On the function and development of spatial structure in layered neural networks

1993 J. Phys. A: Math. Gen. 26 2549

(http://iopscience.iop.org/0305-4470/26/11/008)

J. Phys. A: Math. Gen. 26 (1993) 2549–2571. Printed in the UK

On the function and development of spatial structure in layered neural networks

H J J Jonker†, A C C Coolen‡ and J J Denier van der Gon†

† Utrecht Biophysics Research Institute, University of Utrecht, Princetonplein 5, NL 3584 CC Utrecht, The Netherlands

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

Received 5 January 1992

Abstract. In the first part of this paper we study the relation between spatial structure and information processing properties of layered Ising spin neural networks with lateral interactions. The interactions *between* layers are given by the Hebb rule, the interactions *within* layers by the so-called anti-Hebb rule. Secondly we study the *development* of spatial structure in such networks as the result of an *unsupervized* learning process (now both neurons and interactions are dynamical variables). By calculating the spectrum of the output covariance matrix as a function of the spectrum of the input covariance matrix, we show how the spatial characteristics of the input signals are reflected in the information processing properties of the equilibrated system.

1. Introduction

Biological neural networks share with magnetic systems the property of consisting of a large number of interacting noisy microscopic elements (the neurons), which are more or less similar. This enables the application of statistical mechanics in studying neural networks, if one accepts a reduction in the number of degrees of freedom of individual neurons. Such a reduction cannot be avoided if any analytical progress is to be made in studying networks consisting of more than just a few neurons which interact in a non-trivial way. Furthermore, statistical mechanics simply *shows* that macroscopic features of large interacting particle systems usually do not depend on details of the microscopic elements (universality). Motivated by neuro-anatomical data the aim of this paper is to study analytically the function and development of spatial structure in layered neural networks with lateral interactions. The neural interactions are assumed to evolve in time according to Hebbian-type [1] rules; spatial structure leads to an additional position dependence of interactions, in contrast to the fully connected models studied in the early (pioneering) papers by Hopfield [2] and Amit *et al* [3].

Two classes of models in which neural interactions carry a position dependence have been studied in the literature (apart from a simple organization in uniform layers). The common ingredient of the first class is that position dependence is the result of *random* dilution (either uniformly [4], in a layered context [5] or in a modular context [6]). The so-called extreme dilution (satisfying a certain scaling requirement), which is mostly employed, simplifies mathematical analysis, but is no longer realistic biologically. The second class of models allow for a specific position dependence of the density (or strength) of interactions [7–9], which is the direction we will take. Our learning rules will be as follows: *between* layers interaction modification is proportional to the state correlation of the two neurons

0305-4470/93/112549+23$07.50

involved, *within* layers modification is proportional to *minus* the state correlation of the neurons involved. The latter prescription is often referred to as the *anti-Hebb* rule and has been shown to yield interesting system-theoretical properties [10–13]. Layered Ising spin models with such interactions (without additional spatial structure) remain ergodic in the thermodynamic limit, which allows for exact analytical solutions [14, 15]. In order to finally study the *development* of spatial structure the interactions will have to be treated as dynamic variables (in addition to the neurons), which complicates models considerably. The usual strategy is to resort to the adiabatic limit (which is justified for biological neural networks, since the time-scale for the evolution of interactions is usually much larger than the time-scale of neural processes). Formally one can now proceed. However, in order to go beyond deriving general statements [16–18] or presenting simulation results [13, 19] one will have to *calculate* the state correlations between pairs of neurons that drive the learning rules. Apart from networks with *linear* neurons [20, 21], this can be done for Ising spin models of the type [15].

This paper is organized as follows. In section 2 we study the properties of layered Ising spin neural networks with lateral interactions and spatial structure. The interactions remain constant. In section 3 we allow interactions to evolve in time in an unsupervized manner and study how spatial structure can develop and how the spatial properties of the input signals will be reflected in the information processing properties of the equilibrated system.

2. Spatial structure in layered neural networks with lateral interactions

In this section we study the properties of layered Ising spin neural networks with lateral interactions and spatial structure. Interactions *between* layers are given by the Hebb rule, the interactions *within* layers by the so-called anti-Hebb rule. The spatial structure is imposed by defining the absolute strength of these interactions to be position-dependent. First we derive deterministic evolution equations for a suitably chosen set of (local) order parameters. We show that the system remains ergodic in the thermodynamic limit under certain conditions on the imposed spatial structure, and calculate the macroscopic equilibrium state. Finally we study the relation between spatial structure and information processing properties.

2.1. *Definitions*

The neurons are modelled as Ising spins ($s_i = 1$ if neuron i fires and $s_i = -1$ if it is at rest), arranged in an architecture of two equally large layers. Microscopic configurations of the input neurons and the output neurons will be denoted by the vectors $\boldsymbol{s}^{\mathrm{in}} \in \{-1, 1\}^N$ and $\boldsymbol{s}^{\mathrm{out}} \in \{-1, 1\}^N$ respectively. The output neurons are laterally interconnected via *fixed* interactions J^{oo} and receive additional signals from the input neurons via *fixed* interactions J^{io}.

The states s_i^{in} of the N neurons in the first (input) layer will be 'clamped' (i.e. assumed to be prescribed), whereas the states s_i^{out} of the N neurons in the second (output) layer evolve in time according to a stochastic alignment to local fields (or post-synaptic potentials) $h_i(\boldsymbol{s}^{\mathrm{out}})$:

$$h_i(\boldsymbol{s}^{\mathrm{out}}) \equiv \sum_{j=1, j\neq i}^{N} J_{ij}^{\mathrm{oo}} s_j^{\mathrm{out}} + \sum_{k=1}^{N} J_{ik}^{\mathrm{io}} s_k^{\mathrm{in}} .$$

The probability of finding the output layer at time t in microscopic configuration s^{out} is written as $p_t(s^{\text{out}})$. The stochastic alignment process is governed by a master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(s^{\text{out}}) = \sum_{i=1}^{N} w_i(F_i s^{\text{out}})p_t(F_i s^{\text{out}}) - p_t(s^{\text{out}})\sum_{i=1}^{N} w_i(s^{\text{out}}) \tag{1}$$

where $F_i s^{\text{out}} \equiv (s_1^{\text{out}}, \ldots, -s_i^{\text{out}}, \ldots, s_N^{\text{out}})$ and where the transition rates $w_i(s^{\text{out}})$ (which give the probability per unit time of the state change $s^{\text{out}} \to F_i s^{\text{out}}$) are defined as

$$w_i(s^{\text{out}}) \equiv \tfrac{1}{2}[1 - g(\beta s_i^{\text{out}} h_i(s^{\text{out}}))]. \tag{2}$$

The function $g(x)$ must have the properties $g(-x) = -g(x)$, $g'(x) \geqslant 0$, $g(x) \in [-1, 1]$ and $g'(0) = 1$ (like, for example, $g(x) = \tanh(x)$). The inverse 'temperature' $\beta \equiv 1/T$ is a measure of the amount of stochastic noise in the alignment process.

The neural interactions are in this section assumed to be the result of a supervized Hebbian type learning process of the form

$$\Delta J_{ij}^{\text{io}} \sim s_i^{\text{out}} s_j^{\text{in}} \qquad \Delta J_{ij}^{\text{oo}} \sim -s_i^{\text{out}} s_j^{\text{out}}.$$

The difference between these two prescriptions (Hebbian learning *between* layers and anti-Hebbian learning *within* layers) has a natural biological interpretation and, furthermore, has been shown to generate interesting information processing properties in layered models without spatial structure [15]. During the learning stage p specific input patterns $\chi^\mu \in \{-1, 1\}^N$ have been 'clamped' on the input side, in combination with p corresponding output patterns $\xi^\mu \in \{-1, 1\}^N$ on the output side. Spatial structure in the interactions is embedded by introducing for each neuron pair (i, j) a quantity K_{ij} which represents the number of synapses operating between the axonal endings of j and the dendrites of i (if, for instance, the distance between i and j is too large we put $K_{ij} = 0$). The interaction matrices can now be written as

$$J_{ij}^{\text{io}} \equiv \frac{1}{K} K_{ij}^{\text{io}} \sum_{\mu=1}^{p} \xi_i^\mu \chi_j^\mu \tag{3}$$

$$J_{ij}^{\text{oo}} \equiv -\frac{1}{K} K_{ij}^{\text{oo}} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{4}$$

where $K > 0$ is a constant to be defined later. We will restrict ourselves to the case $K_{ij}^{\text{oo}} = K_{ji}^{\text{oo}}$ $\forall ij$, so the matrix J^{oo} is symmetric.

2.2. *Derivation of macroscopic dynamical laws*

Since the physical location of neurons has become relevant, any macroscopic level of description must involve *locally* defined order parameters. We divide both the input layer and the output layer into n equally large non-overlapping clusters of adjacent neurons, defined by the index sets I_k^{in} and I_k^{out} respectively ($k = 1, \ldots, n$). The number of neurons in each cluster is N/n. Our local order parameters are now chosen to be the familiar overlap parameters (which measure correlation between the microscopic spin variables and the embedded patterns), calculated within the clusters:

$$q_k^\mu(s^{\text{out}}) \equiv \frac{n}{N}\sum_{i \in I_k^{\text{out}}} \xi_i^\mu s_i^{\text{out}} \qquad m_k^\mu(s^{\text{in}}) \equiv \frac{n}{N}\sum_{i \in I_k^{\text{in}}} \chi_i^\mu s_i^{\text{in}}. \tag{5}$$

We will use the following notations: $q^\mu = (q_1^\mu, \ldots, q_n^\mu)$ and $q_k = (q_k^1, \ldots, q_k^p)$. The analysis of the behaviour of networks of this type in terms of local order parameters has been performed in [8,9]. The essential assumptions for this analysis to apply are:

- The number of clusters n is sufficiently small, so that dynamic mean-field theory becomes exact in the thermodynamic limit. For the present model this implies the scaling requirement $n^2 p \ll \sqrt{N}$.
- The typical length scale for variations in the spatial structure parameters K_{ij}^{io} and K_{ij}^{oo} is much larger than the cluster size.

The first assumption has been shown to guarantee [8,9,22] that by taking the thermodynamic limit the macroscopic probability distribution evolves in time according to a Liouville equation. It represents an upper limit to the number of different order parameters that can be expected to evolve in time according to exact dynamic mean-field laws. Using the second assumption, we can write the local fields $h_i(s^{\mathrm{out}})$ in terms of macroscopic quantities only:

$$h_i = \frac{1}{K}\sum_{l=1}^{n}\sum_{\mu=1}^{p}\left\{\sum_{j\in I_l^{\mathrm{in}}} K_{ij}^{\mathrm{io}}\xi_i^\mu \chi_j^\mu s_j^{\mathrm{in}} - \sum_{j\in I_l^{\mathrm{out}}} K_{ij}^{\mathrm{oo}}\xi_i^\mu \xi_j^\mu s_j^{\mathrm{out}}\right\}$$

$$\approx \sum_{\mu=1}^{p}\xi_i^\mu \sum_{l=1}^{n}\left\{A_{kl}m_l^\mu - B_{kl}q_l^\mu\right\} \tag{6}$$

where the constant K in (3), (4) is chosen as $K \equiv N/n$, and where the matrices A and B are defined as

$$A_{kl} \equiv \left(\frac{n}{N}\right)^2 \sum_{i\in I_k^{\mathrm{out}}}\sum_{j\in I_l^{\mathrm{in}}} K_{ij}^{\mathrm{io}} \qquad B_{kl} \equiv \left(\frac{n}{N}\right)^2 \sum_{i\in I_k^{\mathrm{out}}}\sum_{j\in I_l^{\mathrm{out}}} K_{ij}^{\mathrm{oo}}. \tag{7}$$

These matrices represent the spatial structure at the coarse-grained cluster scale. Note that (6) becomes an exact equality if we choose the structure matrices K^{io} and K^{oo} to be constant within clusters.

The evolution in time of the macroscopic state probability

$$P_t(q^1, \ldots, q^p) \equiv \sum_{s^{\mathrm{out}}} p_t(s^{\mathrm{out}}) \prod_{\mu=1}^{p}\delta\left[q^\mu - q^\mu(s^{\mathrm{out}})\right]$$

turns out to be governed by a Liouville equation (on finite time-scales) in the thermodynamic limit $N \to \infty$ (see [8,9] or the review [22]). The evolution in time of the np order parameters $\{q^\mu\}$ thereby becomes deterministic. The laws governing this deterministic evolution are the following set of coupled non-linear differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}q_k^\mu = -q_k^\mu + \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}\xi_i^\mu g\left[\beta\sum_{l=1}^{n}\sum_{\nu=1}^{p}\xi_i^\nu(A_{kl}m_l^\nu - B_{kl}q_l^\nu)\right]. \tag{8}$$

2.3. Equilibrium

The zero-temperature evolution equations are

$$\frac{\mathrm{d}}{\mathrm{d}t}q_k^\mu = -q_k^\mu + \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}\xi_i^\mu \mathrm{sign}\left[\sum_{l=1}^{n}\sum_{\nu=1}^{p}\xi_i^\nu(A_{kl}m_l^\nu - B_{kl}q_l^\nu)\right]. \tag{9}$$

The system (9) can in principle have numerous fixed-point attractors. If we choose, for example, $A_{kl} \equiv 0$ and $B_{kl} \equiv -\delta_{kl}$ we obtain the equations describing n decoupled Hopfield networks [2], each of which is known to have many stable states [3]. Some less trivial choices for B were studied in [8], which sustain the picture of a wealth of fixed-point attractors. This picture, however, changes drastically if there is no positive feedback in the system, i.e. if all eigenvalues of B are positive. In the latter case we will show that there is *only one global attractor* in the system, given by

$$\bar{q}^\mu = B^{-1}Am^\mu \qquad \mu = 1\ldots p. \tag{10}$$

The only additional and necessary condition for this statement to hold is that the system must be able to actually realize the macroscopic state (10): a microscopic configuration s^{out} must exist such that

$$\forall \mu, k: \quad \bar{q}_k^\mu = \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}\xi_i^\mu s_i^{\mathrm{out}}. \tag{11}$$

In order to show that (10) is the unique equilibrium solution of (9) we introduce the deviations $r_k^\mu \equiv q_k^\mu - \bar{q}_k^\mu$ which evolve in time according to

$$\frac{\mathrm{d}}{\mathrm{d}t}r_k^\mu = -r_k^\mu - \bar{q}_k^\mu - \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}\xi_i^\mu \mathrm{sign}[u_k^i]$$

where $u_k^i \equiv \sum_{\rho l}\xi_i^\rho B_{kl}r_l^\rho$. The matrix B is symmetric by definition and positive-definite by assumption. We will expand on the validity of this assumption in section 2.4. The non-negative scalar function E turns out to be a global Lyapunov function for the process (9):

$$E \equiv \tfrac{1}{2}\sum_{\mu kl} r_l^\mu B_{lk} r_k^\mu.$$

The time derivative of E can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}E = \sum_{\mu kl} r_l^\mu B_{lk}\left\{-r_k^\mu - \bar{q}_k^\mu - \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}\xi_i^\mu \mathrm{sign}[u_k^i]\right\}$$

$$= -2E - \sum_k \frac{n}{N}\sum_{i\in I_k^{\mathrm{out}}}|u_k^i|\left\{\bar{s}_i^{\mathrm{out}}\mathrm{sign}[u_k^i] + 1\right\} \leqslant -2E$$

(since $|\bar{s}_i^{\rm out}| = 1$). Apparently E is a Lyapunov function; its minimum is $E = 0$, which is obtained for $r_k^\mu = 0$ $(\forall k\mu)$. This, in turn, immediately implies (10). Note that this result has been derived without making any assumptions with respect to the distribution from which the patterns are drawn.

A first remarkable aspect of the stationary solution (10) is the *linear* dependence of the equilibrium output order parameters $\{q^\mu\}$ on the input order parameters $\{m^\mu\}$, in spite of the fact that the microscopic system is based on binary elements with highly non-linear dynamical rules (there is no noise that might have a linearizing effect). For the fully connected model with negative feedback, similar behaviour has been reported in [14, 15]; the presence of spatial structure turns out not to affect this linearity (provided that B is positive-definite). Note that the derivation shown above does not depend on the distribution from which the patterns are drawn. Equation (10) therefore holds for any choice of the patterns. A second important property is the uniqueness of the macroscopic equilibrium state: the system remains ergodic in the limit $N \to \infty$ and responds to external input in a way similar to how an anti-ferromagnet responds to an external magnetic field. Thirdly the appearance of the inverse of B is peculiar, since it can lead to counter-intuitive system properties. If, for instance, $A = B$ then the equilibrium macroscopic output state will be an exact copy of the macroscopic input state. If input information is spatially spread out by *divergent* connections A_{kl}, this will be counteracted by an appropriate *convergent* collective equilibration process in the output layer, if the lateral output interactions B_{kl} are similarly *divergent*. In section 2.6 we will elaborate on the system behaviour if A and B are different.

Next we will study the effect of noise in the system on the equilibrium solution and its stability. We shall see that introduction of noise can lift the restrictions on the spatial structure (i.e. the smallest eigenvalue of B being positive). For the function $g(x)$ in the definition of the microscopic transition rates of the master equation we choose the *saturation* function, defined as

$$\mathrm{sat}(x) \equiv \begin{cases} x & \text{if } |x| \leqslant 1 \\ \mathrm{sign}(x) & \text{if } |x| > 1\,. \end{cases} \tag{12}$$

This choice is made for computational simplicity; without proof we mention that the results for other choices of g are qualitatively similar (according to simulations). In the noiseless case the macroscopic equilibrium state turned out to be independent of the distribution from which the patterns are drawn. This is no longer true if $T > 0$. To highlight only the essential properties of the $T > 0$ behaviour, we will, however, restrict ourselves to randomly drawn unbiased patterns, so that

$$\forall k: \quad \frac{n}{N}\sum_{i\in I_k^{\rm out}} \phi(\xi_i) \to \langle \phi(\xi)\rangle_\xi \equiv 2^{-p}\sum_{\xi\in\{-1,1\}^p} \phi(\xi) \qquad N \to \infty\,.$$

Inspired by the results in [15] we make the ansatz

$$\bar{q}^\mu \equiv [B + T\mathbf{1}]^{-1} A m^\mu \qquad \mu = 1 \ldots p \tag{13}$$

where $\mathbf{1}$ is the identity matrix $\mathbf{1}_{kl} = \delta_{kl}$. We again study the evolution in time of the variables $r_k^\mu \equiv q_k^\mu - \bar{q}_k^\mu$, for which we find, using (8), (12) and (13)

$$\frac{\mathrm{d}}{\mathrm{d}t} r_k = -r_k - \bar{q}_k - \left\langle \xi\, \mathrm{sat}\left[\xi\cdot(\beta\sum_l B_{kl} r_l - \bar{q}_k)\right]\right\rangle_\xi . \tag{14}$$

The origin ($r_k^\mu = 0\ \forall k, \mu$) is a critical point of the evolution equation (14) if for all k: $\bar{q}_k = \langle \xi \mathrm{sat}[\xi \cdot \bar{q}_k] \rangle_\xi$. This is equivalent to demanding

$$\forall_k \quad \sum_\mu |\bar{q}_k^\mu| \leqslant 1 . \tag{15}$$

Condition (15) is comparable with condition (11) but it is somewhat more severe. By choosing the input vectors $\{m^\mu\}$ small enough, it is nevertheless not difficult to satisfy (15) by virtue of the linear form of (13).

Concentrating on the domain where the saturation function is linear, one can infer from (14) that (13) is at least locally stable if the matrix $\mathbf{1} + \beta B$ is positive-definite. This means that, given a matrix B, local stability of (14) is guaranteed if one chooses

$$T > -\lambda_{\min}(B) . \tag{16}$$

In appendix B we address the question of under what conditions there is *global* stability. It turns out that $T > -2\lambda_{\min}(B)$ is sufficient to guarantee that each initial state converges to (13). Simulations, however, indicate that (16) is already sufficient.

Summarizing, one may conclude that for any coarse-grained structure matrix B one can find a noise level sufficiently large to ensure the system will behave linearly according to (13). The critical noise level depends on the smallest eigenvalue of B.

2.4. *Translation-invariant spatial structure kernels*

We will study in more detail spatial structures K_{ij} that depend on the distance $|i - j|$ only. By construction we can now write $A_{kl} = a(|k - l|)$ and $B_{kl} = b(|k - l|)$; the matrices A and B are symmetric Toeplitz matrices. To investigate whether a structure function b gives rise to a positive definite matrix B (which is the requirement for finding a unique global attractor for $T = 0$), we can make use of a general statement about Toeplitz matrices [23]. Reformulated to our purpose this statement reads:

If $B_{kl} = b(|k - l|)$ and $F(x)$ is defined by

$$F(x) \equiv b(0) + 2\sum_{k=1}^{\infty} b(k)\cos(kx) \tag{17}$$

then all eigenvalues λ of B, denoted by $\lambda(B)$, satisfy the inequalities

$$\min_{x\in[-\pi,\pi]} F(x) \leqslant \lambda(B) \leqslant \max_{x\in[-\pi,\pi]} F(x) .$$

Since $F(x)$ can often be calculated analytically, these inequalities enable us to check whether B is positive-definite. In table 1 we list the function $F(x)$ (17) for some choices of the structure function b.

The impression one obtains from table 1 is that the concavity versus convexity of the structure function b is an important property. The triangular shape represents the intermediate case. These observations are indeed correct: in appendix A we prove that the matrix $B_{kl} = b(|k - l|)$ is positive-definite if b is a positive, monotonically decreasing, concave function. These conditions on b are sufficient, but not necessary. This can be illustrated by studying the Gaussian structure $B_{kl} = \exp(-|k - l|^2/2\sigma^2)$, which is partly

Table 1. Examples of structure functions b and eigenvalue properties of corresponding structure matrices $B_{kl} \equiv b(|k-l|)$; $\lambda(B) \geqslant \min_{x\in[-\pi,\pi[} F(x)$. Note that, with respect to the sign of $\min F(x)$, the shape of b (concavity versus convexity) is more important than the width w. $\theta(x)$ denotes the step function: $\theta(x) = 1$ for $x \geqslant 0$ and $\theta(x) = 0$ otherwise.

Shape	$b(\|k\|)$	$F(x) \equiv b(0) + 2\sum_{k=1}^{\infty} b(k)\cos(kx)$	$\min_{x\in[-\pi,\pi[} F(x)$
Hyperbola	$(1+\|k\|)^{-1}$	$-2\ln\|2\sin(x/2)\|\cos x + (\pi - x)\sin x - 1$	> 0
Exponential[a]	$\exp(-w\|k\|)$	$\dfrac{\sinh w}{\cosh w - \cos x}$	> 0
Triangle[b]	$\left(1 - \dfrac{\|k\|}{w}\right)\theta(w - \|k\|)$	$\dfrac{1-\cos(wx)}{w(1-\cos x)}$	$= 0$
Parabola[b]	$\left(1 - \left[\dfrac{\|k\|}{w}\right]^2\right)\theta(w - \|k\|)$	$\dfrac{\sin(wx)\cos(x/2) - 2w\cos(wx)\sin(x/2)}{2w^2\sin^3(x/2)}$	< 0
Block[b]	$\theta(w - \|k\|)$	$\dfrac{\cos(wx) - \cos[(w+1)x]}{1-\cos x}$	< 0

[a] $w > 0$.
[b] w integer > 1.

convex ($|k-l| < \sigma$) and partly concave ($|k-l| > \sigma$), but nevertheless corresponds to a positive-definite matrix B:

$$\sum_{kl}^{n} x_k B_{kl} x_l = \sum_{kl}^{n} x_k \mathrm{e}^{-(k-l)^2/2\sigma^2} x_l = \sqrt{\frac{2}{\pi\sigma^2}} \int_{-\infty}^{+\infty} \left[\sum_{k=1}^{n} x_k \mathrm{e}^{-(t-k)^2/\sigma^2}\right]^2 \mathrm{d}t > 0$$

(the integral cannot yield 0 because it is not possible to satisfy $\sum_{k=1}^{n} x_k \exp[-(t-k)^2/\sigma^2] = 0$ for all t).

Whether or not our zero-temperature two-layer spatially structured neural network has a unique macroscopic fixed-point attractor, which at the macroscopic level corresponds to performing a nice linear transformation, turns out to depend critically on the shape, or rather on the second derivative, of the function describing the position dependence of the lateral connection density.

2.5. The continuum limit

If the number of clusters n is large (with the restriction $n^2 \ll \sqrt{N}/p$, $N \to \infty$), we can take a continuum limit and replace the cluster labels by continuous position vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. For q_k we write $q(\boldsymbol{x})$, $A_{kl} \to A(\boldsymbol{x},\boldsymbol{y})$ and $B_{kl} \to B(\boldsymbol{x},\boldsymbol{y})$. The evolution equations (8) become

$$\frac{\mathrm{d}}{\mathrm{d}t} q(\boldsymbol{x}) = -q(\boldsymbol{x}) + \left\langle \xi\, g\left[\xi \cdot \beta \left\{ \int_{D^{\mathrm{in}}} A(\boldsymbol{x},\boldsymbol{y}) m(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} - \int_{D^{\mathrm{out}}} B(\boldsymbol{x},\boldsymbol{y}) q(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right\}\right]\right\rangle_{\xi} \tag{18}$$

where D^{in} and D^{out} denote the spatial regions defining the two network layers. In the noiseless case ($T = 0$) the equilibrium value $\bar{q}(\boldsymbol{x})$ is the solution of

$$\int_{D^{\mathrm{out}}} B(\boldsymbol{x},\boldsymbol{y}) q(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \int_{D^{\mathrm{in}}} A(\boldsymbol{x},\boldsymbol{y}) m(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \tag{19}$$

provided that B is positive-definite, i.e. for all $\phi \in L_2(D^{\text{out}})$

$$\int_{D^{\text{out}}} \phi(x)B(x,y)\phi(y)\mathrm{d}x\mathrm{d}y > 0. \tag{20}$$

The proof that the equilibrium state is defined by (19) runs along the same lines as the proof for the discrete case (i.e write down the evolution equation for $r(x) \equiv q(x) - \bar{q}(x)$ and define the Lyapunov function $E \equiv \int \mathrm{d}x\mathrm{d}y\ B(x,y)r(x)\cdot r(y)$, which obeys $(\mathrm{d}/\mathrm{d}t)E \leqslant -2E$).

For translation-invariant and infinitely large systems ($A(x,y) = a(|x-y|)$, $B(x,y) = b(|x-y|)$ and $D^{\text{in}} = D^{\text{out}} \equiv (-\infty,\infty)^d$), the integrals represent convolutions, so that their Fourier transforms factorize. We adopt the following conventions towards notation:

$$\hat{\Phi}(k) = \mathcal{F}[\Phi(x);k] \equiv \int \mathrm{d}x\ \Phi(x)\mathrm{e}^{\mathrm{i}k\cdot x}$$

$$\Phi(x) = \mathcal{F}^{-1}[\hat{\Phi}(k);x] \equiv [2\pi]^{-d}\int \mathrm{d}k\ \hat{\Phi}(k)\mathrm{e}^{-\mathrm{i}k\cdot x}$$

$$(\Phi\otimes\Psi)(x) \equiv \int \mathrm{d}y\ \Phi(x-y)\Psi(y).$$

If we restrict our analysis to order parameter fields and spatial structure kernels which are in the Hilbert space $L_2((-\infty,\infty)^d)$, we can use the factorization $\mathcal{F}[(\Phi\otimes\Psi)(x);k] = \hat{\Phi}(k)\hat{\Psi}(k)$, so that the condition for finding the unique macroscopic fixed point (20) becomes $\hat{b}(k) > 0$ for all k. If this condition is satisfied, the solution of (19) is given by

$$\bar{q}^{\mu}(x) = \mathcal{F}^{-1}\left[\frac{\hat{a}(k)\hat{m}^{\mu}(k)}{\hat{b}(k)};x\right] \qquad \mu = 1\ldots p \tag{21}$$

which provides a simple analytical expression of the macroscopic equilibrium state. If $T > 0$ and $g(x) = \mathrm{sat}(x)$ in (18), then (for random unbiased patterns) the unique macroscopic equilibrium state can be calculated similarly:

$$\bar{q}^{\mu}(x) = \mathcal{F}^{-1}\left[\frac{\hat{a}(k)\hat{m}^{\mu}(k)}{T+\hat{b}(k)};x\right] \qquad \mu = 1\ldots p \tag{22}$$

provided that $T > -2\min_k \hat{b}(k)$.

2.6. *Simulations*

We performed simulations of our model with $N = 1530$ input and output neurons. Furthermore, since at this stage our prime interest is in spatial properties rather than information capacity, we restricted our simulations to $p = 1$. Each layer was divided into $n = 51$ clusters, each consisting of 30 neurons. Since the resulting cluster size is rather modest, fluctuations could not be ignored. Therefore quantitative statements about the simulations are given in the form of *time averages* of the cluster correlations $q_k(s^{\text{out}})$. The (fixed) input state s^{in} was chosen such that at the macroscopic level the input cluster correlations were to acquire a Gaussian shape:

$$m_k(s^{\text{in}}) = \gamma\exp[-(k-\bar{k})^2/2\sigma^2] \tag{23}$$

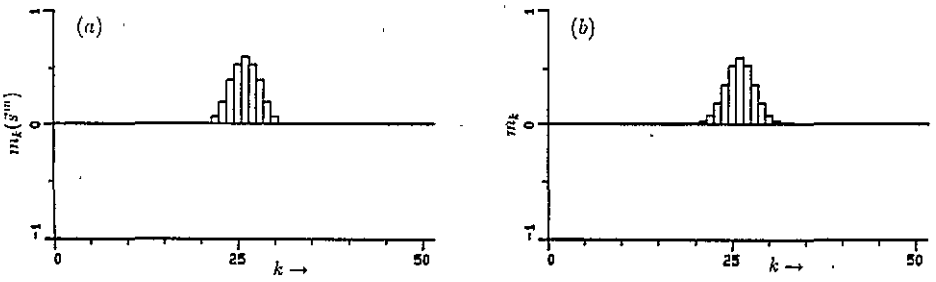

Figure 1. (*a*) Macroscopic input state $\{m_k\}$ as used in simulations (roughly Gaussian). (*b*) Macroscopic input state $\{m_k\}$, equation (23), as used for theoretical predictions.

with $\bar{k} = 26$, $\gamma = 0.6$ and $\sigma = 2$. The actual values of the cluster correlations resulting from the fixed microscopic configuration s^{in} is shown in figure 1(*a*); the idealization (23) (used for calculating the prediction of our theory) is shown in figure 1(*b*).

First we simulated the noiseless case $T = 0$ and chose for the microscopic spatial structure:

$$K_{ij}^{lo} = \exp(-r_a'|i-j|) \qquad K_{ij}^{oo} = \exp(-r_b'|i-j|)$$

with $r_a' = 0.009$ and $r_b' = 0.004$. Figure 2(*a*) shows the macroscopic output state $\bar{q}_k(s^{out})$ in equilibrium resulting from the actual simulations, given the input signals depicted macroscopically in figure 1(*a*). In order to compare this result with the theoretical prediction (10), we first have to calculate the coarse-grained structure matrices A and B defined in (7). For B, for example, one obtains for the present choice of $\{K^{io}, K^{oo}\}$ in an n-cluster arrangement of N spins:

$$B_{kl} = \exp(-r_b'L|k-l|)\left[\frac{\sinh(r_b'L/2)}{L\sinh(r_b'/2)}\right]^2 \qquad (k \neq l)$$

$$B_{kk} = \frac{L\sinh(r_b') + \exp(-r_b'L) - 1}{2L^2\sinh^2(r_b'/2)}$$

with $L \equiv N/n$ (the cluster size). If both $r_b \equiv r_b'L$ and $r_a \equiv r_a'L$ are small (as is the case for the values given above), then A and B are fairly well described by $\exp(-r_a|k-l|)$ and $\exp(-r_b|k-l|)$, respectively. The conditions $r_a \ll 1$ and $r_b \ll 1$ are equivalent to the familiar requirement that the typical length scale of fluctuations in spatial structure must be much larger than the spatial size of the individual clusters. If we insert for A and B the resulting exponentials we can calculate the theoretical prediction (10) for the macroscopic equilibrium state the result is shown in figure 2(*b*). In the same picture we have plotted the result obtained by taking the continuum limit (full curve), as given by (21), with $m(x) = \gamma\exp(-x^2/2\sigma^2)$, $a(|x-y|) = \exp(-r_a|x-y|)$ and $b(|x-y|) = \exp(-r_b|x-y|)$. The corresponding analytical expression for $\bar{q}(x)$ is (note: the spatial dimension of the system is 1)

$$\bar{q}(x) = \gamma e^{-x^2/2\sigma^2}\left\{\frac{r_a}{r_b} + \frac{\sigma(r_b^2 - r_a^2)}{\sqrt{2}r_b}\left[\Phi\left(\frac{r_a\sigma}{\sqrt{2}} + \frac{x}{\sqrt{2}\sigma}\right) + \Phi\left(\frac{r_a\sigma}{\sqrt{2}} - \frac{x}{\sqrt{2}\sigma}\right)\right]\right\}$$

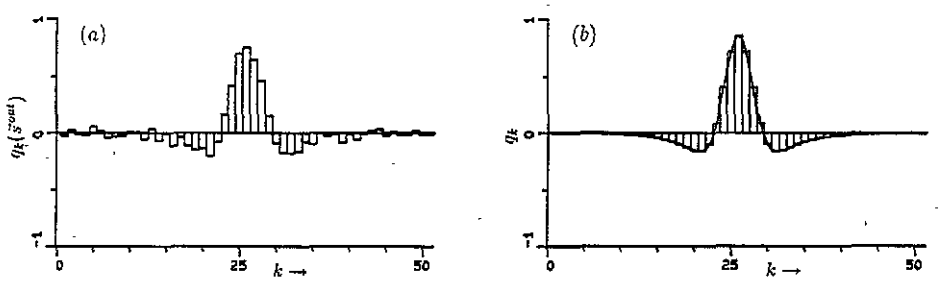

Figure 2. (*a*) Simulation result: macroscopic output state $\{\bar{q}_k\}$ in equilibrium (averaged over 20 flips per neuron) at $T = 0$, given the input of figure 1(*a*) and an exponential structure ($r'_a = 0.009$, $r'_b = 0.004$). (*b*) Theoretical prediction of the macroscopic output state $\{\bar{q}_k\}$ according to the discrete theory (10). Full curve: prediction according to the continuum theory.

with $\Phi(z) \equiv 2^{-1}\sqrt{\pi}\exp(z^2)\mathrm{erfc}(z) \equiv \int_z^\infty \exp(z^2 - t^2)\mathrm{d}t$. If $r_a = r_b$, then the output correlations will indeed be exactly equal to the input correlations. From figure 2 we may conclude that the three results (simulations, discrete theory, continuum theory) are in good mutual agreement.

Next we considered the case $T > 0$ and performed simulations with triangular-shaped spatial structures ($\theta(x)$ is the step function):

$$K_{ij}^{io} = (1 - r'_a|i-j|)\theta(1 - r'_a|i-j|) \qquad K_{ij}^{oo} = (1 - r'_b|i-j|)\theta(1 - r'_b|i-j|)\,.$$

In this case the macroscopic structure matrices are found to be ($k \neq l$):

$$A_{kl} = (1 - r_a|k-l|)\theta(1 - r_a|k-l|) \qquad B_{kl} = (1 - r_b|k-l|)\theta(1 - r_b|k-l|)$$

with $r_a = r'_a L$ and $r_b = r'_b L$ (L again represents the cluster size N/n). For r_a and r_b small the above expressions are good approximations for the diagonal elements $k = l$ as well. Our simulations were performed for $r_a = \frac{1}{6}$ and $r_b = \frac{1}{8}$. In figure 3 we have depicted the simulation result and the theoretical result from the discrete theory (13) for the choice $T = 0.1$; figure 4 indicates the results for the higher noise level $T = 1.0$. According to (22), the continuum theory predicts

$$\bar{q}(x) = \gamma\sigma\sqrt{\frac{2}{\pi}}\int_0^\infty \frac{r_a[1 - \cos(k/r_a)]}{(T/2)k^2 + r_b[1 - \cos(k/r_b)]}\cos(kx)\mathrm{e}^{-\sigma^2k^2/2}\mathrm{d}k\,.$$

Analytical evaluation of this integral is rather complicated, but it shows clearly how increasing the noise level T has a damping effect both on the amplitude and on the spatial oscillations of the macroscopic state (compare figures 3 and 4).

We may conclude that, provided the variations in spatial structure involve length scales which are indeed much larger than the cluster size, both the discrete theory (at the level of individual cluster correlations) and the continuum theory (in which the order parameters have become fields) are in good agreement with simulations (even for relatively small system sizes).

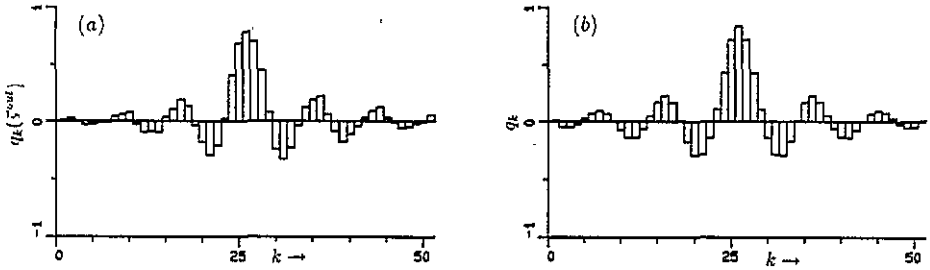

Figure 3. (*a*) Simulation result: macroscopic output state $\{\bar{q}_k\}$ in equilibrium (averaged over 20 flips per neuron) at $T = 0.1$, given the input of figure 1(*a*) and a triangular structure ($r'_a = (6L)^{-1}$, $r'_b = (8L)^{-1}$). (*b*) Theoretical prediction of the macroscopic output state $\{\bar{q}_k\}$ according to the discrete theory (13).

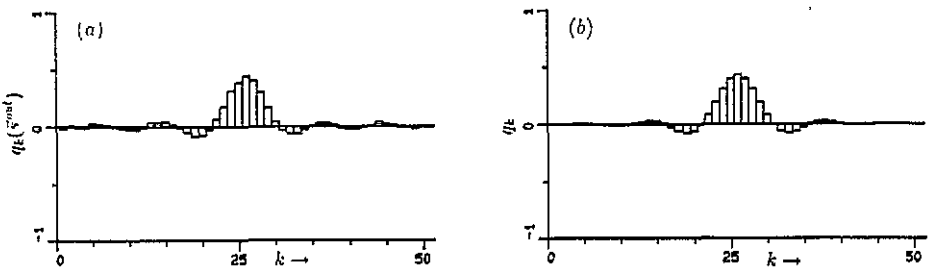

Figure 4. (*a*) Simulation result: macroscopic output state $\{\bar{q}_k\}$ in equilibrium (averaged over 20 flips per neuron) at $T = 1.0$, given the input of figure 1(*a*) and a triangular structure ($r'_a = (6L)^{-1}$, $r'_b = (8L)^{-1}$). (*b*) Theoretical prediction of the macroscopic output state $\{\bar{q}_k\}$ according to the discrete theory (13).

3. On the development of spatial structure by unsupervized learning

Next we turn to the question of how spatial structure can develop in layered Ising spin neural networks with lateral interactions, as the result of an unsupervized learning process, and how the developing structure depends on spatial properties of the information presented in the input layer. Now both neurons and interactions are dynamical variables. Interactions *between* layers evolve in time according to the Hebb rule, the interactions *within* layers according to the so-called anti-Hebb rule. First we show that if the (spatially structured) initial interaction matrices have a separable structure, this structure will be preserved by the learning process. This important property enables us to describe the evolution in time of the interaction matrices on a macroscopic level. We calculate the spectrum of the output covariance matrix in equilibrium as a function of the spectrum of the input covariance matrix and show how the spatial properties of the input signals will be reflected in the information processing properties of the equilibrated system.

3.1. Definitions

Both neurons and neural interactions will now be defined as *dynamical* variables. Neurons are again modelled as Ising spins, arranged in an architecture of two equally large layers.

The output neurons are laterally interconnected via *time-dependent* connections $J^{oo}(t)$ and receive additional signals from the input neurons via *time-dependent* connections $J^{io}(t)$.

The states s_i^{in} of the N neurons in the first (input) layer will be prescribed. The states s_i^{out} of the N neurons in the second (output) layer evolve in time according to a stochastic local field alignment described in section 2.1. However, here the process (1) is non-stationary due to the fact that both the interactions within and between the two layers and (possibly) the prescribed input states s^{in} are time-dependent. We assume the interactions to evolve in time according to the Hebbian [1] learning rules introduced in the previous section:

$$\tau\frac{d}{dt}J_{ij}^{io} \equiv \frac{1}{K}s_i^{out}(t)s_j^{in}(t) - \epsilon J_{ij}^{io} \qquad \tau\frac{d}{dt}J_{ij}^{oo} \equiv -\frac{1}{K}s_i^{out}(t)s_j^{out}(t) - \epsilon J_{ij}^{oo}$$

(with the scaling constant $K > 0$ to be determined later) which are now stochastic non-linear differential equations (since the spin states are stochastic variables). The decay term has been introduced simply to prevent interactions from unbounded growing.

It is not realistic to assume that one can solve the above combined stochastic system of evolving spin states and evolving spin interactions analytically at the microscopic level of individual spins and individual interactions. However, we can exploit the fact that in biological systems the characteristic time-scales involved in the evolution of interactions is much larger than the time-scales of the neuronal dynamics: $\tau \gg 1$. If τ is sufficiently large we may, as far as *neuronal* evolution is concerned, assume the interactions to be constant and replace the above equations by

$$\tau\frac{d}{dt}J_{ij}^{io} \equiv \frac{1}{K}\langle s_i^{out}\rangle s_j^{in} - \epsilon J_{ij}^{io} \tag{24}$$

$$\tau\frac{d}{dt}J_{ij}^{oo} \equiv -\frac{1}{K}\langle s_i^{out}s_j^{out}\rangle - \epsilon J_{ij}^{oo} \tag{25}$$

in which brackets denote averaging over the asymptotic microscopic probability distribution of the process (1), given the *stationary* values $\{J_{ij}^{io}\}$, $\{J_{ij}^{oo}\}$ for the interactions. In other words we restrict ourselves to the adiabatic limit.

3.2. Macroscopic dynamical laws

In the previous section we considered only fixed interaction matrices with a separable (Hebbian) structure, which enabled the derivation of evolution equations for macroscopic order parameters. In the present case we can only choose the *initial* interaction matrices to have a nice structure; we will have to prove that this structure is preserved by the unsupervized learning process (24), (25).

Suppose the neural interactions are of the form

$$J_{ij}^{io} = \frac{n}{N}K_{ij}^{io}\sum_{\mu\nu=1}^{p}\xi_i^\mu \hat{A}^{\mu\nu}\chi_j^\nu$$

$$J_{ij}^{oo} = -\frac{n}{N}K_{ij}^{oo}\sum_{\mu\nu=1}^{p}\xi_i^\mu \hat{B}^{\mu\nu}\xi_j^\nu$$

where, for simplicity, the patterns ξ^μ and χ^μ are drawn at random from a uniform distribution on $\{-1, 1\}^N$ and where the quantities $\{K_{ij}^{io}, K_{ij}^{oo}\}$ represent the spatial structure which is assumed to vary slowly with distance. This form is more general than the one

studied in section 2 (the latter corresponds to choosing $\hat{A}^{\mu\nu} = \hat{B}^{\mu\nu} = \delta_{\mu\nu}$). As in section 2 we define a partitioning of each layer into n clusters $\{I_k^{\rm in}, I_k^{\rm out}\}$ and corresponding coarse-grained matrices, which in the present case carry additional pattern indices

$$A_{kl}^{\mu\nu} \equiv \left(\frac{n}{N}\right)^2 \sum_{i\in I_k^{\rm out}} \sum_{j\in I_l^{\rm in}} K_{ij}^{\rm io} \hat{A}^{\mu\nu} \qquad B_{kl}^{\mu\nu} \equiv \left(\frac{n}{N}\right)^2 \sum_{i\in I_k^{\rm out}} \sum_{j\in I_l^{\rm out}} K_{ij}^{\rm oo} \hat{B}^{\mu\nu}$$

and find that the local field at output site $i \in I_k^{\rm out}$ can be expressed in terms of cluster correlations

$$q_k^\mu(s^{\rm out}) \equiv \frac{n}{N} \sum_{i\in I_k^{\rm out}} \xi_i^\mu s_i^{\rm out} \qquad m_k^\mu(s^{\rm in}) \equiv \frac{n}{N} \sum_{i\in I_k^{\rm in}} \chi_i^\mu s_i^{\rm in}$$

$$h_i \approx \sum_{\mu=1}^{p} \xi_i^\mu \sum_{l=1}^{n} \sum_{\nu=1}^{p} \left\{ A_{kl}^{\mu\nu} m_l^\nu - B_{kl}^{\mu\nu} q_l^\nu \right\} . \tag{26}$$

If variations in the matrices $\{K^{\rm io}, K^{\rm oo}\}$ occur only on length scales much larger than the cluster size, the above relations again become equalities. Since the microscopic dynamical laws rely only on how the local fields depend on the system state, we might as well take as a starting point a situation where spatial structure variations occur only on the cluster scale. Therefore we study the situation where the initial interactions $\{J_{ij}^{\rm io}(0), J_{ij}^{\rm oo}(0)\}$ already have the coarse-grained form

$$i \in I_k^{\rm in},\ j \in I_l^{\rm out}: \quad J_{ij}^{\rm io} = \frac{n}{N} \sum_{\mu\nu=1}^{p} \xi_i^\mu A_{kl}^{\mu\nu} \chi_j^\nu \tag{27}$$

$$i \in I_k^{\rm oo},\ j \in I_l^{\rm oo}: \quad J_{ij}^{\rm oo} = -\frac{n}{N} \sum_{\mu\nu=1}^{p} \xi_i^\mu B_{kl}^{\mu\nu} \xi_j^\nu . \tag{28}$$

In this case we recover (26) directly (which now has become a strict equality). If we choose for the function $g(x)$ in the microscopic transition rates (2) the saturation function (12) and make sure that the scaling requirement $n^2 p \ll \sqrt{N}$ is fulfilled, we find in the limit $N \to \infty$ the (generalized) deterministic laws for the evolution in time of the local order parameters:

$$\frac{\rm d}{{\rm d}t} q_k^\mu = -q_k^\mu + \left\langle \xi^\mu {\rm sat}\left[\beta \sum_{l\rho\nu} \xi^\rho \left(A_{kl}^{\rho\nu} m_l^\nu - B_{kl}^{\rho\nu} q_l^\nu \right) \right] \right\rangle_\xi . \tag{29}$$

The only difference between this result and the laws found in section 2 is that the tensors A and B have become of rank 4 instead of rank 2. Therefore we can take over most of the analysis in section 2 directly (if properly translated). We will write $y = Ax$ for the contraction $y_k^\mu = \sum_{l\nu} A_{kl}^{\mu\nu} x_l^\nu \ \forall k, \mu$ and $\mathbf{1}_{kl}^{\mu\nu} \equiv \delta_{kl}\delta_{\mu\nu}$. The analysis in appendix B, applied to the present case, enables us to conclude: if the input order parameters m are stationary, the unique equilibrium solution of (29) is given by

$$\bar{q} = [B + T\mathbf{1}]^{-1} A m \tag{30}$$

provided that the input m is such that condition (15) has been met and that

$$T > -2\lambda_{\rm min}(B) = -2 \min_x \frac{\sum_{kl\mu\nu} x_k^\mu B_{kl}^{\mu\nu} x_l^\nu}{\sum_{k\mu} (x_k^\mu)^2} . \tag{31}$$

For the large N equilibrium state (30) we can easily calculate, following [15], the *microscopic* equilibrium averages controlling the temporal development (24), (25) of the interactions in the adiabatic limit:

$$i \in I_k^{\text{out}}: \quad \langle s_i^{\text{out}} \rangle = \text{sat}[\beta \sum_{l\rho\nu} \xi_i^\rho \left(A_{kl}^{\rho\nu} m_l^\nu - B_{kl}^{\rho\nu} \bar{q}_l^\nu \right)] = \sum_\mu \xi_i^\mu \bar{q}_k^\mu$$

$$i \neq j: \quad \langle s_i^{\text{out}} s_j^{\text{out}} \rangle = \langle s_i^{\text{out}} \rangle \langle s_j^{\text{out}} \rangle .$$

Note that we have used the fact that (31), in combination with (15), guarantees that in the equilibrium state (30) the saturation function need be evaluated only in its linear regime. If we insert the above expressions into the unsupervized learning rules (24), (25) we obtain

$$i \in I_k^{\text{out}}: \quad \tau \frac{\text{d}}{\text{d}t} J_{ij}^{\text{io}} \equiv \frac{1}{K} \sum_\mu \xi_i^\mu \bar{q}_k^\mu s_j^{\text{in}}(t) - \epsilon J_{ij}^{\text{io}}$$

$$i \in I_k^{\text{out}}, \; j \in I_l^{\text{out}}: \quad \tau \frac{\text{d}}{dt} J_{ij}^{\text{oo}} \equiv -\frac{1}{K} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu \bar{q}_k^\mu \bar{q}_l^\nu - \epsilon J_{ij}^{\text{oo}} .$$

The final ingredient to be added is that the microscopic input configuration s^{in} is chosen at random according to a stationary distribution, subject to the constraint that the input correlations are $\{m_k^\mu\}$. By virtue of the adiabatic limit only the expectation values of the microscopic input variables will determine the evolution of the interactions. If we restrict the allowed inputs according to $\sum_\mu |m_k^\mu| \leqslant 1$ $(\forall k)$ and use the fact that the distribution of the pattern vectors $\{\chi^\mu\}$ is uniform, we can write $\langle s_i^{\text{in}} \rangle = \sum_\nu \chi_i^\nu m_k^\nu$ for $k \in I_k^{\text{in}}$. Inserting these expressions into the learning rules and choosing $K \equiv N/n$ we obtain the main result of this subsection, which is the proof that the structure (27), (28) will be preserved:

$$i \in I_k^{\text{out}}, \; j \in I_l^{\text{in}}: \quad \tau \frac{\text{d}}{dt} J_{ij}^{\text{io}} \equiv \frac{n}{N} \sum_{\mu\nu} \xi_i^\mu \chi_j^\nu \left\{ \bar{q}_k^\mu m_l^\nu - \epsilon A_{kl}^{\mu\nu} \right\}$$

$$i \in I_k^{\text{out}}, \; j \in I_l^{\text{out}}: \quad \tau \frac{\text{d}}{dt} J_{ij}^{\text{oo}} \equiv -\frac{n}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu \left\{ \bar{q}_k^\mu \bar{q}_l^\nu - \epsilon B_{kl}^{\mu\nu} \right\} .$$

3.3. Evolution of interactions described at a macroscopic level

Since the structure (27), (28) of the interaction matrices is preserved by the unsupervized learning process (24), (25) we now have the opportunity to describe the evolution in time of the neural interactions at the level of dynamic order parameters $\{A_{kl}^{\mu\nu}, B_{kl}^{\mu\nu}\}$, in a way similar to how we analysed the neural dynamics. If the input order parameters $\{m_k^\mu\}$ are drawn according to some probability distribution $\mathcal{P}\left[\{m_k^\mu\}\right]$ (which is constant during the learning process) and if we assume τ to be sufficiently large to ensure self-averaging of A and B with respect to $\mathcal{P}$, we find that the evolution in time of the interaction order parameters is governed by

$$\tau \frac{\text{d}}{\text{d}t} A_{kl}^{\mu\nu} = -\epsilon A_{kl}^{\mu\nu} + \langle \bar{q}_k^\mu m_l^\nu \rangle_{\mathcal{P}} \tag{32}$$

$$\tau \frac{\text{d}}{\text{d}t} B_{kl}^{\mu\nu} = -\epsilon B_{kl}^{\mu\nu} + \langle \bar{q}_k^\mu \bar{q}_l^\nu \rangle_{\mathcal{P}} \tag{33}$$

(in which we have denoted averages over $\mathcal{P}[\cdot]$ by $\langle\cdots\rangle_{\mathcal{P}}$). These laws are coupled and non-linear because the output order parameters $\bar{q}$ depend on the input order parameters m *through* the tensors A and B. The parameter τ only gauges the time-scale for the evolution of interactions. In order to suppress notation we will put $\tau \equiv 1$ and introduce the input covariance tensor M and the output covariance tensor Q

$$M_{kl}^{\mu\nu} \equiv \langle m_k^\mu m_l^\nu \rangle_{\mathcal{P}} \qquad Q_{kl}^{\mu\nu} \equiv \langle \bar{q}_k^\mu \bar{q}_l^\nu \rangle_{\mathcal{P}}$$

(which are by definition semi-positive-definite). The differential equations (32), (33) are such that negative eigenvalues of B will vanish at least exponentially. This means that, if there is some amount of noise in the system, sooner or later there will be a moment such that $T > -2\lambda_{\min}(B)$. This can be seen by writing (33) in an integral form:

$$B(t) = B(0)\mathrm{e}^{-\epsilon t} + \int_0^t \mathrm{e}^{\epsilon(t'-t)} Q(t')\mathrm{d}t'$$

(the tensor Q is positive-definite). From the moment when $T > -2\lambda_{\min}(B)$ we know that $\bar{q}$ depends on m, A and B according to (30). The differential equations (32), (33) then become

$$\frac{\mathrm{d}}{\mathrm{d}t}A = -\epsilon A + CM \tag{34}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}B = -\epsilon B + CMC^{\mathsf{T}} \tag{35}$$

where $C \equiv [B + T\mathbf{I}]^{-1}A$. C^{T} denotes the transpose of C in both index pairs (k, l) and (μ, ν), i.e. $(C^{\mathsf{T}})_{kl}^{\mu\nu} \equiv C_{lk}^{\nu\mu}$. Since the distribution $\mathcal{P}$ is stationary, the input covariance tensor M is constant. Note also that the simple relation $Q(t) = C(t)MC^{\mathsf{T}}(t)$ holds.

The equations (34), (35) are similar to the ones derived in [15] for the fully connected model; the only difference is the rank of the tensors A and B (4 instead of 2). In order to analyse the asymptotic behaviour of (34), (35) it turns out that one can follow exactly the route followed in [15]. We will not repeat this analysis but take over the final results. The behaviour of the system will be described in terms of the covariance tensors M and Q. Let λ_i^M denote the ith eigenvalue of M ($i = 1, \ldots, np$). For $t \to \infty$ the eigenvalues λ_i^Q of Q become

$$\lambda_i^Q = (\lambda_i^M - \epsilon T)\theta(\lambda_i^M - \epsilon T) \qquad i = 1, \ldots, pn \tag{36}$$

(in which $\theta(x)$ is the step function). As the temperature T (which turns out to play the role of a 'filtering threshold') is varied, the system undergoes repeated second-order phase transitions at the critical values $T_i \equiv \epsilon^{-1}\lambda_i^M$. The behaviour of the system resembles a Principal Components Analysis [24, 25]: components (eigenvectors) of the input covariance tensor M that correspond to eigenvalues larger than the threshold $\Delta \equiv \epsilon T$ can pass through, whereas components with smaller eigenvalues are suppressed. The extreme cases are $\Delta = 0$ and $\Delta > \lambda_{\max}^M$. For $\Delta = 0$ the system is completely transparent: each component of M passes through unaffectedly. For $\Delta > \lambda_{\max}^M$ all components of M are suppressed (no long-term order will survive in the output layer).

The asymptotic eigenvalues of Q can be readily calculated from the eigenvalues of M according (36). Note, however, that in the eigenvalue problem $\sum_{l\nu} M_{kl}^{\mu\nu} x_l^\nu = \lambda x_k^\mu$ the pattern information (indexed by (μ, ν)) and the spatial information (indexed by (k, l)) are generally intermingled. Furthermore, apart from some special cases, one does not know the eigenvectors corresponding to these eigenvalues.

3.4. Examples

In this section we apply the theoretical results obtained on the asymptotic behaviour of unsupervized learning to specific examples. We calculate the spectrum of the output covariance matrix, given three specific choices for the statistical and spatial properties of the input signals, i.e. for the distribution $\mathcal{P}[\{m_k^\mu\}]$.

Example 1. As a first example we study the case where the relation between spatial properties and pattern indices is constant, so that averages of the type $\langle\Phi\rangle_{\mathcal{P}}$ can be written as

$$\langle\Phi[\{m_k^\mu\}]\rangle_{\mathcal{P}} \equiv \int d\gamma\ \rho(\gamma)\Phi[\{\gamma^\mu f_k^\mu\}] \tag{37}$$

in which the quantities f_k^μ (describing how spatial properties couple to pattern indices) are fixed. The input covariance tensor M, given the choice (37), becomes

$$M_{kl}^{\mu\nu} \equiv \langle m_k^\mu m_l^\nu\rangle_{\mathcal{P}} = f_k^\mu f_l^\nu \Gamma^{\mu\nu} \qquad \Gamma^{\mu\nu} \equiv \int d\gamma\ \rho(\gamma)\gamma^\mu\gamma^\nu .$$

Let us define the p quantities $F^\mu \equiv \sqrt{\sum_k (f_k^\mu)^2}$ and introduce the auxiliary matrix $\hat{\Gamma}$ with eigenvectors ϕ_i and (non-negative) eigenvalues $\lambda_i^{\hat{\Gamma}}$ $(i = 1, \ldots, p)$:

$$\hat{\Gamma}^{\mu\nu} \equiv F^\mu \Gamma^{\mu\nu} F^\nu \qquad \sum_\nu \hat{\Gamma}^{\mu\nu}\phi_i^\nu = \lambda_i^{\hat{\Gamma}}\phi_i^\mu .$$

The eigenvalue problem for M, which acquires the form $f_k^\mu \sum_{l\nu} \Gamma^{\mu\nu} f_l^\nu x_l^\nu = \lambda^M x_k^\mu$, can be solved easily in terms of $\hat{\Gamma}$:

$$\lambda_i^M = \lambda_i^{\hat{\Gamma}} \qquad i = 1, \ldots, p \qquad \text{eigenvectors}: \quad x_{ik}^\mu \equiv \phi_i^\mu f_k^\mu / F^\mu$$

$$\lambda_i^M = 0 \qquad i = p+1, \ldots, pn \qquad \text{eigenspace}: \quad \sum_k f_k^\mu x_k^\mu = 0 \qquad \forall\mu .$$

The eigenvalues of the *output* covariance tensor are therefore given by

$$\lambda_i^Q = [\lambda_i^{\hat{\Gamma}} - \epsilon T]\theta[\lambda_i^{\hat{\Gamma}} - \epsilon T] \qquad i = 1, \ldots, p$$

$$\lambda_i^Q = 0 \qquad i = p+1, \ldots, np .$$

This spectrum turns out not to depend on the details of the parameters f_k^μ, only on the global strength F^μ with which individual input sources γ^μ couple to the system as a whole. This example clearly shows that the presence in the input signals of spatial structure, such that each input source γ^μ couples to the system in a spatially different but *constant* way, does not lead to results other than those obtained for the fully connected network [15]. Note, however, that this does not imply that the network will become fully connected during the learning process. The final interaction structure, which can only be predicted for some special cases (see next example), in general does depend on the choice made for the parameters f_k^μ and, furthermore, on the initial interactions.

Example 2. As a second example we study input covariance tensors with translation invariance. For notational and computational convenience we follow the continuous approach, i.e. the covariance tensor M becomes a translation-invariant kernel $M^{\mu\nu}(x-y)$. An important implication of this is that translation invariance will be preserved by the learning process. If at time t the system is translation-invariant, we may write

$$A^{\mu\nu}(x,y;t)=a^{\mu\nu}(x-y;t) \qquad B^{\mu\nu}(x,y;t)=b^{\mu\nu}(x-y;t)\,.$$

As a consequence the kernel C in (34), (35) will also be translation-invariant, $C^{\mu\nu}(x,y;t)=c^{\mu\nu}(x-y;t)$, so that translation invariance is preserved. If we concentrate on the case $p=1$ (the case $p>1$ will be discussed in example 3), then the indices (μ,ν) can be dropped and the evolution equations (34), (35) simplify in Fourier language to

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{a}(k;t)=-\epsilon\hat{a}(k;t)+\frac{\hat{M}(k)\hat{a}(k;t)}{\hat{b}(k;t)+T} \tag{38}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{b}(k;t)=-\epsilon\hat{b}(k;t)+\hat{M}(k)\left[\frac{\hat{a}(k;t)}{\hat{b}(k;t)+T}\right]^2 . \tag{39}$$

In addition to the spectrum of the output covariance tensor, which now is a translation-invariant kernel $Q(x-y)\equiv(CMC^{\mathrm{T}})(x-y)$, we are now in a position to also calculate the equilibrium spatial structure itself explicitly by solving (38), (39). In Fourier language all relevant operators are diagonal; eigenvalues are simply equal to Fourier components. For the output covariance matrix in equilibrium we find the eigenvalues

$$\lambda^Q(k)\equiv\hat{Q}(k)=\left[\lambda^M(k)-\epsilon T\right]\theta\left[\lambda^M(k)-\epsilon T\right] \qquad \lambda^M(k)\equiv\hat{M}(k)\,. \tag{40}$$

The operation performed by the system at any stage t in the learning process is given by the outcome of the spin relaxation: $q(x)=\int\mathrm{d}y\;c(x-y;t)m(y)$. Therefore the equilibrium Fourier coefficients $\hat{c}(k)$ tell us exactly how the system responds to input signals after the unsupervized learning stage:

$$\hat{c}(k)=\begin{cases}\sqrt{1-\epsilon T/\lambda^M(k)} & \text{if } \epsilon T<\lambda^M(k)\\ 0 & \text{if } \epsilon T\geqslant\lambda^M(k)\;.\end{cases} \tag{41}$$

If, for example, during the learning stage we present (in a one-dimensional system) input states of the type $m(x)=\cos(\omega x-\phi)$, in which the random phase variable ϕ and the random frequency ω are drawn from the distribution

$$\mathcal{P}(\omega,\phi)\equiv\frac{\theta(\phi)\theta(2\pi-\phi)}{[2\pi]^{\frac{1}{2}}\Delta\omega}e^{-\frac{1}{2}(\omega-\bar{\omega})^2/(\Delta\omega)^2}$$

we obtain the translation-invariant input covariance kernel $M(x-y)$:

$$M(x-y)=\tfrac{1}{2}\cos[\bar{\omega}(x-y)]\;\mathrm{e}^{-\frac{1}{2}(\Delta\omega)^2(x-y)^2}$$

$$\lambda^M(k)\equiv\hat{M}(k)=\frac{\sqrt{2\pi}}{4\Delta\omega}\left[\mathrm{e}^{-\frac{1}{2}((k-\bar{\omega})/\Delta\omega)^2}+\mathrm{e}^{-\frac{1}{2}((k+\bar{\omega})/\Delta\omega)^2}\right]$$

The $t\to\infty$ spectrum of the output covariance kernel Q follows directly from (40). In figure 5 we have depicted the eigenvalues $\lambda^M(k)$ and $\lambda^Q(k)$ and the Fourier components $\hat{c}(k)$ (using (41)). This picture clearly shows that after the learning stage the system behaves as a band filter.

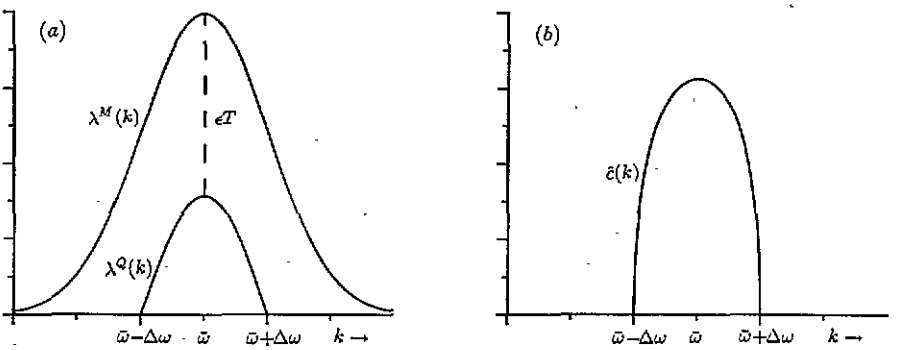

Figure 5. (*a*) Eigenvalue spectrum $\{\lambda^M(k)\}$ of the (translation-invariant) input covariance matrix and eigenvalue spectrum $\{\lambda^Q(k)\}$ of the (translation-invariant) output covariance matrix ($\epsilon T \equiv \lambda^M(\bar{\omega}+\Delta\omega)$, $\bar{\omega} \gg \Delta\omega$). (*b*) Fourier spectrum in equilibrium of the convolution kernel C as given by (41). Since $\tilde{q}(x) = \int dy\ C(x-y)m(y)$ this picture indicates that after the learning stage the system behaves as a band filter. Spatial frequencies k in the input for which $|k-\bar{\omega}| > \Delta\omega$ are suppressed.

Example 3. In our final example we expand on the previous one by considering in a one-dimensional system input signals of the form $m^\mu(x) = \gamma^\mu f^\mu(x-\phi)$, $\mu = 1 \ldots p$, where the $f^\mu(z)$ are even functions, ϕ is randomly drawn from a uniform distribution and γ is randomly drawn from a distribution $\mathcal{P}(\gamma)$. Apart from boundary effects (which we will neglect) the input covariance kernel M is translation-invariant and its Fourier transform is

$$\hat{M}^{\mu\nu}(k) = \Gamma^{\mu\nu}\hat{f}^\mu(k)\hat{f}^\nu(k) \tag{42}$$

where (as with example 1) $\Gamma^{\mu\nu} \equiv \langle\gamma^\mu\gamma^\nu\rangle_{\mathcal{P}}$. In contrast to the situation with example 1, the choice made for the functions $f^\mu(x)$ will now have an important impact on the way the network behaves, by virtue of the k-dependence of $\hat{f}$ in (42).

If, for instance, we choose $f^\mu(z)$ and $f^\nu(z)$ in such a way that $\hat{f}^\mu(k)\hat{f}^\nu(k) = \delta_{\mu\nu}[\hat{f}^\mu(k)]^2$ we find

$$\hat{M}^{\mu\nu}(k) = \delta_{\mu\nu}\left[\hat{f}^\mu(k)\right]^2\langle(\gamma^\mu)^2\rangle_{\mathcal{P}}$$

which implies that correlations between γ^μ and γ^ν with $\mu \neq \nu$ have become irrelevant. If, on the other hand, the input sources γ^μ are projected onto the network according to identical Gaussian blobs, $f^\mu(z) = f(z) \equiv \exp(-\frac{1}{2}x^2/\sigma^2)\ \forall\mu$, one finds

$$\lambda_i^M(k) = \lambda_i^\Gamma\hat{f}^2(k) \qquad i = 1, \ldots, p$$
$$\lambda_i^Q(k) = \left[\lambda_i^\Gamma\hat{f}^2(k) - \epsilon T\right]\theta[\lambda_i^\Gamma - \Delta(k)] \qquad i = 1, \ldots, p$$

with $\Delta(k) \equiv \epsilon T/\hat{f}^2(k)$ and $\{\lambda_i^\Gamma\}$ denoting the eigenvalues of the matrix $\Gamma^{\mu\nu}$. The filtering cut-off $\Delta(k)$, which determines whether or not components of Γ can pass through, has now become frequency-dependent. This means that after the learning stage, the filtering characteristics depend on the spatial presentation of the input signals. If afterwards the input consists solely of high spatial frequencies, a relatively small number of components of Γ can pass through (i.e. only those that have large eigenvalues). In particular, if $k^2 > \log[2\pi\sigma^2\lambda_{\max}^\Gamma/\epsilon T]/\sigma^2$ nothing will pass through at all. If the input is spatially constant (only the Fourier component $k = 0$ is present), $\Delta(k)$ is minimal and a minimal number of components will be suppressed.

4. Discussion

In this paper we have studied the function and development of spatial structure in layered Ising spin neural networks. Interactions *between* layers develop according to the Hebb rule; interactions *within* layers according to the so-called anti-Hebb rule. This architecture has a natural biological interpretation in terms of a layer of excitatory neurons projecting onto a layer of laterally interconnected inhibitory neurons [14, 15]. Spatial structure has been imposed by defining the absolute strength of the interactions to be position-dependent. The input state is assumed to be known; the output state is the result of a relaxation process. The present model is a generalization of the (fully connected) model studied in [14, 15], with which it turns out to share the properties that ergodicity remains unbroken in the thermodynamic limit and that the state correlations of neuron pairs that are relevant for the modification of synaptic interactions can be calculated analytically. These properties enabled us to derive detailed analytical results not only for the case of supervized learning but also for the (more complicated) case of unsupervized learning.

In the case of *supervized* learning the interactions remained fixed. For systems with spatial structure that varies slowly with distance we have been able to calculate at the level of local order parameters the (unique) equilibrium configuration as a function of the input configuration. In equilibrium, under certain temperature-dependent restrictions on the system's spatial structure, the output order parameter configuration turned out to be related to the input order parameter configuration by a *linear* transformation. This transformation, which depends on the temperature and the spatial structure between and within layers, can be calculated exactly and can give rise to counter-intuitive behaviour. Our analysis shows that one function of lateral structure may be to induce, through the *lateral* relaxation process, a spatial convergence of information that has been spread out by spatially divergent interactions *between* the layers.

Next we studied (in the adiabatic limit) the more complicated problem of how spatial structure can develop in layered Ising spin neural networks with lateral interactions, as the result of an *unsupervized* learning process, and how the developing structure depends on spatial properties of the information presented in the input layer. If the (spatially structured) initial interaction matrices have a separable structure, this structure is preserved by the learning process. This enabled us to describe the evolution in time of the interaction matrices on a macroscopic level of local order parameters. Although we were able to predict the final spatial structure explicitly only in some special cases, we could predict the information processing properties of the equilibrated system in all cases. To this end we have calculated the spectrum of the output covariance matrix in equilibrium as a function of the spectrum of the input covariance matrix. The equilibrated system turned out to perform a type of principal component analysis on the macroscopic input signal (the latter consists of position-dependent overlaps with a given set of prototype patterns). The fact that spatial characteristics and vectorial characteristics of the input and output signals are processed simultaneously allows for the development by unsupervized learning of modules that perform tasks like, for instance, band filtering and (spatial) frequency-dependent principal component analysis (we have worked out in detail some specific examples). A nice property with respect to self-organization is that, apart from some (fixed) system parameters, the filtering characteristics are determined by the input the network receives during the unsupervized learning stage.

We regard it an encouraging fact that our spatially structured model, in spite of the extreme non-linearity of its ingredients, turns out to perform a type of principal component analysis, as this appears to be a sensible manner of information processing of physiological

(sensory) input [12, 26, 27]. A natural next step will be to apply our results to realistic situations and find out what they might predict or explain upon choosing in our equations specific well documented biological structures, such as the cerebellum. In order to do so, we will then also have to take into account Dale's law, which, in the language of this paper, states that the interaction matrices must have a specific non-symmetric structure with respect to the signs of the matrix elements. This (biological) constraint has not been taken into account yet; the function of the present paper has been simply to develop intuition and appropriate analytical tools.

Appendix A. Concave structure functions

In this appendix we prove that symmetric Toeplitz matrices B (i.e. matrices of the form $B_{ij} = b(|i-j|)$, $i, j = 1 \ldots n$) are positive-definite if the function b is a positive monotonically decreasing concave function:

$$\forall x \geqslant 0: \qquad b(x) > 0 \qquad \frac{\mathrm{d}}{\mathrm{d}x} b(x) < 0 \qquad \frac{\mathrm{d}^2}{\mathrm{d}x^2} b(x) > 0 .$$

In the following proof the key idea, which we owe to Jan de Boer, consists of defining the auxiliary variables c_n:

$$c_{n-1} \equiv b(n-1) \qquad c_{n-2} \equiv b(n-2) - b(n-1)$$
$$c_k \equiv b(k) - 2b(k+1) + b(k+2) \qquad k = 0 \ldots n-3 .$$

By construction all c_k ($k = 0 \ldots n-1$) are positive. The quantities defining our matrix B can be written in terms of the $\{c_n\}$:

$$b(k) = \sum_{l=k}^{n-2} c_l (l+1-k) + c_{n-1} .$$

If we write the quadratic form $x \cdot Bx$ in terms of the new variables $\{c_n\}$ we obtain

$$x \cdot Bx \equiv \sum_{i,j=1}^{n} x_i x_j b(|i-j|) = \sum_{i,j=1}^{n} x_i x_j \sum_{k=0}^{n-1} b(k) \delta_{k,|i-j|}$$
$$= c_0 \|x\|^2 + \sum_{l=1}^{n-2} c_l x T^{(l+1)} x + c_{n-1} \left[\sum_{i=1}^{n} x_i \right]^2$$

in which

$$T_{ij}^{(w)} \equiv \sum_{k=0}^{w} (w-k) \delta_{w,|i-j|} = w \left[1 - \frac{|i-j|}{w} \right] \theta(w - |i-j|) .$$

The matrices $T^{(w)}$ are Toeplitz matrices with a triangular structure function, which have already been proven to be semi-positive-definite (see section 2.4, table 1). We conclude that $\lambda_{\min}(B) \geqslant c_0 > 0$.

Appendix B. Macroscopic equilibrium for general T

In this appendix we present the proof that for random patterns the expression

$$\bar{q} = [B + T\mathbb{1}]^{-1} Am \tag{B1}$$

defines a *global* (and therefore unique) attractor of the macroscopic set of $T > 0$ evolution equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_k^\mu = -q_k^\mu + \langle \xi^\mu \mathrm{sat}[\beta \xi^\rho \left(A_{kl}^{\rho\nu} m_l^\nu - B_{kl}^{\rho\nu} q_l^\nu \right)] \rangle_\xi \tag{B2}$$

if we impose the conditions

$$\forall_{k \leqslant n}: \qquad \sum_\mu |\bar{q}_k^\mu| \leqslant 1 \qquad \text{and} \qquad T > -2\lambda_{\min}(B)\,. \tag{B3}$$

For notational convenience we have introduced the summation convention: double occurrence of an index means that it is understood to be summed over. To show that (B1) is globally stable we study the variables r_k^μ:

$$r_k^\mu = P_{kl}^{\mu\nu}(q_l^\nu - \bar{q}_l^\nu)$$

where $P \equiv \mathbb{1} + \beta B$. Note that the conditions (B3) guarantee that P is positive-definite. Following [28], we write the evolution equations for $\{r_k^\mu\}$ as linear equations, perturbed by non-linearities:

$$\frac{\mathrm{d}}{\mathrm{d}t} r_k^\mu = -P_{kl}^{\mu\nu} r_l^\nu + P_{kl}^{\mu\nu} \phi_l^\nu(\beta B P^{-1} r)$$

where

$$\begin{aligned}\phi_l^\nu(r) &\equiv (r_l^\nu - \bar{q}_l^\nu) - \langle \xi^\nu \mathrm{sat}[\xi^\rho (r_l^\rho - \bar{q}_l^\rho)] \rangle_\xi \\ &= \langle \xi^\nu \left\{ \xi^\rho (r_l^\rho - \bar{q}_l^\rho) - \mathrm{sat}\left[\xi^\rho (r_l^\rho - \bar{q}_l^\rho)\right] \right\} \rangle_\xi\,.\end{aligned}$$

In the last step we have made use of the fact that we are dealing with random patterns. For all $\{r_k^\mu\}$ the norm $\|\phi(r)\| \equiv \{\phi_l^\nu(r)\phi_l^\nu(r)\}^{1/2}$ obeys the inequality

$$\|\phi(r)\| \leqslant \|r\|\,. \tag{B4}$$

To show this, we define the function $\gamma(a, b)$ of two scalars a, b by

$$\gamma(a, b) \equiv \frac{(a - b) - \mathrm{sat}(a - b)}{a}$$

which can be shown to be bounded according to $0 \leqslant \gamma(a, b) \leqslant 1$ if $a \in R$ and $b \in [-1, 1]$. Because of the first condition in (B3) all $|\xi^\rho \bar{q}_l^\rho| \leqslant 1$, therefore we can write

$$\phi_l^\nu(r) = \Gamma^{\nu\rho}(r) r_l^\rho \qquad \Gamma^{\nu\rho}(r) \equiv \langle \xi^\nu \xi^\rho \gamma_\xi(r, \bar{q}) \rangle_\xi$$

with $0 \leqslant \gamma_\xi(r, \bar{q}) \leqslant 1$. The quadratic form $x^\nu \Gamma^{\nu\rho} x^\rho$ obeys

$$0 \leqslant x^\nu \Gamma_{\nu\rho} x^\rho \equiv \langle (\xi^\nu x^\nu)^2 \gamma_\xi(r, \bar{q}) \rangle_\xi \leqslant \langle (\xi^\nu x^\nu)^2 \rangle_\xi = x^\nu x^\nu\,.$$

The eigenvalues of the (symmetric) matrix Γ are therefore between 0 and 1, which proves statement (B4).

Finally we define the non-negative scalar function E (our candidate for a Lyapunov function):

$$E(r) \equiv r_k^\mu (P^{-1})_{kl}^{\mu\nu} r_l^\nu$$

which is zero only for $r_k^\mu = 0$ $(\forall k\mu)$. The time derivative of $E(r)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} E(r) = -r_k^\mu r_k^\mu + r_k^\mu \phi_k^\mu(\beta B P^{-1} r) \leqslant \{-1 + \|\beta B P^{-1}\|\}\|r\|^2$$

where we have made use of (B4). The function E decreases monotonically as soon as for all eigenvalues $\lambda(B)$: $-1 < \beta\lambda(B)/[1+\beta\lambda(B)] < 1$. These conditions are indeed satisfied if $T > -2\lambda_{\min}(B)$, in which case E will decrease monotonically to 0 and (B1) is a global attractor of the system (B2).

References

[1] Hebb D.O. 1949 *The Organization of Behaviour* (New York: Wiley) p 62
[2] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554–8
[3] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007–18
[4] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167–73
[5] Fassnacht C and Zippelius A 1991 *Network* **2** 63–84
[6] O'Kane D and Treves A 1992 *J. Phys. A: Math. Gen.* **25** 5055–69
[7] Noest A J 1989 *Phys. Rev. Lett.* **63** 1739–42
[8] Coolen A C C 1990 *Statistical Mechanics of Neural Networks* ed L Garrido (Berlin: Springer) pp 318–96
[9] Coolen A C C and Lenders L G V M 1992 *J. Phys. A: Math. Gen.* **25** 2577–92
[10] Kohonen T 1982 *Associative Memory—a System Theoretic Approach* (New York: Springer)
[11] Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158–9
[12] Rubner J and Schulten K 1990 *Biol. Cybern.* **62** 193–9
[13] Földiak P 1990 *Biol. Cybern.* **64** 165–70
[14] Jonker H J J, Coolen A C C and Denier van der Gon J J 1989 *Proc. IEE on Artificial Neural Networks* (London: IEE) p 23–6
[15] Jonker H J J and Coolen A C C 1991 *J. Phys. A: Math. Gen.* **24** 4219–34
[16] Shinomoto S 1987 *J. Phys. A: Math. Gen.* **20** L1305–9
[17] Dong D W and Hopfield J J 1992 *Network* **3** 267–83
[18] Benaim M 1992 *Europhys. Lett.* **19** 241–6
[19] Carlson A 1990 *Biol. Cybern.* **64** 171–6
[20] Rubner J and Tavan P 1989 *Europhys. Lett.* **10** 693–8
[21] Leen T K 1991 *Network* **2** 85–105
[22] Coolen A C C and Sherrington D 1992 *Oxford University Preprint* OUTP-92-49S
[23] Grenander U and Szegö G 1958 *Toeplitz Forms and their Applications* (California: University of California Press) p 62
[24] Lawley D N and Maxwell A E 1971 *Factor Analysis as a Statistical Method* (London: Butterworth)
[25] Oja E 1982 *J. Math. Biol.* **15** 267–73
[26] Sanger T D 1989 *Neural Networks* **2** 459–73
[27] Hancock P J B, Baddeley R J and Smith L S 1992 *Network* **3** 61–70
[28] Khalil H K 1992 *Non-linear Systems* (New York: Macmillan) p 193–211