Learning in Linear Feature-Discovery Networks *

Todd K. Leen

Dept. of Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
Beaverton OR 97006-1999

tleen@cse.ogi.edu

Abstract

We describe the dynamics of learning in unsupervised linear feature-discovery networks that have recurrent lateral connections. Bifurcation theory provides a description of the location of multiple equilibria and limit cycles in the weight-space dynamics.

Introduction

Unsupervised neural network models that recursively estimate principal components of the input correlation have received increased attention in recent literature. This interest is due in part to their utility as encoding devices; among linear transformations for dimension reduction, principal component analysis (PCA) provides the lowest mean square error encoding. A second motivation arises from the fact that the algorithms are based on simple Hebbian learning rules that have analogs in biophysical mechanisms for changing the strength of synapses [2]. Thus the algorithms may provide insight into the organization of biological nervous systems.

In a classic paper, Oja [3] made the remarkable observation that a simple model neuron with an Hebbian adaptation rule develops into a filter for the first principal component of the input distribution. Oja proved convergence of the algorithm by appealing to Ljung's results [4] relating recursive stochastic algorithms to ordinary differential equations.

Several researchers have extended Oja's work, developing networks that perform a complete PCA. Sanger [5] proposed an algorithm that uses a single layer of weights

*This article is adapted from [1].

0-8194-0693-7/91/$4.00

with a set of cascaded feedback projections to force different nodes to filter for different principal components. This architecture singles out a particular node for each principal component. Sanger gave a convergence proof based on Oja's original work.

Oja and Karhunen [6] gave a related algorithm that maps to a three-layer feed-forward network of linear neurons [7]. Simulations show that this algorithm projects inputs onto an orthonormal basis spanning the principal subspace. In a theoretical analysis, Krogh and Hertz [8] showed that the *only* asymptotically stable equilibria of the algorithm correspond to this orthonormal basis. At about the same time that [6] appeared, Williams proposed a closely related algorithm that he called symmetric error-correction [9]. Williams also showed that asymptotically stable equilibrium points correspond to orthogonal weight vectors spanning the principal subspace. Sanger and Oja's algorithms have good convergence properties, and we have used both to reduce the dimension of speech signals for phoneme recognition tasks [10].

In another class of models, nodes are forced to filter for different principal components by a set of lateral connections. These models consist of a single layer of nodes with weights from the input space and additional lateral connections between the nodes. The weights from the inputs evolve according to a Hebbian rule similar to that in Oja's single neuron model. The lateral connections evolve according to an *anti*-Hebbian rule. The idea is that the lateral connections should develop so as to decorrelate the activity of the nodes. The anti-Hebbian rule drives changes in the lateral connections in the proper direction. These networks have two learning rates associated with them, one for the weights from the inputs and another for the lateral connections. As we shall see, the ratio of the two learning rates is a parameter that can bear significantly on the converged state of the network.

Rubner and Schulten [11] constructed a network with cascaded lateral connections; the i^{th} node receives signals from the input and all nodes j with $j < i$. Like Sanger's scheme, this architecture singles out a particular node for each principal component. Rubner and Tavan [12] gave conditions on the two learning rates that insure stability of the equilibrium at which the network performs PCA, but no global convergence proof exists.

Although useful for adaptive signal processing, networks with cascaded lateral connections are biologically implausible. Földiák [13] addressed this, presenting simulations of a network with *full* lateral connectivity. His simulations show that the weights from the input space converge to the subspace spanned by the principal eigenvectors of the input's auto-correlation matrix, and that the resulting network captures as much infor-

mation as standard PCA. However, the converged weights are not (in general) equal to the correlation eigenvectors, and they need not be orthonormal.

It seems more difficult to obtain theoretical convergence results for PCA networks that use lateral connections, than for the case of either Oja's original single-neuron model or Sanger's model. For example, Rubner's theoretical results address linear stability of the PCA solution, rather than global convergence. Földiák's algorithm is even more formidable due to the recurrent lateral connections. The latter algorithm *does* have an equilibrium at which the weights from the input are along the correlation eigenvectors, however the usual linear analysis fails to determine stability. We will consider this in some detail below.

The work that we present in the remainder of this paper is aimed at describing some of the dynamical phenomena that appear in models with recurrent lateral connections. Bifurcation theory locates equilibria that do not have the weight vectors along the correlation eigenvectors, and reveals limit cycle solutions as well.

The models that we present are not meant to replace those discussed above. If you need to perform PCA with a neural network, you are probably better off using either the algorithm in [5], or the algorithm in [7] (if you can live with mixtures of the eigenvectors). However, the models here are of intrinsic interest as they exhibit richer phenomenology and may be more closely related to biological structures. The analyses show that there are parameter regimes in which multiple stable equilibria or limit cycles coexist with the equilibrium at which the networks perform PCA. This situation is analogous to the case of supervised networks with multiple minima of their cost function, and serves as a reminder that local (i.e. linear) stability of a particular equilibrium may not be enough to specify the computation performed by a mature network.

The Single-Neuron Model

In Oja's model [3] the input, $x \in R^N$, is a random vector assumed here to be drawn from a stationary probability density. The vector of weights from the input to the node is denoted ω and the node response y is linear; $y = x \cdot \omega$. The continuous-time, ensemble averaged form of the learning rule is

$$\begin{aligned} \dot{\omega} &= < x\,y > - < y^2 > \omega \\ &= R\omega - (\omega \cdot R\omega)\,\omega \equiv f(\omega) \end{aligned} \tag{1}$$

where $< \ldots >$ denotes the average over the ensemble of inputs, and
$R = < x\, x^T >$ is the input's auto-correlation matrix. The unit-magnitude eigenvectors of R are denoted e_i, $i = 1 \ldots N$ and are assumed to be ordered in decreasing magnitude of the associated eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_N > 0$ (assumed distinct). Equation (1) has equilibria for ω along any of the eigenvectors e_i, but only the equilibria $\omega = \pm e_1$ are stable. The other equilibria are saddle points. In fact Oja shows that the weight vector asymptotically approaches $\pm e_1$. The variance of the node's response is thus maximized and the node acts as a filter for the first principal component of the input.

Extending the Single Neuron Model

To extend the model to a system of $M \leq N$ nodes we consider a set of linear neurons with weight vectors (called the forward weights) $\omega_1 \ldots \omega_M$ connecting each node to the $N-$dimensional input. Without interactions between the nodes in the array, all M weight vectors would converge to $\pm e_1$.

Suppose that the interactions between the cells can be parameterized by some coupling constant (e.g. a learning rate) C. For small values of C, we can treat the coupling as a perturbation and the equations of motion for the entire system take the form

$$\dot{\omega}_i = f_i(\omega_i) + C\, G_i(\omega_1, \ldots, \omega_M, \eta), \quad i = 1, \ldots, M \tag{2}$$

where f_i is a vector field of the same form as in (1) and η are auxiliary state variables that mediate the interactions between the nodes. Additional equations describe the evolution of these auxiliary variables,

$$\dot{\eta}_{ij} = f_{\eta_{ij}}(\omega, \eta, C). \tag{3}$$

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.

We are interested in the equilibria at which the forward weight vectors ω_i are along the M leading eigenvectors of R. We refer to this equilibrium as the PCA equilibrium. Assume that the system (2, 3) has an equilibrium at $X_0 = \{\, \omega_i = e_i, \;\; \eta_{ij} = \eta^0{}_{ij} \,\}$ and that η^0 is continuous in C. At $C = 0$ this equilibrium cannot be stable since the cells with $\omega_i = e_i$ for $i > 1$ will inherit the instabilities of (1).

Under one of two fairly general assumptions the PCA equilibrium will be unstable below some critical value for C. If the linear part of (2, 3) is hyperbolic (no zero or purely imaginary eigenvalues), then the flow near X_0 is structurally stable and the PCA

equilibrium will remain unstable for values of C near zero. [14]. A less restrictive assumption is that at least one of the eigenvalues of the linearization of (2, 3) at X_0 corresponding to the unstable subspace is not repeated. Then this eigenvalue will be continuous in C [15] and again the desired equilibrium will remain unstable for small values of C. Thus under either of these assumptions (hyperbolic fixed point or a non-repeated eigenvalue with positive real part), there is a minimum value for $|C|$ below which the PCA equilibrium is unstable. The two models we consider both exhibit this behavior (see also Rubner et. al. [12]).

We consider two approaches to building interactions that force nodes to filter for different statistical features. In the first approach an internode potential is constructed and the resulting equations of motion are made local by introducing lateral connections that evolve according to an anti-Hebbian rule. For reasons that will become clear, the resulting model is referred to as a minimal coupling scheme. In the second approach the lateral connections are present from the outset, and we directly write equations of motion for the forward weights analogous to (1). The evolution of the lateral connection strengths is given as a simple anti-Hebbian rule.

Minimal Coupling

The response of the i^{th} node in the array is taken to be linear in the input

$$y_i = x \cdot \omega_i. \tag{4}$$

The adaptation of the forward weights was derived from the potential suggested in [16] and modified to give a local learning rule. This requires the introduction of a set of symmetric lateral connections η_{ij}. Details of the derivation appear in [1] and [17] . The learning rules for the system are

$$\begin{aligned} \dot{\omega}_i &= < x y_i > + \sum_{j \neq i}^{M} \eta_{ij} < x y_j > - < y_i^2 > \omega_i \\ &= R\omega_i + \sum_{j \neq i}^{M} \eta_{ij} R\omega_j - (\omega_i \cdot R\omega_i)\, \omega_i. \end{aligned} \tag{5}$$

for the forward weights and

$$\begin{aligned} \dot{\eta}_{ij} &= -d(\eta_{ij} + C < y_i y_j >) \\ &= -d(\eta_{ij} + C\, \omega_i \cdot R\omega_j), \quad \eta_{ii} = 0 \end{aligned} \tag{6}$$

for the lateral connections. Here C and d are free parameters.

Notice that the response of the i^{th} node is given by (4) and is thus independent of the signals carried on the lateral connections. In this sense the lateral signals affect node plasticity but not node response and the coupling is regarded as minimal. This learning rule can also be derived as a low-order approximation to the complete coupling model discussed later.

Stability and Bifurcation

By inspection the weight dynamics given by (5) and (6) have an equilibrium at

$$X_0 \equiv (\omega_i = e_i, \ \eta_{ij} = 0). \tag{7}$$

At this equilibrium the outputs are the first M principal components of input vectors. In suitable coordinates the linear part of the equations of motion break into block diagonal form with any possible instabilities constrained to 3×3 sub-blocks. Details of the stability and bifurcation analyses are given in [17].

The PCA equilibrium X_0 is linearly stable if and only if

$$d > d_0 = \frac{(\lambda_i - \lambda_j)^2 (\lambda_i + \lambda_j)}{\lambda_i^2 + \lambda_j^2} \tag{8}$$

$$C > C_0 = \frac{1}{\lambda_i + \lambda_j}, \quad 1 \le (i,j) \le M. \tag{9}$$

At C_0 or d_0 there is a qualitative change (a bifurcation) in the learning dynamics. At d_0 there is a Hopf bifurcation to oscillating weights. At C_0 there is a bifurcation to multiple equilibria. The bifurcation normal form in the latter case was found by Liapunov-Schmidt reduction [18, for example] performed at the bifurcation point (X_0, C_0).

At (X_0, C_0) there is a supercritical pitchfork bifurcation. Two *unstable* equilibria appear near X_0 for $C > C_0$. At these equilibria the forward weights are mixtures of e_M and e_{M-1} and the lateral connection strengths are non-zero.

The position of *stable* equilibria away from (X_0, C_0) can be found in principal by examining terms of order five and higher in the bifurcation expansion. As a practical alternative we examine the bifurcation from the homogeneous solution, X_h, in which all weight vectors are proportional to e_1 with non-zero η. For a system of two nodes this equilibrium is asymptotically stable provided $d > 0$ and

$$C < C_h \equiv \min \left\{ \begin{array}{c} (\lambda_1 - \lambda_2)/(2\lambda_1\lambda_2) \\ 1/\lambda_1 \end{array} \right\}. \tag{10}$$

If the first condition in (10) corresponds to smaller C_h than the second, then there is a supercritical pitchfork bifurcation at C_h. Two *stable* equilibria emerge from X_h for $C > C_h$. At these stable equilibria, the forward weight vectors are mixtures of the first two correlation eigenvectors and the lateral connection strengths are nonzero.

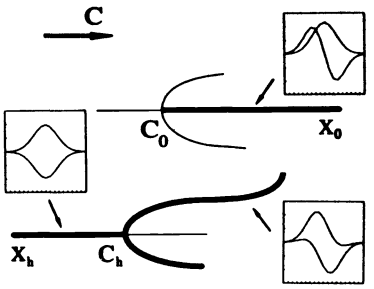

Figure 1: Bifurcation diagram for the minimal model.

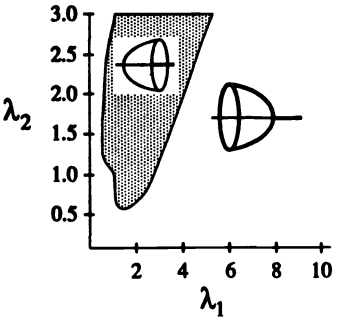

Fig 2: Regions in the (λ_1, λ_2) plane corresponding to supercritical (shaded) and subcritical (unshaded) Hopf bifurcation in the complete model.

The complete bifurcation diagram for a system of two nodes is shown in Fig. 1. The upper portion of the figure shows the bifurcation at (X_0, C_0). The horizontal line represents the PCA equilibrium X_0. This equilibrium is stable (heavy line) for $C > C_0$, and unstable (light line) for $C < C_0$. The subsidiary, unstable, equilibria that emerge from (X_0, C_0) lie on the light, parabolic branches of the top diagram.

The lower portion of Fig. 1 shows the bifurcation from (X_h, C_h) for a system of two nodes. The horizontal line represents the homogeneous equilibrium X_h. This is stable for $C < C_h$ and unstable for $C > C_h$. The stable equilibria consisting of mixtures of the correlation eigenvectors lie on the heavy parabolic branches of the diagram. For networks with more nodes, there are presumably further bifurcations along the supercritical stable branches emerging from (X_h, C_h). Equilibria consisting of varied eigenvector mixtures are observed in simulations.

Each inset in the figure shows equilibrium forward weight vectors for both nodes in a two-node network. These configurations were generated by numerical integration of the equations of motion (5) and (6). We used a correlation matrix corresponding to an ensemble of noise vectors with short-range correlations between the components. Simulations of the corresponding *discrete*, pattern-by-pattern learning rule confirm the

form of the weight vectors shown here.

Full Coupling

In a more conventional coupling scheme, the signals carried on the lateral connections affect the node activities directly. For linear node response, the vector of activities is given by

$$y = (1-\eta)^{-1}\,\omega x \equiv u\,\omega\,x \tag{11}$$

where $y \in R^M$, η is the $M \times M$ matrix of lateral connection strengths and ω is an $M \times N$ matrix whose i^{th} row is the forward weight vector to the i^{th} node. In an analog system, the outputs (11) would be obtained as the equilibrium of the node activation dynamics. We assume that the convergence of the node activation is fast with respect to changes in the weights. The adaptation rules are

$$\dot{\omega} = < yx^T > - Diag(< yy^T >)\,\omega \tag{12}$$

$$\dot{\eta}_{ij} = D\,\eta_{ij} - C < y_i y_j^T >,\ \eta_{ii} = 0, \tag{13}$$

where D and C are constants and $Diag$ sets the off-diagonal elements of its argument equal to zero. The PCA equilibrium, X_0 is linearly stable if

$$D > 0 \tag{14}$$

$$C > C_0 \equiv \frac{D}{\lambda_i + \lambda_j} + \frac{(\lambda_i - \lambda_j)^2}{\lambda_i^2 + \lambda_j^2}. \tag{15}$$

As the parameter D passes through zero, the PCA equilibrium becomes unstable. At $D = 0$ this model reduces to that given by Földiák [13]. The latter model thus operates at a bifurcation point of the system considered here. It is exactly at this point that the linear analysis fails to determine the stability of the PCA equilibrium.

Simulations of two-cell networks with $D = 0$ and initial conditions near the PCA equilibrium show convergence to an equilibrium near, but not identical to X_0. Varying the initial conditions within a neighborhood of X_0, we find different, though nearby equilibria. This suggests that there may be a continuum of stable equilibria near X_0 for $D = 0$.

If, on the other hand, the condition on C is violated, then the network undergoes a Hopf bifurcation leading to oscillations. Depending on the eigenvalue spectrum of the input correlation, this bifurcation may be subcritical (with stable limit cycles near X_0 for $C < C_0$), or supercritical (with unstable limit cycles near X_0 for $C > C_0$). Figure

2 shows the corresponding regions in the (λ_1, λ_2) plane for a network of two nodes with $D = 1$. Simulations show that even in the supercritical regime, stable limit cycles are found for $C < C_0$, and for $C > C_0$ sufficiently close to C_0. This suggests that the complete bifurcation diagram in the super-critical regime is shaped like the bottom of a wine bottle, with only the indentation shown in figure 2.

The model discussed by Rubner [12] also undergoes a Hopf bifurcation for small values of the lateral learning rate. However, the critical rate is bounded above by unity (compare with (15)). In ([17]) we show that the stability criteria in (15) can be ameliorated by allowing the relaxation of the lateral connections to depend on the node activities. This leads to a critical value for C that is bounded above by two. The bifurcation behavior has not been analyzed.

Summary

The primary goal of this study has been to give a theoretical description of learning in feature-discovery models; in particular models that use lateral interactions to ensure that nodes tune to different statistical features. The models presented here have several different limit sets (equilibria and cycles) whose stability and location in the weight space depends on the relative learning rates in the network, and on the eigenvalue spectrum of the input correlation. Bifurcation theory provides a description of the location and determines stability of these different limiting solutions. We expect that these tools can provide insight into similar algorithms.

Acknowledgments

This work was supported by the Office of Naval Research. The author thanks Jenny Orr for carrying out some of the simulations.

References

[1] Todd K. Leen. Dynamics of learning in recurrent feature-discovery networks. In Richard P. Lippmann, John Moody, and David Touretzky, editors, *Advances in Neural Information Processing Systems 3*. Morgan Kauffmann, to appear 1991.

[2] Thomas H. Brown, Edward W. Kairiss, and Claude L. Keenan. Hebbian synapses: Biophysical mechanisms and algorithms. *Annual Review of Neuroscience*, 13:475–511, 1990.

[3] E. Oja. A simplified neuron model as a principal component analyzer. *J. Math. Biology*, 15:267–273, 1982.

[4] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, 22:551–575, 1977.

[5] T. Sanger. An optimality principle for unsupervised learning. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems 1*. Morgan Kauffmann, 1989.

[6] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *J. of Math. Anal. and Appl.*, 106:69–84, 1985.

[7] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[8] Anders Krogh and John A. Hertz. Hebbian learning of principal components. *Unpublished Preprint*, 1990.

[9] Ronald J. Williams. Feature discovery through error-correction learning. Technical Report ICS-8501, Dept. of Cognitive Science, University of California, 1985.

[10] Todd K. Leen, M. Rudnick, and D. Hammerstrom. Hebbian feature discovery improves classifier efficiency. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pages I–51 – I–56, June 1990.

[11] Jeanne Rubner and Klaus Schulten. Development of feature detectors by self-organization: A network model. *Biol. Cyb.*, 62:193–199, 1990.

[12] Jeanne Rubner and Paul Tavan. A self-organizing network for principal component analysis. *Europhysics Lett.*, 20:693–698, 1989.

[13] P. Foldiák. Adaptive network for optimal linear feature extraction. In *Proceedings of the IJCNN*, pages I 401–405, 1989.

[14] M. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, Inc., San Diego, 1974.

[15] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 1983.

[16] A.L. Yuille, D.M. Kammen, and D.S. Cohen. Quadrature and the development of orientation selective cortical cells by Hebb rules. *Biol. Cybern.*, 61:183–194, 1989.

[17] Todd K. Leen. Dynamics of learning in linear feature-discovery networks. *Network : Computation in Neural Systems*, 2:85–105, 1991.

[18] Martin Golubitsky and David Schaeffer. *Singularities and Groups in Bifurcation Theory, Vol. I*. Springer-Verlag, New York, 1984.