SUBMITTED TO NEURAL NETWORKS, APRIL 8, 1994

Feedforward Hebbian Learning with Nonlinear Output Units: A Lyapunov Approach

Todd Troyer

W.M. Keck Center for Integrative Neuroscience
and Department of Pysiology
University of California, San Francisco
San Francisco, CA. 94143
e-mail: todd@phy.ucsf.edu

Abstract

A Lyapunov function is constructed for the unsupervised learning equations of a large class of two layer networks. Units in the output layer are recurrently connected by fixed symmetric weights; only the feedforward connections between layers undergo learning. In contrast to much of the previous work on self-organization of this type, the output units have nonlinear transfer functions. The Lyapunov function is similar in form to that derived by Cohen-Grossberg and Hopfield. Two theorems are proved regarding the location of stable equilibria in the limit of high gain transfer functions. The analysis is applied to the soft competitive learning networks of Amari and Takeuchi.

Keywords:
Lyapunov function, correlational learning, unsupervised learning, competitive learning, saturation, high gain sigmoid, principal component analysis.

Acknowledgements:

The author thanks Morris Hirsch for his inspiration and guidance, and Ken Miller and Daniel Rosen for useful comments on the manuscript. This research was supported in part by the National Science Foundation.

1 Introduction

Many researchers have independently considered unsupervised learning in two layer networks with the same basic architecture and similar learning rules (Amari and Takeuchi, 1978; Intrator, 1992; Linsker, 1988; Malsburg 1973; Miller, 1990; White, 1992; Whitelaw and Cowan, 1981; Xu, 1993). In these networks, the feedforward weights are updated according to an averaged Hebbian learning rule and there is a fixed pattern of connections governing the interaction among output nodes. Most research concerns linear networks where this type of learning has been shown to be closely related to principal component analysis. For reviews see (Hornik and Kuan, 1992; Oja, 1992). Since the dynamics of completely linear systems are either trivial or unbounded, nonlinearities are introduced in the learning equations to bound the growth of the weights.

This article analyzes the "opposite" class of networks in which bounded learning dynamics result from nonlinearities in the output node transfer function; the learning equations are bilinear in the weights and levels of output activity. The chief result is to produce a Lyapunov function[1] for these dynamics. However, the production of a Lyapunov function is only a beginning when analyzing a general class of systems; the existence of such a function proves only that the dynamics is convergent. (See (Hirsch, 1989a) for exact theorems.) Additional insight into the system's behavior must be obtained from either examining the overall structure of the function or analyzing the energy surface for a particular network of interest.

This article takes both approaches. After presenting the general model and introducing notation, we discuss the specific network considered by Amari and Takeuchi in which the recurrent connections take the form of uniform inhibition (Amari and Takeuchi, 1978). When the output node transfer function is a high gain sigmoid, this network performs a form of *soft competitive learning* in which a variable number of units can "turn on" in response to a given input pattern.

In addition to being interesting in its own right, this soft competitive learning network serves as a specific example for illustrating the more general results presented in sections 4–7. The Lyapunov function presented in section 4 is very similar in form to that discovered by Cohen and Grossberg (Cohen and Grossberg, 1983) and Hopfield (Hopfield, 1984) for the activation dynamics of symmetric additive networks; the connection to additive networks is made explicit in section 5. In sections 6 and 7 we prove theorems concerning the location of stable equilibria for the learning in the limit of high gain transfer functions, and in section 8 we apply these results to the Amari-Takeuchi network. The formal proofs of all theorems are reserved for the appendix.

2 Model and Notation

We consider a class of two layer networks in which the output layer is recurrently connected with fixed symmetric weights $T_{ij} = T_{ji}$. The T_{ij} are referred to as *lateral* weights. There

[1] The terms Lyapunov function and energy function will be used interchangeably.

are also *feedforward* weights W_{im} connecting input node m to output node i. Only these feedforward weights undergo learning. Inputs are presented to the network as a fixed pattern of activity x^α across the input layer.[2] For a fixed input vector x^α the following is the system's *activation dynamics*:

$$\dot{y}_i^\alpha = -y_i^\alpha + \sum_j T_{ij} z_j + \sum_m W_{im} x_m^\alpha \ . \tag{1}$$

y_i^α is the activity of the ith output node and $z_i^\alpha = g(y_i^\alpha)$ is its output. We will focus on the case where the squashing function g is a high gain sigmoid, but for now assume only that g is differentiable and non-decreasing.

We would like the dynamics (1) to reliably transform the vector of input activities x^α to a vector z^α of outputs. To ensure that this transformation is well defined, the final state of the recurrent dynamics (1) must depend *only* on the feedforward input and hence be independent of the initial pattern of output activity. So we make the following assumption: for each input pattern x^α, there exists a single globally attracting equilibrium vector y^α, i.e., the activation dynamics are globally asymptotically stable (Hirsch, 1989a). Therefore, each feedforward weight matrix W results in a well-defined vector-valued input/output mapping F_W, where $F_W(x^\alpha) = z^\alpha = g(y^\alpha)$. Note that this mapping F_W is strongly dependent on the choice of lateral connection matrix T and squashing function g.

We make the usual assumption that the weight change is slow compared to the time scale of the activation dynamics and so ignore transients and focus on the equilibrium vectors y^α. The network's *learning dynamics* is obtained by changing the strength of a feedforward weight W_{im} according to the average correlation of activity at input node m and output node i, considered at equilibrium:

$$\begin{aligned} \dot{W}_{im} &= -W_{im} + \sum_\alpha p^\alpha z_i^\alpha x_m^\alpha && (2) \\ z_i^\alpha = g(y_i^\alpha) &= g\left(\sum_j T_{ij} z_j + \sum_m W_{im} x_m^\alpha\right) \ . && (3) \end{aligned}$$

p^α is the probability of presenting pattern x^α and $-W_{im}$ is an exponential decay term added to keep the weights bounded. The goal of this analysis is to determine whether the rule for changing the feedforward weight matrix W results in the network learning some "interesting" computation F_W when presented with a collection of input patterns x^α.

3 An Example: Soft Competitive Learning

One such "interesting" computation is to group the input patterns into clusters or categories. Since the 1960s, many authors[3] have considered this task in a simple model of a category detecting nerve cell. The simplest case consists of a single output cell with N input lines.

[2] Throughout, i and j will index output nodes, m will index input nodes and α and β will index input patterns.

[3] We follow the approach taken by Amari (Amari, 1977; Amari, 1983; Amari and Takeuchi, 1978).

Patterns are presented randomly and the connections undergo Hebbian learning as in (2). What does such a cell learn?

The activation dynamics of this cell is trivial: the output is simply set to the value $z^\alpha = g(\vec{W} \cdot x^\alpha)$. Here $\vec{W}$ is the vector of input weights. If the squashing function for the cell is a high gain sigmoid, i.e., a smooth, invertible approximation to the Heavyside step function, z^α is approximately 0 or 1 depending on whether $W \cdot x^\alpha$ is positive or negative. If $z^\alpha \approx 1$, we say that the cell *detects* pattern x^α. The set of all patterns detected by a cell constitutes the *category* for that cell. Looking at the learning equation for a single pattern,

$$\dot{\vec{W}} = -\vec{W} + z^\alpha x^\alpha \ ,$$

we see that if the cell detects x^α it moves its weight vector $\vec{W}$ toward x^α. Otherwise $\vec{W}$ decays toward 0.[4] It follows that the averaged learning equation results in a final weight vector that points in the direction of the centroid of the patterns it detects.

Now suppose we have a number of cells independently clustering a large data set into categories. The final distribution of categories will depend both on the locations of the patterns and the distribution of initial weight vectors in the input space. Regions with high pattern density will tend to attract more weight vectors, with the possibility that the resulting categories will be highly overlapping or redundant. In order to create a more uniform arrangement of categories, Amari and Takeuchi (Amari and Takeuchi, 1978) proposed a network architecture in which output units uniformly inhibit one another: if a node detects a given input pattern x^α, it sends out a negative signal to all other output nodes making it more difficult to detect x^α. Letting λ denote the strength of the inhibitory connections, we have the following equations:

Activation Dynamics:

$$\dot{y}_i = -y_i - \lambda \sum_j z_j + \sum_m W_{im} x_m^\alpha \ . \tag{4}$$

Learning Dynamics:

$$\dot{W}_{im} = -W_{im} + \sum_\alpha p^\alpha z_i^\alpha x_m^\alpha \tag{5}$$

$$z_i^\alpha = g\left(-\lambda \sum_j z_j + \sum_m W_{im} x_m^\alpha\right) \ . \tag{6}$$

With this example in mind, we return to a consideration of the general network.

4 A Lyapunov Function

The main result of this article is to produce a Lyapunov function $\boldsymbol{L}$ for the learning dynamics (2). The proof and statement of a more general form of the theorem are reserved for the appendix.

[4] A learned threshold is often included for such units, i.e. $z^\alpha = g(W \cdot x^\alpha - \theta)$ and $\dot{\theta} = -\theta + c\sum_\alpha p^\alpha z^\alpha$. The analyses presented herein generalize easily to handle such cases (Troyer, 1991).

Theorem 1 *Assume that the transfer function g is increasing and continuously differentiable and that the activation dynamics (1) has a unique globally attracting, asymptotically stable equilibrium with coordinates y_i^α. Then the following function is a Lyapunov function for the learning dynamics (2), where $z_i^\alpha = g(y_i^\alpha)$ is given by equation (3):*

$$\boldsymbol{L} = \sum_{i,\alpha} \left(-\frac{1}{2} \sum_\beta p^\alpha p^\beta (x^\alpha \cdot x^\beta) z_i^\alpha z_i^\beta - \frac{p^\alpha}{2} \sum_k T_{ij} z_i^\alpha z_j^\alpha + p^\alpha \int^{z_i^\alpha} (g^{-1})'(\xi) d\xi \right) . \tag{7}$$

One important aspect of the function $\boldsymbol{L}$ is that it contains separate terms corresponding to separate components in the network's computation. The first term is quadratic in the outputs $z_i^\alpha = g(y_i^\alpha)$ and depends on the input patterns x^α presented to the network. The second term is also quadratic and depends on the lateral connectivity T within the output layer. The last term depends on the choice of the squashing function g. Thus the structure of the function $\boldsymbol{L}$ leads to a parsing of a problem into its constituent parts.

The other key point concerning the function $\boldsymbol{L}$ is that even though the learning dynamics is defined on the space of feedforward weights W_{im}, $\boldsymbol{L}$ is expressed in terms of the *output coordinates* z_i^α. Changing the weight matrix W results in a change in the input/output mapping F_W. Therefore the dynamics given by (2) and (3) yields trajectories in the space of outputs z_i^α. If we assume g is bounded, the range of g is some open interval I. With L patterns and N output nodes the output space is the open hypercube I^{LN}. Often we will view the dynamics as taking place on this *output space hypercube.*

In fact, we can see the effects of the changing pattern of weights at many different stages in the computation performed by the network. As first suggested by Hirsch (Hirsch, 1989b), we can view the learning dynamics in terms of the total feedforward or *afferent* input a_i^α coming into output node i when the network is presented with input pattern x^α:

$$a_i^\alpha = \sum_m W_{im} x_m^\alpha = W_i \cdot x^\alpha \ .$$

Rewriting equation (2) as a vector equation and taking the inner product with the input pattern vector x^α we have

$$\begin{aligned} \dot{W}_i &= -W_i + \sum_\beta p^\beta x^\beta z_i^\beta \\ \dot{W}_i \cdot x^\alpha &= -W_i \cdot x^\alpha + \sum_\beta p^\beta (x^\alpha \cdot x^\beta) z_i^\beta \\ \dot{a}_i^\alpha &= -a_i^\alpha + \sum_\beta p^\beta (x^\alpha \cdot x^\beta) z_i^\beta \ . \end{aligned} \tag{8}$$

We now have all the ingredients necessary to sketch a proof that $\boldsymbol{L}$ is indeed a Lyapunov function. Viewing the dynamics in output coordinates,

$$\dot{\boldsymbol{L}} = \sum_{i,\alpha} \frac{\partial \boldsymbol{L}}{\partial z_i^\alpha} \dot{z}_i^\alpha \ .$$

But it is easy to show that $\frac{\partial \boldsymbol{L}}{\partial z_i^\alpha} = -p^\alpha \dot{a}_i^\alpha$, so that

$$\dot{\boldsymbol{L}} = -\sum_{i,\alpha} p^\alpha \; \dot{a}_i^\alpha \dot{z}_i^\alpha = -\sum_\alpha p^\alpha \; \dot{a}^\alpha \cdot \dot{z}^\alpha \; .$$

The completion of the proof relies on a nontrivial analysis of the activation dynamics (1). That $\dot{a}^\alpha \cdot \dot{z}^\alpha > 0$ and hence $\dot{\boldsymbol{L}} < 0$ follows from the fact that z^α is an asymptotically stable equilibrium vector for the recurrent dynamics (1) when driven by the total external (feedforward) input a^α. Intuitively this means that a change in the feedforward input vector a^α and the resulting change in the output vector z^α must have similar directions. In particular, if one increases the feedforward input to a single node, the recurrent dynamics will settle to a state in which that unit's output has increased.

5 Relation to Additive Network Equations

The learning dynamics in this two layer network are closely related to the activation dynamics of the standard additive network equations (Cohen and Grossberg, 1983; Hopfield, 1984).[5]

$$\dot{u}^\alpha = -u^\alpha + \sum_\beta S^{\alpha\beta} g(u^\beta) \; . \tag{9}$$

In fact the afferent dynamics (8) is a *generalization* of these additive net equations. As written, (8) and (9) are nearly identical. The chief difference is that in (8) the outputs $z_i^\alpha = g(a_i^\alpha + \sum_j T_{ij} z_j^\alpha)$ are *implicit* functions of the variables a_i^α. But if we set $T_{ij} = 0$ and let $u^\alpha = a_i^\alpha$ and $S^{\alpha\beta} = p^\beta x^\alpha \cdot x^\beta$, (8) reduces to the additive equations (9).

Therefore, the theorems proved in this article can be applied to additive networks and are in some cases inspired by previous work on such networks. However, one must be very careful in that the *interpretation* of the two sets of equations is completely different. The additive network equations describe the changing activity of a collection of nodes connected by a fixed pattern of weights, whereas (8) reflects the change in the the weights due to Hebbian learning in a certain two layer network.

6 Saturation

The existence of a Lyapunov function guarantees that the averaged learning equations (2) will converge to feedforward weight matrix $\hat{W}$.[6] But to answer the question "what has the network learned?" we must construct a network interpretation of the input/output mapping $F_{\hat{W}}$. A class of mappings where such an interpretation is relatively straightforward is one with binary outputs, i.e., every output unit can be characterized as either "on" or "off."

[5](Linsker, 1988) outlines a partial connection between an energy function for Hebbian learning in a certain linear network and the energy function for discrete additive networks discussed in (Hopfield, 1982).

[6]We actually need to show that the equilibria of (2) are isolated (Hirsch, 1989a). This will be true for almost every set of input patterns x^α and lateral connectivity matrix T and will be assumed throughout.

When an output node turns on we will say that node *detects* the pattern presented. Such networks can be seen as separating the input patterns into a collection of categories, where the *category* for each output unit consists of the input vectors that it detects.

Since we would like binary behavior from units with continuous transfer functions, we focus on squashing functions g that are smooth approximations to the Heavyside step function: Let g be strictly increasing and bounded between the values g^+ and g^-. We define $z_i^\alpha = g_\kappa(y_i^\alpha) = g(\kappa y_i^\alpha)$. κ is the *gain* parameter and determines the "steepness" of our approximation, $\frac{\partial z_i^\alpha}{\partial y_i^\alpha} = g'_\kappa = \kappa g'$. We call a node *saturated* if its output $z_i^\alpha = g(y_i^\alpha)$ is near its extreme values g^+ or g^-. For simplicity we set $g^+ = 1$ and $g^- = 0$.

We will say some property of the network holds *after learning* if it holds at a stable equilibrium $\hat{W}$. Hence the outputs *saturate after learning* iff $z_i^\alpha \approx 1$ or 0 at a stable equilibrium for the learning equation (2). Geometrically this means that the equilibrium output lies near some corner of the output space hypercube. We will call such a corner a *stable corner*. In his 1984 paper (Hopfield, 1984), Hopfield proposed a sufficient condition guaranteeing saturation in additive networks.[7] By exploiting the connection outlined in section 5, we can prove the following theorem which generalizes the Hopfield result.

Theorem 2 (Saturation Theorem) [8]

Let g be strictly increasing and bounded. Under the assumptions of the theorem 1, consider the equations

$$\dot{W}_{im} = -W_{im} + \sum_\alpha p^\alpha g_\kappa(y_i^\alpha) x_m^\alpha \tag{10}$$

$$y_i = \sum_j T_{ij} g_\kappa(y_j^\alpha) + \sum_m W_{im} x_m^\alpha \ . \tag{11}$$

Then if $p^\alpha \parallel x^\alpha \parallel^2 + T_{ii} > 0$ for all patterns x^α and for all i, for sufficiently high gain the dynamics saturates after learning. That is $\forall \delta > 0$, $\exists K$ such that if the gain $\kappa > K$ and z_i^α are the coordinates of a stable equilibrium, $\exists$ a corner c of of the closed output space hypercube with $\mid z_i^\alpha - c_i^\alpha \mid < \delta$ for all i and α.

The geometric intuition behind the theorem was first presented in (Hopfield, 1984). Recall that the Lyapunov function $\boldsymbol{L}$ is composed of three terms, the first two are quadratic in the outputs z_i^α, and the last depends on the squashing function g. In the limit of infinite gain is easy to show that this last term serves only to restrict the dynamics to the output space hypercube: it contributes infinite energy at the boundary, but zero energy in the interior. Therefore, to find the local minima of $\boldsymbol{L}$ we need only to focus on its quadratic component which we denote by Q. For saturation we must show that the local minima of Q lie neither in the interior of the hypercube nor in the interior of any of its faces. But this follows from simple eigenvalue arguments if the diagonal elements of the quadratic form Q are strictly negative. Returning to the definition of $\boldsymbol{L}$, we see that the diagonal elements of Q are given by $-(p^\alpha \parallel x^\alpha \parallel^2 + T_{ii})$.

[7] The actual condition proposed in (Hopfield, 1984) is not quite sufficient for saturation and was stated without formal proof.

[8] See also (Hirsch, 1991) for a more general saturation theorem.

7 Ghost Attractor Dynamics

The Saturation Theorem gives a condition guaranteeing that every stable equilibrium is near *some* corner of the output space hypercube. We now focus on *which* corners have a nearby stable equilibrium.

First, rewrite the dynamics (8) in a more general form:

$$\dot{a} = -a + Sz \tag{12}$$

where a and z are vectors indexed by both i and α and z is defined by $z_i^\alpha = g(y_i^\alpha) = g(h_i^\alpha(a))$. Here h_i^α is the function that gives the implicit dependence of y_i^α on the vector a. Notice that if g is a high gain sigmoid, the second term is nearly constant except for those small regions of state space where $h_i^\alpha(a) \approx 0$. Define a *region* $\mathcal{R}$ in the state space of (12) as the inverse image of a single orthant in the space of activations y_i^α, i.e. a set of vectors a such that $h_i^\alpha(a)$ does not change sign for all i and α.

So for each corner of output space c there is a region $\mathcal{R}(c)$ in which $z \approx c$ and the dynamics (12) is well approximated by an exponential decay to some *ghost* attractor $\mathcal{G}(c) = Sc$. The point $\mathcal{G}(c)$ is called a ghost attractor since $\mathcal{G}(c)$ may not lie in the region $\mathcal{R}(c)$. Then as a trajectory crosses the boundary of $\mathcal{R}(c)$, $\mathcal{G}(c)$ "disappears" and is replaced by the ghost attractor of the newly entered region. These dynamics are illustrated in figure 1.

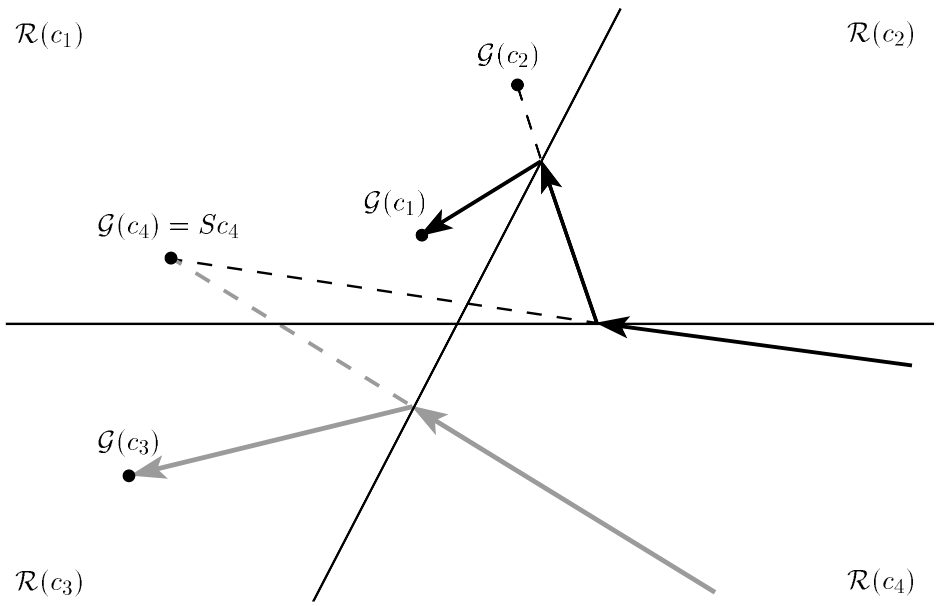

Figure 1. Two orbits for the ghost attractor dynamics (12) with infinite gain $\kappa = \infty$. Both orbits start in region $\mathcal{R}(c_4)$ and hence decay toward the ghost attractor $\mathcal{G}(c_4) = Sc_4$. The gray orbit crosses over into $\mathcal{R}(c_3)$ and approaches the stable equilibrium $\mathcal{G}(c_3) \in \mathcal{R}(c_3)$. The black orbit crosses into $\mathcal{R}(c_2)$ and is attracted to $\mathcal{G}(c_2)$ before arriving at $\mathcal{G}(c_1) \in \mathcal{R}(c_1)$.

From this picture we can see that in the limit of high gain, [9] there is a stable equilibrium near the ghost attractor Sc if and only if Sc lies in the region $\mathcal{R}(c)$. More formally, the following theorem determines which corners of output space are associated with stable equilibria in the case of high gain transfer functions.

Theorem 3 (Ghost Attractor Theorem) *Consider the dynamics (12) and fix a corner c. Suppose that the ghost attractor Sc lies in the interior of some region $\mathcal{R}$. Then for sufficiently high gain κ, there is an attracting output equilibrium vector $\hat{z}$ near c iff $\mathcal{R} = \mathcal{R}(c)$, the region corresponding to c.*

8 Soft Competitive Learning: Analysis

We now return to the competitive learning network of section 3 and apply our results. By theorem 1 we know that the learning procedure described by equations (4) and (5) converges. Since we would like the output nodes to cluster the input patterns into well defined categories, we focus our attention on saturated solutions.

By the saturation theorem we know that given sufficiently high gain, all final states of the system will yield saturated outputs if we require

$$p^\alpha \parallel x^\alpha \parallel^2 > \lambda \ . \tag{13}$$

Define the *significance* of a pattern to be the product of its squared length and its probability of presentation, and recall that λ indicates the strength of the lateral inhibition. Condition (13) says that as long as the significance of all patterns presented to the network is greater than the level of inhibition, the network will parse the inputs into well defined categories. For a given category C let p^C denote the probability of presenting a pattern in C. Then the *centroid* x^C of C is given by $x^C = \frac{1}{p^C} \sum_{\alpha \in C} p^\alpha x^\alpha$. We can use the ghost attractor theorem to search for necessary conditions on solutions that yield saturated outputs.

Theorem 4 *Assume saturation and let C_i be the category detected by cell i after learning. Then under the hypotheses of the Ghost Attractor Theorem, the following is a necessary condition on the collection of categories for sufficiently high gain:*

$$x^\alpha \cdot x^{C_i} > \frac{\lambda Z^\alpha}{p^{C_i}} \tag{14}$$

if $x^\alpha \in C_i$. The opposite inequality must hold if $x^\alpha \notin C_i$. $Z^\alpha = \sum_j z_j^\alpha$ is the total output activity upon presentation of x^α.

Note that saturation implies that for high gain, Z^α is the *number* of cells detecting x^α and hence is the number of categories to which x^α belongs. λZ^α is the total (uniform) inhibitory signal received by all nodes when the network is presented with x^α. So condition

[9] With infinite gain the vector field (12) becomes discontinuous. The properties of such "faulty" vector fields are investigated in (Kidder, 1993).

(14) says that for x^α to belong to a category C_i, the inner product of x^α with the centroid of C_i must be greater than the level of lateral inhibition divided by the probability of presenting some pattern in C_i. The p^{C_i} in the denominator arises from the fact that the weight vector to a particular node i grows only when that node detects some pattern in its category C_i. Hence a node with a larger p^{C_i} will have a larger weight vector and hence will recruit more patterns into its category C_i.

9 Discussion

This article has presented a Lyapunov approach to the unsupervised learning dynamics of a certain class of two layer neural networks. The nonlinearity in the input/output mapping is inherited from recurrent network equations and the correlational learning rule is linear in the weights W_{ij} and outputs z_i^α. The analysis provides a framework for discovering the stable states of such networks and we have used these tools to examine a form of soft competitive learning in detail.

What new insights into the general problem have been gained by this approach?

First, the Lyapunov function $\boldsymbol{L}$ allows one to ascertain the relative importance of the intrinsic connectivity and the external environment for learning. Suppose for simplicity that there are n equally probable input patterns x^α, and let S be the covariance matrix $S^{\alpha\beta} = x^\alpha \cdot x^\beta$. Since $\frac{1}{n}S$ and the lateral connectivity matrix T contribute to $\boldsymbol{L}$ in comparable ways, we can see that if the eigenvalues of $\frac{1}{n}S$ are much larger than the eigenvalues of T, the learning dynamics will be driven chiefly by the environment (constrained by the nonlinearities in the transfer function g). If, however, the eigenvalues of T are much larger than those of $\frac{1}{n}S$, the final states of the system will be largely predetermined and the environment can only affect the learning through such mechanisms as symmetry breaking.

Second, even though the learning dynamics is defined on the space of weight matrices W, at different times it has been useful to think of the dynamics as taking place on the space of afferent inputs a_i^α, the space of activations y_i^α, or on the output space hypercube (z_i^α's). This leads one to view the network computation as a series of coordinate transformations,

$$\begin{array}{ccccccc} W_{im}\text{'s} & \longrightarrow & a_i^\alpha\text{'s} & \longrightarrow & y_i^\alpha\text{'s} & \longrightarrow & z_i^\alpha\text{'s} \\ \multicolumn{3}{r}{\leftarrow \text{Learning Dynamics}} & \Big| & \multicolumn{3}{l}{\text{Activation Dynamics} \rightarrow} \end{array}$$

with the learning dynamics defined to the "left" of the transition from afferent to output activation coordinates, and the activation dynamics to the "right". Given a particular two layer network and correlational learning rule, this suggests that one should determine the "location" of nonlinearity in the problem to determine the most appropriate set of coordinates in which to analyze the dynamics.

References

Amari S. and Takeuchi A. (1978). Mathematical theory on formation of category detecting nerve cells. *Biological Cybernetics*, 29:127–136.

Cohen M. and Grossberg S. (1983). Absolute stability of global pattern formation and parallel memory storage by competetive neural networks. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13:815–826.

Hirsch M.W. (1989a). Convergent activation dynamics in continuous time networks. *Neural Networks*, 2:331–349.

Hirsch M.W. (1989b). On the Amari-Takeuchi theory of category formation. In *Proceedings of the International Joint Conference on Neural Networks in Washington D.C.*, volume 1, pages 297–300.

Hirsch M.W. (1991). Saturated outputs for high-gain, self-exciting networks. *International Journal of Bifurcation and Chaos*, 1:211–217.

Hopfield J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, 79:2554–2558.

Hopfield J.J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science, USA*, 81:3088–3092.

Horn R.A. and Johnson C.R. (1991). *Matrix Analysis.* Cambridge University Press, New York.

Kidder J. (1993). *A Theory of Faulty Dynamics.* PhD thesis, University of California at Berkeley.

Linsker R. (1988). Self-organization in a perceptual network. *Computer*, 21:105–107.

Troyer T. (1991). Dynamics of the Amari-Takeuchi competetive learning model. In *Proceedings of the International Joint Conference on Neural Networks, Seattle.*

A Mathematical Appendix

A.1 General Statement and Proof of Theorem 1

We need a couple of lemmas from linear algebra. Let M_n denote the set of $n \times n$ real matrices and let $H_n \subset M_n$ denote the set of $n \times n$ real symmetric matrices.

Lemma 1 *Let $A, B \in H_n$ with A positive definite. Then the product AB is a diagonalizable matrix with the same number of positive, negative and zero eigenvalues as B.*

Proof: See (Horn and Johnson, 1991), p. 465.

Lemma 2 *Let $C \in H_n$ be positive semi-definite and suppose all the eigenvalues of $D \in M_n$ have positive real part. Then if the product $P = CD$ is symmetric, it is positive semi-definite. Furthermore, if C is positive definite, so is P.*

Proof: By changing coordinates if necessary we can assume that C is a diagonal matrix with the first k entries > 0. Write $CD = P$ in block form:

$$\begin{bmatrix} C_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ 0 & 0 \end{bmatrix}$$

Since P is symmetric, $P_{12} = C_{11}D_{12} = 0$. But C_{11} is positive and diagonal implying $D_{12} = 0$. It follows that the first k components of any eigenvector for D will be an eigenvector for D_{11} with the same eigenvalue. Hence all eigenvalues for D_{11} have positive real part. Writing $C_{11}^{-1}P_{11} = D_{11}$ and applying lemma 1, we have that D_{11} is diagonalizable and P_{11} has all positive eigenvalues. Therefore P is positive semi-definite.

Theorem 1 (General formulation.) *Let $T : \boldsymbol{R}^K \to \boldsymbol{R}^K$ and assume $\hat{g} : \boldsymbol{R} \to \boldsymbol{R}$ is non-decreasing and continuously differentiable. Let $g : \boldsymbol{R}^K \to \boldsymbol{R}^K$ be defined by componentwise application of $\hat{g}$. Suppose that for all $a \in \boldsymbol{R}^K$ the dynamics*

$$\dot{\tilde{y}} = -\tilde{y} + T(g(\tilde{y})\ ) + a = H(a, \tilde{y}) \tag{15}$$

has a unique globally attracting, asymptotically stable equilibrium y. Consider the following differential equation:

$$\dot{a} = -a + C(g(y)\ ) \tag{16}$$

where $C : \boldsymbol{R}^K \to \boldsymbol{R}^K$. Assume that $\exists\ \mathcal{C}^2$ functions $\boldsymbol{C} : \boldsymbol{R}^K \to \boldsymbol{R}$ and $\boldsymbol{T} : \boldsymbol{R}^K \to \boldsymbol{R}$ and a $K \times K$ positive diagonal matrix P such that $\nabla \boldsymbol{C} = PC$ and $\nabla \boldsymbol{T} = PT$. Then the following is a Lyapunov function for (16):

$$\boldsymbol{L} \ = -\boldsymbol{C}(g(y)\ ) - \boldsymbol{T}(g(y)\ ) + \sum_{n=1}^{K} P_{nn} \int^{y_n} \xi g'(\xi)\, d\xi \tag{17}$$

If g is strictly increasing, then $\boldsymbol{L}$ is strict.

Proof: For simplicity we let $z = g(y)$ and let G' be the diagonal matrix with entries $g'(y)$. y is an implicitly defined function of a satisfying

$$H(a, y) = -y + T(g(y)\ ) + a = 0\ .$$

Since y is an asymptotically stable equilibrium for (15), the eigenvalues of the matrix $\left[\frac{\partial H}{\partial y}\right]$ have negative real parts. By the implicit function theorem,

$$\left[\frac{\partial y}{\partial a}\right] = -\left[\frac{\partial H}{\partial y}\right]^{-1}\left[\frac{\partial H}{\partial a}\right] = \left[\frac{\partial(-H)}{\partial y}\right]^{-1}\ .$$

Hence all the eigenvalues of the matrix $\left[\frac{\partial a}{\partial y}\right] = \left[\frac{\partial(-H)}{\partial y}\right]$ have positive real part.

Taking the derivative of $\boldsymbol{L}$ with respect to the vector y and using the chain rule

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial y} &= -\frac{\partial \boldsymbol{C}(g(y)\ )}{\partial y} - \frac{\partial \boldsymbol{T}(g(y)\ )}{\partial y} + PG'y \\
&= -G'P\,(C(z) + T(z) - y) \\
&= -G'P\,(C(z) - a) \\
&= -G'P\dot{a} \\
&= -G'P\left[\frac{\partial a}{\partial y}\right]\dot{y}
\end{aligned}$$

Let $T'(z)$ denote the derivative of T at z. Since $PT = \nabla \boldsymbol{T}$, $PT'(z)$ is the Hessian of $\boldsymbol{T}$ at z and hence can be written as an $M \times M$ symmetric matrix. Because G' and P are both diagonal, so is $G'P$. Therefore

$$G'P\left[\frac{\partial a}{\partial y}\right] = G'P\left[\frac{\partial(-H)}{\partial y}\right] = G'P - G'PT'(z)G'$$

is also a symmetric matrix. It follows from lemma 2 that $G'P\left[\frac{\partial a}{\partial y}\right]$ is positive semi-definite and positive definite if g' is strictly greater than 0. Since

$$\dot{\boldsymbol{L}} = -\dot{y} \cdot \frac{\partial \boldsymbol{L}}{\partial y} = -\dot{y} \cdot \left(G'P\left[\frac{\partial a}{\partial y}\right]\dot{y}\right)$$

we have shown that $\boldsymbol{L}$ is a Lyapunov function and $\boldsymbol{L}$ is strict if g is strictly increasing.

A.2 Proof of the Saturation Theorem

By theorem 1 we know that the function $\boldsymbol{L}$ of equation (7) is a Lyapunov function for the dynamics (2). At a stable equilibrium $\hat{y}$, the gradient of $\frac{\partial \boldsymbol{L}}{\partial y} = 0$ and the matrix of second derivatives of $\boldsymbol{L}$ with respect to y must be positive semi-definite. In particular, the diagonal elements

$$\frac{\partial^2 \boldsymbol{L}}{\partial {y_i^\alpha}^2}(\hat{y}) = p^\alpha g'_\kappa(\hat{y}_i^\alpha)\left(1 - g'_\kappa(\hat{y}_i^\alpha)(p^\alpha \parallel x^\alpha \parallel^2 - T_{ii})\right) \geq 0$$

By our hypotheses $p^\alpha g_\kappa > 0$ and $p^\alpha \parallel x^\alpha \parallel^2 - T_{ii} > 0$. Therefore $g'_\kappa(\hat{y}_i^\alpha) < 1/(p^\alpha \parallel x^\alpha \parallel^2 + T_{ii})$. But since $g'_\kappa(\hat{y}_i^\alpha)$ is bounded, $\kappa \quad g'(\kappa\hat{y}_i^\alpha)$, $g'(\kappa\hat{y}_i^\alpha) \to 0$ as $\kappa \to \infty$. The assumptions on g imply that for any δ, $\exists \epsilon > 0$ such that if $g'(u) < \epsilon$, $z = g(u)$ is within δ of g^+ or g^-. Choosing K large enough to force $g'(\kappa\hat{y}_i^\alpha) < \epsilon$ for all i and α completes the proof.

A.3 Proof of the Ghost Attractor Theorem

Let B^o denote the interior of a set B. For all a in the interior of $\mathcal{R}(c)$, $z(a) \to c$ as $\kappa \to \infty$. Therefore the vector field (12)

$$\dot{a} = -a + Sz \quad \longrightarrow \quad \dot{a} = -a + Sc \tag{18}$$

as $\kappa \to \infty$. If $Sc \in \mathcal{R}(c)^o$ we can choose $K > 0$ and a spherical ball B with $Sc \in B \subset \mathcal{R}(c)^o$ such that if $\kappa > K$ the vector field (12) points inward at all points on the boundary of B. Since we have a Lyapunov function for (12) and have assumed that all equilibria are isolated, there exists an attracting equilibrium in B with corresponding output vector $\hat{z}$ near c.

If $Sc \notin \mathcal{R}(c)^o$ it is easy to see from (18) that $\forall y \in \mathcal{R}(c)^o \; \exists \; i$, α and K such that if the gain $\kappa > K$ then $\dot{a}_i^\alpha \neq 0$. It follows that there is no output equilibrium $\hat{z}$ near c.

A.4 Proof of Theorem 4

We apply the Ghost Attractor Theorem. Fix a corner c, i.e. $c_i^\alpha = 0$ or 1. From equation (8) we have that the ghost attractor $\mathcal{G}(c)$ is given by

$$\mathcal{G}(c)_i^\alpha = \sum_\beta p^\beta x^\alpha \cdot x^\beta c_i^\beta \ . \tag{19}$$

The region $\mathcal{R}(c)$ corresponding to c consists of all afferent input vectors a such that

$$g(a_i^\alpha - \lambda \sum_j z_j^\alpha) = c_i^\alpha \ , \tag{20}$$

where g the Heavyside step function. Substituting (19) into (20) we see that for $\mathcal{G}(c)$ to lie in the region $\mathcal{R}(c)$ we must have

$$\sum_\beta p^\beta x^\alpha \cdot x^\beta c_i^\beta - \lambda \sum_k z_k^\alpha > 0 \tag{21}$$

if $c_i^\alpha = 1$ and the opposite inequality if $c_i^\alpha = 0$. Let C_i be the category detected by node i. By definition $c_i^\beta = 1$ iff $x^\beta \in C_i$. Recall also that we have defined $Z^\alpha = \sum_k z_k^\alpha$. It follows that if $x^\alpha \in C_i$, then

$$\begin{aligned} \sum_{\beta \in C_i} p^\beta x^\alpha \cdot x^\beta &> \lambda Z^\alpha \\ x^\alpha \cdot \sum_{\beta \in C_i} \frac{p^\beta}{P^{C_i}} x^\beta &> \frac{\lambda Z^\alpha}{P^{C_i}} \\ x^\alpha \cdot x^{C_i} &> \frac{\lambda Z^\alpha}{P^{C_i}} \ . \end{aligned}$$

The opposite inequality must hold for $x^\alpha \notin C_i$.