Unsupervised dynamic learning in layered neural networks

1991 J. Phys. A: Math. Gen. 24 4219

(http://iopscience.iop.org/0305-4470/24/17/032)

Unsupervised dynamic learning in layered neural networks

H J J Jonker and A C C Coolen

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

Received 18 December 1990, in final form 10 April 1991

Abstract. We consider a stochastic two-layer neural network of binary neurons in which the connections between the layers are updated according to the Hebb rule, whereas the lateral connections in the output layer are modified according to an anti-Hebb rule. In equilibrium the output overlap is found to be a linear transformation of the input overlap. Next we extend the model by considering learning as a dynamic process, which means that synaptic efficacies as well as neuronal states may vary in time. Despite the coupling of these two variables, we show that in this particular model the behaviour can be well analysed. It turns out that the network filters the information available at the input in such a way that important components of the input data can pass through, whereas components with a low information content are suppressed.

1. Introduction

The *anti-Hebb* rule has been studied in a variety of models with different objectives, e.g. the novelty filter of Kohonen [1], the unlearning principle of Hopfield [2] and the Boltzmann machine [3]. Recent studies [4, 5] consider models of two-layer (input-output) networks in which the connections between the layers are modified according to the standard Hebb rule, whereas the lateral connections within the output layer are modified according to an anti-Hebb rule. In [4] the neurons are modelled as *linear* elements and the lateral connections are organized in a hierarchical way. Due to this architecture and learning process, orthogonal feature detectors arise. In [5] the network considered consists of *binary* neurons which have *symmetric* lateral connections in the output layer. In the learning stage combinations of input-output patterns are clamped on the network. It turned out that in the operational stage an arbitrary input pattern gave rise to an output pattern that correlated with the learnt output patterns in the same way as the input pattern correlated with the learnt input patterns. In this way the network was able to interpolate linearly between a set of basic patterns.

Because this property of interpolating linearly between stored patterns is interesting and possibly physiologically relevant, we generalize these results by studying a more general learning scheme and by introducing noise into the system.

Next we study how learning takes place as an unsupervised dynamic process. In other words, during learning we do not clamp patterns on the output, but we let the network generate the output patterns as a result of the current connections and input. The generated patterns form the basis on which the connections are modified. Actually, the output layer is treated merely as a hidden layer. In the analysis of models with

0305-4470/91/174219+16$03.50

hidden layers one usually encounters fundamental problems owing to the strong coupling between the state of the neurons and the connections in the network (the connections determine the state of the neurons, which in turn affect the connections because of the learning rule). Although there are some exceptions [6, 7], it is generally very hard or even impossible to obtain analytical results.

We will show that in the case of the interpolation model the behaviour of the system can be well analysed. This is possible because the statistical properties of the network are known.

Finally, we will discuss the relation of the model to physiology.

2. Supervised learning

In the model two layers are distinguished: an input layer containing N_i neurons and an output layer containing N_o neurons. The total number of neurons is denoted by $N = N_i + N_o$. The neurons are modelled by Ising spins: if s_i represents the state of neuron i, it either assumes the value +1 (neuron i fires) or the values −1 (neuron i is quiescent). The state of the input neurons and the output neurons will be denoted by the vectors $\boldsymbol{s}^{\text{in}}$ and $\boldsymbol{s}^{\text{out}}$ respectively.

The output neurons are mutually connected via synaptic couplings J^{oo}. In addition they receive input from the input layer via couplings J^{io}. During the learning phase, specific patterns are clamped both on the input and on the output side. The input and output patterns are denoted by $\boldsymbol{\chi}^{(\nu)} \nu = 1 \ldots p$ and $\boldsymbol{\xi}^{(\mu)} \mu = 1 \ldots p$, respectively; p denotes the number of different patterns.

We consider the following supervised learning scheme. When the input–output combination $(\boldsymbol{\chi}^{(\nu)}, \boldsymbol{\xi}^{(\mu)})$ is clamped on the network, the values of the connections are modified according to the following learning rules:

$$\Delta J^{\text{io}}_{ik} = +\frac{1}{N} \xi^{(\mu)}_i \chi^{(\nu)}_k \tag{1}$$

$$\Delta J^{\text{oo}}_{ij} = -\frac{1}{N} \xi^{(\mu)}_i \xi^{(\mu)}_j \qquad \Delta J^{\text{oo}}_{ii} = 0. \tag{2}$$

The first learning rule (1) is the standard (generalized) Hebb rule [8] applied to a two-layer network; the second learning rule (2) could be called an *anti-Hebb* rule.

Let $A_{\mu\nu}$ denote the number of times that $\boldsymbol{\chi}^{(\nu)}$ has been clamped on the input at the same time as $\boldsymbol{\xi}^{(\mu)}$ on the output, and let $B_{\mu\mu}$ denote the total number of times that $\boldsymbol{\xi}^{(\mu)}$ has been clamped on the output. Then, assuming the initial values were zero, we can write the value of the connections after the learning process as:

$$J^{\text{io}}_{ik} = +\frac{1}{N} \sum_{\mu\nu}^{p} \xi^{(\mu)}_i A_{\mu\nu} \chi^{(\nu)}_k$$

$$J^{\text{oo}}_{ij} = -\frac{1}{N} \sum_{\mu\nu}^{p} \xi^{(\mu)}_i B_{\mu\nu} \xi^{(\nu)}_j \qquad J^{\text{oo}}_{ii} = 0.$$

Both A and B are $p \times p$ matrices; note that B is a diagonal matrix here.

After the learning stage an operational stage is considered, during which a certain state $\boldsymbol{s}^{\text{in}}$ is clamped on the input. The initial output state is chosen at random. Evolution in the network takes place by asynchronous updating of the output neurons with the probability w_i that s_i^{out} will change its state (flips):

$$w_i(\boldsymbol{s}^{\text{out}}) = \tfrac{1}{2}(1 - g(\beta h_i^{\text{out}} s_i^{\text{out}})) \tag{3}$$

where $g(x)$ is an arbitrary odd monotonically rising function of x, which is bounded by -1 and $+1$ and for which $g'(0) = 1$. Examples of such functions are $\tanh(x)$ or $\operatorname{erf}(\sqrt{\pi}\, x/2)$. At this stage we will not yet assign a specific function to g. h_i^{out} represents the total post-synaptic input (PSP) of output neuron i and consists of the contributions of the input and output neurons:

$$h_i^{\text{out}} = \sum_{j=1}^{N_o} J_{ij}^{\text{oo}} s_j^{\text{out}} + \sum_{k=1}^{N_i} J_{ik}^{\text{io}} s_k^{\text{in}}. \tag{4}$$

Finally, $\beta \equiv 1/T$, the temperature T being a measure of the amount of noise in the system.

In order to analyse the system at a macroscopic level, we introduce the overlap (or correlation) parameters [9]:

$$\boldsymbol{m} = \frac{1}{N_i} \sum_{k=1}^{N_i} \boldsymbol{\chi}_k s_k^{\text{in}} \qquad \boldsymbol{q} = \frac{1}{N_o} \sum_{j=1}^{N_o} \boldsymbol{\xi}_j s_j^{\text{out}} \tag{5}$$

where we have adopted the notation $\boldsymbol{\chi}_k = (\chi_k^{(1)}, \ldots, \chi_k^{(p)})$, $\boldsymbol{\xi}_j = (\xi_j^{(1)}, \ldots, \xi_j^{(p)})$. Taking $n = N_i/N_o$, the PSP h_i^{out} in (4) can now be rewritten as

$$h_i^{\text{out}} = \boldsymbol{\xi}_i \cdot \left[\frac{n}{n+1} A\boldsymbol{m} - \frac{1}{n+1} B\boldsymbol{q}\right] + \frac{1}{N} \boldsymbol{\xi}_i B \boldsymbol{\xi}_i s_i^{\text{out}}.$$

In appendix A it is shown by taking the limit $N_o \to \infty$, n and p fixed, that the dynamic behaviour of the model is governed by the nonlinear autonomous differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{q} = -\boldsymbol{q} + \lim_{N_o \to \infty} \frac{1}{N_o} \sum_{i=1}^{N_o} \boldsymbol{\xi}_i g\left(\beta \boldsymbol{\xi}_i \cdot \left[\frac{n}{n+1} A\boldsymbol{m} - \frac{1}{n+1} B\boldsymbol{q}\right]\right). \tag{6}$$

Although B is a diagonal matrix having only positive elements, in the following it is sufficient for B to be symmetric and positive definite (thus invertible).

In the limit $\beta \to \infty$ any choice of g boils down to the sign function. If the contribution of the input neurons is considered as an external field, the energy function of this system has the same form as the Hamiltonian of an Ising spin system:

$$E = -\tfrac{1}{2} \sum_{i=1}^{N_o} \sum_{j=1}^{N_o} s_i^{\text{out}} J_{ij}^{\text{oo}} s_j^{\text{out}} - \sum_{i=1}^{N_o} s_i^{\text{out}} \left(\sum_{k=1}^{N_i} J_{ik}^{\text{io}} s_k^{\text{in}} \right).$$

To simplify the equations, we define $\boldsymbol{m}' = nB^{-1}A\boldsymbol{m}$ and $\boldsymbol{x} = \boldsymbol{q} - \boldsymbol{m}'$. Neglecting a constant, we can write the energy:

$$E = \frac{N_o}{2(n+1)} \boldsymbol{x} B \boldsymbol{x}.$$

It is known that under the $T = 0$ dynamics the energy will decrease to a (local) minimum. Now we will show that E decreases to the global minimum 0, if some conditions are imposed on the input overlap $\boldsymbol{m}$. Using (6), differentiation of E with respect to t yields

$$\frac{\mathrm{d}}{\mathrm{d}t} E = -2E - \boldsymbol{m}' \cdot \boldsymbol{\nabla} E - \lim_{N_o \to \infty} \frac{1}{N_o} \sum_{i=1}^{N_o} |\boldsymbol{\xi}_i \cdot \boldsymbol{\nabla} E|. \tag{7}$$

In order to obtain a proper characterization of the equilibrium solution, it is convenient to work with the partition method as introduced in [10]. The set of all indices $i \leqslant N_o$ is divided into subsets I_η defined by

$$I_\eta = \{i \leqslant N_o \,|\, \xi_i = \eta\}.$$

The mean activity of the neurons in subset I_η is given by

$$q_\eta(s^{\text{out}}) = \frac{1}{|I_\eta|} \sum_{i \in I_\eta} s_i^{\text{out}}.$$

For the correlation $q(s^{\text{out}})$ it holds that

$$q(s^{\text{out}}) = \sum_\eta \eta p_\eta q_\eta(s^{\text{out}}) \tag{8}$$

where $p_\eta = |I_\eta| / N_o$ denotes the fraction of indices that belong to subset I_η. If the output patterns are randomly drawn from a certain distribution, p_η corresponds to the probability that $i \in I_\eta$.

Since by definition all $|q_\eta| \leqslant 1$, equation (8) clearly demonstrates that all vectors $q(s^{\text{out}})$ are contained in a (p-dimensional) *limited* space, which we will call D. The shape of D is determined by the quantities p_η which in turn depend on the particular choice of the output patterns.

In fact, any vector $m' \in R^p$ can be written in the form $m' = \sum_\eta \eta p_\eta m'_\eta$, although this does not necessarily imply that all $|m'_\eta| \leqslant 1$. But if the latter condition has been satisfied, then m' is an element of D, which means that m' can be represented by $q(s^{\text{out}})$. In other words, an output configuration s^{out} exists such that $q(s^{\text{out}}) = m'$.

Applying the partition method to (7), we get

$$\frac{\mathrm{d}}{\mathrm{d}t} E = -2E - \sum_\eta p_\eta |\eta \cdot \nabla E| (1 + m'_\eta \operatorname{sgn}(\eta \cdot \nabla E)).$$

If $m' \in D$, that is, if all $|m'_\eta| \leqslant 1$, then

$$\frac{\mathrm{d}}{\mathrm{d}t} E \leqslant -2E. \tag{9}$$

Combining (9) with the fact that $\forall x \neq 0$: $E > 0$ (since B is positive definite), it follows that $E \to 0$, which means that $x \to 0$, and consequently $q \to m'$. So, provided that

$$m \in \{z \in D \,|\, nB^{-1}Az \in D\} \tag{10}$$

the equilibrium solution of (6) is:

$$q = nB^{-1}Am. \tag{11}$$

Note that condition (10) prescribes that the right-hand side of (11) must be within D, the space of vectors attainable by $q(s^{\text{out}})$. If this condition is not satisfied, equation (11) does not hold, since the right-hand side yields a result that cannot be represented by a correlation parameter (for instance values larger than 1). Throughout the rest of this paper, we will assume that m is such that condition (10) is always satisfied.

There are several remarkable aspects about equilibrium solution (11). First of all, the output correlation q appears to be a linear transformation of the input correlation m. This is surprising if one takes into consideration the fact that the network is composed of binary neurons and that there is no noise in the system. Secondly, the solution does not depend on the initial output state, nor does it depend on the choice of the patterns;

for instance, they need not be uncorrelated. Another remarkable aspect is the presence of the inverse of B, which yields interesting features. For instance, if during the learning phase input–output combinations of the form $(\boldsymbol{\chi}^{(\mu)}, \boldsymbol{\xi}^{(\mu)})$ have been presented to the network for every μ just once, both A and B will have become $p \times p$ identity matrices. Consequently, the equilibrium solution will be $\boldsymbol{q} = n\boldsymbol{m}$. If, hereafter, for instance, input–output combination $(\boldsymbol{\chi}^{(1)}, \boldsymbol{\xi}^{(1)})$ is presented once again, both A and B will change, whereas the equilibrium solution will remain the same because of the rescaling property of the inverse operation. In general, this implies that the number of times that input–output combinations $(\boldsymbol{\chi}^{(\mu)}, \boldsymbol{\xi}^{(\mu)})$ are presented to the network does not affect the $T=0$ equilibrium solution, although the connections in the network do change.

Utilizing our knowledge of the equilibrium solution at $T=0$, for small T we can follow a perturbation-theoretical approach by trying a formal expansion of $\boldsymbol{q}$ with respect to T:

$$\boldsymbol{q} = \boldsymbol{q}_0 + T\boldsymbol{q}_1 + T^2\boldsymbol{q}_2 + \ldots$$

$\boldsymbol{q}_0$ is given by (11). Substitution of this expansion in (6) yields:

$$\boldsymbol{q}_0 + T\boldsymbol{q}_1 + T^2\boldsymbol{q}_2 + \ldots = -\lim_{N_o \to \infty} \frac{1}{N_o} \sum_{i=1}^{N_o} \boldsymbol{\xi}_i g\left(\frac{1}{n+1}\boldsymbol{\xi}_i \cdot B[\boldsymbol{q}_1 + T\boldsymbol{q}_2 + T^2\boldsymbol{q}_3 + \ldots]\right).$$

At this point the usual step is to perform a Taylor expansion of g in the neighbourhood of $\boldsymbol{\xi}_i B\boldsymbol{q}_1/(n+1)$ and collect all terms of the same order in T. This results in a recurrence relation for $\boldsymbol{q}_l$ ($l = 1, \ldots$). So, to proceed it is necessary to assign a function to g. For convenience, we opted for the semi-linear function:

$$g(x) = \begin{cases} x & \text{for } |x| \leqslant 1 \\ \operatorname{sgn}(x) & \text{for } |x| > 1. \end{cases} \tag{12}$$

If $\boldsymbol{m}$ is not too large (the particular boundaries will be specified later), g can be regarded as linear, and we find the recurrence relation

$$\boldsymbol{q}_l = -\frac{1}{n+1} QB\boldsymbol{q}_{l+1}$$

where the correlation matrix Q is defined by

$$Q_{\mu\nu} \equiv \frac{1}{N_o} \sum_{i=1}^{N_o} \xi_i^{(\mu)} \xi_i^{(\nu)}.$$

The resulting series for $\boldsymbol{q}$, the convergence of which is controlled by the magnitude of T, can be evaluated:

$$\boldsymbol{q} = n[I + T(n+1)B^{-1}Q^{-1}]^{-1}B^{-1}A\boldsymbol{m} \tag{13}$$

(I denotes the $p \times p$ identity matrix).

By direct substitution of (13) in equation (6), one can readily verify that (13) holds for general T, provided that $\boldsymbol{m}$ is such that $\Sigma_\mu |(Q^{-1}\boldsymbol{q})_\mu| \leqslant 1$. The latter restriction must be imposed because the argument of g may not extend beyond the region where g is linear. Note that the $T=0$ solution is also contained in (13).

At the microscopic level, however, there is a fundamental difference between $T=0$ and $T>0$ when equilibrium has been reached. At $T=0$ there will be no more spin-flips once the network has reached an arbitrary configuration $\boldsymbol{s}^{\text{out}}$ for which $\boldsymbol{q}(\boldsymbol{s}^{\text{out}}) = n B^{-1}A\boldsymbol{m}$. On the other hand, a system with some noise ($T>0$) will perform a sort of random walk in configuration space through the plane of all $\boldsymbol{s}^{\text{out}}$ for which $\boldsymbol{q}(\boldsymbol{s}^{\text{out}})$ satisfies (13). This enables us to define the average configuration $\langle \boldsymbol{s}^{\text{out}} \rangle$. In appendix A an expression is derived which connects the average configuration with the output correlation and the learnt patterns (see (31)):

$$\langle s_i^{\text{out}} \rangle = \boldsymbol{\xi}_i \cdot Q^{-1} \boldsymbol{q}.$$

This quantity will play an important role in the next section.

3. Learning as a dynamic process

In the previous section a clear distinction was made between the learning stage and the operational stage. In the learning stage the connections were adjusted while both input and output layer were forced into an activity pattern. In the operational stage the connections were fixed while the neurons could freely change their states as a result of these connections. In this section we will drop the strict distinction between the two stages and consider a system in which learning and operating happen in an integrated way. After initializing the connections, we will no longer force the output into a pattern, but allow the network itself to generate the output patterns which in turn serve to update the connections.

The essential (and probably plausible) assumption we make is that the values of the connections change on a much larger time-scale than that on which the neurons change their states. With regard to the Hebbian mechanism, this implies that the condition for synaptic enhancement is not simply the *conjunction* of pre- and post-synaptic activity, but that a *correlation* between pre- and post-synaptic activity over a certain period of time is required [11]. Furthermore we introduce a decay factor δ to prevent the connections from unbounded growth [1, 6, 7, 12]. Expressed as a formula:

$$\Delta J_{ik}^{\text{io}} = + \frac{\varepsilon}{N} \langle s_i^{\text{out}} s_k^{\text{in}} \rangle_\tau - \delta J_{ik}^{\text{io}} \tag{14}$$

$$\Delta J_{ij}^{\text{oo}} = - \frac{\varepsilon}{N} \langle s_i^{\text{out}} s_j^{\text{out}} \rangle_\tau - \delta J_{ij}^{\text{oo}} \qquad \Delta J_{ii}^{\text{oo}} = 0 \tag{15}$$

with ε a small constant representing the learning rate and τ a certain period. It should be noted that equations (14) and (15) are in fact generalizations of equations (1) and (2).

Suppose that at $t = 0$ the initial connections are given by

$$J_{ik}^{\text{io}}(0) = + \frac{1}{N} \sum_{\mu\nu} \xi_i^{(\mu)} A_{\mu\nu}^0 \chi_k^{(\nu)} \tag{16}$$

$$J_{ij}^{\text{oo}}(0) = - \frac{1}{N} \sum_{\mu\nu} \xi_i^{(\mu)} B_{\mu\nu}^0 \xi_j^{(\nu)}. \tag{17}$$

A^0 and B^0 may be arbitrary matrices, although B^0 is restricted to the class of symmetric and positive definite matrices. For simplicity we take both input and output patterns mutually uncorrelated.

During the following process, a large number of input configurations is presented at the input. We assume that there is some noise present at the output as well as at the input, but we will take the correlation of the input configurations with the learnt input patterns, $\boldsymbol{m}(t)$, to be constant during a period τ, which is much larger than the time needed for the network to reach equilibrium. If τ is large enough, the average over τ may be replaced by the average as computed in appendix A, equations (31) and (32). This leads to

$$\Delta J^{\mathrm{io}}_{ik} = +\frac{\varepsilon}{N}\sum_{\mu\nu} \xi_i^{(\mu)} q_\mu m_\nu \chi_k^{(\nu)} - \delta J^{\mathrm{io}}_{ik}$$

$$\Delta J^{\mathrm{oo}}_{ij} = -\frac{\varepsilon}{N}\sum_{\mu\nu} \xi_i^{(\mu)} q_\mu q_\nu \xi_j^{(\nu)} - \delta J^{\mathrm{oo}}_{ij} \qquad \Delta J^{\mathrm{oo}}_{ii} = 0.$$

If the parameter ε is relatively small and the number of times that connection modifications take place is large, we can approximate the difference equations by a system of differential equations. The expectation value of the connections at time t can then be written as

$$\overline{J^{\mathrm{io}}_{ik}}(t) = +\frac{1}{N}\sum_{\mu\nu} \xi_i^{(\mu)} A_{\mu\nu}(t) \chi_k^{(\nu)}$$

$$\overline{J^{\mathrm{oo}}_{ij}}(t) = -\frac{1}{N}\sum_{\mu\nu} \xi_i^{(\mu)} B_{\mu\nu}(t) \xi_j^{(\nu)} \qquad \overline{J^{\mathrm{oo}}_{ii}}(t) = 0$$

with $A(t)$ and $B(t)$ given by the differential equations

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} A_{\mu\nu} &= \varepsilon \overline{q_\mu m_\nu} - \delta A_{\mu\nu} \qquad & A(0) &= A^0 \\ \frac{\mathrm{d}}{\mathrm{d}t} B_{\mu\nu} &= \varepsilon \overline{q_\mu q_\nu} - \delta B_{\mu\nu} \qquad & B(0) &= B^0. \end{aligned} \tag{18}$$

Recalling (13), the expression for $\boldsymbol{q}$ as a function of $\boldsymbol{m}$, and the orthogonality of the output patterns ($Q = I$), we get after rescaling the temperature $\tilde{T} = (n+1)T$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} A &= n\varepsilon[B + \tilde{T}I]^{-1} AM - \delta A \qquad & A(0) &= A^0 \\ \frac{\mathrm{d}}{\mathrm{d}t} B &= n^2\varepsilon[B + \tilde{T}I]^{-1} AMA^{\mathrm{T}}[B + \tilde{T}I]^{-1} - \delta B \qquad & B(0) &= B^0. \end{aligned} \tag{19}$$

Here M denotes the covariance matrix of the input defined by $M_{\mu\nu} \equiv \overline{m_\mu m_\nu}$. Note that by definition M is semi-positive definite. For the time being, we will assume that M is also invertible (thus positive definite).

Since the system of equations (19) describes two coupled nonlinear differential equations for $p \times p$ matrices, an analytical expression for A and B at any t is hard to find. In order to get rid of the inverse operations, we introduce matrices $C \equiv n[B + \tilde{T}I]^{-1}A$ and $V \equiv [B + \tilde{T}I]^{-1}$, leading to the system of equations

$$\frac{\mathrm{d}}{\mathrm{d}t} V = \delta V - \tilde{T}\delta V^2 - \varepsilon VCMC^{\mathrm{T}} V \tag{20}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} C = -\varepsilon VCM\left[C^{\mathrm{T}}C - nI + \tilde{T}\frac{\delta}{\varepsilon} M^{-1}\right]. \tag{21}$$

Though still coupled, these equations are much more suitable for numerical integration. Note that C represents the transformation (13) of the input correlation $\boldsymbol{m}(Q = I)$.

As we demonstrate below we are interested in the matrix $Z \equiv C^{\mathrm{T}}C$ rather than C in particular. Since it is still very difficult to find a closed form for $Z(t)$, we will confine ourselves to finding the stationary solution $\hat{Z}$ analytically (see appendix B). It appears that $\hat{Z}$ commutes with M, which implies that $\hat{Z}$ and M can be reduced to the diagonal form by the same base transformation. Let the eigenvalues of M be arranged in increasing order ($\mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_p$). Transformed to a diagonal matrix, $\hat{Z}$ acquires the form (see (41)):

$$n\begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 1-\frac{\tilde{T}\delta}{n\varepsilon}\mu_i^{-1} & & \\ & & & & \ddots & \\ & & & & & 1-\frac{\tilde{T}\delta}{n\varepsilon}\mu_p^{-1} \end{pmatrix} \qquad \mu_{i-1} \leqslant \frac{\tilde{T}\delta}{n\varepsilon} < \mu_i .$$

So, if the temperature $\tilde{T}$ is chosen larger than the critical value $n\varepsilon\mu_p/\delta$, all eigenvalues of $\hat{Z}$ are zero, representing the case where all information has been lost in the system. If $\tilde{T}$ is chosen below $n\varepsilon\mu_1/\delta$, all eigenvalues differ from zero; in this case the p-degrees of freedom present in the input space are all conserved. If the value of $\tilde{T}$ is gradually raised, eigenvalues of $\hat{Z}$ successively become zero and degrees of freedom are lost.

It is interesting to analyse the stationary form of the output covariance matrix i.e. $\overline{q_\mu q_\nu}$ (see appendix B). Reduced to the diagonal form, not necessarily by the same base transformation as that applied to M and $\hat{Z}$, the output covariance matrix reads:

$$n\begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \mu_i-\frac{\tilde{T}\delta}{n\varepsilon} & & \\ & & & & \ddots & \\ & & & & & \mu_p-\frac{\tilde{T}\delta}{n\varepsilon} \end{pmatrix} \qquad \mu_{i-1} \leqslant \frac{\tilde{T}\delta}{n\varepsilon} < \mu_i .$$

So the eigenvalues of the output covariance matrix can be determined by the following procedure. Subtract $\tilde{T}\delta/n\varepsilon$ from an eigenvalue of M; if the result is positive, one finds the corresponding eigenvalue of the output covariance matrix by multiplying the result by n; if the result of the subtraction is negative, then the eigenvalue is zero.

4. Numerical experiments

There are three suitable ways of studying the model, each involving an increasingly higher level: first of all, by carrying out the actual simulations at the microscopic level; secondly, by repeatedly numerically integrating the flow equations (6) for the order parameters and modifying A and B; and lastly, by numerically integrating the system of differential equations (20) and (21). Before we describe each method in detail, quantities must be found which enable us to compare the different methods with each other. As we will see, appropriate quantities are

$$r^{\alpha\beta} \equiv \frac{1}{N_o}\sum_{i=1}^{N_o} \langle s_i^{\mathrm{out}}\rangle^{(\alpha)} \langle s_i^{\mathrm{out}}\rangle^{(\beta)} \tag{22}$$

i.e. the correlation between the average pattern $\langle \boldsymbol{s}^{\text{out}}\rangle^{(\alpha)}$, which is the result of an input correlation $\boldsymbol{m}^{(\alpha)}$, and $\langle \boldsymbol{s}^{\text{out}}\rangle^{(\beta)}$, the result of $\boldsymbol{m}^{(\beta)}$. It is most convenient to choose $\boldsymbol{m}^{(\alpha)}$ and $\boldsymbol{m}^{(\beta)}$ proportional to the unity vectors, $m_\mu^{(\alpha)} \sim \delta_{\mu\alpha}$, $m_\mu^{(\beta)} \sim \delta_{\mu\beta}$, with $\alpha = 1 \ldots p$, $\beta = \alpha \ldots p$.

The recipe for a simulation is as follows. First of all, choose matrices A^0 and B^0 and initialize connections according to (16) and (17). To commence the actual simulation, draw a vector $\boldsymbol{m}$ from a distribution characterized by covariance matrix M. Next, let the network evolve to equilibrium according to an asynchronous updating process (3) with $T>0$. Continue this process after equilibrium has been reached. In order to determine the average configuration, take once in a while a 'snapshot' of the system, that is, store the current configuration. Afterwards the average activity pattern can be computed by summing all configurations stored and dividing the result by the total number. Obviously the larger the number, the better is the approximation of the theoretical average (31). The next step is to update the connections according to (14) and (15) with application of (30). Then draw another $\boldsymbol{m}$ etc.

Analysis of the network at certain times is performed by presenting the network input configurations with correlations $\boldsymbol{m}^{(\alpha)}$ and $\boldsymbol{m}^{(\beta)}$ proportional to the unity vectors and by subsequently computing the corresponding average configurations. Hereafter the quantities of interest $r^{\alpha\beta}$ are calculated by means of (22). The connections are not modified during analysis.

The second method involves the macroscopic level. After A^0 and B^0 have been chosen, a vector $\boldsymbol{m}$ is drawn from the distribution with covariance matrix M. The next step is to numerically integrate the flow equation (6) to find the equilibrium solution of $\boldsymbol{q}$ as a result of $\boldsymbol{m}$ and the current values A and B. Once equilibrium has been reached, A and B are modified according to (18). Then another $\boldsymbol{m}$ is drawn, etc.

The values of $r^{\alpha\beta}$ are measured by studying the output correlation in equilibrium $\boldsymbol{q}^{(\alpha)}$ which results from $\boldsymbol{m}^{(\alpha)}$, $\alpha = 1 \ldots p$. The quantities $r^{\alpha\beta}$ are calculated using (22) and (31) with $Q = I$:

$$r^{\alpha\beta} = \frac{1}{N_{\text{o}}} \sum_{i=1}^{N_{\text{o}}} \boldsymbol{q}^{(\alpha)} \cdot \boldsymbol{\xi}_i \boldsymbol{q}^{(\beta)} \cdot \boldsymbol{\xi}_i = \boldsymbol{q}^{(\alpha)} \cdot \boldsymbol{q}^{(\beta)}. \tag{23}$$

The third method consists of numerically integrating the system of equations (20) and (21). This method can be compared with the previous methods by calculating the quantities $r^{\alpha\beta}$ which follow from equations (23) and (13):

$$r^{\alpha\beta} = \boldsymbol{m}^{(\alpha)} C^{\mathrm{T}} C \boldsymbol{m}^{(\beta)} = \boldsymbol{m}^{(\alpha)} Z \boldsymbol{m}^{(\beta)}.$$

So if $\boldsymbol{m}^{(\alpha)}$ and $\boldsymbol{m}^{(\beta)}$ are taken proportional to the different unity vectors, $r^{\alpha\beta}$ are proportional to the elements $Z_{\alpha\beta}$ of the matrix Z.

We performed two different experiments, each according to the three methods described above. For simplicity we chose $p = 2$. In the figures the values of r^{11}, r^{12} are depicted as a function of time. The values of r^{22} have been omitted because they do not provide extra information: due to the choice of M, in the experiments the values of r^{22} are the same or virtually the same as the values of r^{11}. In both series of experiments the input vectors $\boldsymbol{m}$ were drawn from a distribution with a covariance matrix as given in table 1, having eigenvalues $\mu_1 = 0.08$ and $\mu_2 = 0.18$. The learning rate was $\varepsilon = 0.25$ and the decay $\delta = 0.02$. The layers consisted of an equal number of neurons ($n = 1$).

In our interpretation of the results we will use the following terminology. By event α we mean the presence of input configurations having a correlation with the input

Table 1. Initial values and the covariance matrix M of the input data.

Experiment	A^0	B^0	M
1	$\begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 0.13 & 0.05 \\ 0.05 & 0.13 \end{pmatrix}$
2	$\begin{pmatrix} 5.00 & 0.00 \\ 0.00 & 5.00 \end{pmatrix}$	$\begin{pmatrix} 5.00 & 0.00 \\ 0.00 & 5.00 \end{pmatrix}$	$\begin{pmatrix} 0.13 & 0.05 \\ 0.05 & 0.13 \end{pmatrix}$

patterns given by $\boldsymbol{m}^{(\alpha)}$. Furthermore, we will call the resulting average output configuration, $\langle \boldsymbol{s}^{\text{out}} \rangle^{(\alpha)}$, the internal representation of event α.

In the first experiment (figures 1(a) and 1(b)) we took $\tilde{T} = 0.1$; A^0 and B^0 are given in table 1. Initially, the network's ability to distinguish event 1 from event 2 is rather poor, as the corresponding internal representations are largely correlated. (r^{12} is about the same as r^{11}). However, as time proceeds, the network's ability to discriminate between the two events improves and eventually the internal representations are almost orthogonal (r^{12} is much smaller than r^{11}).

In the second experiment (figures 1(c) and 1(d)) the converse happens. We raised the temperature to $\tilde{T} = 1.1$ and started with $A^0 = B^0 = 5I$ (see table 1). So, initially, the internal representations are completely orthogonal ($r^{12} = 0.0$). During the learning process, however, there is a gradual loss of the ability to distinguish event 1 from 2. In the end, the system can no longer discern whether event 1 or 2 happens ($r^{12} = r^{11}$). This clearly demonstrates how information present at the input is lost at the output. The smallest eigenvalue of M, μ_1, is below the critical value $\tilde{T}\delta / n\varepsilon = 0.088$ and therefore a degree of freedom is suppressed by the system (i.e. the direction of the eigenvector that corresponds to μ_1). This should be contrasted with the first series of experiments, in which eventually the input information is represented at the output in a form that is almost unaffected by the network.

It appears from the experiments that method 2 and 3 are in good agreement. The network simulations, however, show systematic deviations, suggesting that there is more noise in the network than expected. This is due to finite size effects as well as to the practical impossibility of measuring the average configurations without error. Since these serve to update the connections, it is crucial to determine them in good approximation, otherwise artefacts may occur due to the cumulation of errors. On the other hand, increasing the number of samples makes the simulations very time consuming. A possible way (not applied in the simulations presented here) to solve these problems is to restrict the analogue depth of the connections, but we will not elaborate on the technical details as this goes beyond the scope of this paper.

5. Discussion

We have considered a two-layer network consisting of binary neurons. First we studied a supervised learning stage in which the connections were modified according to the learning rules (1) and (2). With regard to physiology it is interesting to submit learning rule (2) to closer inspection. Due to the minus sign one often refers to it as an anti-Hebb rule, implying the reverse of the Hebbian mechanism. A different interpretation is

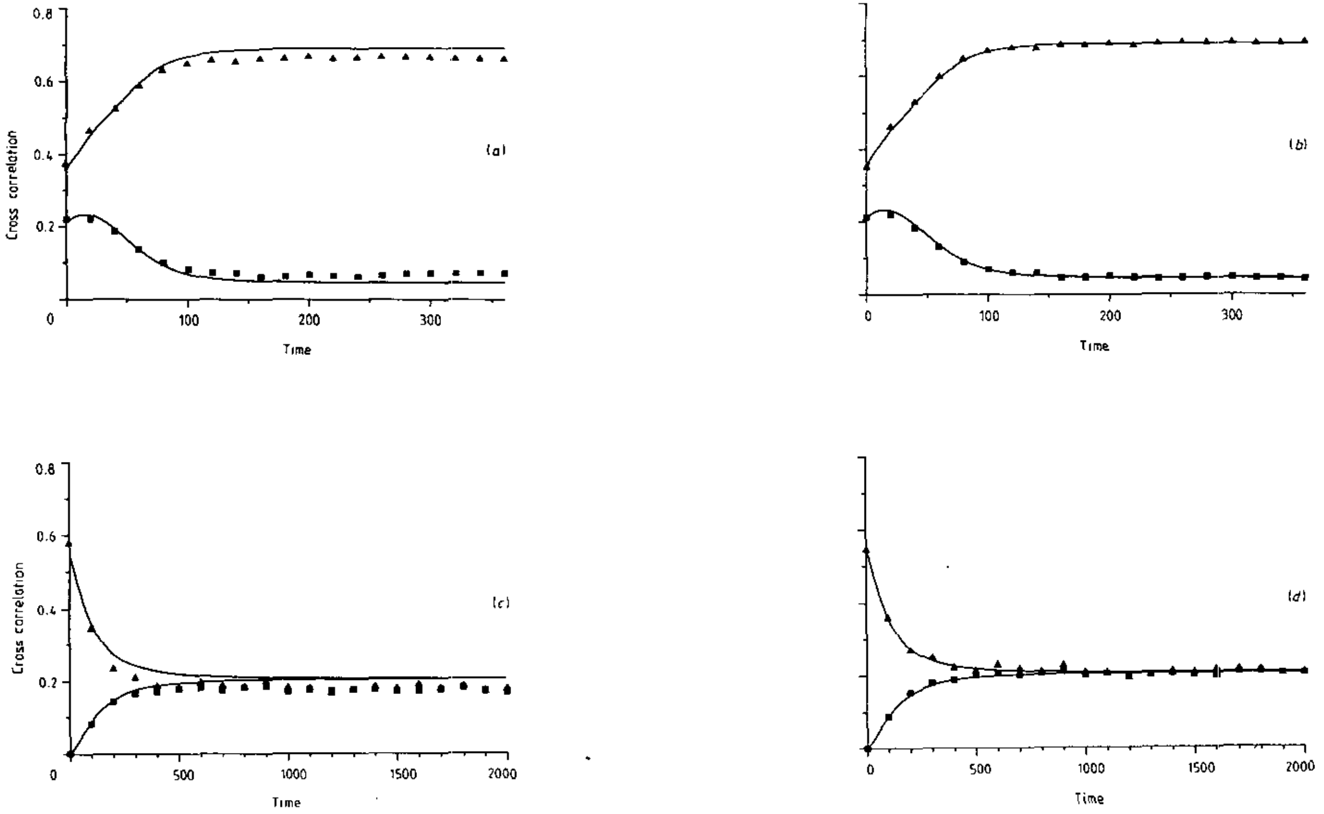

Figure 1. Solid lines indicate the theoretical curves of r^{11} and r^{12}, determined by numerically integrating (20) and (21) (method 3). Data markers indicate the results of network simulations (method 1) or of the flow equations (method 2). Triangles represent the values of r^{11} and squares the values of r^{12}. The values depicted are the average of 10 different computer simulations. Errors in the mean are smaller than or equal to the symbol size. The system parameters were: $\varepsilon = 0.25$, $\delta = 0.02$, $n = 1$ ($N_i = N_o = 60$). Input vectors used for analysis: $\boldsymbol{m}^{(1)} = (0.9, 0.0)$ and $\boldsymbol{m}^{(2)} = (0.0, 0.9)$. In the network simulations the average configurations were determined on the basis of 800 samples. (*a*) Experiment 1, results of network simulations at $\bar{T} = 0.2$. (*b*) Experiment 1, results of flow equations at $\bar{T} = 0.2$. (*c*) Experiment 2, results of network simulations at $\bar{T} = 1.1$. (*d*) Experiment 2, results of flow equations at $\bar{T} = 1.1$.

possible. According to the Hebbian mechanism for synaptic modification [13], a synapse increases its synaptic efficacy if there is synchronous or correlated activity between its pre- and post-synaptic neuron. This implies that if the pre-synaptic neuron is of the inhibitory type and its synaptic efficacy increases, the synapse becomes *more* inhibitory. Therefore in the abstract formulation of the Hebbian mechanism a minus sign automatically appears in the case of a pre-synaptic inhibitory neuron. So if we compose the network such that the input neurons are all of the excitatory type and the output neurons of the inhibitory type, we only need *one* general learning rule based on a Hebbian mechanism, which leads to equations (1) and (2). Although we are inclined to favour one general principle rather than two distinct mechanisms, the hypothesis of modifiable inhibitory synapses does not correspond to the general belief expressed in physiological literature that synaptic plasticity applies solely to excitatory neurons [14, 15]. A reason for this is that experimental data so far have demonstrated Hebbian kind of modification solely for excitatory synapses. It is difficult to tell whether this is due to a lack of experimental data or whether this is a basic property of biological networks. In any case it is likely to be much more difficult to perform an experiment designed to demonstrate the modifiability of inhibitory synapses.

The behaviour of the interpolation model when subjected to a supervised learning stage appeared to be characterized by the linear dependence of the output overlap on the input overlap. This property could be relevant for a biological system since it could serve to represent a continuous multidimensional quantity in a network. In addition it could serve to transfer sensory information to deeper levels in the brain for further processing, without affecting this information. It should be stressed that the network behaves linearly at the macroscopic level whereas the constituent elements are very nonlinear.

Since from a physiological point of view it is probably more plausible that neuronal states as well as synaptic efficacies are variable in time and that the system is not subjected to a supervisor, we considered an unsupervised dynamic learning process. Despite the complicating factor of the strong coupling between the two types of variables, the behaviour of the network could be analytically well described. Apparently the network acts as a filter for the information presented at the input. Eigenvectors of the input covariance matrix which correspond to an eigenvalue below a critical value are suppressed by the system. This critical value can be controlled by setting the system parameters, i.e. the internal noise rate T, the learning rate against decay, and the relative magnitude of the layers. Interestingly, this manner of processing a set of data complies with the well known method of principal components analysis [4, 16–18]. There, one also computes the eigenvalues of the covariance matrix of the input data because the component (eigenvector) that corresponds to a larger eigenvalue is expected to possess a higher information content. The model we described does not perform a principal components analysis, but performs data reduction by suppressing all components that have an information content below a critical value. Physiologically such a system could be very useful for pre-processing the large amount of available sensory data, so as to extract the relevant features and neglect the unimportant details [18].

Acknowledgments

The authors wish to thank Professors J J Denier van der Gon and Th W Ruijgrok for their constructive and stimulating comments.

Appendix A. Derivation of the flow equations

We will try to derive equations for the flow of the order parameters q_μ according to the method proposed in [19]. Let $p(s^{\text{out}}, t)$ denote the probability of finding the output of the system at time t in the state s^{out}. For notational convenience we will omit the superscript 'out' from s^{out}. The master equation for $p(s, t)$ is given by

$$\frac{\partial}{\partial t} p(s, t) = -\sum_{i=1}^{N_o} w_i(s) p(s, t) + \sum_{i=1}^{N_o} w_i(F_i s) p(F_i s, t)$$

where w_i is defined by (3) and F_i is the spin-flip operator

$$F_i(s_1, \ldots, s_i, \ldots s_{N_o}) = (s_1, \ldots, -s_i, \ldots s_{N_o}).$$

The probability that at time t the output state $\boldsymbol{s}$ will have overlap $\boldsymbol{q}(\boldsymbol{s}) = \boldsymbol{q}$ is:

$$P(\boldsymbol{q}, t) = \sum_{\boldsymbol{s}} \delta(\boldsymbol{q} - \boldsymbol{q}(\boldsymbol{s})) p(s, t). \tag{24}$$

Our purpose is to find a differential equation for the distribution $P(\boldsymbol{q}, t)$ using the master equation. In the limit $N_o \to \infty$, p fixed, one finds:

$$\int \mathrm{d}\boldsymbol{q}\, \Psi(\boldsymbol{q}) \frac{\partial}{\partial t} P(\boldsymbol{q}, t) = \int \mathrm{d}\boldsymbol{q}\, P(\boldsymbol{q}, t) \boldsymbol{\nabla}\Psi(\boldsymbol{q}) \left[-\boldsymbol{q} + \lim_{N_o \to \infty} \frac{1}{N_o} \sum_{i=1}^{N_o} \boldsymbol{\xi}_i g(\beta h_i^{\text{out}}) \right] \tag{25}$$

where $\Psi(\boldsymbol{q})$ is an arbitrary test function. Assuming there is no uncertainty in $\boldsymbol{q}(0)$, the initial value of the output correlation, the solution of (25) is

$$P(\boldsymbol{q}, t) = \delta(\boldsymbol{q} - \boldsymbol{q}^*(t)). \tag{26}$$

The equation for the flow itself becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{q}^*(t) = -\boldsymbol{q}^*(t) + \lim_{N_o \to \infty} \frac{1}{N_o} \sum_{i=1}^{N_o} \boldsymbol{\xi}_i g(\beta h_i^{\text{out}}(\boldsymbol{q}^*(t))).$$

The consequence of (26) is that any quantity $r(\boldsymbol{s})$ which can be written as $r(\boldsymbol{q}(\boldsymbol{s}))$ is deterministic. Its value at time t is simply $r(\boldsymbol{q}^*(t))$ without any fluctuation. But a still more detailed result can be obtained.

Define the average of $r(\boldsymbol{s})$ over all possible states $\boldsymbol{s}$ by

$$\langle r(\boldsymbol{s}) \rangle = \sum_{\boldsymbol{s}} r(\boldsymbol{s}) p(s, t). \tag{27}$$

In addition, we define an alternative average, namely the value of $r(\boldsymbol{s})$ averaged over every $\boldsymbol{s}$ that satisfies $\boldsymbol{q}(\boldsymbol{s}) = \boldsymbol{q}$:

$$\langle r(\boldsymbol{s}) \rangle_{\boldsymbol{q}} = \frac{\Sigma_{\boldsymbol{s}}\, r(\boldsymbol{s}) p(s, t) \delta(\boldsymbol{q} - \boldsymbol{q}(\boldsymbol{s}))}{\Sigma_{\boldsymbol{s}}\, p(s, t) \delta(\boldsymbol{q} - \boldsymbol{q}(\boldsymbol{s}))}.$$

Multiplying the right-hand side of (27) by the unit operator $\int \mathrm{d}\boldsymbol{q}\, \delta(\boldsymbol{q} - \boldsymbol{q}(\boldsymbol{s}))$ and using subsequently (24) and (26), we get

$$\langle r(\boldsymbol{s}) \rangle = \int \mathrm{d}\boldsymbol{q}\, P(\boldsymbol{q}, t) \langle r(\boldsymbol{s}) \rangle_{\boldsymbol{q}} = \langle r(\boldsymbol{s}) \rangle_{\boldsymbol{q}^*}$$

Similarly, one can derive

$$\langle r(s, \boldsymbol{q}(\boldsymbol{s})) \rangle = \langle r(s, \boldsymbol{q}(\boldsymbol{s})) \rangle_{\boldsymbol{q}^*} = \langle r(s, \boldsymbol{q}^*(t)) \rangle_{\boldsymbol{q}^*} = \langle r(s, \boldsymbol{q}^*(t)) \rangle \tag{28}$$

which will turn out to be a useful identity.

With use of the master equation expressions can be also found for the expectation value of an arbitrary quantity $r(s)$. Differentiation of (27) with respect to t produces

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle r(\boldsymbol{s})\rangle &= \sum_{i=1}^{N_o}\langle w_i(\boldsymbol{s})[r(F_i\boldsymbol{s})-r(\boldsymbol{s})]\rangle \\ &= \tfrac{1}{2}\sum_{i=1}^{N_o}\langle r(F_i\boldsymbol{s})-r(\boldsymbol{s})\rangle - \tfrac{1}{2}\sum_{i=1}^{N_o} g(\beta h_i^{\mathrm{out}})\langle s_i r(F_i\boldsymbol{s})-s_i r(\boldsymbol{s})\rangle. \qquad (29)\end{aligned}$$

Note that we have applied (28) and that we have made use of the fact that g is an odd function. We will use (29) to compute $\langle s_i\rangle$ in equilibrium. It follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle s_i\rangle = -\langle s_i\rangle + g(\beta h_i^{\mathrm{out}}).$$

Hence, the equilibrium value of $\langle s_i\rangle$ is given by

$$\langle s_i\rangle = g(\beta h_i^{\mathrm{out}}).$$

Similarly, $\langle s_i s_j\rangle (i\neq j)$ can be found:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle s_i s_j\rangle = -2\langle s_i s_j\rangle + \langle s_i g(\beta h_j^{\mathrm{out}})\rangle + \langle s_j g(\beta h_i^{\mathrm{out}})\rangle.$$

In equilibrium we find for $i\neq j$

$$\langle s_i s_j\rangle = \tfrac{1}{2}\langle s_i\rangle g(\beta h_j^{\mathrm{out}}) + \tfrac{1}{2}\langle s_j\rangle g(\beta h_i^{\mathrm{out}}) = \langle s_i\rangle\cdot\langle s_j\rangle. \qquad (30)$$

Note that this result has been derived without assigning a function to g. Taking g according to (12) and using (6) and (13), we find for all $i\neq j$:

$$\langle s_i\rangle = \boldsymbol{\xi}_i\cdot Q^{-1}\boldsymbol{q} \qquad (31)$$

$$\langle s_i s_j\rangle = \boldsymbol{\xi}_i\cdot Q^{-1}\boldsymbol{q}\,\boldsymbol{\xi}_j\cdot Q^{-1}\boldsymbol{q} \qquad (32)$$

provided that $\boldsymbol{q}$ is such that $\Sigma_\mu |(Q^{-1}\boldsymbol{q})_\mu| \leqslant 1$.

Appendix B. Matrix equations

In the derivation of (11) and (13) it was important that B was positive definite. To check whether B remains positive definite during the learning process,we introduce the dummy variable $B^* = B\exp(\delta t)$ and find by (19)

$$\frac{\mathrm{d}}{\mathrm{d}t}B^* = \varepsilon\,\mathrm{e}^{\delta t}CMC^{\mathrm{T}}.$$

The matrix on the right-hand side is at least semi-positive definite since for all vectors $\boldsymbol{x}$: $\boldsymbol{x}\cdot CMC^{\mathrm{T}}\boldsymbol{x} = C^{\mathrm{T}}\boldsymbol{x}\cdot MC^{\mathrm{T}}\boldsymbol{x}\geqslant 0$, because by definition M is semi-positive definite. Consequently, B will remain positive definite if B^0 is positive definite. This implies that V, which is defined by $V=(B+\tilde{T}I)^{-1}$ and which commutes with B, can only have positive eigenvalues. Therefore V is invertible. Similarly one can prove that both B and V remain symmetric.

Using $\mathrm{d}V/\mathrm{d}t = -V(\mathrm{d}B/\mathrm{d}t)V$, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}V = \delta V - \tilde{T}\delta V^2 - \varepsilon VCMC^{\mathrm{T}}V \qquad (33)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}C = -\varepsilon VCM\left[C^{\mathrm{T}}C - nI + \tilde{T}\frac{\delta}{\varepsilon}M^{-1}\right]. \qquad (34)$$

A useful identity follows from (19), namely:

$$n\varepsilon \hat{V}\hat{C}M = \delta\hat{C} \tag{35}$$

where the notation ˆ indicates that we are dealing with a critical point. Applying (35), we find that the critical points of (33) and (34) satisfy

$$\hat{V} - \frac{1}{\tilde{T}}\left(I - \frac{1}{n}\hat{C}\hat{C}^{\top}\right) = 0 \tag{36}$$

$$\hat{C}\left(\hat{C}^{\top}\hat{C} - nI + \frac{\hat{T}\delta}{\varepsilon}M^{-1}\right) = 0. \tag{37}$$

Recalling that $\hat{V}$ is symmetric, we find by virtue of (35) that the matrix of interest $\hat{Z} = \hat{C}^{\top}\hat{C}$ commutes with M:

$$\hat{C}^{\top}\hat{C}M - M\hat{C}^{\top}\hat{C} = \frac{\delta}{\varepsilon n}\hat{C}^{\top}\hat{V}^{-1}\hat{C} - \frac{\delta}{\varepsilon n}(\hat{C}^{\top}\hat{V}^{-1}\hat{C})^{\top} = 0.$$

Therefore $\hat{Z}$ and M can be reduced to the diagonal form by the same base transformation. From (37) it follows that an eigenvalue of $\hat{Z}$ can assume two values:

$$z_i = 0 \qquad \text{or} \qquad z_i = n - \frac{\hat{T}\delta}{\varepsilon\mu_i}$$

where μ_i indicates eigenvalue number i of M. Note that $z_i = 0$ always holds, whereas the second critical value is only valid if $\mu_i > \tilde{T}\delta/n\varepsilon$, since by definition $\hat{Z}$ is semi-positive definite ($\boldsymbol{x} \cdot C^{\top}C\boldsymbol{x} = |C\boldsymbol{x}|^2 \geqslant 0$).

With regard to the stability of the critical points, we consider small perturbations $C = \hat{C} + \rho C'$ and $V = \hat{V} + \rho V'$, $\rho \to 0$. Tedious but straightforward calculation yields the linearized equations ($\rho \to 0$):

$$\frac{\mathrm{d}}{\mathrm{d}t}V' = -\delta V' - \frac{\delta}{n}[\hat{V}C'\hat{C}^{\top} + \hat{C}C'^{\top}\hat{V}] + \mathrm{O}(\rho) \tag{38}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}C' = -\varepsilon\hat{V}C'M\left(\hat{Z} - nI + \frac{\tilde{T}\delta}{\varepsilon}M^{-1}\right) - \frac{\delta}{n}\hat{C}[C'^{\top}\hat{C} + \hat{C}^{\top}C'] + \mathrm{O}(\rho). \tag{39}$$

The latter equation exhibits no V' dependence. Furthermore it follows from (38) that if $C' \to 0$, then $V' \to 0$ because $\delta > 0$. Hence the stability of the system depends solely on (39).

The stability of the extreme case $\hat{C} = 0$, $\hat{V} = \tilde{T}^{-1}I$ can be checked easily. One finds

$$\frac{\mathrm{d}}{\mathrm{d}t}C' = C'\left(\frac{n\varepsilon}{\tilde{T}}M - \delta I\right) + \mathrm{O}(\rho) \qquad \rho \to 0$$

implying stability if $n\varepsilon M/\tilde{T} - \delta I$ is negative definite, that is, if all $\mu_i < \tilde{T}\delta/n\varepsilon$. With regard to the other extreme case, $\hat{Z} = nI - \tilde{T}\delta\varepsilon^{-1}M^{-1}$ and all $\mu_i > \tilde{T}\delta/n\varepsilon$, we consider perturbations on $\hat{Z}$, $Z = \hat{Z} + \rho Z' + \mathrm{O}(\rho^2)$, where $Z' = C'^{\top}\hat{C} + \hat{C}^{\top}C'$:

$$\frac{\mathrm{d}}{\mathrm{d}t}Z' = -\frac{\delta}{n}[\hat{Z}Z' + Z'\hat{Z}] + \mathrm{O}(\rho) \qquad \rho \to 0$$

the solution of which is given by

$$Z'(t) = \mathrm{e}^{-(\delta/n)\hat{Z}t}Z'(0)\,\mathrm{e}^{-(\delta/n)\hat{Z}t}.$$

Clearly $\hat{Z}$ is stable, since in this case all eigenvalues of $\hat{Z}$ are positive, as stated.

But what happens in the intermediate cases? Suppose $z_k = 0$ but $\mu_k > \tilde{T}\delta/n\varepsilon$. We will prove that this solution is not stable. Therefore, consider the direction $\boldsymbol{e}_k$, i.e. the eigenvector of M that corresponds to the eigenvalue μ_k. By (39) and (36) we get

$$\frac{\mathrm{d}}{\mathrm{d}t} C'\boldsymbol{e}_k = \frac{n\varepsilon}{\tilde{T}}\left[\left(\mu_k - \frac{\tilde{T}\delta}{n\varepsilon}\right) I - \frac{\mu_k}{n}\hat{C}\hat{C}^\top\right] C'\boldsymbol{e}_k. \tag{40}$$

Since C' can be chosen arbitrarily, we take C' such that $C'\boldsymbol{e}_k$ is in the kernel of $\hat{C}\hat{C}^\top$, i.e. $\hat{C}\hat{C}^\top C'\boldsymbol{e}_k = \boldsymbol{o}$. It is always possible to find such a direction since the spectrum of eigenvalues of $\hat{C}\hat{C}^\top$ is equal to that of $\hat{C}^\top\hat{C}$. Equation (40) reveals that the critical value $z_k = 0$ cannot be stable if $\mu_k > \tilde{T}\delta/n\varepsilon$.

Summarizing, the only possible equilibrium solution is the matrix $\hat{Z}$ which possesses eigenvalues given by:

$$\begin{aligned} z_i &= 0 && \text{for } 1 \leqslant i < k \\ z_i &= n - \frac{\tilde{T}\delta}{\varepsilon\mu_i} && \text{for } k \leqslant i \leqslant p \end{aligned} \tag{41}$$

where k is the index of the smallest eigenvalue of M that satisfies $\mu_k > \tilde{T}\delta/n\varepsilon$ ($\mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_p$).

Finally, we analyse the covariance matrix of the output:

$$\overline{q_\mu q_\nu} = \sum_{\rho\sigma} \hat{C}_{\mu\rho}\overline{m_\rho m_\sigma}\hat{C}_{\nu\sigma} = (\hat{C}M\hat{C}^\top)_{\mu\nu}.$$

Consider the vector $\boldsymbol{e}'_k \equiv \hat{C}\boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is an eigenvector of $\hat{Z}$ (and consequently of M) corresponding to an eigenvalue $z_k \neq 0$. One finds

$$\hat{C}M\hat{C}^\top\boldsymbol{e}'_k = \mu_k z_k \boldsymbol{e}'_k.$$

Since the rank of $\hat{C}M\hat{C}^\top$ is equal to the rank of $\hat{Z}$, the remaining eigenvalues must be zero. So the eigenvalues of the output covariance matrix are $n(\mu_k - \tilde{T}\delta/n\varepsilon)$ if $\mu_k > \tilde{T}\delta/n\varepsilon$; otherwise they are zero.

References

[1] Kohonen T 1982 *Associative Memory—a System Theoretic Approach* (New York: Springer)

[2] Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158–9

[3] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cognitive Sci.* **9** 147–69

[4] Rubner J and Schulten K 1990 *Biol. Cybern.* **62** 193–9

[5] Jonker H J J, Coolen A C C and Denier van der Gon J J 1989 *Proc. IEE on Artificial Neural Networks* (London: IEE) 23–26

[6] Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617–80

[7] Shinomoto S 1987 *J. Phys. A: Math. Gen.* **20** L1305–9

[8] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554–8

[9] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007

[10] Hemmen van J L and Kühn R 1986 *Phys. Rev. Lett.* **57** 913–6

[11] Brown Th H, Kairiss E W and Keenan C L 1990 *Ann. Rev. Neurosci.* **13** 475–511

[12] Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457–62

[13] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley)

[14] Bear M F, Cooper L N and Ebner F F 1987 *Science* **237** 42–8

[15] diPrisco G V 1984 *Prog. Neurobiol.* **22** 89–102

[16] Lawley D N and Maxwell A E 1971 *Factor Analysis as a Statistical Method* (London: Butterworth)

[17] Oja E 1982 *J. Math. Biology* **15** 267–73

[18] Kühnel H and Tavan P 1990 *Parallel Processing in Neural Systems and Computers* ed R Eckmiller, G Hartmann and G Hauske (Amsterdam: Elsevier) 187–90

[19] Coolen A C C and Ruijgrok Th W 1988 *Phys. Rev.* A **38** 1007–18