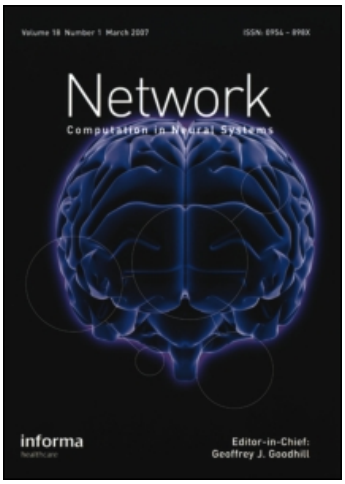
This article was downloaded by: *[OHSU Science & Engineering Library]*
On: *16 April 2009*
Access details: *Access Details: [subscription number 906391935]*
Publisher *Informa Healthcare*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Network: Computation in Neural Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713663148

Dynamics of learning in linear feature-discovery networks

Tood K. Leen [a]
[a] Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Beaverton, OR, USA

Online Publication Date: 01 February 1991

To cite this Article Leen, Tood K.(1991)'Dynamics of learning in linear feature-discovery networks',Network: Computation in Neural Systems,2:1,85 — 105
To link to this Article: DOI: 10.1088/0954-898X/2/1/005
URL: http://dx.doi.org/10.1088/0954-898X/2/1/005

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Network 2 (1991) 85–105. Printed in the UK

Dynamics of learning in linear feature-discovery networks

Todd K Leen

Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, 19600 NW von Neumann Drive, Beaverton, OR 97006-1999, USA

Received 5 September 1990

Abstract. In this paper I address the dynamics of learning in unsupervised neural feature-discovery networks. The models introduced incorporate feedforward connections modified by a Hebb law, and recurrent lateral connections modified by an anti-Hebb law. Conditions for stability of equilibria are derived, and bifurcation theory is used to explore the behaviour near loss of stability. Stability of the equilibria is shown to depend on the learning rates in the system, and on the statistics of the input signal. The bifurcation analyses reveal previously overlooked behaviours, including equilibria that consist of mixtures of the principal eigenvectors of the input autocorrelation, as well as limit cycles. The results provide a more complete picture of adaptation in Hebbian feature-discovery networks.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

1. Introduction

One of the problems faced by both natural and artificial adaptive systems is the construction of efficient representations of the environment. In recent years there has been considerable interest in the notion that local Hebbian adaptation can provide a mechanism for building such representations.

Oja (1982) made a remarkable observation that provides a link between Hebbian learning rules and ideas from signal processing. He showed that a simple model neuron, with a particular Hebbian adaptation rule, develops into a filter for the first principal component of the input distribution. Linsker (1988) extended this idea by suggesting that perceptual systems organize themselves to maximize information transfer. (A thorough theoretical account of Linsker's early simulations is given by MacKay and Miller (1990).)

Several researchers have extended Oja's work, suggesting Hebbian networks that perform a complete principal component analysis (PCA). Oja and Karhunen (1985) discuss an algorithm that maps to a three-layer, feedforward network of linear neurons (Oja 1989). Sanger (1989) proposed an algorithm that uses a set of cascaded feedback projections to the input space, and gives a convergence proof based on Oja's original work. This architecture singles out a particular cell for each principal component. Both of these models have good convergence properties, and are useful for signal encoding applications (Leen *et al* 1990).

In a fundamentally different approach, lateral connections between the nodes are introduced. The adaptation of the lateral connection strengths and the effect of the laterally propagated signals on the nodes are designed to force each node to tune

0954-898X/91/010085+21$03.50 © 1991 IOP Publishing Ltd

to a different statistical feature. Rubner and Schulten (1990) discuss a model with cascaded lateral signal paths; the ith cell receives lateral signals from all cells j with $j < i$. Like Sanger's scheme, this architecture singles out a particular cell for each principal component. In contrast, the models discussed here have no architecturally distinguished cells. This enhances their biological plausibility. Foldiak (1989) employs recurrent lateral connections that develop according to an anti-Hebbian rule. The cells in this algorithm do not, in general, filter for the principal components, but rather for mixtures of the principal components. I show that this arises from a bifurcation that results from the form of the anti-Hebb rule.

From a dynamical viewpoint, algorithms that use lateral connections are poorly understood. The goal of this paper is to help form a more complete picture of feature-discovery models that use lateral signal flow. I discuss two new models, with particular emphasis on their learning dynamics. The models incorporate Hebbian and anti-Hebbian learning, and recurrent lateral connections. The anti-Hebbian rules depart slightly from forms previously given in the literature.

I give stability analyses and derive bifurcation diagrams for the models. Stability analysis shows that the adaptation of the lateral connections needs to be fast relative to the adaptation of the weights from the input, in order for the network to perform PCA. Bifurcation theory provides a description of the behaviour near loss of stability. The bifurcation analyses reveal behaviours that previous researchers have overlooked. These include equilibria in which the weight vectors are combinations of the eigenvectors of the input's autocorrelation, as well as limit cycles. I have used a computer algebra system† to effectively handle stability and bifurcation calculations for the high-dimensional network equations.

In the next section, I make some general observations on stability that hold for a broad class of models. In section 3, I develop a model that treats the lateral connections in an unorthodox way. The signals carried on the lateral connections affect cell plasticity, but not the cell activation. The model in section 4 uses lateral signals in the conventional way with signals affecting the activity of the target node.

2. Extending the single-neuron model

In Oja's (1982) model the input vector, $\boldsymbol{x} \in R^N$, is a random variable drawn from a stationary probability distribution. The vector of synaptic weights is denoted $\boldsymbol{\omega}$, and the postsynaptic response is linear,

$$y = \boldsymbol{\omega} \cdot \boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\omega} \tag{1}$$

where the superscript T denotes transpose. The discrete rule for the adaptation of the synaptic strengths is

$$\delta \boldsymbol{\omega} = \gamma [\boldsymbol{x} y - y^2 \boldsymbol{\omega}] \tag{2}$$

where $\delta \boldsymbol{\omega}$ is the change in the weight vector in reponse to the input $\boldsymbol{x}$, and γ is the learning rate. The first term in (2) is the usual Hebbian correlation term. The second term in (2) is an active decay which prevents the magnitude of the weight vector from diverging.

† MATHEMATICA version 1.2, Copyright 1988 and 1989, Wolfram Research Inc.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

Averaging over the ensemble of inputs, and passing to the continuous time form (Ljung 1977, and references therein) leaves

$$\dot{\omega} = \langle x\, y \rangle - \langle y^2 \rangle\, \omega \tag{3}$$

where $\langle \ldots \rangle$ indicates the ensemble average. Substituting (1) into (3) leaves

$$\dot{\omega} = Q\omega - (\omega \cdot Q\omega)\,\omega \equiv f(\omega) \tag{4}$$

where $Q = \langle x x^{\mathrm{T}} \rangle$ is the autocorrelation of the input patterns. I denote the (unit-magnitude) eigenvectors of Q by $e_i, i = 1, \ldots, N$. The corresponding eigenvalues are assumed to be ordered as $\lambda_1 > \lambda_2 > \cdots > \lambda_N > 0$.

Equation (4) has equilibria when the weight vector is along, or opposite, any of the eigenvectors of the autocorrelation. Only the equilibrium at $\omega = \pm e_1$ is stable; the other equilibria are saddle points. In fact, Oja shows that the weight vector asymptotically approaches $\pm e_1$. The variance of the cell's response is thus maximized, and the cell acts as a filter for the first principal component of the input distribution.

2.1. Extending to multiple cells

The goal is to extend this scheme to a system of $M \leq N$ cells whose weight vectors converge to the leading M eigenvectors of the autocorrelation. Consider a set of M linear neurons with weight vectors $\omega_1, \ldots, \omega_M$ connecting each to the N-dimensional input space. The model is built by replicating (4) for each cell and adding suitable interaction terms designed to force the weight vectors to converge to different eigenvectors of the input's autocorrelation.

If the interactions between cells are parametrized by a coupling constant (e.g. a learning rate) then there will be a critical value for this parameter below which the desired equilibrium will lose stability. To show this I will assume that the interactions between the cells are weak and treat them as a perturbation. The equations of motion then take the form

$$\dot{\omega}_i = f_i(\omega_i) + C\, G_i(\omega_1, \ldots, \omega_M, \eta) \qquad i = 1, \ldots, M \tag{5}$$

where f_i is a vector field of the same form as in (4), C is a coupling constant which specifies the strength of the interactions, and G_i carries the interactions between the cells. The η are auxiliary state variables appearing in the interactions. Additional differential equations describe the evolution of the auxiliary variables,

$$\dot{\eta}_{ij} = f_{\eta_{ij}}(\omega, \eta, C). \tag{6}$$

In the following sections the auxiliary variables will take the form of the strengths of lateral couplings between the cells. The constant C will appear as a learning rate associated with these lateral couplings.

I further assume that at $C = 0$ the system (5), (6) has an equilibrium at $\omega_i = e_i$, $\eta_{ij} = \eta^0_{ij}$. This equilibrium cannot be stable since the cells with $\omega_i = e_i$ for $i > 1$ will inherit the instabilities of (4). Assuming that η^0 is continuous in C, one can show that the instabilities will persist for small values of C. Consequently, for any such model there is a minimum value for $|C|$ below which the equilibrium is unstable. The two models we consider both exhibit this behaviour (see also Rubner and Schulten (1990)).

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

3. Minimal coupling

In this section interactions between the cells are introduced in a minimal fashion. The signals on the lateral connections will mediate synaptic plasticity, but do not directly affect cell responses to the input signal. This model is obtained by writing a potential for the single-neuron model of section 2, and augmenting this single-neuron potential with interaction terms designed to orthogonalize the weight vectors. The interactions are made local by introducing lateral couplings between the cells. This derivation leads naturally to an anti-Hebbian adaptation rule for the lateral connections.

3.1. Potential formulation

I extend the single-cell system to an array of M cells, with weight vectors ω_i, $i = 1, \ldots, M$, and linear responses $y_i = \omega_i \cdot x$. The derivation is built on a potential formulation. The model in (4) performs hill-climbing on the variance of the cell response. Thus the Hebb term, $Q\omega$, in (4) is recovered by taking minus the gradient of

$$U(\omega) = -\tfrac{1}{2}\langle y^2 \rangle = -\tfrac{1}{2}\,\omega \cdot Q\omega \tag{7}$$

with respect to ω.

Following Yuille *et al* (1989) (see also Fuchs and Haken 1988) I introduce an interaction potential that penalizes correlations between the cell responses

$$I = \sum_{i \neq j} \tfrac{1}{2}\langle y_i y_j \rangle^2 = \sum_{i \neq j} \tfrac{1}{2}(\omega_i \cdot Q\omega_j)^2. \tag{8}$$

This form elevates the potential in regions of the weight space where the cell responses are correlated or anticorrelated. The total potential is given by combining (8) with M copies of (7),

$$\begin{aligned} U_{\text{tot}} &= \sum_{i=1}^{M} U(\omega_i) \;+\; C\,I \\ &= -\tfrac{1}{2}\sum_{i=1}^{M}(\omega_i \cdot Q\omega_i) \;+\; \tfrac{1}{2}C\sum_{i \neq j}(\omega_i \cdot Q\omega_j)^2 \end{aligned} \tag{9}$$

where C is a coupling constant (cf subsection 2.1).

A vector field given by the gradient of U_{tot} will lead to unbounded weight vectors. In order to avoid this divergence, I introduce a decay term of the form given in (4). The equations of motion for the ω are

$$\begin{aligned} \dot{\omega}_i &= -\nabla_{\omega_i} U - (\omega_i \cdot Q\omega_i)\,\omega_i \\ &= Q\omega_i - C\sum_{j \neq i}(\omega_i \cdot Q\omega_j)\,Q\omega_j - (\omega_i \cdot Q\omega_i)\,\omega_i. \end{aligned} \tag{10}$$

To help interpret (10), I rewrite it in terms of ensemble averages over the input patterns and the response of the ith cell, $y_i = \omega_i \cdot x$. Thus

$$\begin{aligned} \dot{\omega}_i &= \langle x\,y_i \rangle \;-\; C\sum_{j \neq i}\langle y_i y_j \rangle \langle x\,y_j \rangle \;-\; \langle y_i^2 \rangle \omega_i \\ &= \left\langle x\left[y_i - C\sum_{j \neq i}\langle y_i y_j \rangle\, y_j\right]\right\rangle \;-\; \langle y_i^2 \rangle \omega_i. \end{aligned} \tag{11}$$

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The first term on the right-hand side of (11) drives changes in the weight vector according to the correlation between the the input signal x and the modified response,

$$\tilde{y}_i \equiv y_i - C \sum_{j \neq i} \langle y_i y_j \rangle y_j. \tag{12}$$

The jth term of the sum in (12) is the response y_j gated by a factor proportional to the correlation between cells j and i. This correlation is *not* computed locally within the network. Furthermore the response of the jth cell is not locally available to cell i.

3.2. Local realization

The most natural way to localize these computations is to provide a set of lateral connections between the cells. For this model the cell activities are computed only from the signals carried on the weights ω_i as in (1). The signals carried on the lateral connections will mediate synaptic plasticity without directly influencing cell response.

I introduce a symmetric matrix of $M(M-1)/2$ distinct lateral connection strengths,

$$\begin{aligned} &\eta_{ij} \quad i,j = 1,\ldots,M, \quad \neq j \\ &\eta_{ii} = 0 \end{aligned}$$

and require that these equilibriate to $-C\langle y_i y_j \rangle$. The simplest dynamics to carry out this relaxation is

$$\begin{aligned} \dot{\eta}_{ij} &= -d\,(\eta_{ij} + C\,\langle y_i y_j \rangle) \\ &= -d\,(\eta_{ij} + C\,\omega_i \cdot Q\,\omega_j) \end{aligned} \tag{13}$$

where d is a rate constant. This form captures the notion that the lateral connections develop to oppose correlations, $\langle y_i y_j \rangle$, between the cell responses.

Finally, I make the substitution $C\,\omega_i \cdot Q\,\omega_j \rightarrow -\eta_{ij}$ in (10) to obtain the equations of motion for the forward weights,

$$\dot{\omega}_i = Q\,\omega_i + \sum_{j \neq i} \eta_{ij}\,Q\,\omega_j - (\omega_i \cdot Q\,\omega_i)\,\omega_i. \tag{14}$$

Equations (13) and (14) are the adaptation dynamics for the system of M cells with weights ω_i from the input space and lateral connections η_{ij}.

3.3. Stability and bifurcation behaviour

The stability of the desired equilibrium is dependent on the free parameters, in accord with the discussion in subsection 2.1. The primary result is that the adaptation of the lateral connections needs to be *fast* relative to the adaptation of the forward weights in order for the system to perform a principal components analysis. Beyond the stability analysis, I derive bifurcation diagrams for the system.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

3.3.1. Stability. By inspection (13) and (14) have an equilibrium at

$$X_0 \equiv \{\omega_i = e_i,\ i = 1,\ldots,M;\ \eta_{ij} = 0\} \qquad \forall\, C. \tag{15}$$

To treat the stability of this equilibrium it is convenient to expand the ω_i in the basis of eigenvectors of Q, writing the components as

$$\omega_i^j \equiv \omega_i \cdot e_j.$$

Next, collect all variables into a single coordinate vector and arrange the components as

$$\begin{aligned} X = [&(\omega_1^2, \omega_2^1, \eta_{12}), (\omega_1^3, \omega_3^1, \eta_{13}), \ldots (\omega_{M-1}^M, \omega_M^{M-1}, \eta_{M-1,M}), \\ &\{\omega_1^1, \omega_2^2, \ldots, \omega_M^M\}, (\omega_1^{M+1}, \ldots, \omega_1^N, \ldots, \omega_M^N)]. \end{aligned} \tag{16}$$

This vector contains $M(M-1)/2$ triplets of the form $(\omega_i^j, \omega_j^i, \eta_{ij})$, M components ω_i^i, and $M(N-M)$ components in the last block. The components in the last block are the projections of the weight vectors onto the eigenvectors of Q *orthogonal* to the M-dimensional principal component subspace. In this notation, the equilibrium (15) at which the network performs PCA is at

$$X_0 = [(0,0,0), \ldots, (0,0,0), \{1,1,\ldots\}, (0,0,\ldots)]. \tag{17}$$

I write the equations of motion (13), (14) in shorthand as

$$\dot{X} = F(X, C) \tag{18}$$

with $F(X,C)$ defined by the components of the right-hand sides of (13) and (14) arranged in the same order as the coordinate vector (16). The linear part of $F(X,C)$ determines asymptotic stability of X_0. At X_0 the linear part takes the block-diagonal form (see appendix A)

$$\mathrm{D}F_0 \equiv \left(\frac{\partial F}{\partial X}\right)\Big|_{X_0} = \begin{Bmatrix} [\mathcal{M}_{12}] & & & & \\ & [\mathcal{M}_{13}] & & & \\ & & \ddots & & \\ & & & \{A\} & \\ & & & & (B) \end{Bmatrix} \tag{19}$$

where the 3×3 sub-blocks, $[\mathcal{M}_{ij}]$, $i < j$, are of the form

$$\mathcal{M}_{ij} = \begin{bmatrix} \lambda_j - \lambda_i & 0 & \lambda_j \\ 0 & \lambda_i - \lambda_j & \lambda_i \\ -C\,d\,\lambda_j & -C\,d\,\lambda_i & -d \end{bmatrix} \tag{20}$$

and $\{A\}$ and (B) are diagonal matrices of the form

$$A = \begin{Bmatrix} -2\lambda_1 & & & \\ & -2\lambda_2 & & \\ & & \ddots & \\ & & & -2\lambda_M \end{Bmatrix} \tag{21}$$

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

and

$$B = \begin{pmatrix} \lambda_{M+1} - \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_N - \lambda_1 & & & & \\ & & & \lambda_{M+1} - \lambda_2 & & & \\ & & & & \ddots & & \\ & & & & & \lambda_N - \lambda_2 & \\ & & & & & & \ddots \\ & & & & & & & \lambda_{M+1} - \lambda_M & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & \lambda_N - \lambda_M \end{pmatrix}. \tag{22}$$

Concentrate for the moment on the lower two blocks A and B. Since the eigenvalues of Q are ordered according to $\lambda_1 > \lambda_2 > \cdots$, the elements in these blocks are all negative. Thus the invariant subspaces corresponding to these blocks are stable eigenspaces. The block B is of particular interest. This block corresponds to perturbations out of the principal component subspace. These perturbations are always damped, indicating that the space spanned by the principal eigenvectors is asymptotically stable.

The only possible instabilities arise in the 3×3 sub-blocks $\mathcal{M}_{ij}$. These blocks define invariant subspaces of DF_0, and each can be considered separately. Thus the stability problem for the entire recurrent network reduces to the consideration of a set of 3×3 matrices.

Applying the Routh–Hurwitz (Willems 1970) conditions to the characteristic equation for $\mathcal{M}_{ij}$ shows that the equilibrium X_0 is asymptotically stable provided

$$d > d_{ij_0} \equiv \frac{(\lambda_i - \lambda_j)^2 (\lambda_i + \lambda_j)}{\lambda_i^2 + \lambda_j^2} \tag{23}$$

$$C > C_{ij_0} \equiv \frac{1}{\lambda_i + \lambda_j}. \tag{24}$$

These conditions must be satisfied for all choices of the indices (i, j). The critical values C_0 and d_0 depend on the eigenvalue spectrum of Q, so the stability of the equilibrium is dependent on the second-order statistics of the input signal. Note that for a given value of C, (24) is apt to be violated for networks with a large number of cells since C_0 increases with decreasing (λ_i, λ_j). In particular the lowest value of C_{ij_0} occurs for the largest (i, j). This is the minimum value of C for which the desired equilibrium is stable.

3.3.2. Bifurcation behaviour. I want to locate the equilibria

$$F(X(C), C) = 0 \tag{25}$$

near (X_0, C_0) and determine their stability. A direct solution of (25) is intractable. Instead I use the Liapunov–Schmidt reduction (Golubitsky and Schaeffer 1984) to replace the high-dimensional system (25) with a low-dimensional system

$$g(z, C) = 0 \tag{26}$$

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

which is easily solved. The reduced system (26) is equivalent to (25) in the sense that the zeros of g are in one-to-one correspondence with the zeros of F, and the stability of the bifurcating equilibria can be inferred from g. In this sense the reduced function completely characterizes the bifurcation. The reduction is accomplished by means of a perturbation expansion about the bifurcation point (X_0, C_0). Details are given in appendix B.

To describe the bifurcation from X_0, I assume that the stability condition (24) is violated for a *single* pair of indices (i,j). At $C = C_{ij0}$, $\mathcal{M}_{ij}$ has a simple zero eigenvalue. The corresponding eigenvector of DF_0 is denoted v_r. The reduced function is a real-valued function of the scalar variables z and C, where z is the displacement from X_0 along v_r. The equilibrium X_0 corresponds to $z = 0$.

The perturbation expansion shows that, to third order in z, the reduced function is

$$g(z,C) = -d\,(\lambda_i + \lambda_j)(C - C_0)\,z + d\,z^3 + \ldots \tag{27}$$

which is the normal form for a supercritical pitchfork bifurcation. Generically, the reduced function is of the form

$$g(z,C) = \alpha\,(C - C_0)\,z + \beta\,z^2 + \mu\,z^3 + \ldots .$$

The term quadratic in z is missing from (27) by virtue of inversion symmetries in the system. For example, in a network of two cells, equations (13) and (14) and the equilibrium (X_0) are invariant under the transformation

$$\begin{aligned} &\omega_1^k \rightarrow -\omega_1^k \qquad k = 2, \ldots, N \\ &\omega_2^1 \rightarrow -\omega_2^1 \\ &\eta_{12} \rightarrow -\eta_{12}. \end{aligned} \tag{28}$$

The reduced function, $g(z,C)$, inherits the inversion symmetry. Consequently the term quadratic in z is absent. (There are analogous symmetries for larger networks, although the full symmetry group for arbitrary networks is not known at present.)

The bifurcation diagram (the solution set of $g(z,C) = 0$) is shown in the upper portion of figure 1. The branch corresponding to the equilibrium X_0 is stable for $C > C_0$ and unstable for $C < C_0$. Two *unstable* branches are present for $C > C_0$. At the equilibria on the unstable branches the forward weight vectors are mixtures of e_i and e_j, and the lateral connection η_{ij} is non-zero. Calculations indicate that the form of this bifurcation is independent of both the number of nodes, M, in the network as well as the dimension, N, of the input space.

The position of stable equilibria away from (X_0, C_0) can be inferred from terms in the bifurcation expansion of order z^5 and higher, or alternatively as follows. For simplicity I consider the case of two cells. I examine the homogeneous solution for which both weight vectors are proportional to the principal eigenvector,

$$X_h \equiv \left\{ (\omega,\eta) \mid \omega_1 = \pm\omega_2 = \frac{1}{\sqrt{1 + C\lambda_1}}\,e_1,\ \eta_{12} = \mp\frac{C\lambda_1}{1 + C\lambda_1} \right\}.$$

This is asymptotically stable provided

$$C < C_h \equiv \min \left\{ \begin{matrix} (\lambda_1 - \lambda_2)/(2\lambda_1\lambda_2) \\ 1/\lambda_1 \end{matrix} \right\}. \tag{29}$$

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

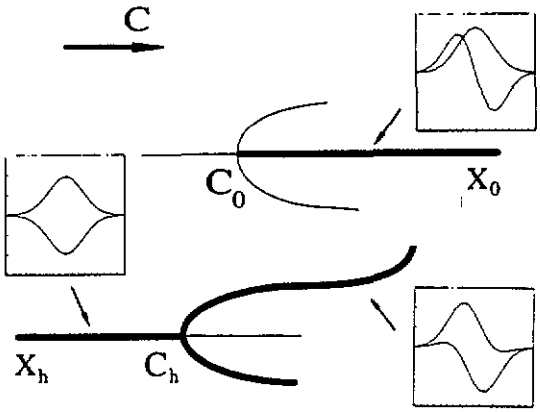


Figure 1. Bifurcation diagram for the minimal model. Heavy lines are stable branches and light lines are unstable branches. Insets show the receptive fields corresponding to equilibria on the stable branches.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

If the first condition in (29) is violated, then there is a supercritical pitchfork bifurcation, (bottom portion of figure 1). (I have not determined the normal form under violation of the second condition of (29).) The equilibria on the *stable* bifurcating branches are mixtures of e_1 and e_2 with non-zero η_{12}. These branches presumably join the unstable supercritical branches of the bifurcation at (X_0, C_0).

Finally if the condition on d in (23) is violated for a single pair of indices (i, j) while the condition on C in (24) is satisfied, then the equilibrium X_0 loses stability through a Hopf bifurcation. A pair of complex-conjugate eigenvalues of DF_0 cross the imaginary axis at

$$\pm \mathrm{i}\Lambda_0 = \pm \mathrm{i}\,(\lambda_i - \lambda_j)\sqrt{C\,(\lambda_i + \lambda_j) - 1}$$

and the forward weights and lateral connection strengths will begin to oscillate.

3.3.3. Numerical results. Numerical integration of (13) and (14) for the case of two cells confirms the picture developed above. For large C, the weight vectors converge to the leading eigenvectors of the autocorrelation. For $C < C_h$, the weight vectors collapse to the leading eigenvector. Values of C are found at which the weight vectors converge to mixtures of the eigenvectors of the correlation.

As an example, the insets in figure 1 show the receptive fields (ω_1 and ω_2) corresponding to the stable branches of the bifurcation diagram. The two curves in each diagram depict the converged forward weight vectors for the two cells in the network. Each weight vector has 19 components. The magnitude of each of the components is plotted on the ordinate as a function of the component number along the abscissa.

These configurations were generated by a correlation matrix corresponding to an ensemble of N-dimensional noise vectors with short-range correlations between the components. The elements of the correlation matrix are given by

$$Q_{ij} = \exp\left[-\alpha\,(i-j)^2\right] \exp\left[-\beta\left(\frac{i+j-(N+1)}{2}\right)^2\right] \qquad \alpha, \beta > 0, \quad i, j = 1, \ldots, N.$$

According to the first exponential factor, correlations drop off with separation between the components. The second factor further reduces correlations between components

at the edges of the input field. This corresponds to a decreased synaptic density at the edges of the input field. Similar correlation matrices are used, for example, by Yuille *et al* (1989). For the simulations given here, $N = 19$, $\alpha = 0.125$, $\beta = 0.04$. The first four eigenvalues of the resulting correlation matrix are (3.908, 2.185, 1.221, 0.682). For the simulations in figure 1, $d = 2.5$, $d_0 = 0.902$.

The bifurcation diagram in figure 1 was further confirmed by numerical integration, sweeping the coupling strength up and back down through the bifurcation points. Figure 2 shows the cosine of the angle between the two weight vectors, as a function of the coupling strength C. At the lowest values of C the weight vectors are opposite one another, corresponding to X_h. As C is increased this configuration becomes unstable and the angle between the weight vectors begins to close. At the highest values of C, the weight vectors are orthogonal, corresponding to X_0. As for the simulations in figure 1, $d = 2.5$ while $d_0 = 0.902$†.

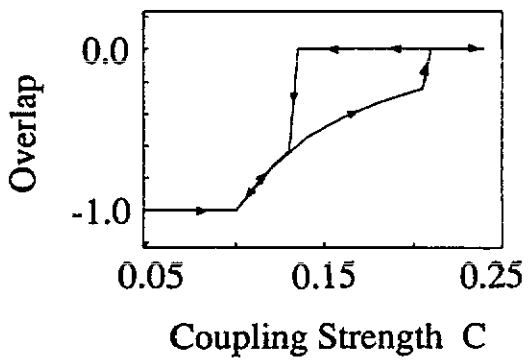


Figure 2. Hysteresis curve obtained by tracing out the bifurcation diagram of figure 1 for a two-cell model.

For networks with more than two cells, there are presumably additional bifurcations along the branches emanating from (X_h, C_h). Simulations show various mixed states. For example, figure 3 shows receptive field configurations generated by a three-cell model. The critical coupling value for these simulations is $C_0 = 0.294$. Figure 3(a) shows the receptive fields for $C = 1.0$. These are the eigenfunctions of the input correlation. Figure 3(b) shows the receptive fields generated at $C = 0.28$. One of the nodes has converged to the leading eigenfunction (full line), while the other two nodes have converged to mixtures of the second and third correlation eigenfunctions. Figure 3(c) shows the receptive fields generated at $C = 0.18$. Two of the nodes have converged to the leading eigenfunction, while the third has converged to the second eigenfunction. Again $d = 2.5$, while $d_0 = 2.2$.

9.4. Activity-dependent adaptation

The conditions on the stability of X_0 suggest that the adaptation rule for the lateral connection strengths can be modified to enhance the stability of X_0. Examining (23) and (24), it is clear that the critical value for $C \times d$ is bounded above by unity. Furthermore the relaxation rate required by (23) is bounded above by

$$\overline{d_0} \equiv \lambda_i + \lambda_j$$

† The destabilization of X_0 occurs somewhat below C_0, presumably due to numerical integration errors.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

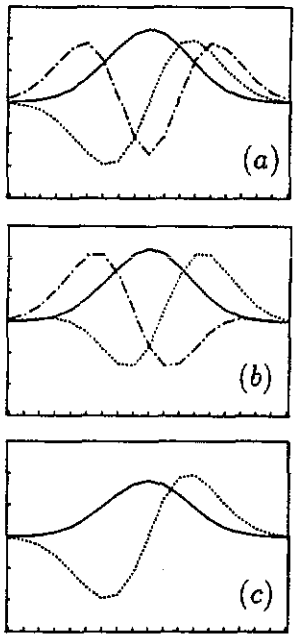


Figure 3. Receptive fields generated in a three-cell simulation with $C_0 = 0.294$. (*a*) Weight vectors at $C = 1.0$ are the eigenvectors of the autocorrelation. (*b*) At $C = 0.28$, $\omega_1 = e_1$, $\omega_2 = 0.477\, e_3 - 0.768\, e_2$, $\omega_3 = 0.477\, e_3 + 0.768\, e_2$. (*c*) At $C = 0.18$, $\omega_1 = \omega_3 = 0.766\, e_1$, $\omega_2 = 0.999\, e_2$.

which is the sum of the node response variances at the equilibrium point X_0.

This suggests that the stability may be improved by weighting the relaxation rate of the lateral connections by the variance of the cell responses. I make this change, rewriting (13) as

$$\begin{aligned}\dot{\eta}_{ij} &= -\langle y_i^2 + y_j^2 \rangle\, \eta_{ij} - C \langle y_i\, y_j \rangle \\ &= -(\omega_i \cdot Q\, \omega_i + \omega_j \cdot Q\, \omega_j)\, \eta_{ij} - C\, \omega_i \cdot Q\, \omega_j. \end{aligned} \tag{30}$$

With this change, the critical 3×3 sub-blocks of DF_0 become

$$\mathcal{M}_{ij} = \begin{bmatrix} \lambda_j - \lambda_i & 0 & \lambda_j \\ 0 & \lambda_i - \lambda_j & \lambda_i \\ -C\lambda_j & -C\lambda_i & -(\lambda_i + \lambda_j) \end{bmatrix} \tag{31}$$

and the conditions for stability of the equilibrium X_0 reduce to

$$C > 1 \tag{32}$$

which no longer depends on the spectrum of the autocorrelation. Note that when (32) is violated, DF_0 has $M(M-1)/2$ null eigenvectors.

4. Complete coupling

The model presented in the previous section uses the lateral connections to mediate plasticity, but not cell responses. In this section I discuss a model that uses the lateral connections in the usual sense. Conditions for stability are given, as well as the behaviour near the loss of stability. I also suggest an enhancement similar to that in subsection 3.4, that provides a more robust algorithm.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

4.1. Equations of motion

I define the $M \times N$ matrix of forward weights, ω, connecting the input space to the cell array. The ith row of ω is the weight vector, ω_i, to the ith cell of the array. The matrix η has the same structure as in section 3. In addition, $y \in R^M$ denotes the vector of cell responses. The components of y are denoted y_i, $i = 1, \ldots, M$.

The response of a cell is given by the sum of the forward-propagated signals and the laterally propagated signals,

$$y = \omega x + \eta y.$$

Solving for y leaves

$$y = (1 - \eta)^{-1} \omega x \equiv u \omega x. \tag{33}$$

In an analog VLSI or biological system the matrix inversion in (33) would be calculated implicitly through the *node* dynamics (assuming convergent activation dynamics (Hirsch 1989)†). For implementation in a digital system, or for simulation without explicit node dynamics, some form of direct matrix inversion would need to be calculated. Simulations show that a truncated series expansion of $u(\eta)$ is a viable alternative.

The ensemble-averaged, continuous time form of the adaptation rule for the forward weights takes the form

$$\dot{\omega}_i = \langle y_i x \rangle - \langle y_i y_i \rangle \omega_i$$

or equivalently

$$\begin{aligned} \dot{\omega} &= \langle y x^{\mathrm{T}} \rangle - \mathrm{Diag} \langle y y^{\mathrm{T}} \rangle \omega \\ &= u \omega Q - \mathrm{Diag} (u \omega Q \omega^{\mathrm{T}} u^{\mathrm{T}}) \omega \end{aligned} \tag{34}$$

where Diag takes the diagonal elements of its argument. The lateral connection strengths develop according to

$$\begin{aligned} \dot{\eta}_{ij} &= d \eta_{ij} - C \langle y_i y_j \rangle = d \eta_{ij} - C \langle y y^{\mathrm{T}} \rangle_{ij} \\ &= d \eta_{ij} - C (u \omega Q \omega^{\mathrm{T}} u^{\mathrm{T}})_{ij} \qquad 1 \le i \ne j \le M. \end{aligned} \tag{35}$$

4.2. Stability and bifurcation

This system has an equilibrium at

$$X_0 \equiv \{ \omega_i = e_i, i = 1, \ldots, M; \ \eta_{ij} = 0 \} \qquad \forall C. \tag{36}$$

In order to address the stability of the equilibrium, I follow the treatment of subsection 3.3 and expand the rows of ω in the basis of eigenvectors of Q, writing the components as

$$\omega_i^j \equiv \omega_i \cdot e_j.$$

† For example, the system $dy/dt = (1/\tau)(-y + \omega x + \eta y)$ has a globally attracting fixed point at $y = u \omega x$ provided the matrix $(1 - \eta)$ is positive definite.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The variables are regrouped as in (16) and the equations of motion are written as in (18).

To perform the stability and bifurcation calculations, I expand u as a power series in η

$$u \simeq 1 + \eta + \eta^2 + \eta^3 + \dots . \tag{37}$$

The first-order term is sufficient for the stability calculation. The terms through order η^3 are required to address the bifurcation.

As in subsection 3.3.1, the linear part of the vector field at the equilibrium breaks into block diagonal form with any instabilities constrained to 3×3 sub-blocks. These sub-blocks take the form

$$\mathcal{M}_{ij} = \begin{bmatrix} \lambda_j - \lambda_i & 0 & \lambda_j \\ 0 & \lambda_i - \lambda_j & \lambda_i \\ -C\lambda_j & -C\lambda_i & d - C(\lambda_i + \lambda_j) \end{bmatrix}. \tag{38}$$

The stability conditions read

$$d > 0 \tag{39}$$

$$C > C_0 \equiv \frac{d}{(\lambda_i + \lambda_j)} + \frac{(\lambda_i - \lambda_j)^2}{(\lambda_i^2 + \lambda_j^2)}. \tag{40}$$

It is useful to digress briefly and discuss the form of the adaptation rule (35) for η in relation to the stability condition (39). Previous authors (Foldiak 1989, Rubner and Schulten 1990) have used an anti-Hebbian rule of the form

$$\dot{\eta}_{ij} = -\langle y_i y_j \rangle. \tag{41}$$

The present development shows that for a system with full lateral connectivity, the PCA equilibrium may not be stable without the linear term. Equation (39) shows that removing the term linear in η can lead to an instability.

If the stability condition on C is violated, then the network undergoes a Hopf bifurcation, leading to oscillating solutions. To see this, calculate the characteristic polynomial of $\mathcal{M}_{ij}$,

$$P(L) = L^3 + (C(\lambda_i + \lambda_j) - d)\, L^2 + [C(\lambda_i^2 + \lambda_j^2) - (\lambda_i - \lambda_j)^2]\, L + d(\lambda_i - \lambda_j)^2. \tag{42}$$

At $C = C_0$, the roots of (42) are

$$L_0 = \frac{-(\lambda_i - \lambda_j)^2 (\lambda_i + \lambda_j)}{\lambda_i^2 + \lambda_j^2} \tag{43}$$

$$L_\pm = \pm \mathrm{i} \sqrt{\frac{d(\lambda_i^2 + \lambda_j^2)}{\lambda_i + \lambda_j}} \equiv \pm \mathrm{i}\, \Lambda_0. \tag{44}$$

The pair of roots in (44) are pure imaginary provided $d > 0$, so DF_0 develops a pair of pure imaginary eigenvalues at $C = C_0$.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The conditions for a non-degenerate Hopf bifurcation are satisfied. If (40) is violated for a ***single*** pair of indices (i, j) then DF_0 has only a single pair of eigenvalues on the imaginary axis. Second, these eigenvalues cross the imaginary axis with non-zero speed as C passes through C_0. To verify the crossing condition, calculate the rate of change of the real part of the complex-conjugate eigenvalues at C_0,

$$\mathrm{Re}\left[\frac{\mathrm{d}\,L_{\pm}(C_0)}{\mathrm{d}\,C}\right] = \mathrm{Re}\left[-\frac{\partial P/\partial C}{\partial P/\partial L}\bigg|_{L_{\pm},C_0}\right]$$
$$= \frac{-d\Lambda_0^2(\lambda_i^2+\lambda_j^2)(\Lambda_0^2+(\lambda_i-\lambda_j)^2)}{2\,[d^2(\lambda_i-\lambda_j)^4+\Lambda_0^6]} < 0 \qquad (45)$$

which confirms that X_0 is stable for $C > C_0$. Thus there is a non-degenerate Hopf bifurcation at (X_0, C_0).

I applied the technique given by Golubitsky and Schaeffer (1984) to determine whether the bifurcation is super- or subcritical. The calculations were carried out with a computer algebra package and the results corroborated by independent numerical analysis of the bifurcation equations at specific values of (λ_i, λ_j), and by numerical integration of the equations of motion. I carried out the calculations for the case of two cells in a two-dimensional input space (five degrees of freedom).

I find that the direction of the bifurcation depends on the eigenvalues λ_1 and λ_2 appearing in $\mathcal{M}_{12}$. The expression for the function that determines the direction of the bifurcation is rather complicated, and the results are best displayed graphically. Figure 4 shows the regions corresponding to sub- and supercritical bifurcation for $d = 1$. The shaded region in the plot corresponds to values of (λ_1, λ_2) for which the bifurcation is supercritical, with unstable periodic orbits near X_0 for $C > C_0$. The unshaded region corresponds to subcritical bifurcation with stable periodic orbits near X_0 for $C < C_0$.

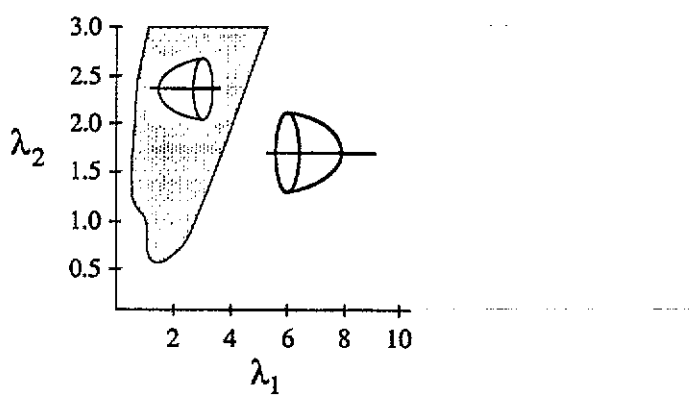


Figure 4. Regions in the (λ_1, λ_2) plane corresponding to supercritical (shaded) and subcritical (unshaded) Hopf bifurcations in the complete model.

Simulations show that even for values of (λ_1, λ_2) corresponding to a supercritical bifurcation, there are stable periodic orbits for $C < C_0$. This suggests that the complete bifurcation diagram in the supercritical regime is shaped like the bottom of a wine bottle, only the indentation of which is shown in figure 4.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

4.3. Activity-dependent adaptation

As in subsection 3.4, the stability of the PCA equilibrium is enhanced by modulating the adaptation rate of the lateral connections according to the cell response variances. I make this change, replacing (35) with

$$\begin{aligned}\dot{\eta}_{ij} &= \langle y_i^2 + y_j^2 \rangle \, \eta_{ij} - C \langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \rangle_{ij} \\ &= [\, (u\omega Q\omega^{\mathrm{T}} u^{\mathrm{T}})_{ii} + (u\omega Q\omega^{\mathrm{T}} u^{\mathrm{T}})_{jj} \,] \, \eta_{ij} \\ &\quad - C(u\omega Q\omega^{\mathrm{T}} u^{\mathrm{T}})_{ij} \qquad 1 \le i \ne j \le M. \end{aligned} \tag{46}$$

With this change, the critical sub-blocks of $\mathrm{D}F_0$ take the form

$$\mathcal{M}_{ij} = \begin{bmatrix} \lambda_j - \lambda_i & 0 & \lambda_j \\ 0 & \lambda_i - \lambda_j & \lambda_i \\ -C\lambda_j & -C\lambda_i & (1-C)(\lambda_i + \lambda_j) \end{bmatrix} \tag{47}$$

and the condition for stability of the equilibrium X_0 reads

$$C > 1 + \frac{(\lambda_i - \lambda_j)^2}{\lambda_i^2 + \lambda_j^2} \tag{48}$$

which, unlike (40) is bounded above by 2.

Lastly in an artificial system, it is desirable to avoid calculating the matrix inversion $(1-\eta)^{-1}$ that appears in the equation for the output activation (33). Simulations in which this inverse is approximated by the first two terms of its series expansion

$$u \approx 1 + \eta$$

provide reasonably good convergence.

5. Discussion

The aim of this paper has been to give a theoretical description of learning in feature-discovery models; in particular those that use lateral interactions to ensure that cells tune to different features in the environment. Towards this end, I have introduced two neural models for linear feature discovery that are based on a combination of Hebbian learning, and recurrent lateral connections that develop according to an anti-Hebbian learning rule. Both models employ anti-Hebbian learning rules that include a term linear in the lateral connections, thus departing from forms previously given in the literature.

The minimal model in section 3 treats the lateral connections in a rather unorthodox way. The signals carried on the lateral connections affect plasticity but *not* the target cell activation. This is advantageous for digital implementation. In a model that uses lateral signal flow to affect cell activation, the values of the cell activities require time to equilibriate. Here the activities are calculated from the forward signals alone, and this can be done in a single machine cycle. Note that this form of coupling can also be derived as a low-order approximation to the model in section 4

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

by expanding the $(\boldsymbol{y}\boldsymbol{x}^{\mathrm{T}})$ term of (34) in a power series in η, and retaining only the linear terms.

Both models have equilibria at which the network performs PCA. There are critical values of the learning rates (C and d) below which this equilibrium is unstable. The bifurcation analyses and simulations show that both models support several solutions.

The minimal model has stable secondary equilibria in which the forward weight vectors are combinations of the eigenvectors of the input autocorrelation. Furthermore there is a range of the coupling strength C over which both equilibria are stable (see figure 1). Within this range, the filter properties of the mature cells will depend on the initial conditions. The complete model has solutions in which the forward weight vectors are oscillating combinations of the eigenvectors of the input autocorrelation. For the complete model to be useful in this regime, learning would have to be turned off after an initial period of adaptation.

Both models expand on earlier work on Hebbian feature discovery and principal component analysis. In the limit of fast relaxation of the lateral weights (large d), the minimal model reverts to the cortical cell model of Yuille *et al* (1989) that motivated our derivation. Foldiak (1989) discusses an algorithm which, like the models discussed here, uses full lateral connectivity. However, the adaptation rule that he uses for the lateral connections has no linear term, in contrast with the anti-Hebb rule given here. As discused in section 4 removing the linear term can result in an instability of the PCA equilibrium; the model without the linear term operates at a bifurcation point of the model discussed here.

This study has shown the utility of bifurcation theory in the analysis of unsupervised learning algorithms. However, several questions remain unanswered. The bifurcations of the secondary equilibria in the minimal model, or the possibility of bifurcations from limit cycles to tori in the complete model, is yet to be explored. The introduction of activity-dependent adaptation rates in subsection 3.4 leads to an improvement in the conditions for stability of the PCA equilibrium; compare (32) with (24). However, the (degenerate) bifurcation at the critical point has not been explored. I expect that, like the simpler model, the augmented model will have stable equilibria that coexist with the PCA equilibrium.

Finally, models that perform PCA employ cells with *linear* postsynaptic response. Apart from the winner-take-all interactions used in competitive learning schemes, the role of nonlinear postsynaptic response in systems with Hebbian adaptation is relatively unexplored.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

Acknowledgments

The author thanks Professor Dan Hammerstrom and Dr Bill Baird for lively discussion. David Roe helped automate the bifurcation calculations and Vince Weatherill worked on the figures. The reviewers and several of my colleagues have provided valuable comments on the manuscript. This work was supported by the Office of Naval Research under contracts N00014-88-K-0329 and N00014-90-1349 and by DARPA grant MDA 972-88-J-1004.

Appendix A. Stability calculations

This appendix provides details of the stability calculations of section 3. To proceed, I recast the equations of motion (13) and (14) in the basis of eigenvectors of Q, writing

the projection of ω_i onto e_j as ω_i^j,

$$\dot{\omega}_i^j = \lambda_j \omega_i^j + \sum_{m \neq i} \eta_{im} \lambda_j \omega_m^j - \left(\sum_m \lambda_m (\omega_i^m)^2 \right) \omega_i^j \tag{A1}$$

$$\dot{\eta}_{ij} = -d \left(\eta_{ij} + C \sum_m \lambda_m \omega_i^m \omega_j^m \right) \qquad 1 \leq i, j \leq M. \tag{A2}$$

The PCA equilibrium is at $X_0 = \{\omega_i^j = \delta_{ij}, \eta_{ij} = 0\}$.

The derivatives appearing in the linearization are easily evaluated. First,

$$\left. \frac{\partial \dot{\omega}_i^j}{\partial \omega_k^l} \right|_{X_0} = \delta_{ik} \delta_{jl} (\lambda_j - \lambda_i) - 2 \delta_{ij} \delta_{il} \delta_{ik} \lambda_i. \tag{A3}$$

Next,

$$\left. \frac{\partial \dot{\omega}_i^j}{\partial \eta_{kl}} \right|_{X_0} = \delta_{ik} \delta_{jl} \lambda_j \qquad 1 \leq i, j \leq M. \tag{A4}$$

The derivatives of the η terms are

$$\left. \frac{\partial \dot{\eta}_{ij}}{\partial \omega_k^l} \right|_{X_0} = -dC [\lambda_l (\delta_{ik} \delta_{jl} + \delta_{jk} \delta_{il})] \qquad 1 \leq i, j \leq M. \tag{A5}$$

and

$$\frac{\partial \dot{\eta}_{ij}}{\eta_{kl}} = -d \delta_{ik} \delta_{jl} \qquad 1 \leq i, j \leq M. \tag{A6}$$

The critical 3×3 sub-blocks of DF_0 are given by the Jacobian matrices

$$\mathcal{M}_{ij} = \frac{\partial(\dot{\omega}_i^j, \dot{\omega}_j^i, \dot{\eta}_{ij})}{\partial(\omega_i^j, \omega_j^i, \eta_{ij})} = \begin{bmatrix} \lambda_j - \lambda_i & 0 & \lambda_j \\ 0 & \lambda_i - \lambda_j & \lambda_i \\ -dC\lambda_j & -dC\lambda_i & -d \end{bmatrix}. \tag{A7}$$

The terms in the sub-block $\{A\}$ (21) of DF_0 are from the derivatives of $\dot{\omega}_i^i$. From (A3) the only non-zero terms are

$$\frac{\partial \dot{\omega}_i^i}{\partial \omega_i^i} = -2\lambda_i. \tag{A8}$$

Finally the sub-block (B) (22) contains derivatives of $\dot{\omega}_i^J$ with $1 \leq i \leq M$ and $J > M$. From (A4) it is clear that all the derivatives with respect to η will vanish. The only remaining terms are

$$\frac{\partial \dot{\omega}_i^J}{\partial \omega_i^J} = \lambda_J - \lambda_i < 0. \tag{A9}$$

Matrix elements outside the block diagonals are easily seen to vanish. This confirms the form of DF_0 given in subsection 3.3.1. A similar calculation gives the analogous form in subsection 4.2.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The block-diagonal form of DF_0 simplifies the stability calculations since the sub-blocks $\mathcal{M}_{ij}$ are the only parts where an instability can arise. The characteristic polynomial for $\mathcal{M}_{ij}$ is

$$P(L) = L^3 + L^2 d + L[d C(\lambda_i^2 + \lambda_j^2) - (\lambda_i - \lambda_j)^2] + d(\lambda_i - \lambda_j)^2[C(\lambda_i + \lambda_j) - 1]. \tag{A10}$$

I applied the Routh-Hurwitz conditions (Willems 1970) to (A10) to derive the stability conditions given in subsection 3.3. As a check, it is straightforward to verify that $P(L)$ develops a simple zero root at $C = C_{ij_0} \equiv 1/(\lambda_i + \lambda_j)$. This is the value of C at which the pitchfork develops. To verify the condition on d, observe that $P(L)$ develops a pair of pure imaginary roots

$$\pm i(\lambda_i - \lambda_j)\sqrt{C(\lambda_i + \lambda_j) - 1}$$

at $d = d_{ij_0} \equiv (\lambda_i - \lambda_j)^2(\lambda_i + \lambda_j)/(\lambda_i^2 + \lambda_j^2)$.

Appendix B. Bifurcation calculations

The bulk of the bifurcation calculations were performed with a symbolic manipulation program, used both interactively and by running code written explicitly for this study. I used the Liapunov-Schmidt reduction to find the bifurcation normal form. As an alternative, one could perform a centre manifold reduction at the bifurcation point. Here I sketch the calculations for the bifurcation from the PCA equilibrium in the minimal model. More details on the Liapunov-Schmidt reduction can be found in Golubitsky and Schaeffer (1984).

The equations of motion for the weights

$$X = F(X, C)$$

have an equilibrium at X_0. I assume that the stability conditions (24) are violated for a *single* pair of indices (i, j). At $C = C_0$, one of the eigenvalues of $\mathcal{M}_{ij}$ becomes zero. The right and left eigenvectors of DF_0 corresponding to the zero eigenvalue are denoted v_r and v_l respectively.

The reduction proceeds as follows. Let S denote the $[MN + M(M-1)/2]$-dimensional configuration space of variables (ω, η). Define the projection $E : S \to \text{Range}\, DF_0$. Points in the configuration space are given coordinates as

$$X = X_0 + z\, v_r + W$$

with $W \in (\text{Ker}DF_0)^{\perp}$. Using the implicit function theorem, the equation

$$E\, F(X_0 + z\, v_r + W, C) = 0 \tag{B1}$$

is solved for $W(z, C)$ in a neighbourhood of the bifurcation point (X_0, C_0). Note that since $F(X_0, C_0) = 0$, $W(0, C_0) = 0$.

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The reduced function is given by

$$\begin{aligned} g(z,C) &= v_l \cdot (1-E)\, F(X_0 + z\, v_r + W(z,C), C) \\ &= v_l \cdot F(X_0 + z\, v_r + W(z,C), C). \end{aligned} \tag{B2}$$

The zero set of F is in one-to-one correspondence with the zero set of g. The latter is the bifurcation diagram for the system.

Stability of equilibria is determined from the sign of $\partial g/\partial z$. Consider an equilibrium $g(z(C),C) = 0$. If $v_l \cdot v_r > 0$ and $\partial g(z(C),C)/\partial z < 0$, then the equilibrium is asymptotically stable, and unstable if $\partial g(z(C),C)/\partial z > 0$.

In practice (B1) is solved for the Taylor series expansion of W. The terms in the series are then substituted into a series expansion of (B2). As an example I give the reduction for a network of $M = 2$ cells in a three-dimensional input space (seven degrees of freedom). Calculations show that the results generalize to larger nets. The coordinates are ordered as

$$X = [\omega_1^2,\ \omega_2^1,\ \eta_{12},\ \omega_1^1,\ \omega_2^2,\ \omega_1^3,\ \omega_2^3]$$

with the equilibrium at

$$X_0 = [0,0,0,1,1,0,0].$$

The linear part of the vector field at X_0 is

$$DF_0 = \begin{bmatrix} \lambda_2 - \lambda_1 & 0 & \lambda_2 & & & & \\ 0 & \lambda_1 - \lambda_2 & \lambda_1 & & & & \\ -C\,d\,\lambda_2 & -C\,d\,\lambda_1 & -d & & & & \\ & & & -2\lambda_1 & & & \\ & & & & -2\lambda_2 & & \\ & & & & & \lambda_3 - \lambda_1 & \\ & & & & & & \lambda_3 - \lambda_2 \end{bmatrix} \tag{B3}$$

At C_0 the right and left eigenvectors of DF_0 corresponding to eigenvalue zero are given by

$$v_r = \left[\frac{\lambda_2}{(\lambda_1 - \lambda_2)},\ \frac{\lambda_1}{(\lambda_2 - \lambda_1)},\ 1,\ 0,\ 0,\ 0,\ 0\right] \tag{B4}$$

and

$$v_l = \left[\frac{d\lambda_2}{(\lambda_1 + \lambda_2)(\lambda_1 - \lambda_2)},\ -\frac{d\lambda_1}{(\lambda_1 + \lambda_2)(\lambda_1 - \lambda_2)},\ -1,\ 0,\ 0,\ 0,\ 0\right]. \tag{B5}$$

The coefficients in the series expansion of W that are required to identify the bifurcation type are found by differentiating (B1) and solving the resulting expression for the coefficients. I find

$$W_c = [0,0,0,0,0,0,0] \tag{B6}$$

and

$$W_{zz} = \left[0,0,0,\frac{-\lambda_1^3 + \lambda_1^2\lambda_2 - \lambda_2^3}{\lambda_1(\lambda_1 - \lambda_2)^2},\ \frac{-\lambda_1^3 + \lambda_1\lambda_2^2 - \lambda_2^3}{\lambda_2(\lambda_1 - \lambda_2)^2},0,0\right]. \tag{B7}$$

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

The terms in the series expansion of the reduced function needed to identify the bifurcation are found by differentiating (B2),

$$g_c = v_l \cdot F_c(X_0, C_0) = 0 \tag{B8}$$

$$g_{zz} = v_l \cdot D^2 F_0[v_r, v_r] \tag{B9}$$

$$g_{zc} = v_l \cdot (DF_c \cdot v_r + D^2 F_0[v_r, W_c]) \tag{B10}$$

$$g_{zzz} = v_l \cdot (D^3 F_0[v_r, v_r, v_r] + 3\, D^2 F_0[v_r, W_{zz}]) \tag{B11}$$

The contraction of the second derivative of F with v_r is given by

$$D^2 F_0[v_r, v_r] = \left[0, 0, 0, \frac{-2(\lambda_1^3 - \lambda_1^2\lambda_2 + \lambda_2^3)}{(\lambda_1 - \lambda_2)^2}, \frac{-2(\lambda_1^3 - \lambda_1\lambda_2^2 + \lambda_2^3)}{(\lambda_1 - \lambda_2)^2}, 0, 0\right]. \tag{B12}$$

Notice that $v_l \cdot D^2 F[v_r, v_r]$ vanishes identically so that g_{zz} must vanish. The inversion symmetry (28) underlies the absence of the quadratic term.

Evaluating the terms in (B10) and (B11), I find

$$g_{zc} = -d(\lambda_1 + \lambda_2)$$
$$g_{zzz} = 6d.$$

To third order, the reduced function is thus

$$\begin{aligned} g &= g_{cz}(C - C_0)\, z + \frac{1}{3!}\, g_{zzz}\, z^3 + \dots \\ &= -d(\lambda_1 + \lambda_2)(C - C_0)\, z + d\, z^3. \end{aligned} \tag{B13}$$

For $C > C_0$, the roots of g are at $z_s = \pm\sqrt{(C - C_0)(\lambda_1 + \lambda_2)}$, and $z = 0$. For $C < C_0$ the only root of g is at $z = 0$. Since X_0 corresponds to $z = 0$, the latter is stable for $C > C_0$, and unstable for $C < C_0$. By exchange of stability, the secondary equilibria z_s should be unstable. As a check note that for $d > d_0$, $v_r \cdot v_l > 0$. Then,

$$\partial g/\partial z = 3d\, z^2 - (C - C_0)\, d(\lambda_1 + \lambda_2).$$

At $z = 0$, $\partial g/\partial z < 0$ for $C > C_0$ indicating stability of X_0. At z_s, $\partial g/\partial z > 0$ indicating instability.

References

Foldiak P 1989 Adaptive network for optimal linear feature extraction *Proc. Int. Joint Conference on Neural Networks* (Washington, DC, June 1989) (Piscataway, NJ: IEEE) pp I-401–I-405

Fuchs A and Haken H 1988 Pattern recognition and associative memory as dynamical processes in a synergetic system *Biol. Cybern.* **60** 17–22

Golubitsky M and Schaeffer D 1984 *Singularities and Groups in Bifurcation Theory* vol I (New York: Springer)

Hirsch M 1989 Convergent activation dynamics in continuous time networks *Neural Networks* **2** 331–49

Leen T K, Rudnick M and Hammerstrom D 1990 Hebbian feature discovery improves classifier efficiency *Proc. Int. Joint Conference on Neural Networks* (Washington, DC, June 1989) (Piscataway, NJ: IEEE) pp I-51–I-56

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009

Linsker R 1988 Self-organization in a perceptual network *Computer* (March) 105–17
Ljung L 1977 Analysis of recursive stochastic algorithms *IEEE Trans. Auto. Control* **AC-22** 551–75
MacKay D J and Miller K D 1990 Analysis of Linsker's application of Hebbian rules to linear networks *Network* **1** 257–97
Oja E 1982 A simplified neuron model as a principal component analyzer *J. Math. Biology* **15** 267–73
—— 1989 Neural networks, principal components, and subspaces *Int. J. Neural Systems* **1** 61–8
Oja E and Karhunen J 1985 On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix *J. Math. Anal. Appl.* **106** 69–84
Rubner J and Schulten K 1990 Development of feature detectors by self-organization: a network model *Biol. Cybern.* **62** 193–9
Sanger T 1989 An optimality principle for unsupervised learning *Advances in Neural Information Processing Systems* ed D S Touretsky (San Mateo, CA: Morgan Kauffmann)
Willems J L 1970 *Stability Theory of Dynamical Systems* (New York: Wiley)
Yuille A L, Kammen D M and Cohen D S 1989 Quadrature and the development of orientation selective cortical cells by Hebb rules *Biol. Cybern.* **61** 183–94

Downloaded By: [OHSU Science & Engineering Library] At: 11:16 16 April 2009