Pattern selectivity in neural networks as a means of understanding basin structures

1993 J. Phys. A: Math. Gen. 26 2901

(http://iopscience.iop.org/0305-4470/26/12/027)

J. Phys. A: Math. Gen. 26 (1993) 2901–2919. Printed in the UK

Pattern selectivity in neural networks as a means of understanding basin structures

A Rau†, K Y M Wong‡ and D Sherrington†

† Department of Physics, University of Oxford, Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, UK

‡ Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Received 16 February 1993

Abstract. We study the problem of learning and retrieving for a pair of correlated patterns within an extensive number of uncorrelated patterns, for networks where learning may be treated as an optimization process with respect to an arbitrary cost function. This formulation is then applied to several specific examples, where we study the pattern *selectivity* of these systems (i.e. their ability to differentiate correlated patterns) and investigate the process of basin shrinking with increasing loading levels. We define and discuss several different retrieval phases, whose existence depends on the competitive interplay of the loading level and the pattern correlation. Discussion of asymptotic retrieval is restricted to dilute asymmetric networks.

1. Introduction

Neural networks consisting of simple processing units are able to perform various computational tasks such as generalization, categorization, optimization, learning or functioning as associative memory [1]. Specifically, the retrieval of stored binary patterns in associative memories has been studied in great depth. It is based on the notion that cognitive acts are identified with the network state approaching an attractor in the long-term behaviour.

In the earliest works the *concepts* or patterns to be memorized were stochastically uncorrelated or unbiased. Even so, such patterns give rise to competing instructions during the learning and retrieving stages of the network, especially when the network is required to store many patterns. On the other hand, there are many applications in which *correlated* patterns are stored in neural networks, and competitive effects are expected to be even more serious.

It has been demonstrated that in the Hopfield model of neural networks [2, 3], the presence of competition between the stored patterns can give rise to spurious states which plague the dynamics of the system. When two patterns are correlated with each other, there may exist a *confused* state in which the system can retrieve the common features of the patterns, but be unable to differentiate the individual patterns. This is quite a universal feature, and has been demonstrated in both fully connected [4] and dilute asymmetric [5] neural networks.

However, although correlated patterns are easily confused in Hopfield networks in which the learning rule is the Hebb rule, it is interesting to study whether the selectivity of systems (i.e. their ability to differentiate correlated patterns) can be increased by considering a broader class of learning rules. It has been shown that the conventional Hebb rule

0305-4470/93/122901+19$07.50

is equivalent to optimizing an appropriate cost function [6,7]. In the present work we generalize the treatment of correlated patterns to other learning rules which can be considered as the result of the application of optimization procedures, thereby clarifying the notion of pattern selectivity in attractor neural networks (ANN) for different learning rules—a problem highly relevant to pattern recognition applications.

Depending on the application of the network, one may require high or low pattern selectivity. When one wants to retrieve single stored patterns out of a set of possibly correlated ones, networks with high selectivity are needed. On the other hand, there are situations which require patterns to be mixed, so that more complex computations can be carried out [8,9].

The analysis of pattern selectivity for different learning rules allows one to study the differences in structure of attractor basins. The Hebb rule, for example, has a low pattern selectivity [5], which may be attributed to its wide but imperfect (with respect to the pattern to be retrieved) basin structures. Confusion can be understood as the interference of basins, which is more likely in the case of wide basins. On the other hand, we will demonstrate that the maximally stable network (MSN) [10] and the pseudo-inverse network (PIN) [11, 12] have high pattern selectivities, which may be attributed to deep perfect basin structures.

The treatment of the retrieval dynamics in networks storing arbitrarily many correlated patterns with distributed (i.e. not constant) [13] stabilities may be quite involved. We shall therefore restrict ourselves to examples with only two non-orthogonal patterns among an extensive number of uncorrelated ones, in which case we can derive a general formalism to treat *optimized* networks, and show for specific choices of the cost functions that the iterative maps for the overlaps in an asymmetrically highly diluted network yield a very rich behaviour. Preliminary results have been presented in [14].

Our paper is structured in the following way: in section 2 we derive iterative maps for the two correlated patterns, which are determined by appropriate aligning field distributions in the network. We discuss the role of learning as an optimization process and give a general method of deriving the coupled aligning field distributions, built on our work on uncorrelated patterns [7]. In section 3 we analyse the iterative maps for the Hebb rule, the pseudo-inverse rule and the maximally stable network with tools from non-linear dynamics. In section 4 we summarize our work and point out possible directions for future research.

2. The model

2.1. The general dynamics

We consider a network of N neurons, each being fed on average by C others through the synaptic connections $\{J_{ij}$ from neuron j to $i\}$. The network is trained to store $p \equiv \alpha C$ binary patterns $\xi_j^\mu = \pm 1$, where $j = 1, \ldots, N$ and $\mu = 1, \ldots, p$. Two patterns, say 1 and 2, are correlated with $\boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^2 = Q$, where we use a normalized scalar product definition: $\boldsymbol{a} \cdot \boldsymbol{b} \equiv \sum_i a_i b_i / N$. The other patterns are uncorrelated among themselves and with the first two. In order to make our model more tractable and to gain insight into the domains of attraction, we will restrict our analysis to a highly and asymmetrically diluted network [5] in the thermodynamic limit with $N \to \infty$, $C \to \infty$ and $\lim_{N\to\infty} \ln N / \ln C = \infty$, so that the retrieval behaviour can be readily analysed in terms of iterative coupled maps of two macroscopic order parameters.

We consider synchronous dynamics (where all neurons are updated at the same time) or asynchronous dynamics (where the neurons are randomly chosen and updated) according

to the updating rule

$$S_i(t+1) = \operatorname{sgn}\left(\frac{1}{\sqrt{C}}\sum_j J_{ij}S_j(t) + T\tau(t)\right) \tag{1}$$

where $\tau(t)$ is a Gaussian random variable of mean 0 and width 1. T is a measure of the strength of the stochastic noise (temperature).

As a first step we generalize the one-step dynamics derived earlier [15–17] for a network whose configuration has a macroscopic overlap with only one pattern.

We are interested in the dynamics of network states having a macroscopic overlap with two correlated patterns, 1 and 2. Our system can then be described in terms of the two (overlap) order parameters m_s and m_d

$$m_s \equiv \frac{1}{NQ_+}\sum_{j\in\mathcal{S}} S_j\xi_j^1 \qquad m_d \equiv \frac{1}{NQ_-}\sum_{j\in\mathcal{D}} S_j\xi_j^1 \tag{2}$$

where $Q_\pm \equiv (1 \pm Q)/2$ and the vector $\boldsymbol{S}$ describes the state of the system at a certain instance; here we have introduced the notion of two sets of sites, one, labelled $\mathcal{S}$ (for 'same'), for the sites j where $\xi_j^1 = \xi_j^2$ and another, $\mathcal{D}$ (for 'different'), for the sites j where $\xi_j^1 = -\xi_j^2$. Subscripts s/d are used to refer to macroscopic parameters associated with the two sets $\mathcal{S}/\mathcal{D}$ respectively.

For synchronous dynamics, we can express the average order parameters (2) at time $t+1$ by

$$m_{s/d}(t+1) = \langle\langle \operatorname{sgn}(h^1(t) + T\tau(t))\rangle_\tau\rangle_{\boldsymbol{S}(0),\mathcal{S}/\mathcal{D}} \tag{3}$$

where $\langle\ldots\rangle_\tau$ indicates the average over the retrieval noise, $h_i^\mu(t)$ is the local field at site i relative to ξ_i^μ, defined by $h_i^\mu(t) \equiv \xi_i^\mu \sum_j' J_{ij}S_j(t)/\sqrt{C}$ where the prime on the summation signifies that it is restricted to sites feeding that under consideration, and $\langle\ldots\rangle_{\boldsymbol{S}(t),\mathcal{S}/\mathcal{D}}$ denotes the average over the configurations $\boldsymbol{S}(t)$ for sites i within the sets $\mathcal{S}/\mathcal{D}$ commensurate with $m_{s/d}(t)$. We assume that the site average can be replaced by an ensemble average over the patterns and in (3) and hereafter we omit the site subscript i.

Since the $S_j(t)$ are independent random binary variables, we can apply the central limit theorem. One can easily see then that h^1 is a Gaussian distributed variable with mean (signal)

$$\mu \equiv \sqrt{Q_+}\Lambda_s m_s(0) + \sqrt{Q_-}\Lambda_d m_d(0) \tag{4}$$

and variance (noise)

$$\sigma^2 \equiv 1 - Q_+(m_s(0))^2 - Q_-(m_d(0))^2 + T^2 \tag{5}$$

where we have introduced the normalized aligning fields $\Lambda_{s/d} \equiv \sum_{j\in\mathcal{S}/\mathcal{D}}' \xi^1 J_j\xi_j^1/\sqrt{CQ_\pm}$ for the two sets of sites.

With an appropriate choice of $\{J_{ij}\}$, the iterative dynamics will result in a movement of the state configuration of the network towards an attractor. The attractor may be of retrieval, non-retrieval or spurious character. Using the feature of a highly diluted network that the state correlations between different time steps remain microscopic [5], the above analysis

suffices for any two immediately sequential time steps and we get the dynamical equation for synchronous dynamics

$$m_{s/d}(t+1) = f_{s/d}(m_s(t), m_d(t))$$

$$\equiv \int \mathrm{d}\Lambda_s \, \mathrm{d}\Lambda_d \, \rho_{s/d}(\Lambda_s, \Lambda_d) \int \mathrm{D}\tau \int \frac{\mathrm{d}h^1}{\sqrt{2\pi}\sigma} \operatorname{sgn}(h^1 + T\tau)$$

$$\times \exp\left[-\frac{1}{2}\left(\frac{h^1 - \mu(t)}{\sigma(t)}\right)^2\right]$$

$$= \left\langle \operatorname{erf}\left(\frac{\sqrt{Q_+}\Lambda_s m_s(t) + \sqrt{Q_-}\Lambda_d m_d(t)}{\sqrt{2(1 - Q_+ m_s^2(t) - Q_- m_d^2(t) + T^2)}}\right)\right\rangle_\Lambda \tag{6}$$

where $Dx \equiv \mathrm{d}x \exp(-x^2/2)/\sqrt{2\pi}$ and $\langle \ldots \rangle_\Lambda$ is a shorthand for the Λ average over the distribution of aligning fields

$$\rho_{s/d}(\Lambda_s, \Lambda_d) = \left\langle \delta\left(\Lambda_s - \sum_{j\in\mathcal{S}}{}' \xi^1 J_j \xi_j^1 / \sqrt{C Q_+}\right) \delta\left(\Lambda_d - \sum_{j\in\mathcal{D}}{}' \xi^1 J_j \xi_j^1 / \sqrt{C Q_-}\right)\right\rangle_{\xi, \mathcal{S}/\mathcal{D}} \tag{7}$$

$\langle \ldots \rangle_{\xi, \mathcal{S}/\mathcal{D}}$ indicates an average with respect to the quenched patterns for the output sites in the sets $\mathcal{S}/\mathcal{D}$ respectively. The corresponding dynamical equations for asynchronous dynamics are

$$\frac{\mathrm{d}m_{s/d}}{\mathrm{d}t} = f_{s/d}(m_s, m_d) - m_{s/d} \tag{8}$$

and lead to the same fixed point structure as (6).

2.2. *The aligning field distributions*

Learning involves the modification of the synapses J. It can be effected by dynamical processes leading to the minimization of a cost function $E(J)$ in the space of all possible $\{J\}$. For example, for a specified set of patterns $\{\xi^\mu\}$, $E(J)$ may be minimized explicitly by means of a 'simulated annealing' procedure in which $E(J)$ is considered as a 'Hamiltonian' in the space of $\{J\}$, and a 'thermal distribution' at an effective annealing temperature T_a is achieved by an appropriate algorithm, such as the Langevin dynamics

$$\frac{\partial J}{\partial t} = -\nabla_J E(J) + \eta(t) \tag{9}$$

where $\eta(t)$ is white noise whose variance scales as T_a, and T_a is taken gradually to zero.

We are interested in statistically relevant properties and thus in averages over the specific pattern sets. We shall also concern ourselves only with separable cost functions which can be expressed as

$$E(J) = -\sum_\mu g(\Lambda^\mu) \tag{10}$$

where the $g(\Lambda)$ are performance functions (to be maximized, hence the minus sign in (10)) and the Λ^μ are the aligning fields

$$\Lambda^\mu \equiv \frac{1}{N}\sum_j{}' \xi^\mu J_j \xi_j^\mu. \tag{11}$$

We also restrict discussions to the case of synapses constrained only by having real values and satisfying the spherical rule

$$\sum_j{}' J_j^2 = C. \tag{12}$$

Wong and Sherrington [7, 18] have shown how the results of statistical averaging over systems minimizing such separable $E(J)$ can be obtained analytically via an extension of the replica method of Gardner and Derrida [10, 19]. The original work on uncorrelated patterns [7, 18] can be generalized to the present case. An outline of the new aspects of the analysis is given in appendix A.

As long as the number of correlated patterns is finite (does not scale with C), the overall aligning field distribution

$$\rho(\Lambda) = \left\langle \delta\left(\Lambda - \frac{1}{\sqrt{C}}\xi \sum_j{}' J_j \xi_j\right)\right\rangle_\xi \tag{13}$$

is unaffected by the correlated patterns in the thermodynamic limit and is given by [7, 18]

$$\rho(\Lambda) = \int \mathrm{D}t\,\delta(\Lambda - \lambda(t)) \tag{14}$$

where $\lambda(t)$ is the value of λ which maximizes $g(\lambda) - (\lambda - t)^2/2\gamma$ and γ is given implicitly by

$$\alpha^{-1} = \int \mathrm{D}t\,(\lambda(t) - t)^2 \tag{15}$$

where $\alpha = p/C$, p being the number of patterns stored. The joint aligning field distributions $\rho_{s/d}(\Lambda_s, \Lambda_d)$ associated with the correlated patterns are given by

$$\rho_{s/d}(\Lambda_s, \Lambda_d) = \int \mathrm{D}t_s \mathrm{D}t_d\,\delta(\Lambda_s - \lambda_s(t_s, t_d))\delta(\Lambda_d - \lambda_d(t_s, t_d)) \tag{16}$$

where $\lambda_s(t_s, t_d)$ and $\lambda_d(t_s, t_d)$ are given through maximizing the following function

$$F = g\left(\sqrt{Q_+}\lambda_s + \sqrt{Q_-}\lambda_d\right) + g\left(\pm\sqrt{Q_+}\lambda_s \mp \sqrt{Q_-}\lambda_d\right) - \frac{1}{2\gamma}(\lambda_s - t_s)^2 - \frac{1}{2\gamma}(\lambda_d - t_d)^2 \tag{17}$$

and the '$\pm, \mp$' refer to the construction of $\rho_{s/d}$ respectively. The stationarity criterion for the maximum reads

$$\begin{aligned} t_s &= \lambda_s - \gamma\sqrt{Q_+}\left[g'\left(\sqrt{Q_+}\lambda_s + \sqrt{Q_-}\lambda_d\right) \pm g'\left(\pm\sqrt{Q_+}\lambda_s \mp \sqrt{Q_-}\lambda_d\right)\right] \\ t_d &= \lambda_d - \gamma\sqrt{Q_-}\left[g'\left(\sqrt{Q_+}\lambda_s + \sqrt{Q_-}\lambda_d\right) \mp g'\left(\pm\sqrt{Q_+}\lambda_s \mp \sqrt{Q_-}\lambda_d\right)\right]. \end{aligned} \tag{18}$$

The above result (15) is based on the assumption that the volume of maximum performance (or minimum energy) shrinks to a point in the space of interactions as the annealing temperature is reduced to zero. For certain performance functions, however, there exist some parameter regions in which the volume of maximum performance does not shrink to a point at zero temperature, and this corresponds to the network being below its critical storage capacity. Criticality is characterized by the condition that $\gamma \to \infty$. All the cases of criticality discussed explicitly below refer to this case.

2.3. Specific cost functions

In this paper, we shall concentrate on three types of network which have received considerable attention recently [20]. The performance functions corresponding to these networks have been discussed in [21].

The first network of interest is referred to as the Hebbian network. Its performance function is $g^{\text{Hebb}}(\Lambda) = \Lambda$ and for uncorrelated patterns it leads to the same Gaussian distributed aligning field [7], as the more conventional Hopfield–Hebb rule $J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu / \sqrt{\alpha N}$.

The second kind of network is one constructed by maximizing the performance function

$$g^{\text{PIN}}(\Lambda) = -\tfrac{1}{2}(\Lambda - \kappa)^2 \tag{19}$$

where κ is a freely adjustable parameter. If we increase κ to its maximum value κ_{PIN} such that error-free retrieval is still possible, the aligning field distribution of the network takes the same form as the one generated by the pseudo-inverse rule, which is characterized by constant stabilities. Hence we refer to this as the pseudo-inverse network (PIN). We restrict discussion to $\kappa = \kappa_{\text{PIN}}$ where the parameter $\gamma \to \infty$.

In the third kind of network, we consider a cost function which ensures that all aligning fields are larger than a certain value κ

$$g^{\text{MSN}}(\Lambda) = \Theta(\Lambda - \kappa). \tag{20}$$

In this case, if we increase κ to its maximum value κ_{MSN} such that error-free retrieval is still possible the aligning field distribution of the network takes the same form as the one generated by the maximum stability rule. Hence we refer to it as the maximally stable network (MSN). Again we restrict discussion to criticality, $\kappa = \kappa_{\text{MSN}}$.

3. Analysis of specific iterative maps

Asymptotically in time the system reaches an attractor or limit cycle. The attractors are given by the stable fixed points $m^*_{s/d}$ of the iterative maps which satisfy the equations

$$m_s^* = f_s(m_s^*, m_d^*) \qquad m_d^* = f_d(m_s^*, m_d^*). \tag{21}$$

There are three structurally different types of attractors and corresponding phases: (i) *retrieval attractors* with both $m_s^* \neq 0$ and $m_d^* \neq 0$; (ii) *non-retrieval attractors* with $(m_s^*, m_d^*) = (0, 0)$; (iii) *confused attractors* with $m_s^* \neq 0$ but $m_d^* = 0$.

In fact, the dynamical equations (21) are invariant under the following transformations,

$$\begin{pmatrix} m_s \\ m_d \end{pmatrix} \to \begin{pmatrix} -m_s \\ m_d \end{pmatrix} \qquad \begin{pmatrix} m_s \\ m_d \end{pmatrix} \to \begin{pmatrix} m_s \\ -m_d \end{pmatrix} \tag{22}$$

as expected from the fact that the labelling of the two patterns and their complements are arbitrary for unbiased patterns. It is therefore sufficient to analyse the iterative maps in the first quadrant, where $0 \leqslant m_{s/d} \leqslant 1$. The following results are given for this sector.

3.1. The Hebb rule

Let us first consider the Hebb rule. Substituting $g^{\text{Hebb}}(\Lambda)$ into (18), we find

$$\rho_{s/d}^{\text{Hebb}} = \frac{1}{2\pi} \exp\left\{ -\frac{\Lambda_{d/s}^2}{2} - \frac{1}{2}\left(\Lambda_{s/d} - 2\sqrt{Q_{\pm}/\alpha}\right)^2 \right\} \tag{23}$$

which leads to iterative maps of the form

$$f_{s/d}^{\text{Hebb}}(m_s, m_d) = \text{erf}\left(\frac{\sqrt{2} Q_{\pm} m_{s/d}}{\sqrt{\alpha(1+T^2)}} \right). \tag{24}$$

As these maps decouple their analysis is straightforward, yielding the three phases: (i) retrieval phase for $\alpha < \alpha_-$; (ii) confused phase for $\alpha_- < \alpha < \alpha_+$; (iii) non-retrieval phase for $\alpha > \alpha_+$, where $\alpha_{\pm}(Q, T) \equiv 8Q_{\pm}^2/[\pi(1+T^2)]$. In both retrieval and confused phases the basins of attraction are maximal (that is they cover all of m-space).

Figure 1 shows the phase diagram for the Hebb rule at $T = 0$. The effect of a finite temperature is, in this case, only a rescaling of the phase boundaries by the factor $[1+T^2]^{-1}$. We see that for α near $2/\pi$ a small correlation Q between the patterns is already sufficient to create confusion. Therefore the pattern selectivity of such a system is poor, which supports the idea that the Hebb rule leads to wide but imperfect basins of attraction.

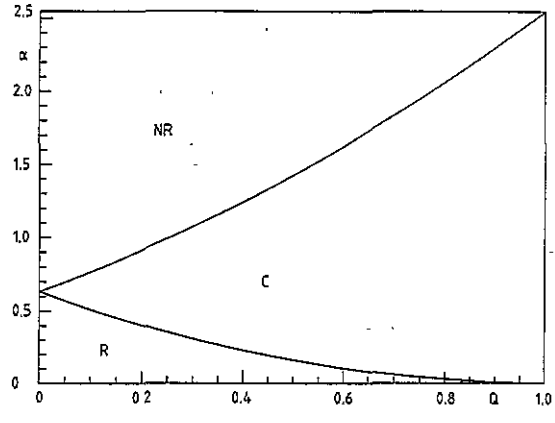

Figure 1. Phase diagram of the Hebbian network in the $Q-\alpha$ space. Phases as described in the text.

The iterative maps for the case of $T = 0$ and for Glauber dynamics at finite temperatures have been derived previously by Derrida *et al* [5]. Fontanari and Köberle [4] analysed the retrieval properties of a fully connected network trained with the Hebb rule and found qualitatively the same results; with the onset of the confused phase occurring for slightly larger $Q \approx 0.3$ and slightly lower $\alpha/\alpha_c \approx 0.7$.

3.1.1. Selectivity for highly correlated patterns. It is interesting to investigate the network's behaviour when it stores patterns which are strongly correlated. This question is equivalent to the issue of where the phase line $\alpha_{\pm}(Q, T=0)$ ends for Q close to 1. Let us consider storing just two patterns (i.e. $\alpha = 2/C$) and determine how similar these two patterns may be and yet still be distinguishable by the network.

Assume the two patterns differ on average at d sites, i.e. $Q = 1 - 2d/C$. The number of input sites in the set $\mathcal{D}$ then follows a Poisson distribution. Since one can easily show

that $m_s^* = 1$ is a stable fixed point of the dynamics, the iterative map of interest with respect to possible confusion is f_d given by

$$f_d(m_d) = \left\langle \operatorname{sgn}\left(\sum_{j\in\mathcal{D}}{}' \xi_j^1 S_j\right)\right\rangle_S. \tag{25}$$

We know that $S_j = \pm\xi_j^1$ with probability $(1 \pm m_d)/2$. Thus for $d \ll C$ we can write

$$f_d(m_d) = \sum_{n=0}^{\infty} \frac{\mathrm{e}^{-d}}{n!} d^n \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1-m_d}{2}\right)^k \left(\frac{1+m_d}{2}\right)^{n-k} \operatorname{sgn}(n-2k). \tag{26}$$

One can see that $f_d(0) = 0$ so that $m_d = 0$ is a fixed point of the dynamics. For it to be a *stable* fixed point the following constraint has to be fulfilled

$$\left.\frac{\partial f_d}{\partial m_d}\right|_{m_d=0} < 1. \tag{27}$$

A brief mathematical analysis yields that for $d < 1.849$ or $Q > 1 - 3.70/C$ this condition is not fulfilled and the Hebb network is unable to distinguish the two patterns. This result is the same as that in Boolean networks for high correlations Q [22].

3.2. The pseudo-inverse rule

As a second example we will consider the pseudo-inverse performance function. In order to invert (18), first we note that $\partial g^{\mathrm{PIN}}(\Lambda)/\partial\Lambda = -\Lambda + \kappa$. This implies that the function $\lambda(t)$ is given by $\lambda(t) = (t + \gamma\kappa)/(1 + \gamma)$, and approaches a single valued function $\lambda(t) = \kappa_{\mathrm{PIN}}$ at criticality, $\gamma = \infty$. This corresponds to an aligning field distribution of constant stability and from (15), we find that α is related to κ_{PIN} via $\kappa_{\mathrm{PIN}} = \sqrt{\alpha^{-1} - 1}$ for $\alpha < 1$. The field distribution (16) can thus be derived,

$$\rho_{s/d}^{\mathrm{PIN}} = \delta\left(\Lambda_{s/d} - \frac{\kappa_{\mathrm{PIN}}}{\sqrt{Q_\pm}}\right)\delta(\Lambda_{d/s}) \tag{28}$$

and the iterative maps reduce to

$$f_{s/d}^{\mathrm{PIN}}(m_s, m_d) = \operatorname{erf}\left(\frac{\kappa_{\mathrm{PIN}} m_{s/d}}{\sqrt{2(1 - Q_+ m_s^2 - Q_- m_d^2 + T^2)}}\right). \tag{29}$$

If one does not insist on the criticality condition $\gamma \to \infty$, it is possible to increase κ beyond κ_{PIN}. Although this introduces error in the aligning field distribution, it has been shown that the storage capacity of the attractor network can be increased, and imperfect retrieval can be extended to a loading level of $\alpha \approx 1.08$ [21]. Nevertheless, we restrict our analysis to the regime of $\kappa = \kappa_{\mathrm{PIN}}$, i.e. $\alpha < 1$.

In the following we restrict analysis of (29) to $T = 0$. The retrieval attractor $(m_s, m_d) = (1, 1)$ is stable for all values of Q for $\alpha < 1$. To consider the stability of the non-retrieval fixed point $(m_s, m_d) = (0, 0)$, we perform a series expansion of (29) around it. We obtain

$$f_{s/d} = \sqrt{\frac{2}{\pi}}\kappa m_{s/d} + \sqrt{\frac{2}{\pi}}\kappa\left(\frac{1 \pm Q}{4} - \frac{\kappa^2}{6}\right) m_{s/d}^3 + \sqrt{\frac{2}{\pi}}\kappa\left(\frac{1 \pm Q}{4}\right) m_{s/d} m_{d/s}^2 + \cdots. \tag{30}$$

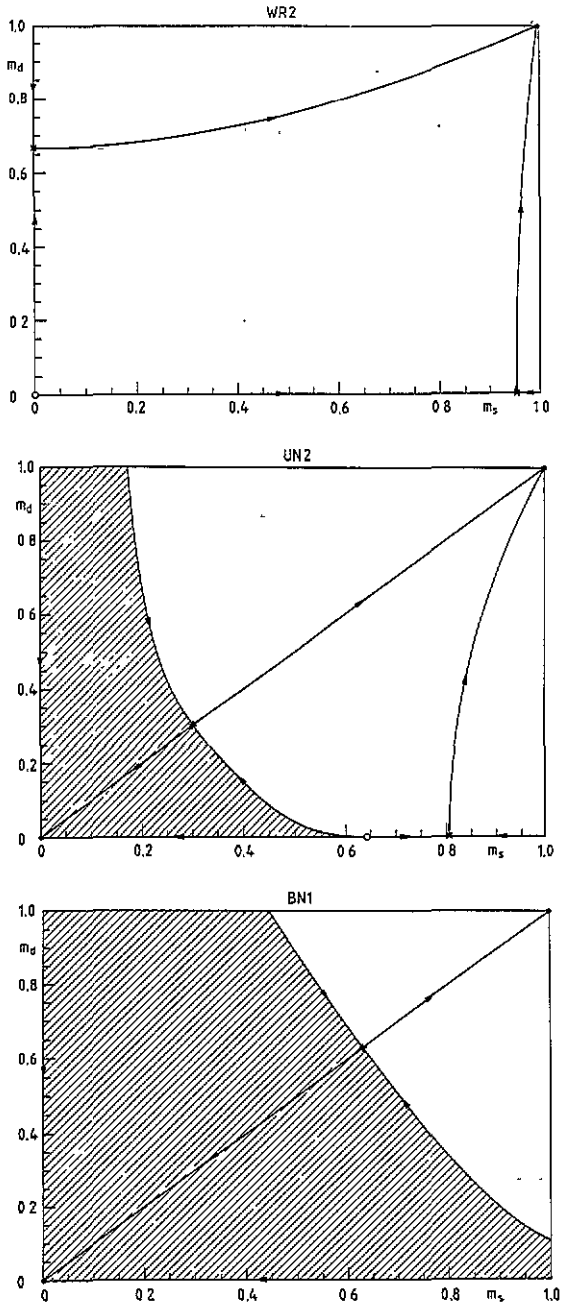

Figure 2. Typical flow diagrams in the m_s–m_d space for the PIN; we show only the first quadrant. The flow lines correspond to basin boundaries and valley bottoms. In figures 2 and 5, circles refer to repellors, crosses to saddle points and full dots to attractors. The basin of the non-retrieval attractor is shaded, and that of the confused attractor is dotted.

The first-order term implies that the non-retrieval fixed point is an attractor for $\sqrt{2/\pi}\kappa < 1$, i.e. for $\alpha > \alpha_b \equiv 2/(2+\pi) \approx 0.39$; and for $\alpha < \alpha_b$ the non-retrieval fixed point becomes a repeller.

The basin structure of the system is determined by the existence and location of the other fixed points. For $\alpha < \alpha_b$ there are saddle points located on both the m_s- and m_d-axes, leading to basin edges as illustrated in figure 2(*a*).

By considering the third-order terms in (30) as well, we observe that for $\alpha > \alpha_b$, a saddle point bifurcates from the non-retrieval fixed point (figure 2(*b*)). This saddle point lies on the boundary separating the retrieval and non-retrieval basins. In the (upper) vicinity of the phase line $\alpha = \alpha_b$, this point is located at $(m_s, m_d) = (m_1, m_1)$ with $m_1^2 \equiv 2(1 - \sqrt{2/\pi\kappa})/(1 - \pi/6) \propto (\alpha - \alpha_b)$.

Furthermore, above the phase line $\alpha = \alpha_b$ and for $Q > Q_T \equiv \pi/3 - 1 \approx 0.047$, a repeller on the m_s-axis bifurcates from the non-retrieval fixed point (figure 2(*b*)). This repeller also lies on the boundary separating the retrieval and non-retrieval basins. In the (upper) vicinity of the point (α_b, Q_T), this repeller is located at $(m_2, 0)$ where $m_2^2 \equiv 4(1 - \sqrt{2/\pi\kappa})/(Q + 1 - \pi/3) \propto (\alpha - \alpha_b)/(Q - Q_T)$. For higher values of α, this repeller merges with the saddle point on the m_s-axis, leading to a discontinuous widening of the non-retrieval basin along the m_s-axis and a change of flow behaviour from that of figure 2(*b*) to figure 2(*c*). The phase line of this discontinuous transition in the α–Q space is determined by the merging of the double solutions of the equation $m = \mathrm{erf}(m\kappa/\sqrt{2(1 - Q_+m^2)})$, i.e. it happens exactly when $(\partial/\partial m)\,\mathrm{erf}(m\kappa/\sqrt{2(1 - Q_+m^2)})$.

The regions of existence of the attractors and repellers are shown in figure 3. The phase lines separate the space into three distinct phases for $\alpha < 1$, and their corresponding basin structures are shown in figures 2(*a*)–(*c*) respectively.

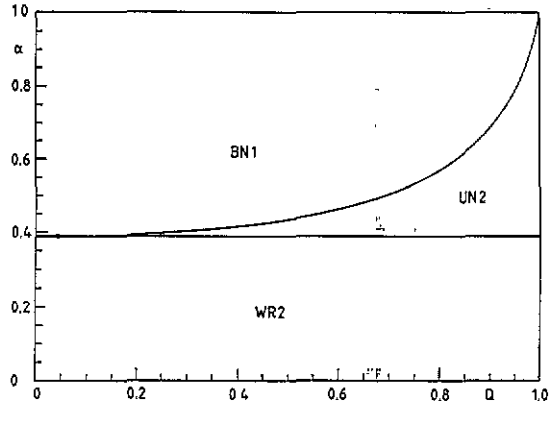

Figure 3. Phase diagram of the PIN in the α–Q space. Phases as described in the text.

(*a*) WR2—wide, perfect retrieval (two valleys)—with $(m_s^*, m_d^*) = (1, 1)$. The basin of attraction is maximal with two 'valleys'† in flow-space. This occurs for $\alpha < \alpha_b$.

(*b*) UN2—unilaterally narrow retrieval (two valleys)—with $(m_s^*, m_d^*) = (1, 1)$ and $(0, 0)$ coexisting. The retrieval basin is narrowed for $\alpha > \alpha_b$ such that retrieval is not possible for an initial state near the m_d-axis. However, selective retrieval of the correlated patterns is still possible for initial states near (but not on) the m_s-axis; hence the narrowing is 'unilateral'.

(*c*) BN1—bilaterally narrow retrieval (one valley)—with $(m_s^*, m_d^*) = (1, 1)$ and $(0, 0)$ coexisting, but the retrieval basin is narrowed everywhere along *both* the m_s- and m_d-axes; hence 'bilaterally' narrowed. This occurs for sufficiently high values of α.

† When we use topographic terminology we do so in analogy with classical motion under gravity. Note that here and below the number of valleys refers to the retrieving section of phase space only.

The transition from wide retrieval to the other two phases is continuous, whereas the transition from the unilaterally narrow to the bilaterally narrow phase is discontinuous. Note that no confused phase is present. Only when the network state starts exactly in a confused ($m_d = 0$) initial configuration will it remain confused.

One can interpret the shrinking of the retrieval basins and the evolution of the basin boundaries as the result of two competing physical tendencies. On the one hand, increasing the storage level α tends to stabilize the non-retrieval state. On the other hand, increasing Q reduces the Hamming distance between the retrieval attractors on either side of the m_s-axis ($(m_s^*, m_d^*) = (1, 1)$ and $(m_s^*, m_d^*) = (1, -1)$). As the basins associated with the two patterns approach each other, they cooperatively lower their barrier on the m_s-axis. This effect is particularly strong for higher values of m_s. Whereas for Hebbian networks the interference of the patterns is so severe that the confused attractor is observed in a wide range of parameter space, for PIN the interference is less excessive but the cooperative effect of the correlated patterns is still manifested in the flow behaviour.

One may distinguish two regimes of Q. For $Q < Q_T$ the 'barrier' along the m_s-axis is high. On increasing α through α_b the non-retrieval fixed point is able to extend its attracting power along both the m_s- and m_d-axes, and the system changes from wide to bilaterally narrow retrieval directly, the saddle points on the m_s- and m_d-axes moving in towards (0, 0) and then up the diagonal. However, for $Q > Q_T$ the barrier along the m_s-axis is lowered by the cooperative effect of the correlated patterns. As a result, the attracting power of the non-retrieval attractor is reduced along the m_s-axis, when compared with the m_d-axis, and the system has an intermediate phase of unilaterally narrow retrieval. When these two competitive forces are of comparable strength, we find a tricritical point with an additional line of discontinuous transitions joining the line of continuous transitions.

3.3. The maximally stable network

For the MSN, the performance function is $g^{\mathrm{MSN}}(\Lambda) = \Theta(\Lambda - \kappa)$. Using (15), the maximal stability κ_{MSN} is given by $\alpha^{-1} = \int_{-\infty}^{\kappa_{\mathrm{MSN}}} \mathrm{D}t(\kappa_{\mathrm{MSN}} - t)^2$. Combining (6) and (16) the retrieval map reduces to

$$f_{s/d}(m_s, m_d) = \int_{\mathcal{R}_{s/d}} \mathrm{D}t_s \mathrm{D}t_d \,\mathrm{erf}\left(\frac{\sqrt{Q_+}\lambda_s m_s + \sqrt{Q_-}\lambda_d m_d}{\sqrt{2(1 - Q_+ m_s^2 - Q_- m_d^2 + T^2)}}\right) \tag{31}$$

where $\mathcal{R}_{s/d}$ represents the functions λ_s, λ_d in terms of t_s, t_d used in the evaluation of $f_{s/d}$ respectively. They are given in appendix B. The double integrals in (31) can be further reduced to single ones if one devises for each sector of integration a suitable rotation in the (t_s, t_d) plane. This completes the formulation of the iterative maps. The analysis of (31) will, however, only be given for $T = 0$.

As shown in figures 4 and 5, the phase structure in the space of α and Q is very rich. Thus we first describe the four phase lines separating the various phases. Line 1 is defined by the bifurcation of the non-retrieval fixed point along the m_d-axis as α is reduced (as illustrated in the change from figure 5(b) to figure 5(a)) and is given by $\sqrt{2Q_-/\pi}\langle\lambda_d\rangle_d = 1$, where we use the shorthand $\langle f\rangle_{s/d} \equiv \int_{\mathcal{R}_{s/d}} Dt_s Dt_d f(t_s, t_d)$. Similarly line 2 is defined as the bifurcation of the non-retrieval fixed point along the m_s-axis as α is changed (as illustrated in the change from figure 5(b) to figure 5(c), or from figure 5(f) to figure 5(g), or from figure 5(e) to figure 5(f)), and is given by $\sqrt{2Q_+/\pi}\langle\lambda_s\rangle_s = 1$.

Line 3 indicates the bifurcation of a saddle point on the m_s-axis in the (transverse) m_d-direction (as illustrated in the change from figure 5(c) to figure 5(d), or from figure 5(b) to

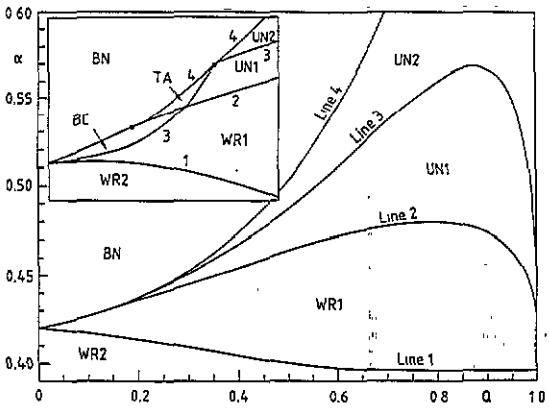

Figure 4. Phase diagram of the MSN in the $Q-\alpha$ space. Phases as described in the text. The large inset in the left upper corner is a sketchy enlargement of the critical region for small Q. Lines 1 to 4 correspond to the sequence of phase lines as observed for increasing α and high Q.

figure 5(f)). Furthermore, the transitional behaviour across this line is determined by the value of $\partial f_d/\partial m_d$ at the saddle point $(m, 0)$, where $\langle \mathrm{erf}(\sqrt{Q_+}\lambda_s m/\sqrt{2(1-Q_+m^2)})\rangle_s = m$, and

$$\frac{\partial f_d}{\partial m_d} = \sqrt{\frac{2Q_-}{\pi(1-Q_+m^2)}}\left\langle \lambda_d \exp\left(-\frac{Q_+m^2\lambda_s^2}{2(1-Q_+m^2)}\right)\right\rangle_d . \tag{32}$$

The saddle point is stable in the transverse direction if $\partial f_d/\partial m_d \leqslant 1$, unstable if otherwise. On line 3, $\partial f_d/\partial m_d = 1$.

Line 4 describes the discontinuous annihilation of a pair of fixed points on the m_s-axis (as illustrated in the change from figure 5(d) or figure 5(g) to figure 5(e)), i.e. where two non-zero solutions of the equation $m = \langle \mathrm{erf}(\sqrt{Q_+}m\lambda_s/\sqrt{2(1-Q_+m^2)})\rangle_s$ merge and disappear. In other words $\partial f_s/\partial m_s = 1$, where

$$\frac{\partial f_s}{\partial m_s} = \frac{1}{1-Q_+m^2}\sqrt{\frac{2Q_+}{\pi(1-Q_+m^2)}}\left\langle \lambda_s \exp\left(-\frac{Q_+m^2\lambda_s^2}{2(1-Q_+m^2)}\right)\right\rangle_s . \tag{33}$$

Line 1, 2 and 3 meet at $(\alpha_1, Q_1) \approx (0.42, 0)$. Lines 2 and 4 are analogues of the two phase lines in the PIN. Therefore they join at a tricritical point corresponding to (α_b, Q_T) in the case of PIN. Specifically, the tricritical point for the MSN is given by $(\alpha_2, Q_2) \approx (0.429, 0.118)$, which is determined by the vanishing of both the third-order coefficient $(\langle\lambda_s\rangle_s - \langle\lambda_s^3\rangle_s/3)$ and the first-order coefficient $(-1+\sqrt{2Q_-/\pi}\langle\lambda_s\rangle_s)$ of an expansion of $(f_s - m_s)$ in the m_s-direction around the origin.

It is interesting to note that lines 3 and 4 touch each other at the tetracritical point $(\alpha_3, Q_3) \approx (0.435, 0.180)$. In the vicinity of this point, line 3 divides the region below line 4 into three phases: (i) both fixed points on the m_s-axis are stable in the transverse (here m_d-) direction (figure 5(g)); (ii) the fixed point with the smaller value of m is transversely stable and the other unstable (figure 5(c)); (iii) both fixed points are transversely unstable (figure 5(d)). It is important to note that since the retrieval functions are continuous in m_s and m_d, regions with behaviour (i) and (iii) must be separated by a region with behaviour (ii). On line 4, however, the fixed points merge, and a direct transition between behaviour (i) and (iii) is possible. This accounts for the existence of the tetracritical point.

Finally, we comment that lines 2 and 3 intersect each other at the point $(\alpha_4, Q_4) \approx (0.431, 0.14)$ and the 'endpoints' for $Q = 1$ of lines 1 through 4 are given by $\alpha \approx 0.396$ (line 1), $\alpha \approx 0.42$ (line 2), $\alpha \approx 0.44$ (line 3) and $\alpha = 2$ (line 4).

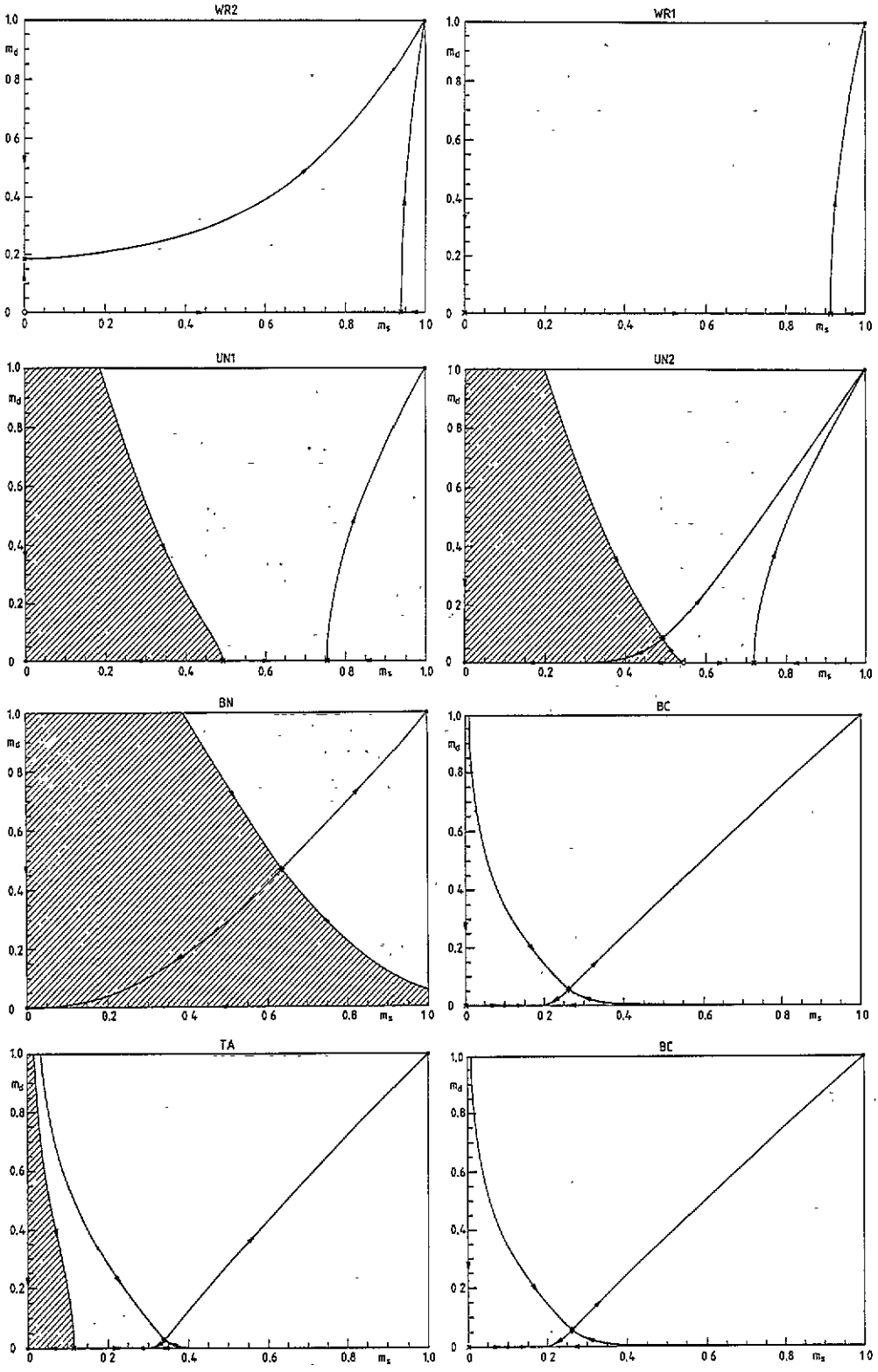

Figure 5. Typical flow diagrams in the m_s–m_d space for the MSN; description as in figure 2.

We can thus distinguish the following eight phases, examples of which are shown in figures 5(*a*)–(*g*) respectively: (*a*) WR2, wide retrieval (two valleys); (*b*) WR1, wide retrieval (one valley); (*c*) UN1, unilaterally narrow retrieval (one valley); (*d*) UN2, unilaterally narrow retrieval (two valleys); (*e*) BN, bilaterally narrow retrieval (one valley); (*f*) BC, bilaterally confused (one valley); (*g*) TA, triple attractor (one valley); (*h*) NR, non-retrieval.

The non-retrieval attractor is stable above line 2, i.e. in the UN1, UN2, BN and TA phases. The retrieval attractor is stable in all phases for $\alpha < 2$. However, we note that in the BC and TA phases, we also find stable attractors on the m_s-axis corresponding to stable confused states. The phases WR2, UN2 and BN have a similar structure to the three phases in PIN. Comparing the phase diagrams of MSN and PIN, the analogues of the phases WR1, UN1, BC and TA of the MSN can be considered as degenerate and lying on the line $\alpha = \alpha_b$ in the PIN.

As in the PIN, the phase behaviour is governed by the competition between the attracting power of the non-retrieval fixed point and the cooperative effect of the correlated patterns. This may be observed in the phase changes undergone by the system from the WR2 phase at low α to the BN1 phase at high α. At high Q, the phase change follows the sequence WR2 $\rightarrow$ WR1 $\rightarrow$ UN1 $\rightarrow$ UN2 $\rightarrow$ BN $\rightarrow$ NR. Since the attracting power of the non-retrieval attractor along the m_s-axis is partially offset by the cooperative attracting power of the correlated patterns, network states in the vicinity of the m_s-axis can be distinguishably retrieved up to a relatively high value of α, which corresponds to the first four in the sequence of six phases above. On reducing Q, however, the cooperative attracting power of the correlated patterns is weaker, and different scenarios for increasing α become possible. One finds first the sequences WR2 $\rightarrow$ WR1 $\rightarrow$ UN1 $\rightarrow$ TA $\rightarrow$ BN $\rightarrow$ NR, then WR2 $\rightarrow$ WR1 $\rightarrow$ BC1 $\rightarrow$ TA $\rightarrow$ BN $\rightarrow$ NR and then WR2 $\rightarrow$ WR1 $\rightarrow$ BC $\rightarrow$ BN $\rightarrow$ NR. In these sequences, respectively only the first three, two and two phases can have network states in the vicinity of the m_s-axis distinguishably retrieved.

It is interesting to note that the confused attractors only exist in the BC and TA phases, which are very small in extent, and can be considered as mere transient phases between the major phases. This confirms that the MSN has a high degree of selectivity, and confused states are apparently not favoured.

4. Conclusion

We have studied in detail the process of basin shrinking and the retrieval of a pair of correlated patterns among many uncorrelated patterns for increasing loading levels and different degrees of correlation. We have investigated the issue of pattern selectivity for the Hebbian, the pseudo-inverse and the maximally stable networks, viewed as networks designed to optimize appropriate performance functions. The analysis can, in principle, be extended to other networks, optimizing arbitrary performance functions. We have found three different types of attractors: non-retrieval, retrieval and confused attractors, together with interesting features in the phase diagrams.

The behaviour of the system is the result of two competing factors, namely the correlation Q between the special pair of patterns and the overall pattern storage ratio α. Generally speaking, higher values of α imply more interference on the memory states and favour non-retrieval. On the other hand, higher values of Q imply that the retrieval attractors of the two marked patterns become closer in the network configuration space, and either the confused attractor (as in the Hebbian network) or the retrieval attractors (as in PIN or MSN) are favoured. The evolution of the various attracting powers in the parameter space gives rise to the phenomena of basin encroaching, shrinking, splitting and wedging [23], resulting in very rich phase diagrams.

As illustration of encroaching, take, for example, the transitions of the triple attractor phase (TA) in the MSN. On increasing Q, the attracting power of the retrieval attractors increases along the m_τ-axis, as they cooperatively squeeze down the basins in this region. This results in the basin of the confused attractor being encroached by the retrieval attractors, leading to the unilaterally narrow retrieval phase (UN1). On increasing α, the non-retrieval attractor is favoured, and it encroaches on the neighbouring confused attractor, resulting in the bilaterally narrow retrieval phase (BN). On reducing α, the non-retrieval attractor is weakened and is encroached by the neighbouring confused attractor, yielding the bilaterally confused phase (BC).

An example of basin shrinking is found in the transition from UN2 to BN where the basin of the non-retrieval attractor expands at the expense of the retrieval attractor on increasing α (or reducing Q). The transition from UN1 to UN2 on increasing α is an example of basin splitting. Here the retrieval basin splits from one valley to two valleys because of the uneven development of the strength of the attractors in different regions of the network configuration space. Within the UN2 phase, we also expect the non-retrieval basin to wedge into the retrieval basin along the m_s-axis, since the basin boundary meets this axis at a repeller (were the meeting point a saddle point, the basin boundary would be normal to the axis). On increasing α, the tip of this wedge-shaped basin finally merges with another saddle point on the axis, resulting in a transition to the BN phase.

Perhaps the most striking observation is the effect of pattern correlation on the PIN and MSN when compared with the Hebbian network. This is due to the fact that the Hebbian network consists of wide and mutually interfering basins, whereas the PIN and MSN have deep and less interfering basins. When the distance between the two correlated retrieval attractors in the configuration space is reduced on increasing Q, the two basins in the Hebbian network interfere severely with each other. Thus the *confused attractor* is favoured at high values of Q. On the other hand, the two basins in PIN or MSN are deep and less interfering. Instead of weakening the barrier between them on increasing Q, they cooperatively squeeze its width. Thus the *retrieval attractors* are favoured at high values of Q.

However, we remark that the present analysis has been restricted to the regime $Q_- \sim \mathrm{O}(C^0)$. For extremely high pattern correlation, say $Q_- \sim \mathrm{O}(C^{-1})$, the present analysis has to be modified. The stability parameters in the PIN and MSN may deviate from the values given in section 3 which are applicable to lower pattern correlations, and confused states may exist in these systems. In fact, in the $p \to 0$ limit, we may consider the case of storing just two correlated patterns as we did for the Hebbian network at the end of section 3.1. In this case the configuration of the weights J_{ij} becomes identical to that of the Hebbian network, namely that for $i \in s$, $J_{ij} = 1/\sqrt{Q_+ C}$ for $j \in S$ and $J_{ij} = 0$ for $j \in D$, whereas for $i \in D$, $J_{ij} = 1/\sqrt{Q_- C}$ for $j \in D$ and $J_{ij} = 0$ for $j \in S$. The analysis for the Hebbian network case thus shows that both the PIN and MSN become confused for $Q > 1 - 3.70/C$. We expect that for general values of α all retrieving networks will exhibit confusion in the region of sufficiently high pattern correlation $Q_- \sim \mathrm{O}(C^{-1})$.

On the whole, this comparative study shows that both PIN and MSN have much higher selectivity than the Hebbian network. Recently, the retrieval properties of the MSN and the Hebbian network have been compared [24]. It was found that because of the different basin structures, the MSN has higher retrieval precision, higher storage capacity, simpler attractors, but lower associativity and robustness. The high selectivity of the MSN is consistent with, and complementary to, these observations. It should therefore be interesting—although numerically involved—to consider the dependence of pattern selectivity on training noise as considered by Wong and Sherrington [7], since the Hebbian and the MSN correspond to the two extreme limits of the training noise level. Arbitrary training noise levels would

allow an understanding of intermediate regimes between the Hebb and the Gardner rules, possibly casting further light onto the problem of confusion in ANN. Also, such a training noise variation might be one possibility to obtain a network of tunable selectivity (although others relying on modification at the training stage might be preferred).

However, we should point out the relevance of the weight space structure when comparing the selectivity of different learning rules. Consider a node i with $\xi_i^1 = -\xi_i^2$ (i.e. it belongs to the set $\mathcal{D}$). Using (28) for PIN, we note that the aligning field contribution Λ_d from those input neurons which can differentiate between the two patterns is equal to $\kappa_{\mathrm{PIN}}/\sqrt{Q_-}$, which can be made arbitrarily large when the patterns become more and more correlated. (This applies to the case when $Q_- \ll 1$ but Q_- is still $\mathcal{O}(C^0)$, which has been considered in this paper; the case of Q_- being nearer to 0 is currently under consideration.) Similarly for the MSN, the aligning field Λ_d for nodes $i \in \mathcal{D}$ is given by $\max\{t_d, \kappa/\sqrt{Q_-}\}$ in the limit $Q_- \ll 1$. Thus the fact that the PIN and MSN do not get confused at high correlation is related to the ability of the network to make the aligning field arbitrarily large at sites where ξ_i^1 and ξ_i^2 are different. This is possible for a spherically constrained J, since it can take any position in the weight space on the unit sphere. However, if the weight space is more constrained, as for example if J is only allowed to take Ising components ($J_j = \pm 1$), the corresponding MSN may have different behaviour at high Q.

The above analysis has been restricted to deterministic retrieval. An extension to probabilistic ($T \neq 0$) retrieval would be straightforward. Discussion of asymptotic retrieval has also been restricted to the case of dilute asymmetric networks. Equation (3) continues to apply to a first-step iteration even in highly connected cases but further analysis requires a different method.

Finally, we discuss the pattern selectivity of networks in the presence of external fields. In our recent work [25,26] we have investigated the effects of external fields on the retrieval properties of highly diluted ANN with general classes of learning rules. We showed that external fields could increase the size of basins of attraction and improve the retrieval quality. Here we propose another application for the use of external fields. As we have just seen, a system trained with the Hebb rule is confused in certain regimes, i.e. it loses the ability to distinguish correlated patterns. To avoid this during retrieval one could apply an external field having finite overlap with pattern one or two at the sites for which the pattern bits are different (or at least a subset of them).

Acknowledgments

We acknowledge financial support from the SERC. One of us (AR) would like to thank the Studienstiftung des deutschen Volkes and Corpus Christi College, Oxford for the award of two scholarships. We appreciate motivating discussions with Eytan Domany at the initial stage of this problem and acknowledge further fruitful discussions with Timothy Watkin and Richard Penney.

Appendix A

Here we will outline briefly the derivation of the aligning field distributions. We want to train our network so that its specific choice of interactions minimizes the energy function

$$E(J) = -\sum_{\mu} g(\Lambda^{\mu}). \tag{A.1}$$

To do so we introduce, following [19], a Gibbs measure on the configuration space

$$d\mu(J) = dJ\,\delta(J\cdot J - 1)e^{-\beta E(J)} \tag{A.2}$$

where β is the inverse of the annealing temperature which will later on be decreased to 0 to settle the system in its energy minimum. Using replicas to calculate the pattern averaged free energy we get

$$\rho_{s/d}(\Lambda_s,\Lambda_d) = \lim_{\substack{n\to 0\\ \beta\to\infty}} \left\langle \int \left(\prod_\alpha d\mu(J_\alpha)\right)\delta(\Lambda_s - \Lambda_s^a)\delta(\Lambda_d - \Lambda_d^a)\right\rangle_{\xi,\mathcal{S}/\mathcal{D}} \tag{A.3}$$

where $\alpha \in \{1,\dots,a,\dots,n\}$ is the replica *index* (not to be confused with the storage capacity), and $\Lambda^\alpha_{s/d} \equiv \sum'_{j\in\mathcal{S}/\mathcal{D}} \xi^1 J_j^\alpha \xi_j^1/\sqrt{C Q_\pm}$. Here we average over the couplings with the properly normalized Boltzmann weights and then over the possible realizations of ξ for the sets $\mathcal{S}/\mathcal{D}$ of sites correspondingly.

The replica calculation follows along the lines of [18]. One can write the field distribution as

$$\begin{aligned}\rho_{s/d}(\Lambda_s,\Lambda_d) = \lim_{n\to 0} \mathrm{Extr}_{q_{\alpha\beta},\hat q_{\alpha\beta},\hat\epsilon_\alpha}\\ \times\left\{\exp C\left(-\sum_{\alpha<\beta} q_{\alpha\beta}\hat q_{\alpha\beta} + G_J + \alpha G_\Lambda\right)\left[\int\left(\prod_\alpha \frac{d\hat\lambda_s^\alpha\, d\lambda_s^\alpha}{2\pi}\frac{d\hat\lambda_d^\alpha\, d\lambda_d^\alpha}{2\pi}\right)\right.\right.\\ \times\exp\left\{-\frac{1}{2}\sum_\alpha[(\hat\lambda_s^\alpha)^2 + (\hat\lambda_d^\alpha)^2] - \sum_{\alpha<\beta} q_{\alpha\beta}(\hat\lambda_s^\alpha\hat\lambda_s^\beta + \hat\lambda_d^\alpha\hat\lambda_d^\beta)\right.\\ +\sum_\alpha \beta\left[g\left(\sqrt{Q_+}\lambda_s^\alpha + \sqrt{Q_-}\lambda_d^\alpha\right) + g\left(\pm\left(\sqrt{Q_+}\lambda_s^\alpha - \sqrt{Q_-}\lambda_d^\alpha\right)\right)\right]\\ \left.\left.+\,i\sum_\alpha(\lambda_s^\alpha\hat\lambda_s^\alpha + \lambda_d^\alpha\hat\lambda_d^\alpha)\right\}\delta(\Lambda_s - \lambda_s^a)\delta(\Lambda_d - \lambda_d^a)\right]\Bigg\}.\end{aligned} \tag{A.4}$$

The functions G_J and G_Λ are given by

$$G_J = \ln\left[\int\left(\prod_\alpha dJ_\alpha\right)\exp\left\{-\sum_\alpha \hat\epsilon_\alpha(J_\alpha^2 - 1) + \sum_{\alpha<\beta}\hat q_{\alpha\beta}J_\alpha J_\beta\right\}\right] \tag{A.5}$$

$$G_\Lambda = \ln\left[\int\left(\prod_\alpha \frac{d\lambda^\alpha\, d\hat\lambda^\alpha}{2\pi}\right)\exp\left\{\sum_\alpha[\beta g(\lambda^\alpha) + i\lambda^\alpha\hat\lambda^\alpha - \tfrac{1}{2}(\hat\lambda^\alpha)^2] - \sum_{\alpha<\beta} q_{\alpha\beta}\hat\lambda^\alpha\hat\lambda^\beta\right\}\right] \tag{A.6}$$

but in the $n\to 0$ limit the first exponential term in (A.4) vanishes leaving only the integral to be evaluated. Within the framework of replica symmetry and in the limit $\beta\to\infty$ we get the result reported in equations (15)–(17).

Appendix B

In order to find $\rho_{s/d}$ for the MSN we have to maximize the function (17) with respect to λ_s and λ_d. This maximization leads to four different regions in the $t_s - t_d$ space. For m_s, we find the mapping $\mathcal{R}_s$ to be

$$
\begin{aligned}
&\text{(i)} \begin{cases} (\lambda_s, \lambda_d) = (t_s, t_d) \\ \text{for } \sqrt{Q_+}t_s + \sqrt{Q_-}t_d > \kappa;\ \sqrt{Q_+}t_s - \sqrt{Q_-}t_d > \kappa \end{cases} \\
&\text{(ii)} \begin{cases} (\lambda_s, \lambda_d) = (\sqrt{Q_-}[\sqrt{Q_-}t_s + \sqrt{Q_+}t_d] + \sqrt{Q_+}\kappa, \sqrt{Q_+}[\sqrt{Q_-}t_s + \sqrt{Q_+}t_d] \\ \qquad -\sqrt{Q_-}\kappa) \\ \text{for } \sqrt{Q_-}t_s + \sqrt{Q_+}t_d > \sqrt{Q_-/Q_+}\kappa;\ \sqrt{Q_+}t_s - \sqrt{Q_-}t_d < \kappa \end{cases} \\
&\text{(iii)} \begin{cases} (\lambda_s, \lambda_d) = (\sqrt{Q_-}[\sqrt{Q_-}t_s - \sqrt{Q_+}t_d] + \sqrt{Q_+}\kappa, -\sqrt{Q_+}[\sqrt{Q_-}t_s - \sqrt{Q_+}t_d] \\ \qquad +\sqrt{Q_-}\kappa) \\ \text{for } \sqrt{Q_-}t_s - \sqrt{Q_+}t_d > \sqrt{Q_-/Q_+}\kappa;\ \sqrt{Q_+}t_s + \sqrt{Q_-}t_d < \kappa \end{cases} \\
&\text{(iv)} \begin{cases} (\lambda_s, \lambda_d) = (\kappa/\sqrt{Q_+}, 0) \\ \text{for } \sqrt{Q_-}t_s + \sqrt{Q_+}t_d < \sqrt{Q_-/Q_+}\kappa;\ \sqrt{Q_-}t_s - \sqrt{Q_+}t_d < \sqrt{Q_-/Q_+}\kappa. \end{cases}
\end{aligned}
\tag{B.1}
$$

The region boundaries are shown in figure B1(a).

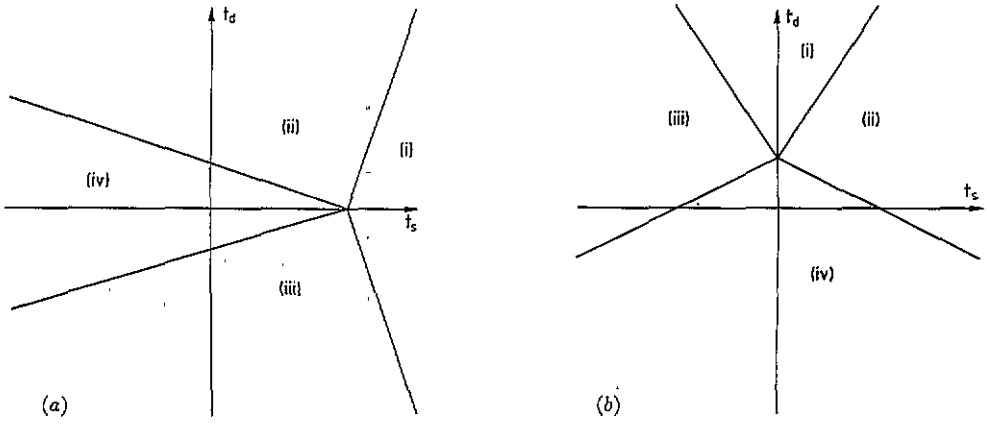

Figure A1. Regions of integration for the $t_s - t_d$ space.

For $\mathcal{R}_d$ we find

$$
\begin{aligned}
&\text{(i)}\ \begin{cases} (\lambda_s, \lambda_d) = (t_s, t_d) \\ \text{for } \sqrt{Q_+}t_s + \sqrt{Q_-}t_d > \kappa;\ \sqrt{Q_+}t_s - \sqrt{Q_-}t_d < -\kappa \end{cases} \\
&\text{(ii)}\ \begin{cases} (\lambda_s, \lambda_d) = (\sqrt{Q_-}[\sqrt{Q_-}t_s + \sqrt{Q_+}t_d] - \sqrt{Q_+}\kappa, \sqrt{Q_+}[\sqrt{Q_-}t_s + \sqrt{Q_+}t_d] \\ \qquad + \sqrt{Q_-}\kappa) \\ \text{for } \sqrt{Q_-}t_s + \sqrt{Q_+}t_d > \sqrt{Q_+/Q_-}\kappa;\ \sqrt{Q_+}t_s - \sqrt{Q_-}t_d > -\kappa \end{cases} \\
&\text{(iii)}\ \begin{cases} (\lambda_s, \lambda_d) = (\sqrt{Q_-}[\sqrt{Q_-}t_s - \sqrt{Q_+}t_d] + \sqrt{Q_+}\kappa, -\sqrt{Q_+}[\sqrt{Q_-}t_s - \sqrt{Q_+}t_d] \\ \qquad + \sqrt{Q_-}\kappa) \\ \text{for } \sqrt{Q_-}t_s - \sqrt{Q_+}t_d < -\sqrt{Q_+/Q_-}\kappa;\ \sqrt{Q_+}t_s + \sqrt{Q_-}t_d < \kappa \end{cases} \\
&\text{(iv)}\ \begin{cases} (\lambda_s, \lambda_d) = (0, \kappa/\sqrt{Q_-}) \\ \text{for } \sqrt{Q_-}t_s + \sqrt{Q_+}t_d > -\sqrt{Q_+/Q_-}\kappa;\ \sqrt{Q_-}t_s + \sqrt{Q_+}t_d < \sqrt{Q_+/Q_-}\kappa. \end{cases}
\end{aligned}
\tag{B.2}
$$

The region boundaries are shown in figure B1(b).

It should be noted that (λ_s, λ_d) is continuous at the region boundaries.

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)
[2] Hopfield J J 1982 *Proc. Natl Acad. Sci. (USA)* **79** 255
[3] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[4] Fontanari J F and Köberle R 1988 *J. Phys. A: Math. Gen.* **21** 2477
[5] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[6] Peretto P 1988 *Neural Networks* **1** 309
[7] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175
[8] Yedidia J S 1989 *J. Phys. A: Math. Gen.* **22** 2265
[9] Sompolinsky H 1986 *Phys. Rev.* A **34** 2571
[10] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[11] Kohonen T 1984 *Self-Organization and Associative Memory* (Berlin: Springer)
[12] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique* **46** L359
[13] Pázmándi F and Geszti T 1990 *Europhys. Lett.* **13** 673
[14] Rau A, Wong K Y M and Sherrington D 1992 *Europhys. Lett.* **17** 649
[15] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[16] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[17] Krauth W, Mézard M and Nadal J-P 1988 *Complex Systems* **2** 387
[18] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659
[19] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[20] Abbott L 1990 *Network* **1** 105
[21] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715
[22] Wong K Y M and Sherrington D 1989 *J. Phys. A: Math. Gen.* **22** 2233
[23] Wong K Y M and Campbell C 1992 *J. Phys. A: Math. Gen.* **25** 2227
[24] Wong K Y M and Sherrington D 1992 *Physica* A **185** 453
[25] Rau A and Sherrington D 1990 *Europhys. Lett.* **11** 499
[26] Rau A, Sherrington D and Wong K Y M 1991 *J. Phys. A: Math. Gen.* **24** 313