CONTRIBUTED ARTICLE

Feedforward Hebbian Learning with Nonlinear Output Units: a Lyapunov Approach

TODD W. TROYER

University of California

(Received 11 May 1994; revised and accepted 10 April 1995)

Abstract—A Lyapunov function is constructed for the unsupervised learning equations of a large class of neural networks. These networks have a single layer of adjustable connections; units in the output layer are recurrently connected with fixed symmetric weights. The constructed 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 functions, Correlational learning, Unsupervised learning, Competitive learning, Saturation, High gain sigmoid, Principal component analysis.

1. INTRODUCTION

Many researchers have independently considered unsupervised learning in a large class of neural networks that share the same basic architecture and similar learning rules (Malsburg, 1973; Amari & Takeuchi, 1978; Whitclaw & Cowan, 1981; Miller, 1990; Intrator, 1992; White, 1992; Xu, 1993). In these networks, a single layer of feedforward weights are updated according to an averaged Hebbian learning rule; there is a fixed pattern of lateral connections governing the interaction among output nodes.1 Most research has focussed on linear networks where Hebbian 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 (Karhunen & Joutsensalo, 1994; Plumbley, 1995).

This paper 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 function2 for these dynamics. The production of a Lyapunov function is only a beginning when analyzing a general class of systems, however, since the existence of such a function proves only that the dynamics is convergent (Hirsch, 1989a). 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 paper takes both approaches. After presenting the general model and introducing notation, we discuss the specific network considered by Amari and Takeuchi in which the lateral connections take the form of uniform inhibition (Amari & 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

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.

Requests for reprints should be sent to Todd W. Troyer, W.M. Keck Center for Integrative Neuroscience, University of California, San Francisco, 513 Parnassus Ave. Box 0444, San Francisco, CA 94143, USA.

1 Földiák (1990), Leen (1991), Matsuoka and Kawamoto (1994) consider similar networks with learned lateral connections.

2 The terms Lyapunov function and energy function will be used interchangeably.
to that discovered by Cohen and Grossberg (1983) and 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. A preliminary account of some of these results has been reported previously (Troyer, 1991).

2. MODEL AND NOTATION

We consider a class of networks with two layers of nodes 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 are also feedforward weights \( W_{m} \) 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 \( x^{\alpha} \) across the input layer.4 For a fixed input vector \( x^{\alpha} \) the following is the system's activation dynamics:

\[
\dot{y}_{i}^{\alpha} = -y_{i}^{\alpha} + \sum_{j} T_{ij} x_{j}^{\alpha} + \sum_{m} W_{im} x_{m}^{\alpha},
\]

(1)

\( y_{i}^{\alpha} \) is the activity of the \( i \)th 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 \( x^{\alpha} \) to a vector \( z^{\alpha} \) 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 of independent of the initial pattern of output activity. So we make the following assumption: for each input pattern \( x^{\alpha} \), there exists a single globally attracting equilibrium vector \( y^{\alpha} \), 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 \( y^{\alpha} \). 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:

\[
\dot{W}_{im} = -W_{im} + \sum_{\alpha} p^{\alpha} z_{i}^{\alpha} x_{m}^{\alpha}
\]

(2)

\[ z_{i}^{\alpha} = g(y_{i}^{\alpha}) = g \left( \sum_{j} T_{ij} x_{j}^{\alpha} + \sum_{m} W_{im} x_{m}^{\alpha} \right). \]

(3)

\( p^{\alpha} \) is the probability of presenting pattern \( x^{\alpha} \) 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 \( x^{\alpha} \).

3. AN EXAMPLE: SOFT COMPETITIVE LEARNING

One such "interesting" computation is to group the input patterns into clusters or categories. Since the 1960s, many authors5 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. 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 \( z^{\alpha} = g(\tilde{W} \cdot x^{\alpha}) \). Here \( \tilde{W} \) 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 Heaviside step function, \( x^{\alpha} \) is approximately 0 or 1 depending on whether \( \tilde{W} \cdot x^{\alpha} \) is positive or negative. If \( x^{\alpha} \approx 1 \), we say that the cell detects pattern \( x^{\alpha} \). The set of all patterns detected by a cell constitutes the category for that cell. Looking at the learning equation for a single pattern,

\[
\dot{\tilde{W}} = -\tilde{W} + z^{\alpha} x^{\alpha},
\]

we see that if the cell detects \( x^{\alpha} \) it moves its weight vector \( \tilde{W} \) toward \( x^{\alpha} \). Otherwise \( \tilde{W} \) decays toward 0.5

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

4 We follow the approach taken by Amari (1977, 1983), Amari and Takeuchi (1978).

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

---

Throughout, \( i \) and \( j \) will index output nodes, \( m \) will index input nodes and \( \alpha \) and \( \beta \) will index input patterns.
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 (Masuda, 1993). In order to create a more uniform arrangement of categories, Amari and Takeuchi (1978) proposed a network architecture in which output units uniformly inhibit one another; if a node detects a given input pattern \( x^\alpha \), it sends out a negative signal to all output nodes (including itself) making it more difficult to detect \( x^\alpha \). Letting \( \lambda \) denote the strength of the inhibitory connections, we have the following equations:

**Activation dynamics:**

\[
y_i^\alpha = y_i^\alpha - \lambda \sum_j y_j^\alpha + \sum_m W_{im} x_m^\alpha. \tag{4}
\]

**Learning dynamics:**

\[
W_{im} = W_{im} + \sum_\alpha \rho^\alpha z_i^\alpha x_m^\alpha \tag{5}
\]

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

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

### 4. A LYAPUNOV FUNCTION

The main result of this paper is to produce a Lyapunov function \( L \) for the learning dynamics (2). The proof and statement of a more general form of the theorem are reserved for the Appendix.

**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 \( y_i^\alpha \). Then the following function is a Lyapunov function for the learning dynamics (2), where \( z_i^\alpha = g(y_i^\alpha) \) is given by eqn (3):

\[
L = \sum_{i, \alpha} \left( -\frac{1}{2} \sum_\beta \rho^\beta (x^\alpha \cdot x^\beta) z_{i}^\beta z_i^\alpha \right) - \frac{\rho^\alpha}{2} \sum_j T_{ij} z_i^\alpha z_j^\alpha + \rho^\alpha \int g^{-1}(\xi) d\xi. \tag{7}
\]

One important aspect of the function \( L \) 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 \( x^\alpha \) 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 \( L \) leads to a parsing of a problem into its constituent parts.

The other key point concerning the function \( L \) is that even though the learning dynamics is defined on the space of feedforward weights \( W_{im} \), \( L \) is expressed in terms of the **output coordinates** \( z_i^\alpha \). 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 \( z_i^\alpha \). 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 \( I_{L,N} \). Often we will view the dynamics as taking place in 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 suggested by Hirsch (1989b), we can view the learning dynamics in terms of the total feedforward or **afferent input** \( a_i^\alpha \) coming into output node \( i \) when the network is presented with input pattern \( x^\alpha \):

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

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

\[
\dot{W}_i = -W_i + \sum_\beta \rho^\beta x_i^\alpha z_i^\beta
\]

\[
W_i \cdot x^\alpha = -W_i \cdot x^\alpha + \sum_\beta \rho^\beta (x^\alpha \cdot x^\beta) z_i^\beta
\]

\[
\dot{a}_i^\alpha = -a_i^\alpha + \sum_\beta \rho^\beta (x^\alpha \cdot x^\beta) z_i^\beta. \tag{8}
\]

We now have all the ingredients necessary to sketch a proof that \( L \) is indeed a Lyapunov function. Viewing the dynamics in output coordinates, we have

\[
L = \sum_{i, \alpha} \frac{\partial L}{\partial z_i^\alpha} z_i^\alpha.
\]

But it is easy to show that \( \partial L / \partial z_i^\alpha = -p^\alpha \dot{z}_i^\alpha \), so that

\[
L = -\sum_{i, \alpha} p^\alpha \dot{z}_i^\alpha = -\sum_\alpha p^\alpha \dot{\bar{x}}^\alpha.
\]

The completion of the proof relies on a nontrivial analysis of the activation dynamics (1). That \( \dot{\bar{x}}^\alpha \cdot \bar{x}^\alpha > 0 \) and hence \( L < 0 \) follows from the fact that \( \bar{x}^\alpha \) is an asymptotically stable equilibrium vector.
for the recurrent dynamics (1) when driven by the total external (feedforward) input \( a^a \). Intuitively this means that a change in the feedforward input vector \( a^a \) and the resulting change in the output vector \( z^a \) 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 network are closely related to the activation dynamics of the standard additive network equations (Cohen & Grossberg, 1983; Hopfield, 1984).\textsuperscript{6}

\[
\dot{a}^a = -a^a + \sum_b S^{ab} g(u^b). \tag{9}
\]

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 \( x^a_i = g(a^a_i + \sum_j T_{ij} z^a_j) \) are implicit functions of the variables \( a^a_i \). But if we set \( T_{ij} = 0 \) and let \( a^a = a^a \) and \( S^{ab} = p^{ab} x^a \cdot x^b \), (8) reduces to the additive equations (9).

Therefore, the theorems proved in this paper can be applied to additive networks and are in some cases inspired by previous work on such networks. However, one must be 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 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 some feedforward weight matrix which we denote \( W^r \).\textsuperscript{7} But to answer the question “what has the network learned?” we must construct a network interpretation of the input/output mapping \( F_{y^a} \). 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”. 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 to which it responds.

Since we would like binary behavior from units with continuous transfer functions, we focus on squashing functions \( g \) that are smooth approximations to the Heaviside step function: let \( g \) be strictly increasing and bounded between the values \( g^+ \) and \( g^- \). We define \( z^a_i = g_r(y^a_i) = g(\kappa y^a_i) \); \( \kappa \) is the gain parameter and determines the “steepness” of our approximation,

\[
\frac{\partial z^a_i}{\partial y^a_i} = g' = \kappa g'.
\]

We call a node saturated if its output \( z^a_i = g(y^a_i) \) 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 \( W^r \). Hence the outputs saturate after learning if \( z^a_i \approx 1 \) 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 proposed a sufficient condition guaranteeing saturation in additive networks.\textsuperscript{8} By exploiting the connection outlined in Section 5, we can prove the following theorem which generalizes the Hopfield result.

**Theorem 2** (Saturation Theorem). Let \( g \) be strictly increasing and bounded. Under the assumptions of Theorem 1, consider the equations

\[
W_{im} = -W_{im} + \sum_a p^{a_k} g_k(y^a_i) x^a_m \tag{10}
\]

\[
y^a_i = \sum_j T_{ij} g_k(y^a_j) + \sum_m W_{im} x_m^a. \tag{11}
\]

Then if \( p^{a_k} x^a \cdot x^a \geq T_{ii} > 0 \) for all patterns \( x^a \) 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 \( z^a_i \) are the coordinates of a stable equilibrium, \( \exists \kappa \) a corner \( c \) of the closed output space hypercube with \( |z^a_i - c^a_i| < \delta \) for all \( i \) and \( a \).

The geometric intuition behind the theorem was first presented in Hopfield (1984). Recall that the Lyapunov function \( L \) is composed of three terms, the first two are quadratic in the outputs \( z^a_i \), and the last

---

\textsuperscript{6} Liniker (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).

\textsuperscript{7} 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 \( x^a \) and lateral connectivity matrix \( T \) and will be assumed throughout.

\textsuperscript{8} The actual condition proposed in Hopfield (1984) is not quite sufficient for saturation and was stated without formal proof.

\textsuperscript{9} See also Hirsch (1991) for a more general saturation theorem.
depends on the squashing function \( g \). In the limit of infinite gain it is easy to show that this last term serves only to restrict the dynamics to the interior of the output space hypercube: it contributes infinite energy at the boundary, but zero energy in the interior. Therefore, to find the local minima of \( L \) 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 \( L \), we see that the diagonal elements of \( Q \) are given by \(- \sum \alpha a^\alpha \).

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:

\[
\dot{a} = -a + Sz
\]

(12)

where \( a \) and \( z \) are vectors indexed by both \( i \) and \( \alpha \) and \( z \) is defined by \( z^\alpha = g(y^\alpha) = g(h^\alpha(a)) \). Here \( h^\alpha \) is the function that gives the implicit dependence of \( y^\alpha \) 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 \( h^\alpha(a) \approx 0 \). Define a region \( R \) in the state space of (12) as the inverse image of a single orthant in the space of activations \( y^\alpha \), i.e., a set of vectors \( a \) such that \( h^\alpha(a) \) does not change sign for all \( i \) and \( \alpha \).

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

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

\[ R(c_1) \quad g(c_2) \]
\[ R(c_3) \quad g(c_4) = Sc \]

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

**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 \( R \). Then for sufficiently high gain \( \kappa \), there is an attracting output equilibrium vector \( z \) near \( c \) iff \( R = R(c) \), 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 eqns (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

\[
\sum a^\alpha ||x^\alpha||^2 > \lambda.
\]

(13)

Define the significance of a pattern to be the product of its squared length and its probability of presentation, and recall that \( \lambda \) 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.

---

10 With infinite gain the vector field (12) becomes discontinuous. The properties of such vector fields are investigated in Lewis and Glass (1992) and Kidder (1993).
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:

\[
x^\alpha \cdot x^\beta > \frac{\lambda Z^\alpha}{p^C_i}
\]

if \( x^\alpha \in C_i \). The opposite inequality must hold if \( x^\alpha \notin C_i \), \( Z^\alpha = \sum_j x^\alpha_j \) is the total output activity upon presentation of \( x^\alpha \).

Note that saturation implies that for high gain, \( Z^\alpha \) is the number of cells detecting \( x^\alpha \) and hence is the number of categories to which \( x^\alpha \) belongs; \( \lambda Z^\alpha \) is the total (uniform) inhibitory signal received by all nodes when the network is presented with \( x^\alpha \). So condition (14) says that for \( x^\alpha \) to belong to a category \( C_i \), the inner product of \( x^\alpha \) 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 paper has presented a Lyapunov approach to the unsupervised learning dynamics of a certain class of 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 \( z_i^\alpha \). 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 \( L \) 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 \( x^\alpha \), and let \( S \) be the covariance matrix \( S^\alpha = x^\alpha \cdot x^\beta \). Since \( \frac{1}{n} S \) and the lateral connectivity matrix \( T \) contribute to \( L \) 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 \( d^\alpha_i \), the space of activations \( y^\alpha_i \), or on the output space hypercube \( (z^\alpha_i) \). This leads one to view the network computation as a series of coordinate transformations,

\[
W_{im} \rightarrow d^\alpha_i \rightarrow y^\alpha_i \rightarrow z^\alpha_i
\]

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 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.

NOMENCLATURE

- \( \alpha, \beta \) index for input patterns
- \( d^\alpha_i \) afferent input from \( \alpha \)th pattern to \( i \)th output node
- \( c \) corner of the output space hypercube
- \( C_i \) category detected by cell \( i \)
- \( F_W \) input output mapping:

\[
F_W(x^\alpha) = g(y^\alpha_i) = z_i^\alpha
\]

- \( g \) transfer or squashing function: \( z_i^\alpha = g(y_i^\alpha) \)
- \( \mathcal{S}(c) \) ghost attractor corresponding to corner \( c \)
- \( i, j \) index for output nodes
- \( \kappa \) transfer function gain parameter
- \( L \) Lyapunov function
- \( \lambda \) strength of lateral inhibition in Amari–Takeuchi network
- \( m \) index for input nodes
- \( M \) covariance matrix: \( M^{\alpha \beta} = x^\alpha \cdot x^\beta \)
- \( p^\alpha \) probability of presenting pattern \( \alpha \)
- \( p^C \) probability of presenting some pattern in category \( C \)
- \( \mathcal{B}(c) \) region corresponding to corner \( c \)
- \( S \) generic weight matrix
- \( T_{ij} \) lateral weight from node \( j \) to node \( i \)
- \( W \) feedforward weight matrix at equilibrium
- \( W_{im} \) feedforward weight from node \( m \) to node \( i \)
- \( x^\alpha_m \) output of \( m \)th input node for \( \alpha \)th pattern
- \( W_{xc} \) vectors with components \( W_{im}, x_m^\alpha \)
- \( x^C \) centroid of category \( C \)
- \( y^\alpha_i \) activity of \( i \)th output node for \( \alpha \)th input pattern
- \( z_i^\alpha \) output of \( i \)th output node for \( \alpha \)th input pattern
- \( y^\alpha, z^\alpha, d^\alpha \) vectors with components \( y^\alpha_i, z^\alpha_i, d^\alpha_i \)
- \( Z^\alpha \) total output activity for \( \alpha \)th input pattern
REFERENCES


APPENDIX

A1. 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 $B$.


**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:

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

Since $P$ is symmetric, $P_{22} = C_{11}D_{11} = 0$. But $C_{11}$ is positive and diagonal implying $D_{11} > 0$. 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 a positive real part. Writing $C_{11}P_{11} = D_{11}$ 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 $T : R^k \rightarrow R^k$ and assume $\hat{g} : R \rightarrow R$ is nondecreasing and continuously differentiable. Let $g : R^k \rightarrow R^k$ be defined by componentwise application of $\hat{g}$. Suppose that for all $a \in R^k$ the dynamics

$$
\dot{y} = \hat{y} + T(\hat{y}) + a = H(a, y)
$$

(A1)

has a uniquely globally attracting, asymptotically stable equilibrium $y$.

Consider the following differential equation:

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

(A2)

where $C : R^k \rightarrow R^k$. Assume that $\exists \xi^2$ functions $C : R^k \rightarrow R$ and $T : R^k \rightarrow R$ and $a$ a $k \times k$ positive diagonal matrix $P$ such that $\nabla C = PC$ and $\nabla T$ = PT. Then the following is a Lyapunov function for (A2):

$$
L = -C(g(y)) - T(\hat{y}) + \sum_{x=1}^{k} \int_{\xi}^{\hat{\xi}} \hat{g}(\xi) d\xi.
$$

(A3)

If $g$ is strictly increasing, then $L$ 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

$$
H(a, y) = -y + T(\hat{y}) + a = 0.
$$

Since $y$ is an asymptotically stable equilibrium for (A1), the eigenvalues of the matrix

$$
\frac{\partial H}{\partial y}
$$
have negative real parts. By the implicit function theorem,
\[
\left[ \frac{\partial y}{\partial a} \right] = -\left[ \frac{\partial f}{\partial y} \right]^{-1} \frac{\partial f}{\partial a} = -\left[ \frac{\partial (\mathbf{H} y)}{\partial y} \right]^{-1} \frac{\partial (\mathbf{H} y)}{\partial a} .
\]

Hence all the eigenvalues of the matrix \[ \left[ \frac{\partial y}{\partial a} \right] = \left[ \frac{\partial \mathbf{H} y}{\partial a} \right] \] have positive real parts.

Taking the derivative of \( \mathbf{L} \) with respect to the vector \( y \) and using the chain rule
\[
\frac{\partial \mathbf{L}}{\partial y} = -\frac{\partial \mathbf{C}(g(y))}{\partial y} - \frac{\partial \mathbf{T}(g(y))}{\partial y} + \mathbf{P} \mathbf{g}' y
\]
\[
= -\mathbf{G}' \mathbf{P} (\mathbf{C}(z) + \mathbf{T}(z) - y)
\]
\[
= -\mathbf{G}' \mathbf{P} (\mathbf{C}(z) - \mathbf{a})
\]
\[
= -\mathbf{G}' \mathbf{P} \left( \frac{\partial \mathbf{a}}{\partial y} \right) \mathbf{y}.
\]

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

Therefore
\[
\mathbf{G}' \mathbf{P} \left( \frac{\partial \mathbf{a}}{\partial y} \right) = \mathbf{G}' \mathbf{P} \left( \frac{\partial (\mathbf{H} y)}{\partial y} \right) = \mathbf{G}' \mathbf{P} = \mathbf{G}' \mathbf{PT}'(z) \mathbf{G}'
\]
is also a symmetric matrix. It follows from lemma 2 that
\[
\mathbf{G}' \mathbf{P} \left( \frac{\partial \mathbf{a}}{\partial y} \right)
\]
is positive semi-definite and positive definite if \( \mathbf{g}' \) is strictly greater than 0. Since
\[
-\mathbf{G}' \mathbf{P} \left( \frac{\partial \mathbf{a}}{\partial y} \right) = -\mathbf{y}
\]
\[
\mathbf{L} = -\mathbf{y}
\]
we have shown that \( \mathbf{L} \) is a Lyapunov function and \( \mathbf{L} \) is strict if \( \mathbf{g} \) is strictly increasing.

A2. Proof of the Saturation Theorem

By Theorem 1 we know that the function \( \mathbf{L} \) of eqn (7) is a Lyapunov function for the dynamics (2). At a stable equilibrium \( \mathbf{y} \), the gradient of \( \partial \mathbf{L} / \partial y = 0 \) and the matrix of second derivatives of \( \mathbf{L} \) with respect to \( y \) must be positive semi-definite. In particular, the diagonal elements
\[
\frac{\partial^2 \mathbf{L}}{\partial y^2}(\mathbf{y}) = p^* g'_c(\mathbf{y}) (1 - g_c(\mathbf{y}) p^* \| \mathbf{x}^* \|^2 - T_i) \geq 0.
\]

By our hypotheses \( p^* g'_c > 0 \) and \( p^* \| \mathbf{x}^* \|^2 - T_i > 0 \). Therefore
\[
g'_c(\mathbf{y}) < 1/(p^* \| \mathbf{x}^* \|^2 - T_i). \]
But since \( g'_c(\mathbf{y}) \) is bounded, \( \kappa g'_c(\mathbf{y}) \), \( g'_c(\mathbf{y}) \to 0 \) as \( \kappa \to \infty \). The assumptions on \( g \) imply that for any \( \delta > 0 \) such that if \( g'(u) < \epsilon \), \( z = g(u) \) is within \( \delta \) of \( g^+ \) or \( g^- \). Choosing \( \kappa \) large enough to force \( g'(\kappa y') < \epsilon \) for all \( i \) and \( \alpha \) completes the proof.

A3. Proof of the Ghost Attractor Theorem

Let \( B' \) denote the interior of a set \( B \). For all \( a \) in the interior of \( \mathcal{A}(c) \), \( z(a) \to c \) as \( \kappa \to \infty \). Therefore the vector field (12)
\[
\dot{a} = -a + S c \quad \dot{a} = -a + S c \quad \text{(A4)}
\]
as \( \kappa \to \infty \). If \( \mathcal{S}_c \in \mathcal{A}(F(1 > (c)^c) \) we can choose \( K > 0 \) and a spherical ball \( B \) with \( \mathcal{S}_c \in B \subset \mathcal{A}(c)^c \) 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 \( z \) near \( c \).

If \( \mathcal{S}_c \notin \mathcal{A}(c)^c \) it is easy to see from (A4) that \( \forall y \in \mathcal{A}(c)^c \exists t, \alpha \) and \( \kappa \) such that if the gain \( \kappa > K \) then \( \dot{a} \neq 0 \). It follows that there is no output equilibrium \( z \) near \( c \).

A4. Proof of Theorem 4

We apply the Ghost Attractor Theorem. Fix a corner \( c \), i.e., \( c_i' = 0 \) or 1. From eqn (8) we have that the ghost attractor \( \mathcal{A}(c) \) is given by
\[
\mathcal{A}(c)_c = \sum_{j \neq c} p^j x^j \cdot x^j = \mathcal{A}(c)_c
\]
The region \( \mathcal{A}(c) \) corresponding to \( c \) consists of all afferent input vectors \( a \) such that
\[
g \left( a_i - \lambda \sum_i c_i' \right) = c_i',
\]
where \( g \) is the Heaviside step function. Substituting (A5) into (A6) we see that for \( \mathcal{A}(c) \) to lie in the region \( \mathcal{A}(c) \) we must have
\[
\sum_{j \neq c} p^j x^j \cdot x^j - \lambda \sum_i c_i' > 0
\]
if \( c_i' = 1 \) and the opposite inequality if \( c_i' = 0 \). Let \( C_i \) be the category detected by node \( i \). By definition \( c_i' = 1 \) if \( x^i \in C_i \). Recall also that we have defined \( Z^j = \sum_j z^j \). It follows that if \( x^i \in C_i \), then
\[
\sum_{j \in C_i} p^j x^j \cdot x^j > \lambda Z^j
\]
\[
x^i \cdot x^i > \lambda Z^j.
\]
The opposite inequality must hold for \( x^i \notin C_i \).