants of the variable space are taken into account, basins of attraction are further divided into valleys, each associated with an attractor line. All transients are possible. These regions will be called regions of *non-axial attraction*.

We interpret the transition from axial to non-axial attraction, and the accompanying splitting of the basins of attraction into valleys, as a result of the uneven attracting power of the competing attractors in the state configuration space. Because of the symmetry of the dynamical equation (2.18), the dynamical behaviour along the a axis is independent of the aligning field distribution, and hence independent of α; it is entirely determined by the weight bias. Similarly, the dynamical behaviour along the m axis is entirely independent of the weight bias. On the other hand, the dynamical behaviour off the axes of symmetry depends on the attracting power of both the retrieval and uniform attractors.

In the wide retrieval region, for example, consider the transition from the single valley regime (WR1) to the multi-valley one (WR2). For $\alpha < \alpha^*$, the retrieval attractor is generally strong, whereas the uniform attractor is strongest along the a axis, since the dynamics along this axis is independent of α. When B increases, the uniform basin expands along the a axis, pushing the saddle point to a lower value of a^*. However, in the neighbourhood of this point, the attracting power of the uniform attractor rapidly weakens when one moves away from the axis. Thus for sufficiently large B, the a saddle point turns into a repeller, accompanied by a saddle point bifurcating from it.

Alternatively, this picture of competing attraction can be described in terms of a landscape analogy. This is useful although strictly speaking, the flow cannot be described as the gradient of a potential function, since it is not irrotational (i.e. $\partial \dot{m}/\partial a \neq \partial \dot{a}/\partial m$). If we associate a potential barrier with the basin boundary, then the increasing strength of the retrieval attractor off the a axis tends to lower the barrier at the non-axial region. When the potential at a non-axial point on the barrier becomes lower than that at the a saddle point, the a saddle point turns into a local maximum, while the saddle point shifts to a non-axial position. The axial attractor line turns into a valley boundary, and the narrower uniform basin divides into two valleys, whereas the wider retrieval basin divides into three.

Using similar arguments, the transition between WR1 and WR2 can also be achieved by decreasing α at constant B. Furthermore, the transition between WU1 and WU2 is analogous, except that the roles played by the retrieval and uniform attractors are reversed.

4.4. Basin wedging

There is still another competitive effect when an axial saddle point turns into a repeller. Since the dynamical equations are symmetric with respect to the a and m axes, the basin boundaries should also meet the axes at the same inclination on both sides of the axes of symmetry. In particular, when the basin boundaries intersect the axes of symmetry at a saddle point, they should be normal to each other. This is because the only two flow directions that can pass through a saddle point lie along the eigenvectors of the stability matrix, which are normal at this point of symmetry.

However, when the axial saddle point turns into a repeller, the basin boundary is not necessarily normal to the axes of symmetry, since the dynamic flow can emanate from the repeller in all directions. The angle of intersection depends on the relative attracting powers of the neighbouring attractors sharing the basin boundary.

Figure 4 demonstrates the effects of varying α and B in the three regions of non-axial attraction. For example, in the WR2 region in figure 4(a), the retrieval attractor is increasingly favoured when α or B decreases. The basin boundary becomes increasingly inclined in favour of the retrieval basin, squeezing the uniform attractor

into a narrow wedged shape. The behaviour in the WU2 region is analogous in figure 4(*b*), except that the role of the retrieval and uniform attractors are reversed. The same explanation applies to the duopoly region in figure 4(*c*), in which the wedged shape basin is the uniform basin for low B, but the retrieval basin for high B.

4.5. Negative weight bias

In contrast to the case of positive B, the behaviours of synchronous and asynchronous updating are quite different for negative B. For synchronous updating we see that if $B \to -B$ then f_+ and f_- in (2.18) are interchanged. Consequently the uniform state $(a, m) = (\pm 1, 0)$ will be a cyclic attractor of period 2 giving states of all 1's and all -1's at each alternative time step. The approach to this attractor will involve oscillations about the m axis. The retrieval and non-retrieval attractors still exist as in the case of positive weight bias B, but the approaches to attractors will also involve oscillations about the m axis. Apart from this oscillatory behaviour the plots for synchronous updating are identical to their positive B counterparts. Thus there are no non-retrieval attractors for $B \leqslant -B^*$ and $\alpha > \alpha^*$.

For negative B and asynchronous updating the uniform attractors disappear. Thus only the retrieval and non-retrieval attractors exist and the non-retrieval state is stable even for $B \leqslant -B^*$. It is interesting to compare the dynamics of networks with weight biases $\pm B$. The dynamical equations (2.6) imply that at the state (m_+, m_-),

$$\dot{m}_\pm(-B) + m_\pm = \dot{m}_\mp(B) + m_\mp \tag{4.0}$$

which in turn implies that $\dot{m}(B) = \dot{m}(-B)$ and $\dot{a}(B) + \dot{a}(-B) + 2a = 0$. Hence the boundaries of transience $\mathrm{d}m/\mathrm{d}t = 0$ are identical for the two networks, whereas the boundaries $\mathrm{d}a/\mathrm{d}t = 0$ are in general different.

5. Gauge symmetry and dynamical thresholds

There is an important difference between the static storage and the dynamical retrieval of patterns, namely the issue of gauge invariance. If we are merely interested in whether a sign-constrained neural network stabilizes a pattern in one time step when the *correct* pattern is presented, then it has been argued [9] that the storage capacity of the network is the same for all combination of weight signs, *provided* that the stored patterns are random and unbiased. This gauge invariance property is the consequence of the following simple argument. If the weight sign of a synapse J_{ij} is flipped, the network would have the same storage capacity if the p pattern bits ξ_j^μ are all flipped. Since ξ_j^μ are random and unbiased, ξ_j^μ and $-\xi_j^\mu$ have the same probability of occurence in the original pattern ensemble. Hence the pattern-averaged storage capacity is not modified by the flipping of weight signs.

In fact this perceptron storage property also extends to the case of biased patterns. Suppose we consider the case of biased input patterns $\{\xi_j^\mu\}$, with mean activity a, mapped onto output patterns, $\{\zeta^\mu\}$. Furthermore, suppose that the condition for the stability of stored patterns is modified to $\zeta^\mu \sum_j J_j(\xi_j^\mu - a)/\sqrt{C} > \kappa$, then although the *microscopic* gauge invariant argument for unbiased patterns does not hold, there is nevertheless a *statistical* gauge invariant argument which ensures that the storage capacity is independent of the weight bias. This is because when we flip the sign of a weight J_j, the network would have the same storage capacity if the p quantities

$(\xi_j^\mu - a)$ also flip their signs. Now although $-(\xi_j^\mu - a)$ does not have the same microsopic distribution of $(\xi_j^\mu - a)$, they nevertheless have the same statistical mean (i.e. zero) and variance (i.e. $1 - a^2$), which are the only significant quantities in determining the extent of pattern interference in the thermodynamic limit.

The microscopic gauge invariance argument continues to hold for the dynamics of unbiased patterns provided that the initial input state is restricted to unbiased configurations, i.e. $a^t = 0$. This is evident from the dynamical equation (2.6), which is independent of the weight bias M for $a^t = 0$. However, it does not hold when the input states are biased. Indeed, our study has demonstrated that the attractor structure and phase diagrams depend on the weight sign bias. We have seen that the uniform attractor becomes significant for a very small fractional weight sign imbalance, of the order $O(C^{-1/2})$. For fractional weight sign imbalances of the order $O(1)$, the retrieval basin possesses a repeller on the axis, and has an extremely thin wedged shape. Variations with a small activity away from the m axis rapidly drive the system towards the uniform attractors – potentially a very undesirable feature of the network.

In fact, it is possible to compensate for this instability against fluctuations in the activity by introducing a dynamical threshold in the updating function of each neuron. This dynamical threshold restrains the averaged activity so that it does not approach one of the uniform states. It also restores a statistical gauge invariance similar to that for the static storage of biased patterns outlined above. Dynamical thresholds have also been proposed in optical neural networks [3].

The dynamical threshold can be introduced by subtracting λa^t from each S_j^t in the updating function, i.e.

$$S_i^{t+1} = \operatorname{sgn}\left[\frac{1}{\sqrt{C}} \sum_j J_{ij}(S_j^t - \lambda a^t)\right] \tag{5.0}$$

where $a^t \equiv N^{-1} \sum_j S_j^t$ is the instantaneous activity of the network and the quantity $\lambda a^t \sum_j J_{ij}/\sqrt{C}$ can be considered as the dynamical threshold. As a consequence the retrieval functions (2.18) are now modified to

$$f_\pm(a, m) = \int \mathrm{d}\Lambda \rho(\Lambda) \operatorname{erf}\left(\frac{m\Lambda \pm aM(1-\lambda)}{\sqrt{2(1-m^2-a^2)}}\right). \tag{5.1}$$

By putting $\lambda = 1$, we see that the network dynamics becomes independent of M, or the weight bias B, restoring the statistical gauge invariance to the network dynamics. The uniform attractors are entirely eliminated (and the dynamics becomes similar to that outlined in section 4.1).

In general, for an arbitrary value of λ, the dynamically thresholded network has an effective weight bias of $M(1-\lambda)$. Thus by suitably adjusting the value of λ, it is possible to shift the attractor structure of the network to different regions of behaviour in figure 1 for constant α. For example, if the network initially lies in region TA, it is possible to suppress the spurious non-retrieval attractor by introducing a sufficiently negative λ, so that only retrieval and uniform attractors are present. This enables Shinomoto's cognitive feature [6], which was mentioned in section 1, to be implemented for a much wider range of the weight sign bias.

6. Discussion

We have found three types of attractors in dilute networks with a bias in the weight-signs: retrieval, non-retrieval and uniform attractors. The properties of these attractors, their basin sizes and stability are the competitive result of two macroscopic factors: the storage ratio α and weight bias B. This results in the six regions of attractor and transient behaviours in figure 1, which can be described in terms of basin encroaching, shrinking, splitting and wedging as the attractors compete with each other.

Interestingly, this picture of competitive attractor behaviour is also present in the retrieval of two patterns each having variable aligning field strengths [13]. In fact if we consider, in our case, the uniform state as one of stored patterns with $\xi_j^\mu = +1$, then $M = \sum_j J_{ij}/\sqrt{C}$ is exactly the aligning field for this uniform pattern. The only difference between our network and the two-pattern network in [13] is that the aligning field distribution for the stored pattern is not a delta function; otherwise the two systems have identical dynamical equations. Furthermore, their flow diagrams (figure 1(a)–(d) in [13]) are similar to figures 3(a), (c), (e) and (f) respectively, and their phase diagram (figure 3 in [13]) corresponds to figure 1 if we take into account the fact that α decreases with increasing average aligning field.

Another example of competitive effects is present in the model with an asymmetric Hebbian rule [14]. There the attracting power of the attractors depend on the storage level α and the threshold H (instead of the weight bias), and the basins of attraction can be adjusted by tuning these two parameters.

More generally, we believe that similar competitive effects are present in *any* dynamical systems with multiple attractors. In particular, models of attractor neural network associate each stored pattern with an attractor. Consequently, phases of competitive attractor behaviours can be mapped out as the relative strengths of the attractors are varied.

We have also found interesting transient behaviour. The normal behaviour within the retrieval and uniform basins are retrieval and uniform transients respectively. However, there exist non-retrieval transients in the vicinity of the basin boundary alongside this normal behaviour, i.e. the m or a component of the network state first moves away from the fixed point before eventually approaching it. This effect, which is absent in the dynamics of network states with one component, is most marked for states off the axes of symmetry. It is observed in all regions of figure 1 except part of the triple attractor region. In the triple attractor region, the interplay of the attractors result in a rich transient behaviour, which will be reported elsewhere [18]. When the two attractor strengths are stongly imbalanced, active transients (i.e. the a component increasing) also exist in the vicinity of the basin boundary in the regions of non-axial attraction.

These transient behaviours are again manifestations of the competition between the attractors. Strong attractors are not only able to capture large regions of network state within their own basin, but they are also able to modify the transient evolution of network states in their neighbouring basins.

Our model can be easily generalized to the case of perceptron networks with a fraction of sign-constrained weights [19,20]. If s is the fraction of weights which are unconstrained, we arrive at the same retrieval functions (2.18), except that α and M have to be replaced by $\alpha/(1+s)$ and $M(1-s)/\sqrt{1+s}$ respectively. The attractor structure can therefore be obtained by an appropriate rescaling of the axes in figure 1.

The presence of unconstrained weights reduces the aligning field of the uniform state. In the limit $s \to 1$, the network becomes completely unconstrained, and the uniform attractor becomes destabilized.

Finally, we comment on the biological relevance of neural network models with sign-constrained weights. There appears to be no evidence that individual synapses can switch from one type to the other on any timescale. The subcase $J_{ij}g_j \geqslant 0$ corresponds to the same sign constraint for all the efferent synapses belonging to a neuron (i.e. each neuron would be solely excitatory or inhibitory in its effect). This functional unity of individual neurons is similar to Dale's rule [2]. For many neuronal cells such a functional unity appears to be correct. A specific neurotransmitter is released (either excitatory, such as glutamate or aspartate, or inhibitory such as GABA) with a common effect on all follower cells. However, there are exceptions to this picture. Firstly, neuronal cells exist which release multiple transmitters (among invertebrates such neurons have been identified in Aplysia and there is also some evidence for these neurons among vertebrates). Secondly, the sign of a synapse is not determined by the transmitter but by the properties of the receptors on the postsynaptic cell [20] and there exist neurotransmitters which could have different excitatory or inhibitory effects on the postsynaptic cells. Consequently, though a functional unity (either excitatory or inhibitory) is a common feature of most neurons it is not true in complete generality.

Nevertheless, our study of the dynamical properties of sign-constrained networks is still biologically relevant in a number of ways. Our analysis only assumes that the synapses *feeding* a neuron obey a particular distribution of weight signs. The results are independent of any weight sign dependence of the *transmitting* neurons. Besides, inhibitory interneurons [22] play an important role in the brain and their effects on other neurons could be modelled as alterations of thresholds. It is possible that this mechanism could be related to the dynamical thresholds we have discussed in section 5. Finally, the attractor structures we have found are still present in a rescaled phase diagram even if the synapses are only partially sign-constrained, showing that such features are quite universal.

Acknowledgments

We thank Eytan Domany for discussions and Edmund Rolls for comments on the discussion in section 7. We especially thank David Sherrington for discussions and encouragement in revising the manuscript. This work is partially supported by the McDonnell–Pew Centre of Cognitive Neuroscience. Computation facilities are partially provided by the University of London Computer Centre. Financial support is provided by the Science and Engineering Research Council of UK.

References

[1] Eccles J C 1964 *The Physiology of Synapses* (Berlin: Springer)
[2] Dale H H 1935 *Proc. Roy. Soc. Med.* **28** 319
[3] Jang J-S, Jung S-W, Lee S-Y and Shin S-Y 1988 *Optics Lett.* **13** 248
[4] Shouval H, Shariv I, Grossman T, Friesem A A and Domany E 1991 *Int. J. Neural Systems* **1** 355
[5] Shariv I and Friesem A A 1989 *Optics Lett.* **14** 485
[6] Shinomoto S 1987 *Biol. Cybern.* **57** 1977

[7] Campbell C and Karwatzki J M 1989 *Neuro '89* (EC2, Nimes) 41
[8] Amit D J, Wong K Y M and Campbell C 1989 *J. Phys. A: Math. Gen.* **22** 2039
[9] Amit D J, Campbell C and Wong K Y M 1989 *J. Phys. A: Math. Gen.* A **22** 4687
[10] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[11] Campbell C and Wong K Y M 1990 *Statistical Mechanics of Neural Networks (Lecture Notes in Physics 368)* ed L Garrido (Berlin: Springer) 237
[12] Kohring G A 1989 Co-existence of global and local attractors in neural networks *Preprint* Bonn University
[13] Pázmándi F and Geszti T 1989 *J. Phys. A: Math. Gen.* **22** 5117
[14] Gardner E, Mertens S and Zippelius A 1989 *J. Phys. A: Math. Gen.* **22** 2009
[15] Kepler T and Abbott L F 1988 *J. Physique* **49** 1657
[16] Krauth W, Nadal J-P and Mezard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
[17] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[18] Wong K Y M and Campbell C 1991 Competitive transients in attractor neural networks *in preparation*
[19] Kanter I and Eisenstein E 1990 *J. Phys. A: Math. Gen.* **23** L935
[20] Nadal J-P 1990 *Network* **1** 463–6
[21] Kandel E R 1976 *Cellular Basis of Behaviour* (San Francisco: Freeman)
[22] Treves A and Rolls E T 1990 *Statistical Mechanics of Neural Networks (Lecture Notes in Physics 368)* ed L Garrido (Berlin: Springer) p 81