

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
UNIVERSITY OF NEW MEXICO

Gradient-Like Dynamics in Neural Networks

James W. Howse, Chaouki T. Abdallah, and Gregory L. Heileman
Department of Electrical and Computer Engineering
University of New Mexico, Albuquerque, NM 87131

UNM Technical Report No. EECE93-001

Initial Draft Completed: August 10, 1992
First Major Revision Completed: December 21, 1992
Second Major Revision Completed: April 27, 1993
Current Date: May 14, 1993

Abstract

This report presents a formalism that enables the dynamics of a broad class of neural networks to be understood. A number of previous works have analyzed the Lyapunov stability of neural network models. This type of analysis shows that the excursion of the solutions from a stable point is bounded. The purpose of this work is to present a model of the dynamics that also describes the phase space behavior as well as the structural stability of the system. This is achieved by writing the general equations of the neural network dynamics as a gradient-like system. In this paper some important properties of gradient-like systems are developed and then it is demonstrated that a broad class of neural network models are expressible in this form. [†]

[†]Acknowledgments: This research was supported by a grant from Boeing Computer Services under Contract W-300445.

Chapter 1

Introduction

In this paper we propose a formalism that allows three critical issues in the study of unsupervised neural networks to be analyzed. The first important issue is Lyapunov stability. It is important to establish conditions which guarantee that the node activities and connection weights converge to some equilibrium state of the network. The second important issue is the way in which the network stores information. This involves determining the nature of the equilibrium states in the network. The third important issue is the structural stability. This property determines whether a model can be made into a similarly functioning device, or whether the model can be simulated at a different level of precision (e.g. 8-bit vs. 16-bit). In order to do this, it is important to have some guarantee that small changes in the network parameters do not affect its general behavior.

Addressing all three of these concerns in a general neural network model can be quite difficult. In [3] the first of these problems is addressed by proving that a class of networks with a general equation for the node activation dynamics is Lyapunov stable when the weights are constant and symmetric. As shown in [9] many neural network models that do *not* include learning can be put in this general form. In [14], the aforementioned work is extended by using a similar equation for the node activation dynamics to prove the Lyapunov stability of networks with a number of different weight update rules. A different approach, which addresses all three of the issues discussed above, is taken in [21]. Specifically, some properties of a class of dynamical systems called gradient-like systems are derived and then used to explain some of the dynamics of the Hopfield network. This paper extends the results in [21] by proving additional properties of gradient-like systems as well as allowing the incorporation of weight update in the gradient-like system formulation.

Gradient systems are a mathematically well studied class of dynamical systems. For such

systems, results have been derived to address all three of the above concerns. We show in this paper that most of the desirable properties of gradient systems are possessed by the more general class of gradient-like systems. We also demonstrate that many existing neural network models can be formulated as gradient-like systems. By contrast, few neural networks can be written as gradient systems. This formalism allows any dynamical system which can be cast as a gradient-like system to be analyzed with respect to its Lyapunov stability, phase space behavior, and structural stability.

Lyapunov stability is used to determine whether most trajectories move toward or away from a given equilibrium. If an equilibrium state is Lyapunov stable, then any trajectory started in a given neighborhood of the equilibrium must have a bounded excursion from that equilibrium. Phase space behavior on the other hand, is used to evaluate the specific structure of the equilibria. The phase space is the space consisting of all state variables, and the collection of paths that the system state traverses in this space is called the phase space behavior of the system. It shows, for instance, whether the equilibrium state is a point, a periodic cycle, or some more complex behavior. Finally, structural stability is used to demonstrate whether small changes in system parameters change the qualitative system behavior. For example, in a structurally stable system the position of the equilibrium states in phase space remain similar under small variations of the system parameters.

Chapter 2

Mathematical Formalism

Gradient systems are a very well studied class of dynamic systems. The central idea of this section is to formulate a generalization of the gradient system. This generalization, which is called a gradient-like system, will be shown to possess most of the properties of a gradient system. In the first part of this section the properties of gradient systems will be reviewed. The intuitive behavior of such systems will be examined in terms of these properties. In the second part of the section, the properties of gradient-like systems will be presented and proved.

2.1 Properties of Gradient Systems

A gradient system is one in which the time derivative of the states $\dot{\boldsymbol{x}}$ depends on the gradient of a scalar function $V(\boldsymbol{x})$. Intuitively, the behavior of gradient systems can be understood by realizing that the states of the system may travel in only two ways. They may move *downward* along the surface of $V(\boldsymbol{x})$ following the line of steepest descent, or they may remain *constant*. The function $V(\boldsymbol{x})$ is a scalar function referred to as the gradient potential function. It is a mapping of the form $V : \mathcal{U} \rightarrow \mathbb{R}$ which is required to be twice continuously differentiable, where \mathcal{U} is an open set such that $\mathcal{U} \subset \mathbb{R}^n$. This means that both the first and second derivatives of $V(\boldsymbol{x})$ must

be continuous functions. Gradient dynamics are described by the equation

$$\begin{aligned} \dot{\mathbf{x}} &= -\nabla_{\mathbf{x}} V(\mathbf{x}) = -\mathbf{f}(\mathbf{x}), \\ \therefore \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{pmatrix} &= - \begin{pmatrix} \frac{\partial V}{\partial x_1} \\ \frac{\partial V}{\partial x_2} \\ \vdots \\ \frac{\partial V}{\partial x_n} \end{pmatrix}. \end{aligned} \tag{2.1}$$

The first theorem and its corollary show that trajectories always move toward smaller values of $V(\mathbf{x})$. The only exception to this is at equilibrium points where a trajectory must remain constant.

Theorem 2.1. $\dot{V}(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in U$ and $\dot{V}(\mathbf{x}) = 0$ if and only if \mathbf{x} is an equilibrium point of equation (2.1) (i.e. $\dot{\mathbf{x}} = 0$).

Theorem 2.1 can be used to prove the following corollary.

Corollary 2.1. If $\tilde{\mathbf{x}}$ is an isolated minimum of $V(\mathbf{x})$ then $\tilde{\mathbf{x}}$ is an asymptotically stable equilibrium point of equation (2.1).

Notice that *not* every equilibrium point of equation (2.1) is a local minimum of $V(\mathbf{x})$. However, the previous two results show that every isolated local minimum of $V(\mathbf{x})$ is an asymptotically stable equilibrium point of equation (2.1). This means that if a trajectory starts within some neighborhood of $\tilde{\mathbf{x}}$, eventually it will reach $\tilde{\mathbf{x}}$. This is a purely *local* result, it does not guarantee that *every* trajectory will converge to some equilibrium point. The next theorem gives a geometric description of the flow of the gradient system.

Theorem 2.2. For a system given by equation (2.1) the trajectories at regular points (i.e. points where $\dot{V}(\mathbf{x}) \neq 0$) are orthogonal to the level surfaces of $V(\mathbf{x})$ (i.e. surfaces where $V(\mathbf{x}) = k$, $k \in \mathbb{R}$). Nonregular points are equilibria of the system.

This is a formal statement of the intuitive behavior of the system that was made previously.

The trajectories of a gradient system can only remain constant at an equilibrium point and must move toward smaller values of $V(\mathbf{x})$ at all other points. A *recurrent* trajectory is one that returns to within an arbitrarily small neighborhood of its point at some later time. These two statements imply that for a gradient system, only a trajectory started at an equilibrium point is recurrent. Trajectories started at all other points will move away from those points and not

return. Theorem 2.3 identifies the possible recurrent trajectories in phase space. It is stated in terms of the nonwandering set. Conceptually a nonwandering point lies on or near trajectories which eventually return to within a specified distance of themselves. The set of all such points is the nonwandering set. As given in [10], the mathematical definition of the nonwandering set is the following.

Definition 2.1. A point \mathbf{q} is nonwandering for the trajectory $\phi_t(\mathbf{x}_0) \equiv \mathbf{x}(\mathbf{x}_0, t)$ if for every neighborhood \mathcal{S} of \mathbf{q} and $T > 0$ there exists $t > T$ such that $\{\phi_t(\mathcal{S})\} \cap \mathcal{S} \neq \emptyset$. The set of all such points for all $\mathbf{x}_0 \in \mathcal{K}$ where $\mathcal{K} \subset \mathbb{R}^n$ is the nonwandering set for $\{\phi_t(\mathcal{K})\}$.

Notice that $\{\phi_t(\mathcal{S})\}$ is the set of all solution states at time t , denoted $\{\mathbf{x}(\mathbf{x}_0, t)\}$, which have initial conditions such that $\mathbf{x}_0 \in \mathcal{S}$. So the point \mathbf{q} is a nonwandering point if at least *one* trajectory started in the neighborhood \mathcal{S} , returns to \mathcal{S} at an arbitrarily large value of time. For example, if \mathbf{q} is a point on an unstable limit cycle, only one trajectory from any neighborhood of \mathbf{q} returns to that neighborhood, but this is sufficient to make \mathbf{q} a nonwandering point. Definition 2.1 is used in the following theorem.

Theorem 2.3. *The nonwandering set for the system given by equation (2.1) contains only the equilibrium points of the system.*

The nonwandering set defines a weak concept of recurrence. So Theorem 2.3 states that eventually, the only recurrent trajectories in the phase space are the equilibrium points. This theorem implies that a gradient system has no periodic orbits, no homoclinic orbits (i.e. an orbit connecting a saddle point to itself) or in fact any sort of periodic asymptotic behavior. Intuitively, all of these behaviors would require that the system be able to move both up and down hill along $V(\mathbf{x})$, or remain constant at arbitrary values of $V(\mathbf{x})$, neither of which can occur in a gradient system.

All trajectories of a gradient system must move downhill along $V(\mathbf{x})$. Therefore all trajectories must end at a stable equilibrium point or go to infinity. Likewise all trajectories must begin at an unstable equilibrium point or at infinity. Theorem 2.4 identifies these asymptotic trajectories in phase space as $t \rightarrow \pm\infty$. This theorem is stated in terms of α -limit and ω -limit sets. The ω -limit set is the set of points that all trajectories go to as $t \rightarrow \infty$. The α -limit set is the set of points that all trajectories come from as $t \rightarrow -\infty$. From [11] the definition of the ω -limit set is as follows.

Definition 2.2. A point \mathbf{q} is an ω -limit point of the trajectory $\phi_t(\mathbf{x}_0) \equiv \mathbf{x}(\mathbf{x}_0, t)$ if there exists a sequence $t_n \rightarrow \infty$ such that $\lim_{t_n \rightarrow \infty} \phi_{t_n}(\mathbf{x}_0) \rightarrow \mathbf{q}$. The set of all such points for all $\mathbf{x}_0 \in \mathcal{K}$ where $\mathcal{K} \subset \mathbb{R}^n$ is the ω -limit set for $\{\phi_t(\mathcal{K})\}$.

Letting the sequence be $t_n \rightarrow -\infty$ in the previous definition yields the definition of an α -limit set. The point \mathbf{q} is an ω -limit point if the distance between \mathbf{q} and at least *one* trajectory becomes arbitrarily small at an arbitrarily large value of time. Definition 2.2 is used in the following theorem.

Theorem 2.4. *Let the point \mathbf{y} be an α -limit point or an ω -limit point of a trajectory of equation (2.1). Then \mathbf{y} is an equilibrium point.*

The α -limit and ω -limit sets define a weak concept of repelling and attracting sets. The trajectories approach these sets asymptotically. Therefore this theorem implies that if the equilibrium points are isolated, then the solution state must either go to an equilibrium point or to infinity. Note that infinity can *not* be a member of either the α -limit or ω -limit sets.

For a gradient system the nonwandering set and the union of the α -limit and ω -limit sets are both equal to the same set, namely the set of equilibrium points. In general this is *not* the case. The following example illustrates this and clarifies the difference between a nonwandering point and a limit point.

Example 2.1. [24] Consider the system defined by the differential equations

$$\begin{aligned} \dot{r} &= r(1-r), \\ \dot{\theta} &= \sin^2 \theta + (1-r)^3. \end{aligned} \tag{2.2}$$

Clearly this system is defined in polar coordinates. The nonwandering set and the α -limit and ω -limit sets are most easily seen in the phase space of the system. This is shown in Figure 2.1. It is important to note that once a trajectory reaches either the point $(1, 0)$ or $(-1, 0)$, it will remain there indefinitely. In other words, none of the trajectories that started outside the circle can ever cross the x -axis. The origin is the only α -limit point of the system because all points inside the circle $r = 1$ lie on a trajectory which gets arbitrarily close to the origin after an arbitrarily long negative time. The circle $r = 1$ is the ω -limit set for every point in the plane, except the origin, because all points lie on a trajectory which gets arbitrarily close to some portion of the circle after an arbitrarily long positive time. The nonwandering set contains the points $(0, 0)$, $(1, 0)$ and $(-1, 0)$ because only trajectories started at these three points will eventually return to these points. The three black dots in Figure 2.1 mark the locations of the nonwandering points. So any trajectory started on $r = 1$ will eventually leave the neighborhood of the starting point and *not* return, with the exception of trajectories started at $(1, 0)$ and $(-1, 0)$.

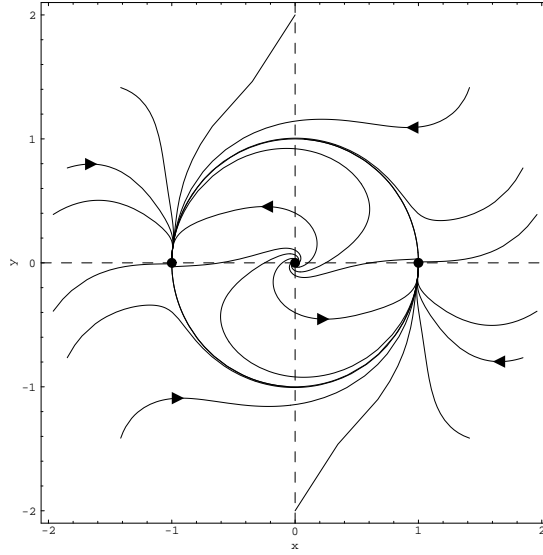


Figure 2.1: Phase space of the system in Equation (2.2)

As this example shows the nonwandering set and the limit set (i.e. the union of the α -limit and ω -limit sets) do not have to contain the same elements. The fundamental difference between limit points and nonwandering points is that while a limit point is a point which at least one trajectory eventually becomes arbitrarily close to (for time running both forward and backward), a nonwandering point is a point which at least one trajectory, *started nearby*, eventually becomes arbitrarily close to (for time running forward only). No idea of recurrence is embodied in the definition of a limit point, while it is fundamental to the definition of a nonwandering point.

The next theorem uses the idea of an isolated equilibrium point. This is defined as follows. The equilibrium solutions of any equation $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ may be of two types. One possibility is that a solution of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is a single point in \mathbb{R}^n . In this event the equilibrium solution is a point and is said to be *isolated*. The other possibility is that a solution of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is a relation which represents a subspace of \mathbb{R}^n . In this case the solution set contains an infinite number of points and the equilibrium solution is *non-isolated*. The following lemma, proved in [25] describes a way to test whether an equilibrium point is isolated.

Lemma 2.1. *Consider an equilibrium point $\bar{\mathbf{x}}$ of the system in equation (2.1). Define the Jacobian $\mathbf{J}_G(\bar{\mathbf{x}})$ as*

$$\mathbf{J}_G(\bar{\mathbf{x}}) = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right] \Big|_{\bar{\mathbf{x}}}. \quad (2.3)$$

If $\mathbf{J}_G(\bar{\mathbf{x}})$ is nonsingular then $\bar{\mathbf{x}}$ is an isolated equilibrium point.

Note that for a gradient system $\mathbf{J}_G(\bar{\mathbf{x}})$ is always a square matrix. Therefore $\mathbf{J}_G(\bar{\mathbf{x}})$ is nonsingular if it has no zero eigenvalues. Collectively the set of equilibrium solutions for any given system may contain both isolated and non-isolated solutions. The next theorem states the conditions under which *all* trajectories will go to some equilibrium point.

Theorem 2.5. *Consider the dynamic system given by equation (2.1). Suppose that the set*

$$\mathcal{M}_c = \{\mathbf{x} \in \mathbb{R}^n : V(\mathbf{x}) \leq c\} \quad (2.4)$$

is compact (i.e. closed and bounded) for every $c \in \mathbb{R}$. Then every solution of the system $\mathbf{x}(t)$ is defined for all $t \geq 0$. Suppose that the system has a finite number of isolated equilibrium points $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m$. Then for every solution $\mathbf{x}(t)$, the limit $\lim_{t \rightarrow \infty} \mathbf{x}(t)$ exists and equals one of the equilibrium points.

There are many possible ways to constrain $V(\mathbf{x})$ so that the set \mathcal{M}_c is compact. One way is to make $V(\mathbf{x})$ bounded below, $V(\mathbf{x}) \geq \delta$ for all $\mathbf{x} \in \mathbb{R}$, and radially unbounded, $V(\mathbf{x}) \rightarrow \infty$ as $\|\mathbf{x}\| \rightarrow \infty$.

Not only is the set of orbits which the trajectories can approach restricted but the way in which those orbits are approached is also limited. This is shown in Theorem 2.6.

Theorem 2.6. *At every equilibrium point $\bar{\mathbf{x}}$ of equation (2.1), the linearized system $\dot{\mathbf{x}} = [\mathbf{J}_G(\bar{\mathbf{x}})] \mathbf{x}$ has real eigenvalues. The Jacobian $\mathbf{J}_G(\bar{\mathbf{x}})$ is*

$$\mathbf{J}_G(\bar{\mathbf{x}}) = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right] \Big|_{\bar{\mathbf{x}}} = \frac{\partial}{\partial \mathbf{x}} [-\nabla_{\mathbf{x}} V(\mathbf{x})] \Big|_{\bar{\mathbf{x}}}. \quad (2.5)$$

Furthermore $\mathbf{J}_G(\bar{\mathbf{x}})$ is diagonalizable and has eigenvectors which form a complete orthonormal set.

In this paper, an improper node is defined as any equilibrium point where the Jacobian is *not* diagonalizable. Since $\mathbf{J}_G(\bar{\mathbf{x}})$ is always diagonalizable, no improper nodes exist for a gradient system. Because the eigenvalues of $\mathbf{J}_G(\bar{\mathbf{x}})$ are real, only three types of equilibrium points are possible: a proper stable point (i.e. a sink), a proper unstable point (i.e. a source), or a saddle point. Conceptually this means that within a small enough neighborhood of any equilibrium point, the shape of any trajectory must be a hyperbola, a parabola, or a line. For a general discussion of equilibrium point analysis in nonlinear systems see [24].

A vector field \mathcal{X} is defined as the mapping $\mathcal{X} : \mathbb{R} \rightarrow \mathbb{R}^n$. A vector field is *structurally stable* if the significant features of the phase space remain unchanged by the addition of a sufficiently small vector field. More specifically, the direction in which points flow along the trajectories as time

increases, and the way in which trajectories approach the equilibrium points and closed orbits, remains the same in the presence of sufficiently small perturbations. For example, an equilibrium point that trajectories are diverging from along a spiral path will remain so for a small enough perturbation of the vector field. Hence a system having one equilibrium point of this type is structurally stable.

Before studying the structural stability of gradient-like systems, it is useful to briefly review the mathematical history of structural stability. It was proven in [19] that structural stability was a generic property of systems on 2-dimensional manifolds. This means that a randomly chosen 2-dimensional system has an *infinitely small* probability of *not* being structurally stable. Therefore almost all 2-dimensional systems are structurally stable. This very convenient property was shown *not* to extend to systems on manifolds with dimension ≥ 3 in [17, 22]. An exception to this was shown in [18], where it was demonstrated that almost all possible gradient systems are structurally stable regardless of manifold dimension. Although structural stability is not generic on manifolds with dimension ≥ 3 , in [18] a class of vector fields was presented which are always structurally stable. This class of vector fields, called Morse-Smale, is defined as follows.

Definition 2.3. Let \mathcal{M}^n be a compact manifold of dimension n . Let \mathcal{X} be a C^r (i.e. continuously differentiable to order r) vector field from the set of all such vector fields on \mathcal{M}^n . \mathcal{X} is a *Morse-Smale* vector field if

1. the critical elements (i.e. equilibrium solutions and closed orbits) of \mathcal{X} are all *hyperbolic*, and are finite in number;
2. if σ_1 and σ_2 are critical elements of \mathcal{X} then $\mathcal{W}^s(\sigma_1)$ is *transverse* to $\mathcal{W}^u(\sigma_2)$;
3. the *nonwandering set* $\Omega(\mathcal{X})$ is equal to the union of the critical elements of \mathcal{X} .

Conceptually a critical element is *hyperbolic* if all nearby trajectories either converge to or diverge from the critical element at an exponential or greater rate. Two subspaces \mathcal{Z}^1 and \mathcal{Z}^2 of a third space \mathcal{Z} are *transverse* if their sum $\mathcal{Z}^1 + \mathcal{Z}^2$ is the entire space \mathcal{Z} . This means that any point in \mathcal{Z} can be decomposed into a point in \mathcal{Z}^1 and a point in \mathcal{Z}^2 . For example, a line and a plane are transverse in \mathbb{R}^3 if they intersect at a nonzero angle. By contrast, two lines cannot be transverse in \mathbb{R}^3 . The stable manifold $\mathcal{W}^s(\sigma_1)$ is the subspace containing those trajectories which converge to σ_1 . The unstable manifold $\mathcal{W}^u(\sigma_2)$ is the subspace containing those trajectories which diverge from σ_2 . The nonwandering set $\Omega(\mathcal{X})$ is discussed in Definition 2.1.

The following theorem gives the conditions under which a gradient system is structurally stable.

Theorem 2.7. *A gradient system is structurally stable if and only if every equilibrium point of equation (2.1) is hyperbolic (i.e. the Jacobian \mathbf{J}_G has no eigenvalues with zero real part) and if all stable and unstable manifolds intersect transversally.*

Conceptually requiring that every equilibrium point be hyperbolic means in a gradient system that all equilibrium points are isolated. This is generally *not* the case. Notice that the two conditions given are necessary and sufficient, hence every structurally stable gradient system must satisfy the definition of a Morse-Smale system.

2.2 Properties of Gradient-Like Systems

Some of the useful properties of gradient systems are also possessed by a more general class of dynamic systems, which will be referred to as gradient-like systems. Some properties of gradient-like systems are discussed in [21], these results are extended in this paper. Gradient-like system dynamics are described by the equation

$$\begin{aligned} \dot{\mathbf{x}} &= -\mathbf{P}(\mathbf{x}) [\nabla_{\mathbf{x}} V(\mathbf{x})] = \mathbf{g}(\mathbf{x}), \\ \therefore \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{pmatrix} &= - \begin{pmatrix} p_{11} \frac{\partial V}{\partial x_1} + p_{12} \frac{\partial V}{\partial x_2} + \cdots + p_{1n} \frac{\partial V}{\partial x_n} \\ p_{21} \frac{\partial V}{\partial x_1} + p_{22} \frac{\partial V}{\partial x_2} + \cdots + p_{2n} \frac{\partial V}{\partial x_n} \\ \vdots \\ p_{n1} \frac{\partial V}{\partial x_1} + p_{n2} \frac{\partial V}{\partial x_2} + \cdots + p_{nn} \frac{\partial V}{\partial x_n} \end{pmatrix}, \end{aligned} \quad (2.6)$$

where $V(\mathbf{x})$ is the gradient potential function defined in equation (2.1). If the matrix $\mathbf{P}(\mathbf{x})$ is symmetric and positive definite (i.e. $\mathbf{y}^T \mathbf{P}(\mathbf{x}) \mathbf{y} > 0 \quad \forall \quad \mathbf{y} \neq \mathbf{0}$) for all values of \mathbf{x} , then equation (2.6) will be called a gradient-like system. Notice that the gradient system in equation (2.1) is a special case of equation (2.6) in which $\mathbf{P}(\mathbf{x})$ is the identity matrix.

The first theorem and its corollary are used to show the asymptotic stability of all isolated local minima of $V(\mathbf{x})$.

Theorem 2.8. *Suppose that the matrix $\mathbf{P}(\mathbf{x})$ is positive definite. Then $\dot{V}(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in \mathcal{U}$, and $\dot{V}(\mathbf{x}) = 0$ if and only if \mathbf{x} is an equilibrium point of equation (2.6) (i.e. $\dot{\mathbf{x}} = \mathbf{0}$).*

Proof: Using the chain rule

$$\dot{V}(\mathbf{x}) = [\nabla_{\mathbf{x}} V(\mathbf{x})]^T \dot{\mathbf{x}} = -[\nabla_{\mathbf{x}} V(\mathbf{x})]^T \mathbf{P}(\mathbf{x}) [\nabla_{\mathbf{x}} V(\mathbf{x})]. \quad (2.7)$$

Since $\mathbf{P}(\mathbf{x})$ is positive definite, by definition $\dot{V}(\mathbf{x}) = 0$ only when $\nabla_{\mathbf{x}}V(\mathbf{x}) = \mathbf{0}$, otherwise $\dot{V}(\mathbf{x}) < 0$. The equilibrium points of equation (2.6) are the solutions of the equation

$$\dot{\mathbf{x}} = \mathbf{0} = -\mathbf{P}(\mathbf{x})[\nabla_{\mathbf{x}}V(\mathbf{x})]. \quad (2.8)$$

Since $\mathbf{P}(\mathbf{x})$ is invertible for *all* values of \mathbf{x} , equation (2.8) implies that $\nabla_{\mathbf{x}}V(\mathbf{x}) = \mathbf{0} \iff \dot{\mathbf{x}} = \mathbf{0}$. \blacklozenge

Note that since $\mathbf{P}(\mathbf{x})$ is positive definite, it does not contribute any equilibrium points to the system. LaSalle's Theorem is required to prove Corollary 2.2. Following the form of [15] this theorem is stated as follows.

Lemma 2.2 (LaSalle). *Consider the system*

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}), \quad (2.9)$$

where $\mathbf{g} : \mathcal{U} \rightarrow \mathbb{R}^n$ is a locally Lipschitz mapping from $\mathcal{U} \subset \mathbb{R}^n$ to \mathbb{R}^n . Let \mathcal{M} be a compact set such that every solution of equation (2.9) which starts in \mathcal{M} remains in \mathcal{M} for all future time. Let $V : \mathcal{M} \rightarrow \mathbb{R}$ be a continuously differentiable function such that $\dot{V}(\mathbf{x}) \leq 0$ in \mathcal{M} . Let \mathcal{E} be the set of all points in \mathcal{M} where $\dot{V}(\mathbf{x}) = 0$. Let \mathcal{I} be the largest invariant set in \mathcal{E} . Then every solution starting in \mathcal{M} approaches \mathcal{I} as $t \rightarrow \infty$.

LaSalle's Theorem and Theorem 2.8 are used to prove the following corollary.

Corollary 2.2. *If $\tilde{\mathbf{x}}$ is an isolated local minimum of $V(\mathbf{x})$ then $\tilde{\mathbf{x}}$ is an asymptotically stable equilibrium point of equation (2.6).*

Proof: From calculus, $\tilde{\mathbf{x}}$ is a minimum of $V(\mathbf{x})$ if $\nabla_{\mathbf{x}}V(\tilde{\mathbf{x}}) = \mathbf{0}$ and $\nabla_{\mathbf{x}}^2V(\tilde{\mathbf{x}}) > \mathbf{0}$. Since $\nabla_{\mathbf{x}}V(\mathbf{x}) = \mathbf{0} \iff \dot{\mathbf{x}} = \mathbf{0}$, therefore every isolated local minimum of $V(\mathbf{x})$ is an equilibrium point of equation (2.6). Since $\tilde{\mathbf{x}}$ is an isolated local minimum of $V(\mathbf{x})$, there exists some neighborhood \mathcal{W} of $\tilde{\mathbf{x}}$ such that $V(\mathbf{x}) > V(\tilde{\mathbf{x}})$ for all $\mathbf{x} \in \mathcal{W} : \mathbf{x} \neq \tilde{\mathbf{x}}$. In this neighborhood, a Lyapunov function $Z(\mathbf{x})$ can be defined as $Z(\mathbf{x}) = V(\mathbf{x}) - V(\tilde{\mathbf{x}})$. For all $\mathbf{x} \in \mathcal{W}$, $Z(\mathbf{x}) > 0$ unless $\mathbf{x} = \tilde{\mathbf{x}}$, where $Z(\mathbf{x}) = 0$. Since $V(\tilde{\mathbf{x}})$ is a constant, its time derivative is zero. Hence the time derivative of $Z(\mathbf{x})$ is

$$\dot{Z}(\mathbf{x}) = -[\nabla_{\mathbf{x}}V(\mathbf{x})]^T \mathbf{P}(\mathbf{x}) [\nabla_{\mathbf{x}}V(\mathbf{x})], \quad (2.10)$$

and by Theorem 2.8, $\dot{Z}(\mathbf{x}) < 0$ for all $\mathbf{x} \in \mathcal{W}$ except at $\mathbf{x} = \tilde{\mathbf{x}}$ where $\dot{Z}(\mathbf{x}) = 0$. Therefore the point $\tilde{\mathbf{x}}$ is asymptotically stable by LaSalle's Theorem. \blacklozenge

Again notice that *not* every equilibrium point of equation (2.6) is a local minimum of $V(\mathbf{x})$. Since $\mathbf{P}(\mathbf{x})$ does not contribute any equilibrium points to the system, it can not contribute any stable

minima to the system. Recall that a trajectory started within some neighborhood of $\tilde{\mathbf{x}}$, eventually reaches $\tilde{\mathbf{x}}$.

The next two theorems describe the phase space behavior of a gradient-like system. Theorem 2.9 identifies the possible recurrent trajectories in phase space. This theorem is stated in terms of the nonwandering set of points in phase space. Definition 2.1 is used in the following theorem.

Theorem 2.9. *The nonwandering set for the system given by equation (2.6) contains only the equilibrium points of the system.*

Proof: $\dot{V}(\mathbf{x})$ describes the change in $V(\mathbf{x})$ along the system trajectories. Since $\dot{V}(\mathbf{x}) \leq 0$ for the system in equation (2.6), the solution state trajectories can only move toward smaller values of $V(\mathbf{x})$, or remain at the same value of $V(\mathbf{x})$. Since the trajectories cannot move toward larger values of $V(\mathbf{x})$, the only nonwandering points are those for which $V(\mathbf{x})$ is constant, in other words $\dot{V}(\mathbf{x}) = 0$. From Theorem 2.8, $\dot{V}(\mathbf{x}) = 0$ if and only if \mathbf{x} is an equilibrium point of equation (2.6). \blacklozenge

This theorem shows that for gradient-like systems, as for gradient systems, the only recurrent trajectories, at large times, are the equilibrium points. So a gradient-like system has no periodic trajectories, no homoclinic trajectories (i.e. a trajectory connecting a saddle point to itself), or any sort of periodic asymptotic behavior.

Theorem 2.10 identifies the asymptotic trajectories in phase space as $t \rightarrow \pm\infty$. This theorem is stated in terms of α -limit and ω -limit sets. Definition 2.2 is used in the following theorem.

Theorem 2.10. *Let \mathbf{y} be an α -limit point or an ω -limit point of the system in equation (2.6). Then \mathbf{y} is an equilibrium point of the system.*

Proof: Let \mathbf{y} be an ω -limit point. Then by definition the point \mathbf{y} is an element of the set

$$\mathcal{L} = \left\{ \mathbf{y} \in \mathcal{U} : \lim_{t_n \rightarrow \infty} \phi_{t_n}(\mathbf{x}_0) \rightarrow \mathbf{y} \right\}. \quad (2.11)$$

Define the positively invariant set \mathcal{P} such that $\phi_t(\mathbf{x}_0) \in \mathcal{P}$ for all $0 \leq t < \infty$. Since $V(\mathbf{x})$ is continuously differentiable, therefore it is continuous. Since $\dot{V}(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in \mathcal{P}$, therefore $V(\mathbf{x}) = \xi$ where ξ is the greatest lower bound of the set $\{V(\phi_t(\mathbf{x}_0)) : t \geq 0\}$. Since every point of \mathcal{L} is a limit of points in \mathcal{P} , and \mathcal{P} is closed in \mathcal{U} it follows that $\mathcal{L} \subset \mathcal{P}$. If $\mathbf{y} \in \mathcal{L}$, then $\phi_t(\mathbf{y})$ is defined for all $t \geq 0$ because \mathcal{P} is positively invariant. Since $\phi_{t_n}(\mathbf{x}_0) \rightarrow \mathbf{y}$ and $t_1 < t_2 < \dots$, therefore $\phi_t(\mathbf{y})$ is defined for all $t \in [-t_n, 0], n = 1, 2, \dots$. Because $-t_n \rightarrow -\infty$, $\phi_t(\mathbf{y})$ is defined for all $t \leq 0$. If $\phi_{t_n}(\mathbf{x}_0) \rightarrow \mathbf{y}$ then $\phi_{t_n+s}(\mathbf{x}_0) \rightarrow \phi_s(\mathbf{y})$ for all $s \in \mathbb{R}$. Since $t_n \rightarrow \infty$ both \mathbf{y} and $\phi_s(\mathbf{y})$ must be contained in \mathcal{L} for all $t \in \mathbb{R}$. Hence

$V(\mathbf{y}) = \xi$ for all $\mathbf{y} \in \mathcal{L}$. This implies that $\dot{V}(\mathbf{y}) = 0$ for all $\mathbf{y} \in \mathcal{L}$. By Theorem 2.8 all $\mathbf{y} \in \mathcal{L}$ are equilibrium points. An analogous proof applies to α -limit points. \blacklozenge

So this theorem shows that for a gradient-like system the union of the asymptotically repelling and attracting sets is simply the set of equilibrium points, again in analogy to a gradient system. So if the equilibrium points are isolated, then the solution state must either go to an equilibrium point or to infinity.

The next theorem states a restriction on $V(\mathbf{x})$ that insures that *every* solution state must go to one of the equilibrium points. The proof of this theorem requires the following two lemmas, which are proved in [12].

Lemma 2.3. *Let $\mathbf{g}(\mathbf{x})$ be locally Lipschitz on a domain $\mathcal{U} \subset \mathbb{R}^n$, and let \mathcal{M} be a compact subset of \mathcal{U} . Let $\mathbf{x}_0 \in \mathcal{M}$ and suppose that every solution of*

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.12}$$

lies entirely in \mathcal{M} . Then there is a unique solution $\mathbf{x}(t)$ that is defined for all $t \geq 0$.

It is also shown in [12] that a function $\mathbf{g}(\mathbf{x})$ which is continuously differentiable on a domain \mathcal{U} is locally Lipschitz on that domain.

Lemma 2.4. *Let the set \mathcal{I} in LaSalle's Theorem consist of a finite number of isolated equilibrium points. Then $\lim_{t \rightarrow \infty} \mathbf{x}(t)$ exists and equals one of the equilibrium points.*

Lemma 2.3 and Lemma 2.4 are used to prove the following theorem.

Theorem 2.11. *Consider the dynamic system given by equation (2.6). Suppose that the set*

$$\mathcal{M}_c = \{\mathbf{x} \in \mathbb{R}^n : V(\mathbf{x}) \leq c\} \tag{2.13}$$

is compact (i.e. closed and bounded) for every $c \in \mathbb{R}$. Then every solution of the system $\mathbf{x}(t)$ is defined for all $t \geq 0$. Suppose that the system has a finite number of isolated equilibrium points $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m$. Then for every solution $\mathbf{x}(t)$, the limit $\lim_{t \rightarrow \infty} \mathbf{x}(t)$ exists and equals one of the equilibrium points.

Proof: Since c is allowed to take on *any* real value, every value of \mathbf{x} for which $V(\mathbf{x})$ is defined is contained by some \mathcal{M}_c . Therefore every solution starts in a set \mathcal{M}_c with $V(\mathbf{x}_0) \leq c$. Since $\dot{V}(\mathbf{x}) \leq 0$ in \mathcal{M}_c the solution $\mathbf{x}(t)$ remains in \mathcal{M}_c for all $t \geq 0$. Since \mathcal{M}_c is compact, by Lemma 2.3, $\mathbf{x}(t)$ is unique and defined for all $t \geq 0$.

Let \mathcal{E} be the set of all points in \mathcal{M}_c where $\dot{V}(\mathbf{x}) = 0$. By Theorem 2.8, \mathcal{E} contains only the equilibrium points of the system. A set \mathcal{D} is invariant if for each \mathbf{x}_0 in \mathcal{D} , $\phi_t(\mathbf{x}_0) \equiv \mathbf{x}(\mathbf{x}_0, t)$

is defined and in \mathcal{D} for all $t \in \mathbb{R}$. Hence the largest invariant set \mathcal{I} in \mathcal{E} consists of all equilibrium points of the system. By LaSalle's Theorem, $\mathbf{x}(t) \rightarrow \mathcal{I}$ as $t \rightarrow \infty$. Since the equilibrium points are isolated, by Lemma 2.4, $\mathbf{x}(t) \rightarrow \mathbf{q}_i$ for some $\mathbf{q}_i \in \mathcal{I}$. \blacklozenge

One condition that guarantees that \mathcal{M}_c is bounded for all values of c , is to make $V(\mathbf{x})$ radially unbounded, $V(\mathbf{x}) \rightarrow \infty$ as $\|\mathbf{x}\| \rightarrow \infty$. This will not insure that \mathcal{M}_c is closed.

Theorem 2.12 identifies how the asymptotic solution states are approached. It shows that the behavior of a gradient-like system near an equilibrium point is well-defined. The following lemma, proved in [5], is needed to prove the theorem.

Lemma 2.5. *Given two real valued, symmetric matrices \mathbf{A} and \mathbf{B} , where \mathbf{B} is also positive definite. The eigenvalue equation*

$$|\mathbf{A} - \lambda\mathbf{B}| = 0 \tag{2.14}$$

has only real roots. Furthermore, the matrix $\mathbf{D} = \mathbf{B}^{-1}\mathbf{A}$ has three properties: real eigenvalues, a complete set of orthonormal eigenvectors, and diagonalizability.

Conceptually this lemma defines a class of *non-symmetric* matrices \mathbf{D} , which are normal [23] and possess real eigenvalues. This lemma is now used to prove the following theorem.

Theorem 2.12. *At every equilibrium point $\bar{\mathbf{x}}$ of equation (2.6), the linearized system $\dot{\mathbf{x}} = [\mathbf{J}_T(\bar{\mathbf{x}})]\mathbf{x}$ has real eigenvalues, where the Jacobian matrix $\mathbf{J}_T(\bar{\mathbf{x}})$ is defined as*

$$\mathbf{J}_T(\bar{\mathbf{x}}) = \left\{ \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \right\} \Big|_{\bar{\mathbf{x}}}. \tag{2.15}$$

Furthermore $\mathbf{J}_T(\bar{\mathbf{x}})$ is diagonalizable and has eigenvectors which form a complete orthonormal set.

Proof: The Jacobian of a gradient-like system can be written as

$$\mathbf{J}_T(\bar{\mathbf{x}}) = \left\{ \mathbf{P}(\mathbf{x}) \left[\frac{\partial [\nabla_{\mathbf{x}}V(\mathbf{x})]}{\partial \mathbf{x}} \right] + \left[\frac{\partial \mathbf{P}(\mathbf{x})}{\partial \mathbf{x}} \right] \nabla_{\mathbf{x}}V(\mathbf{x}) \right\} \Big|_{\bar{\mathbf{x}}}. \tag{2.16}$$

The term $\partial [\nabla_{\mathbf{x}}V(\mathbf{x})] / \partial \mathbf{x}|_{\bar{\mathbf{x}}}$ is the Jacobian of the gradient system given by equation (2.1). It will be denoted by $\mathbf{J}_G(\bar{\mathbf{x}})$ hereafter. Notice that the term $[\partial \mathbf{P}(\mathbf{x}) / \partial \mathbf{x}] \nabla_{\mathbf{x}}V(\mathbf{x})$ is a matrix, and that the quantity $\partial \mathbf{P}(\mathbf{x}) / \partial \mathbf{x}$ is a tensor. Since the quantity $\nabla_{\mathbf{x}}V(\mathbf{x})$ is zero at *any* equilibrium point of equation (2.6), the matrix $[\partial \mathbf{P}(\mathbf{x}) / \partial \mathbf{x}] \nabla_{\mathbf{x}}V(\mathbf{x})$ is always the zero matrix at these points. Therefore the Jacobian of the linearized system is always given by

$\mathbf{J}_T(\bar{\mathbf{x}}) = \mathbf{P}(\mathbf{x})\mathbf{J}_G(\mathbf{x})|_{\bar{\mathbf{x}}}$. Since $\mathbf{P}(\mathbf{x})$ is positive definite, its inverse $\mathbf{P}^{-1}(\mathbf{x})$ is also positive definite. Therefore the eigenvalues of \mathbf{J}_T are given by the roots of

$$|\mathbf{P}(\bar{\mathbf{x}})| |\mathbf{J}_G(\bar{\mathbf{x}}) - \lambda\mathbf{P}^{-1}(\bar{\mathbf{x}})| = 0. \quad (2.17)$$

The determinant $|\mathbf{P}(\bar{\mathbf{x}})|$ is some positive quantity which will not affect the eigenvalues of the system. Note that $\mathbf{P}^{-1}(\bar{\mathbf{x}})$ is positive definite and $\mathbf{J}_G(\bar{\mathbf{x}})$ is symmetric. Therefore by Lemma 2.5, $\mathbf{J}_T(\bar{\mathbf{x}})$ has the properties in the theorem statement. \blacklozenge

Since $\mathbf{J}_T(\bar{\mathbf{x}})$ is always diagonalizable, no improper nodes exist for a gradient-like system. Because the eigenvalues of $\mathbf{J}_T(\bar{\mathbf{x}})$ are real, only three types of equilibrium points are possible: a proper stable point (i.e. a sink), a proper unstable point (i.e. a source), or a saddle point. This is identical to the result for gradient systems.

The next theorem establishes the conditions for which a gradient-like system is structurally stable. To prove the structural stability of gradient-like systems the following theorem, proved in [18], is needed.

Lemma 2.6 (Palis-Smale). *Given that the number of equilibrium points and closed orbits is finite, then the vector field \mathcal{X} is structurally stable if and only if it is Morse-Smale.*

This theorem can be used to prove the following result.

Theorem 2.13. *The system of equation (2.6) is structurally stable if and only if every equilibrium point is hyperbolic and all stable and unstable manifolds intersect transversally.*

Proof: From Theorem 2.9, the nonwandering set of equation (2.6) contains only the equilibrium points of the system. For any structurally stable system on a compact manifold, all equilibrium solutions are hyperbolic [16]. On a compact manifold, the number of hyperbolic points must be finite. Hence there are only a finite number of equilibrium solutions for any structurally stable systems in the form of equation (2.6). Therefore by Lemma 2.6 all gradient-like systems are structurally stable if and only if they are Morse-Smale. \blacklozenge

Conceptually, requiring that every equilibrium point be hyperbolic means in a gradient-like system that all equilibrium points are isolated. This is *not* generally the case. Notice that the two conditions given are necessary and sufficient, hence every structurally stable gradient-like system must satisfy the definition of a Morse-Smale system. It remains to show that the set of structurally stable gradient-like systems is an open and dense subset of the set of all gradient-like systems. This would imply that almost all gradient-like systems are structurally stable.

Intuitively the trajectories of the systems in equations (2.1) and (2.6) both move downhill along the surface described by $V(\mathbf{x})$. In equation (2.1) the “laws of motion” state that the trajectories must follow the line of steepest descent along $V(\mathbf{x})$. In equation (2.6) the matrix $\mathbf{P}(\mathbf{x})$ specifies the “laws of motion” for the trajectories. Stipulating that $\mathbf{P}(\mathbf{x})$ be positive definite means that the trajectories must still move downhill along $V(\mathbf{x})$. So the trajectories of the system in equation (2.6) are a smooth distortion of those in equation (2.1) with $\mathbf{P}(\mathbf{x})$ specifying the transformation. These properties can be used to analyze any system whose dynamics can be expressed in gradient-like form. In the next section a general class of neural networks that allow for weight update will be cast in the form of gradient-like systems.

Chapter 3

Neural Networks

The central idea of this section is to show that many neural network models from the literature can be framed as gradient-like systems. First the results of [3] for fully connected networks with constant weights will be reviewed. Then it will be shown that such networks can be written as gradient-like systems. This result will then be extended to show that fully connected networks with weight update are also gradient-like systems. It will then be shown that networks with Hebbian weight update, anti-Hebbian weight update, and differential Hebbian weight update are special cases of this result. Lastly it will be demonstrated that this formalism can be extended to include networks which are not fully connected, such as multilayer networks, and networks which provide higher order components in the output and correlation terms.

3.1 Review of Lyapunov Function Results

In this section the results of [3] will be reviewed. Consider a fully connected network containing p nodes, where no weight update (i.e learning) occurs. It has been shown in [3] that if the network can be written in the form

$$\dot{x}_i = a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) \right], \quad (3.1)$$

then there exists a Lyapunov function

$$V(\mathbf{x}) = - \sum_{i=1}^p \int_0^{x_i} b_i(\zeta) d'_i(\zeta) d\zeta + \frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p c_{ij} d_i(x_i) d_j(x_j), \quad (3.2)$$

if the following conditions hold:

1. The matrix \mathbf{C} is symmetric and all $c_{ij} \geq 0$.
2. (a) The function $a_i(\xi)$ is continuous $\forall \xi \geq 0$.
 (b) The function $b_i(\xi)$ is continuous $\forall \xi > 0$.
3. (a) The function $a_i(\xi) > 0 \forall \xi > 0$.
 (b) The function $d_i(\xi) \geq 0 \forall \xi \in \mathbb{R}$.
4. The function $d_i(\xi)$ is differentiable and monotonically non-decreasing $\forall \xi \geq 0$.
5. $\limsup_{\xi \rightarrow \infty} [b_i(\xi) - c_{ii}d_i(\xi)] < 0 \forall i = 1, 2, \dots, p$.
6. Either
 - (a) $\lim_{\xi \rightarrow 0^+} b_i(\xi) = \infty$;
 - or
 - (b) $\lim_{\xi \rightarrow 0^+} b_i(\xi) < \infty$ while $\int_0^\epsilon \frac{d\xi}{a_i(\xi)} = \infty$ for some $\epsilon > 0$.

Consider a closed and bounded set \mathcal{M} of the activities \mathbf{x} such that conditions 1 - 6 hold. The Lyapunov function of equation (3.2) can be shown to apply everywhere in the bounded region specified by \mathcal{M} because in that set

$$\dot{V}(\mathbf{x}) = [\nabla_{\mathbf{x}} V(\mathbf{x})]^T \dot{\mathbf{x}} = - \sum_{i=1}^p a_i(x_i) d_i'(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) \right]^2 \leq 0. \quad (3.3)$$

Assuming that the stated conditions 1 - 6 hold, the above result is always true for any overall activation \mathbf{x} in \mathcal{M} . LaSalle's Theorem can then be used to show that all trajectories within \mathcal{M} asymptotically converge to the largest invariant set \mathcal{I} contained in the set

$$\mathcal{E} = \left\{ \bar{\mathbf{x}} \in \mathbb{R}^p : \dot{V}(\bar{\mathbf{x}}) = 0 \forall \bar{\mathbf{x}} \geq 0 \right\}. \quad (3.4)$$

If the functions $d_i(x_i)$ are strictly increasing then the set \mathcal{E} consists of only the equilibrium points of equation (3.1). In this case LaSalle's Theorem shows that the equilibrium points are asymptotically stable because $\dot{V}(\mathbf{x}) = 0$ *only* when $\dot{\mathbf{x}} = 0$ elsewhere $\dot{V}(\mathbf{x}) < 0$.

Some of the conditions imposed in this analysis are rather difficult to understand. Probably the most difficult are the restriction that $c_{ij} \geq 0$, and conditions 5 and 6. These conditions are used to prove that x_i is both lower and upper bounded and that the lower bound is zero. This

result is then used to prove that $a_i(x_i)$ is always positive for all *possible* system states x_i . Since x_i can only be positive, condition 3a guarantees this. It is also used to show that the Lyapunov function in equation (3.2) is bounded. The integral term is bounded because the x_i is bounded, and the scalar product term is bounded because the functions $d_j(x_j)$ are continuous functions of bounded variables.

3.2 Gradient-Like Formulation of the Constant Weight Case

Now it will be shown that a fully connected network with constant weights can be written as a gradient-like system. For later convenience, the notation presented in [14] will be used for the remainder of the paper. In this notation, the dynamics of a fully connected network of p nodes with no weight update are

$$\dot{x}_i = -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) \right]. \quad (3.5)$$

For these dynamics the negative of the Lyapunov function in equation (3.2) satisfies the conditions for a gradient potential function. Observation of equation (3.3) shows that $\nabla_{\mathbf{x}} [-V(\mathbf{x})]$ is given by

$$\begin{pmatrix} \frac{\partial V(x_1)}{\partial x_1} \\ \frac{\partial V(x_2)}{\partial x_2} \\ \vdots \\ \frac{\partial V(x_p)}{\partial x_p} \end{pmatrix} = \begin{pmatrix} d'_1(x_1) \left[b_1(x_1) - \sum_{j=1}^p c_{1j} d_j(x_j) \right] \\ d'_2(x_2) \left[b_2(x_2) - \sum_{j=1}^p c_{2j} d_j(x_j) \right] \\ \vdots \\ d'_p(x_p) \left[b_p(x_p) - \sum_{j=1}^p c_{pj} d_j(x_j) \right] \end{pmatrix}. \quad (3.6)$$

From equation (3.6) it is apparent that the system of equation (3.5) can be written as the

gradient-like system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_p \end{pmatrix} = - \begin{pmatrix} \frac{a_1(x_1)}{d_1'(x_1)} & 0 & \cdots & 0 \\ 0 & \frac{a_2(x_2)}{d_2'(x_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{a_p(x_p)}{d_p'(x_p)} \end{pmatrix} \begin{pmatrix} d_1'(x_1) \left[b_1(x_1) - \sum_{i=1}^p c_{1i} d_i(x_i) \right] \\ d_2'(x_2) \left[b_2(x_2) - \sum_{i=1}^p c_{2i} d_i(x_i) \right] \\ \vdots \\ d_p'(x_p) \left[b_p(x_p) - \sum_{i=1}^p c_{pi} d_i(x_i) \right] \end{pmatrix}, \quad (3.7)$$

$$\therefore \quad \dot{\mathbf{x}} = -\mathbf{P}(\mathbf{x}) [\nabla_{\mathbf{x}} V(\mathbf{x})]. \quad (3.8)$$

The matrix $\mathbf{P}(\mathbf{x})$ is a $(p \times p)$ diagonal matrix. It is positive definite if $a_i(x_i)$ is positive definite (i.e. $a_i(x_i) > 0$ for all $x_i > 0$ and $a_i(x_i) = 0$ only if $x_i = 0$) and $d_i'(x_i)$ is monotonically increasing (i.e. $d_i'(x_i) > 0$).

In [2, 9] two general forms for node activation dynamics are reviewed. They are referred to as additive and multiplicative node dynamics. In the next two sections it will be demonstrated that constant weight networks possessing either type of node activation dynamics can be formulated as gradient-like systems.

3.2.1 Application to an Additive Network

First consider the case of additive node dynamics. A network with p nodes whose activities are governed by an additive equation and where no learning takes place is described by the general equation

$$\left(\frac{1}{\epsilon_i} \right) \dot{x}_i = -\mathcal{A}_i x_i + \mathcal{B}_i \left[I_i + \sum_{j=1}^p F_{ij} Z_{ij}^+ k_j(x_j) \right] - \mathcal{C}_i \left[J_i + \sum_{j=1}^p G_{ij} Z_{ij}^- l_j(x_j) \right]. \quad (3.9)$$

In this equation the term $-\mathcal{A}_i x_i$ is passive decay which causes x_i to go to zero if the other two terms are zero. The constant \mathcal{A}_i determines the rate of decay. The term $\mathcal{B}_i [I_i + \sum_{j=1}^p F_{ij} Z_{ij}^+ k_j(x_j)]$ is the positive (e.g. excitatory) feedback which tries to increase x_i . Finally $\mathcal{C}_i [J_i + \sum_{j=1}^p G_{ij} Z_{ij}^- l_j(x_j)]$ is the negative (e.g. inhibitory) feedback which tries to decrease x_i . The excitatory and inhibitory connection weights to the i th node from the j th node are given by Z_{ij}^+ and Z_{ij}^- respectively. All of the connection weights incident to a specific node do not have to be given equal consideration. For instance, the connections from nodes that are physically closer to the given node may be

considered more important than those from nodes which are farther away. If all of the connection weights into a node are viewed as a field of values in weight space, then this field is sampled by some function. The sample values applied to the excitatory and inhibitory connections to the i th node from the j th node are given by F_{ij} and G_{ij} respectively.

In this paradigm the constants \mathcal{A}_i , \mathcal{B}_i , and \mathcal{C}_i , the sampling values F_{ij} and G_{ij} , the inputs I_i and J_i , and the connection weights Z_{ij}^+ and Z_{ij}^- are always required to be positive. Furthermore the functions $k_j(x_j)$ and $l_j(x_j)$ must yield positive values for all values of x_j . Each node in the network has two sections, one to process the excitatory signals and the other to process the inhibitory signals. Actually there is no loss of generality in writing equation (3.9) with only one set of sampling values. This can be seen by writing the inhibitory connection weights as $Z_{ij}^- G_{ij}/F_{ij} \equiv \tilde{Z}_{ij}^-$. Hence equation (3.9) can be rewritten as

$$\left(\frac{1}{\epsilon_i}\right) \dot{x}_i = -\mathcal{A}_i x_i + [\mathcal{B}_i I_i - \mathcal{C}_i J_i] + \sum_{j=1}^p [\mathcal{B}_i Z_{ij}^+ k_j(x_j) - \mathcal{C}_i \tilde{Z}_{ij}^- l_j(x_j)] F_{ij}. \quad (3.10)$$

If each node is restricted to having only one output function $h_i(x_i)$ for both excitatory and inhibitory signals then the resulting equation is

$$\left(\frac{1}{\epsilon_i}\right) \dot{x}_i = -\mathcal{A}_i x_i + [\mathcal{B}_i I_i - \mathcal{C}_i J_i] + \sum_{j=1}^p [\mathcal{B}_i Z_{ij}^+ - \mathcal{C}_i \tilde{Z}_{ij}^-] F_{ij} h_j(x_j). \quad (3.11)$$

The form of equation (3.11) can be simplified if it is written with respect to a single set of inputs $K_i \equiv \mathcal{B}_i I_i - \mathcal{C}_i J_i$ and a single set of connection weights $W_{ij} \equiv F_{ij} [\mathcal{B}_i Z_{ij}^+ - \mathcal{C}_i \tilde{Z}_{ij}^-]$ both of which can take positive or negative values. So the final form of the additive node activation dynamics is

$$\left(\frac{1}{\epsilon_i}\right) \dot{x}_i = -\mathcal{A}_i x_i + K_i + \sum_{j=1}^p W_{ij} h_j(x_j). \quad (3.12)$$

In this equation $-\mathcal{A}_i x_i$ is a passive decay term which causes x_i to go to zero if the remaining terms are zero. The constant \mathcal{A}_i determines the rate of decay. The function $h_j(x_j)$ is the output function of the j th node, and the input to the i th node is K_i . The connection weight *to* the i th node *from* the j th node is W_{ij} . In equation (3.12) the inputs K_i and the connection weights W_{ij} may both take positive or negative values. Equation (3.12) can be written in the form of

equation (3.7) by using the substitutions

$$\begin{aligned}
a_i(x_i) &= \epsilon_i, \\
b_i(x_i) &= \mathcal{A}_i x_i - K_i, \\
c_{ij} &= W_{ij}, \\
d_j(x_j) &= h_j(x_j).
\end{aligned} \tag{3.13}$$

It is obvious from equation (3.13) that $a_i(x_i)$ is positive for any value of x_i . This formulation makes it clear that *any* additive network can be written as the gradient-like system of equation (3.23) by imposing two conditions: the matrix $[W_{ij}]$ must be symmetric, and the function $h_j(x_j)$ must be twice continuously differentiable as well as monotonically increasing (i.e. $h'_i(x_i) > 0$).

3.2.2 Application to a Multiplicative Network

Next consider the case of multiplicative node dynamics. A one layer network with p nodes whose activities are governed by a multiplicative equation and where no learning takes place is described by the equation

$$\dot{x}_i = -\mathcal{A}_i x_i + (\mathcal{B}_i - \mathcal{D}_i x_i) \left[I_i + \sum_{j=1}^p F_{ij} Z_{ij}^+ k_j(x_j) \right] - (\mathcal{C}_i + \mathcal{E}_i x_i) \left[J_i + \sum_{j=1}^p G_{ij} Z_{ij}^- l_j(x_j) \right] \tag{3.14}$$

The parameters in equation (3.14) are the same as in equation (3.9). For any system in the form of equation(3.14) the activation levels \mathbf{x} of all of the nodes must belong to the bounded set

$$\mathcal{M} = \left\{ x_i \in \mathbb{R} : -\frac{\mathcal{C}_i}{\mathcal{E}_i} \leq x_i \leq \frac{\mathcal{B}_i}{\mathcal{D}_i} \quad \forall i = 1, \dots, p \right\}. \tag{3.15}$$

This bound is only valid if the initial value of the activation $\mathbf{x}(t_0)$ is also contained in the set \mathcal{M} .

Noticing that the activation x_i is bounded below by $-\mathcal{C}_i/\mathcal{E}_i$ and that the function $a_i(x_i)$ must be positive for all positive x_i suggests the variable change

$$y_i = \mathcal{C}_i + \mathcal{E}_i x_i \implies \dot{y}_i = \mathcal{E}_i \dot{x}_i. \tag{3.16}$$

Using this transformation the multiplicative network of equation (3.14) can be written in the

general form given by equation (3.5) where $a_i(x_i)$, $b_i(x_i)$, c_{ji} , and $d_i(x_i)$ are given by

$$\begin{aligned}
a_i(x_i) &= \frac{1}{\mathcal{E}_i} (\mathcal{C}_i + \mathcal{E}_i x_i), \\
b_i(x_i) &= -\frac{1}{\mathcal{C}_i + \mathcal{E}_i x_i} \left[\mathcal{A}_i \mathcal{C}_i + \mathcal{E}_i \mathcal{B}_i I_i + \mathcal{D}_i \mathcal{C}_i I_i + (\mathcal{E}_i \mathcal{B}_i + \mathcal{D}_i \mathcal{C}_i) \sum_{j=1}^p F_{ij} Z_{ij}^+ k_j(x_j) \right] \\
&\quad + \left[\mathcal{A}_i + \mathcal{D}_i I_i + \mathcal{E}_i J_i - \mathcal{D}_i \sum_{j=1}^p F_{ij} Z_{ij}^+ l_j(x_j) - \mathcal{E}_i \sum_{j=1}^p G_{ij} Z_{ij}^- k_j(x_j) \right], \\
c_{ij} &= -\mathcal{D}_i F_{ij} Z_{ij}^+ - \mathcal{E}_i G_{ij} Z_{ij}^-, \\
d_j(x_j) &= k_j(x_j) + l_j(x_j).
\end{aligned} \tag{3.17}$$

In order to cast this system into the form of a gradient-like system, the function $a_i(x_i)$ must be positive for all permissible values of x_i to insure that the matrix $\mathbf{P}(\mathbf{x})$ is positive definite in equation (3.8). One way to insure this is to require that all 6 conditions in section 3.1 are satisfied. These conditions can be satisfied by imposing the following constraints.

1. (a) The matrices \mathbf{Z}^+ and \mathbf{Z}^- are symmetric and all elements are $Z_{ij}^+ \geq 0$, $Z_{ij}^- \geq 0$.
(b) The matrices \mathbf{F} and \mathbf{G} are symmetric and all elements are $F_{ij} \leq 0$, $G_{ij} \leq 0$.
(c) $\mathcal{D}_1 = \mathcal{D}_2 = \dots = \mathcal{D}_p$;
 $\mathcal{E}_1 = \mathcal{E}_2 = \dots = \mathcal{E}_p$.
2. (a) The inputs I_i and J_i are continuous.
(b) The functions $k_j(x_j)$ and $l_j(x_j)$ are differentiable (\implies continuous).
3. (a) The initial activation \mathbf{x} must be in the set \mathcal{M} .
(b) The functions $k_j(x_j)$ and $l_j(x_j)$ are given by $k_j(x_j) \geq 0$, $l_j(x_j) \geq 0 \forall x_j \in \mathbb{R}$.
4. The functions $k_j(x_j)$ and $l_j(x_j)$ are differentiable and monotone increasing $\forall x_j \geq 0$.
5. (a) The inputs I_i and J_i are $I_i > 0$, $J_i > 0$.
(b) *Eventually* the slopes are $k_j'(x_j) < 1$, $l_j'(x_j) < 1$ and they remain that way (i.e. the increase in $k_j(x_j)$ and $l_j(x_j)$ becomes slower than linear for large values of x_j).
6. (a) If the functions $k_j(x_j)$ and $l_j(x_j)$ *do not* go to zero as x_j goes to zero or if they go to zero slower than linearly, then (a) holds.
(b) If the functions $k_j(x_j)$ and $l_j(x_j)$ go to zero faster than linearly then (b) holds.

Notice that condition 6 is true no matter how the functions $k_j(x_j)$ and $l_j(x_j)$ are selected. The restrictions on \mathcal{D}_i and \mathcal{E}_i in condition 1 can be explained intuitively in the following way. The

activation x_i of the i th node is upper and lower bounded by $\mathcal{B}_i/\mathcal{D}_i$ and $-\mathcal{C}_i/\mathcal{E}_i$ respectively. The time constants, which determine the speed at which x_i approaches these upper and lower bounds, are given by $1/\mathcal{D}_i$ and $1/\mathcal{E}_i$ respectively. Hence condition 1 states that the rate, at which a node approaches the upper or lower bound of its activation, is identical for all nodes in the network. Notice that the rate at which the upper and lower bounds are approached may be different. Also *each* node may have a different upper and lower activation bound, since \mathcal{B}_i and \mathcal{C}_i may be set differently for every node. It should be noted that multiplication of the inputs, I_i and J_i , and the functions, $k_j(x_j)$ and $l_j(x_j)$, by appropriate constants will remove this restriction. The restriction imposed on the slopes of $k_j(x_j)$ and $l_j(x_j)$ by condition 5 can not be removed if the system described by equation (3.14) is to be a stable system. However, given the computational advantages of choosing a sigmoidal output function proven in [8], this hardly seems a serious limitation.

3.3 Gradient-Like Formulation of the Updated Weight Case

This section will demonstrate that a fully connected network *with* weight update can be formulated as a gradient-like system. Consider a fully connected network of p nodes with a general weight update rule. In other words, a given node is connected to every other node including itself. Following the form in [14] the dynamics of such a network can be written as

$$\begin{aligned} \dot{x}_i &= -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) \right], \quad i = 1, \dots, p, \\ \dot{c}_{ij} &= f_{ij}(x_i, x_j, c_{ij}), \quad i, j = 1, \dots, p. \end{aligned} \tag{3.18}$$

Equation (3.18) can then be written in a more compact matrix-vector form

$$\begin{aligned} \dot{\mathbf{x}} &= -\mathbf{A}(\mathbf{x}) [\mathbf{b}(\mathbf{x}) - \mathbf{C}\mathbf{d}(\mathbf{x})], \\ \dot{\mathbf{C}} &= \mathbf{F}(\mathbf{x}, \mathbf{C}). \end{aligned} \tag{3.19}$$

In this equation \mathbf{x} is the p dimensional vector of node activities, $\mathbf{A}(\mathbf{x})$ is a $(p \times p)$ dimensional *diagonal* matrix, $\mathbf{b}(\mathbf{x})$ is a p dimensional vector, \mathbf{C} is the $(p \times p)$ dimensional matrix of connection weights, and $\mathbf{d}(\mathbf{x})$ is the p dimensional vector of node output functions. In the weight matrix \mathbf{C} , the row number of a given entry denotes the node that the connection is incident *to* while the column entry indicates the node that the connection is incident *from*. In order to cast the system

of equation (3.19) into the form of a gradient-like system, choose the gradient potential function

$$V(\mathbf{x}, \mathbf{C}) = -\frac{1}{2}\mathbf{d}(\mathbf{x})^T \mathbf{C} \mathbf{d}(\mathbf{x}) + \left[\sum_{k=1}^p \int_0^{x_k} d'_k(\zeta_k) b_k(\zeta_k) d\zeta_k \right] + L(\mathbf{C}). \quad (3.20)$$

In this equation $L(\mathbf{C})$ is a scalar function which determines part of the weight update rule. Since $V(\mathbf{x}, \mathbf{C})$ must be twice continuously differentiable, $L(\mathbf{C})$ must also be twice continuously differentiable. Specific choices for $L(\mathbf{C})$ will be given in the following sections. At this point it is useful to define the vector

$$\mathbf{u} = [x_1, x_2, x_3, \dots, x_p, c_{11}, c_{12}, c_{13}, \dots, c_{pp}]^T. \quad (3.21)$$

Notice that \mathbf{u} contains $(p + p^2)$ elements. Using this notation in equation (3.20), the gradient $\nabla_{\mathbf{u}} V(\mathbf{u})$ becomes

$$\nabla_{\mathbf{u}} V(\mathbf{u}) = \begin{pmatrix} \frac{\partial V(\mathbf{u})}{\partial x_1} \\ \vdots \\ \frac{\partial V(\mathbf{u})}{\partial x_p} \\ \frac{\partial V(\mathbf{u})}{\partial c_{11}} \\ \vdots \\ \frac{\partial V(\mathbf{u})}{\partial c_{pp}} \end{pmatrix} = \begin{pmatrix} d'_1(x_1) \left[b_1(x_1) - \sum_{j=1}^p c_{1j} d_j(x_j) \right] \\ \vdots \\ d'_p(x_p) \left[b_p(x_p) - \sum_{j=1}^p c_{pj} d_j(x_j) \right] \\ -\frac{1}{2} d_1(x_1) d_1(x_1) + \frac{\partial L(\mathbf{C})}{\partial c_{11}} \\ \vdots \\ -\frac{1}{2} d_p(x_p) d_p(x_p) + \frac{\partial L(\mathbf{C})}{\partial c_{pp}} \end{pmatrix}. \quad (3.22)$$

Note that \mathbf{C} must be symmetric in order to obtain this result. The notation $\Delta[h_{11}, h_{22}, \dots, h_{pp}]$ will be used to denote a $(p \times p)$ diagonal matrix with the listed elements along the diagonal. Using this notation the entire system can be written as the gradient-like system

$$\dot{\mathbf{u}} = -\Delta \left[\frac{a_1(x_1)}{d'_1(x_1)}, \dots, \frac{a_p(x_p)}{d'_p(x_p)}, 2, \dots, 2 \right] [\nabla_{\mathbf{u}} V(\mathbf{u})]. \quad (3.23)$$

The diagonal matrix $\Delta[\cdot]$ is $\mathbf{P}(\mathbf{u})$. The notation means that the first p diagonal elements of $\mathbf{P}(\mathbf{u})$ are $a_i(x_i)/d'_i(x_i)$ where $i = 1, \dots, p$, and the remaining p^2 diagonal elements are the constant 2. By identifying equation (3.23) with equation (3.18) it can be shown that the weight update dynamics are

$$\dot{c}_{ij} = -2 \frac{\partial L(\mathbf{C})}{\partial c_{ij}} + d_i(x_i) d_j(x_j). \quad (3.24)$$

Notice that the second term in this equation is the correlation term of the Hebbian learning rule. This means that a proper choice of $L(\mathbf{C})$ allows the commonly used Hebbian weight update rule to be instantiated. The next five sections demonstrate that a number of neural network paradigms can be written as gradient-like systems.

3.3.1 Application to Multilayer Networks

This formalism can be used to describe layered networks. Typically the nodes in a given layer are connected to those nodes in the layers immediately above and below the given layer. Also the nodes within a given layer may be connected to one another. This structure can be formulated by decomposing the activation vector \mathbf{x} and the connection weight matrix \mathbf{C} into

$$\mathbf{x} = \begin{pmatrix} \mathbf{v}_1 \\ \text{---} \\ \mathbf{v}_2 \\ \text{---} \\ \mathbf{v}_3 \\ \text{---} \\ \mathbf{v}_4 \\ \text{---} \\ \vdots \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{R}_{1 \rightarrow 1} & \mathbf{E}_{2 \rightarrow 1} & \mathbf{O} & \mathbf{O} & \cdots \\ \mathbf{E}_{1 \rightarrow 2} & \mathbf{R}_{2 \rightarrow 2} & \mathbf{E}_{3 \rightarrow 2} & \mathbf{O} & \cdots \\ \mathbf{O} & \mathbf{E}_{2 \rightarrow 3} & \mathbf{R}_{3 \rightarrow 3} & \mathbf{E}_{4 \rightarrow 3} & \cdots \\ \mathbf{O} & \mathbf{O} & \mathbf{E}_{3 \rightarrow 4} & \mathbf{R}_{4 \rightarrow 4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (3.25)$$

The vector \mathbf{v}_k represents the node activation in the k th layer. The submatrix $\mathbf{R}_{k \rightarrow k}$ denotes the connection weights between nodes within the k th layer, while $\mathbf{E}_{l \rightarrow k}$ denotes the connection weights from the l th layer to the k th layer. Note that $\mathbf{E}_{l \rightarrow k}$ does not have to be square. Since \mathbf{C} must be symmetric, therefore $\mathbf{R}_{k \rightarrow k} = \mathbf{R}_{k \rightarrow k}^T$ and $\mathbf{E}_{l \rightarrow k} = \mathbf{E}_{k \rightarrow l}^T$. If $\mathbf{R}_{k \rightarrow k}$ is a matrix of constants, then the connection weights between nodes within the k th layer are fixed, they are not part of the network dynamics. If $\mathbf{R}_{k \rightarrow k}$ is the zero matrix \mathbf{O} , then the nodes within layer k are not connected. Because a layered network is not fully connected, the vector \mathbf{u} will contain less than $(p + p^2)$ elements in this case.

In formulating a multilayer network, blocks of the connection matrix \mathbf{C} were set to zero to represent nodes that were not connected to one another and to constant matrices to represent nodes that had static connections to one another. This same idea can be applied to individual pairs

of opposing connection weights (i.e. c_{ij} and c_{ji}). Any pair of opposing weights can be removed by setting the desired connection weight values, c_{ij} and c_{ji} , to zero. The same connections can be made fixed by setting the weight values to the desired constant weight. In either case, the weights c_{ij} and c_{ji} must also be removed from the state vector \mathbf{u} .

3.3.2 Application to Hebbian Learning

The use of the Hebbian weight update rule in neural network models has been widely studied. Some of the properties of networks using Hebbian dynamics are presented in [1, 7, 13]. This choice of weight update rule can be shown to fit into the gradient-like dynamics formalism. Following the form in [14], these networks have dynamics described by the differential equations

$$\dot{x}_i = -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) \right] \quad i = 1, \dots, p, \quad (3.26a)$$

$$\dot{c}_{ij} = -\gamma_{ij} c_{ij} + \lambda_{ij} d_i(x_i) d_j(x_j) \quad i, j = 1, \dots, p. \quad (3.26b)$$

The term $-\gamma_{ij} c_{ij}$ is a passive decay term where γ_{ij} is a constant which determines the decay rate. The constant λ_{ij} determines the growth rate of the connection weight c_{ij} if the nodes at both ends of the connection are active. The matrices containing all such constants are $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ respectively. In order to instantiate the Hebbian learning rule into the gradient-like system of equation (3.23), let the gradient potential function be given by

$$V(\mathbf{u}) = -\frac{1}{2} \mathbf{d}(\mathbf{x})^T \mathbf{C} \mathbf{d}(\mathbf{x}) + \left[\sum_{k=1}^p \int_0^{x_k} d'_k(\zeta_k) b_k(\zeta_k) d\zeta_k \right] + \frac{1}{4} \mathbf{1}^T \left[\mathbf{\Gamma} \circ \mathbf{\Lambda}^{-1} \circ \mathbf{C} \circ \mathbf{C} \right] \mathbf{1}. \quad (3.27)$$

In equation (3.27) note that $\mathbf{1}$ is a p dimensional vector whose elements are all 1. Also the operation \circ denotes the *Schur* product which is defined as $[\mathbf{A} \circ \mathbf{B}]_{ij} = a_{ij} b_{ij}$. Choose the diagonal matrix $\mathbf{P}(\mathbf{u})$ to be

$$\mathbf{P}(\mathbf{u}) = \Delta \left[\frac{a_1(x_1)}{d'_1(x_1)}, \dots, \frac{a_p(x_p)}{d'_p(x_p)}, 2\lambda_{11}, 2\lambda_{12}, 2\lambda_{13}, \dots, 2\lambda_{pp} \right]. \quad (3.28)$$

In order for $\mathbf{P}(\mathbf{u})$ to be positive definite, the matrix $\mathbf{\Lambda}$ must contain strictly positive values. Also note that the last term in equation (3.27) is the function $L(\mathbf{C})$. Additionally, the weight matrix \mathbf{C} learned by the Hebbian rule must be symmetric. The necessary conditions for this to occur

can be found by solving for the equilibrium values of the weights c_{ij} and c_{ji} :

$$\begin{aligned}\dot{c}_{ij} = -\gamma_{ij}c_{ij} + \lambda_{ij}d_i(x_i)d_j(x_j) = 0 &\implies c_{ij} = \frac{\lambda_{ij}}{\gamma_{ij}}d_i(x_i)d_j(x_j), \\ \dot{c}_{ji} = -\gamma_{ji}c_{ji} + \lambda_{ji}d_j(x_j)d_i(x_i) = 0 &\implies c_{ji} = \frac{\lambda_{ji}}{\gamma_{ji}}d_j(x_j)d_i(x_i).\end{aligned}\tag{3.29}$$

Clearly if the matrices $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ are symmetric, then the equilibrium values of c_{ij} and c_{ji} are identical. Strictly speaking the weight matrix must be symmetric at *all* points along the trajectories in order for equation (3.23) to hold. Given that $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ are symmetric, this will be true if the initial conditions of c_{ij} and c_{ji} are the same. A reasonable physical interpretation of this situation is that there is a single bidirectional link between any two different nodes, rather than two unidirectional ones.

3.3.3 Application to Anti-Hebbian Learning

In some applications it is desirable for the outputs of nodes in the same layer to be as uncorrelated as possible. Conceptually this allows each node to code roughly independent features of the input. To decorrelate two nodes, have the weight connecting them decrease when both nodes are active simultaneously. This is referred to as *anti-Hebbian* learning and can be written as

$$\dot{c}_{ij} = -\gamma_{ij}c_{ij} - \lambda_{ij}d_i(x_i)d_j(x_j) \quad i, j = 1, \dots, p.\tag{3.30}$$

A feedforward network employing this learning rule was investigated in [4]. In this case the output layer nodes were connected to one another by weights which were updated by the anti-Hebbian learning rule. The output layer was fully connected to the input layer via weights updated by the Hebbian rule. It is shown in [4] that such a network with linear output functions $d_i(x_i)$ performs a principle component analysis on the input. The equilibrium value of the connection weight c_{ij} in equation (3.30) is given by

$$c_{ij} = -\frac{\lambda_{ij}}{\gamma_{ij}}d_i(x_i)d_j(x_j).\tag{3.31}$$

The same equilibrium point can be obtained using the alternate weight update rule

$$\dot{c}_{ij} = \gamma_{ij}c_{ij} + \lambda_{ij}d_i(x_i)d_j(x_j) \quad i, j = 1, \dots, p.\tag{3.32}$$

A *feedback* network using the version of the anti-Hebbian learning rule given by equation (3.32) can be implemented in the present formalism. Recall from equation (3.25) that the overall connection matrix \mathbf{C} can be decomposed into blocks where $\mathbf{R}_{k \rightarrow k}$ represents the *intra*layer connections in

the k th layer and $\mathbf{E}_{l \rightarrow k}$ represents the *interlayer* connections between the l th and k th layers. The matrix $\mathbf{\Gamma}$ can be similarly decomposed into the form

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{1 \rightarrow 1}^R & \mathbf{\Gamma}_{2 \rightarrow 1}^E & \mathbf{O} & \cdots \\ \mathbf{\Gamma}_{1 \rightarrow 2}^E & \mathbf{\Gamma}_{2 \rightarrow 2}^R & \mathbf{\Gamma}_{3 \rightarrow 2}^E & \cdots \\ \mathbf{O} & \mathbf{\Gamma}_{2 \rightarrow 3}^E & \mathbf{\Gamma}_{3 \rightarrow 3}^R & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (3.33)$$

The submatrices $\mathbf{\Gamma}_{k \rightarrow k}^R$ are the parts of $\mathbf{\Gamma}$ which multiply the portions of \mathbf{C} which contain the *intralayer* weights. The anti-Hebbian rule in equation (3.32) can be implemented by making the components of $\mathbf{\Gamma}_{k \rightarrow k}^R$ negative.

3.3.4 Application to Differential Hebbian Learning

In [14], networks which respond to the rate of change of the node output are introduced. The dynamics of such networks can be written as

$$\dot{x}_i = -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) - \sum_{j=1}^p c_{ij} \dot{d}_j(x_j) \right] \quad i = 1, \dots, p, \quad (3.34a)$$

$$\dot{c}_{ij} = -\gamma_{ij} c_{ij} + \lambda_{ij} d_i(x_i) d_j(x_j) + \rho_{ij} \dot{d}_i(x_i) \dot{d}_j(x_j) \quad i, j = 1, \dots, p. \quad (3.34b)$$

This network can not be written as a gradient system, but a network with similar qualitative behavior can be established. Consider the gradient potential function

$$\begin{aligned} V(\mathbf{x}, \mathbf{C}) = & -\frac{1}{2} \mathbf{d}(\mathbf{x})^T \mathbf{C} \mathbf{d}(\mathbf{x}) - \frac{1}{2} \dot{\mathbf{d}}(\mathbf{x})^T \mathbf{C} \dot{\mathbf{d}}(\mathbf{x}) + \left[\sum_{k=1}^p \int_0^{x_k} d'_k(\zeta_k) b_k(\zeta_k) d\zeta_k \right] \\ & + \frac{1}{4} \mathbf{1}^T \left[\mathbf{\Gamma} \circ \mathbf{A}^{-1} \circ \mathbf{C} \circ \mathbf{C} \right] \mathbf{1}. \end{aligned} \quad (3.35)$$

Choosing the same matrix $\mathbf{P}(\mathbf{u})$ as in equation (3.28) yields dynamics given by

$$\dot{x}_i = -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij} d_j(x_j) - \left(\frac{\dot{d}_i(x_i)}{d'_i(x_i)} \right) \sum_{j=1}^p c_{ij} \dot{d}_j(x_j) \right] \quad i = 1, \dots, p, \quad (3.36a)$$

$$\dot{c}_{ij} = -\gamma_{ij} c_{ij} + \lambda_{ij} d_i(x_i) d_j(x_j) + \lambda_{ij} \dot{d}_i(x_i) \dot{d}_j(x_j) \quad i, j = 1, \dots, p. \quad (3.36b)$$

Clearly the connection weight dynamics of equations (3.34b) and (3.36b) are virtually identical. The behavior of the third term in the activation dynamics is not the same. The bracketed (\cdot) portion of the third term in equation (3.36a) can be rewritten as

$$\frac{\dot{d}'_i(x_i)}{d'_i(x_i)} = \left[\frac{d''_i(x_i)}{d'_i(x_i)} \right] \dot{x}_i. \quad (3.37)$$

Conceptually the quantity $d''_i(x_i)/d'_i(x_i)$ can be viewed as a measure of the radius of the circle needed to approximate $d_i(x_i)$ in the neighborhood of x_i . So this quantity will be zero at points of inflection, where the curvature of $d_i(x_i)$ changes direction. Its magnitude will be greatest where $d_i(x_i)$ is flattest since the curvature is smallest in these regions. Assuming that the functions $d_i(x_i)$ are sigmoidal, then the third term in equation (3.36a) will have the greatest effect on the node activation x_i when the magnitude of x_i is large (i.e. $d_i(x_i)$ is near saturation) or when x_i is changing rapidly (i.e. \dot{x}_i is large).

3.3.5 Application to Higher Order Networks

Networks of the type given by equation (3.26) provide only a linear expansion of the vector $\mathbf{d}(\mathbf{x})$. This is due to the $\sum_{j=1}^p c_{ij}d_j(x_j)$ term in the node activation dynamics. Also simple Hebbian learning can only capture first order correlations with the $\lambda_{ij}d_i(x_i)d_j(x_j)$ learning term. In [6] and [20] networks that allow higher order expansions and correlations are discussed. The dynamics of a quadratic example of these networks can be expressed as

$$\dot{x}_i = -a_i(x_i) \left[b_i(x_i) - \sum_{j=1}^p c_{ij}d_j(x_j) - \sum_{j=1}^p \sum_{k=1}^p e_{ijk}d_j(x_j)d_k(x_k) \right] \quad i = 1, \dots, p, \quad (3.38a)$$

$$\dot{c}_{ij} = -\gamma_{ij}c_{ij} + \lambda_{ij}d_i(x_i)d_j(x_j) \quad i, j = 1, \dots, p, \quad (3.38b)$$

$$\dot{e}_{ijk} = -\kappa_{ijk}e_{ijk} + \mu_{ijk}d_i(x_i)d_j(x_j)d_k(x_k) \quad i, j, k = 1, \dots, p. \quad (3.38c)$$

In order to put this network in the form of the gradient-like system in equation (3.23), select the gradient potential function

$$\begin{aligned} V(\mathbf{x}, \mathbf{C}, \mathbf{E}) = & -\frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p c_{ij}d_i(x_i)d_j(x_j) - \frac{1}{3} \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p e_{ijk}d_i(x_i)d_j(x_j)d_k(x_k) \\ & + \sum_{i=1}^p \int_0^{x_i} b_i(\zeta)d'_i(\zeta)d\zeta + \frac{1}{4} \sum_{i=1}^p \sum_{j=1}^p \frac{\gamma_{ij}}{\lambda_{ij}} c_{ij}^2 + \frac{1}{6} \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p \frac{\kappa_{ijk}}{\mu_{ijk}} e_{ijk}^2, \end{aligned} \quad (3.39)$$

define the state vector \mathbf{u} as

$$\mathbf{u} = [x_1, x_2, x_3, \dots, x_p, c_{11}, c_{12}, c_{13}, \dots, c_{pp}, e_{111}, e_{112}, e_{113}, \dots, e_{ppp}]^T, \quad (3.40)$$

and let the matrix $\mathbf{P}(\mathbf{u})$ be

$$\mathbf{P}(\mathbf{u}) = \Delta \left[\begin{array}{c} \frac{a_1(x_1)}{d_1'(x_1)}, \dots, \frac{a_p(x_p)}{d_p'(x_p)}, 2\lambda_{11}, 2\lambda_{12}, 2\lambda_{13}, \dots, 2\lambda_{pp}, \\ 3\mu_{111}, 3\mu_{112}, 3\mu_{113}, \dots, 3\mu_{ppp} \end{array} \right]. \quad (3.41)$$

This same formalism can be extended to systems with this form of any order.

Chapter 4

An Example Simulation

This section provides a simulation of a very simple neural network. The simulations will be used to illustrate the way in which the various properties of gradient-like systems appear in the dynamical behavior of a neural network. The example network, illustrated in Figure 4.1, consists of two nodes, two weights and an external input. The system uses additive node activation dynamics

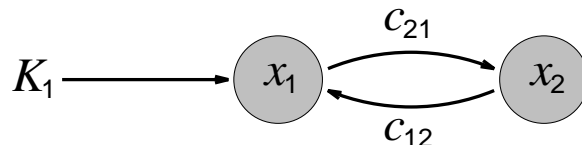


Figure 4.1: Configuration of example network

and Hebbian weight update dynamics as discussed in Sections 3.3.2 and 3.2.1. Equations (3.12) and (3.26b) describe the dynamics of such a network in general. The dynamic equations for this particular example are

$$\begin{aligned}
 \dot{x}_1 &= -\epsilon_1 (x_1 - K_1) + \epsilon_1 c_{12} \tanh(\mathcal{G}_2 x_2), \\
 \dot{x}_2 &= -\epsilon_2 x_2 + \epsilon_2 c_{21} \tanh(\mathcal{G}_1 x_1), \\
 \dot{c}_{12} &= -c_{12} + \tanh(\mathcal{G}_1 x_1) \tanh(\mathcal{G}_2 x_2), \\
 \dot{c}_{21} &= -c_{21} + \tanh(\mathcal{G}_2 x_2) \tanh(\mathcal{G}_1 x_1),
 \end{aligned} \tag{4.1}$$

where \mathcal{G}_1 and \mathcal{G}_2 are constants used to specify the steepness of the output function. The larger the values of \mathcal{G}_1 and \mathcal{G}_2 , the closer the output function becomes to a binary thresholding function. As shown in Sections 3.3.2 and 3.2.1, this network has gradient-like dynamics. For the simulation

results which follow, the values of the constants and the input are

$$\mathcal{G}_1 = 3, \quad \mathcal{G}_2 = 3, \quad \epsilon_1 = 10, \quad \epsilon_2 = 10, \quad K_1 = 50. \quad (4.2)$$

For the values given in Equation (4.2), there are three equilibrium points for the network. These equilibrium solutions are given in Table 4.1. An example of the way in which a trajectory

	state variables			
	x_1	x_2	c_{12}	c_{21}
Equilibrium #1	50.9899	0.994902	0.994902	0.994902
Equilibrium #2	50.9899	-0.994902	-0.994902	-0.994902
Equilibrium #3	50	0	0	0

Table 4.1: Equilibrium solutions for $K_1 = 50$

approaches these points with respect to time is shown in Figure 4.2. In this figure the plot labeled

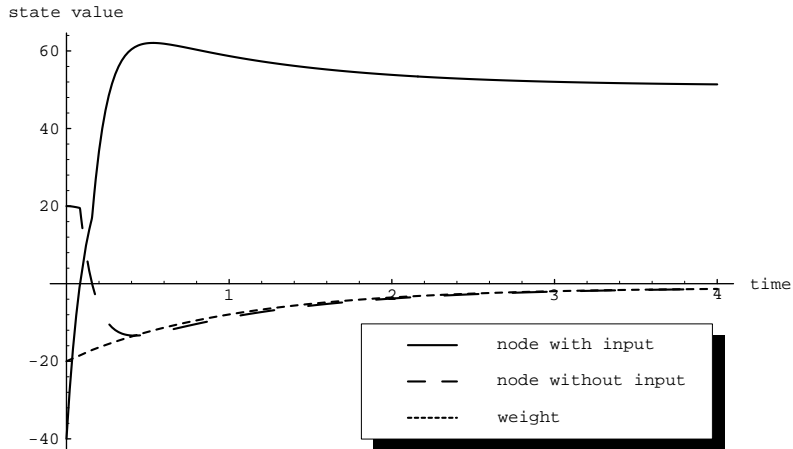


Figure 4.2: Change of the state variables over time

“node with input” represents the node activation values for x_1 , the plot labeled “node without input” shows the node activations for x_2 , and the plot labeled “weight” can be either c_{12} or c_{21} . The way in which the trajectories approach these points with respect to one another can be seen in a phase space diagram. Since this network has gradient-like dynamics, by Theorem 2.9 and Theorem 2.10 the phase space cannot contain any periodic orbits. All trajectories must go to one of the three equilibrium points or to infinity. Unfortunately, the phase space of this system is 4-dimensional, which can not be drawn. However because the weights, c_{12} and c_{21} , are identical at all points in time, a 3-dimensional section of the phase space will show most of the relevant

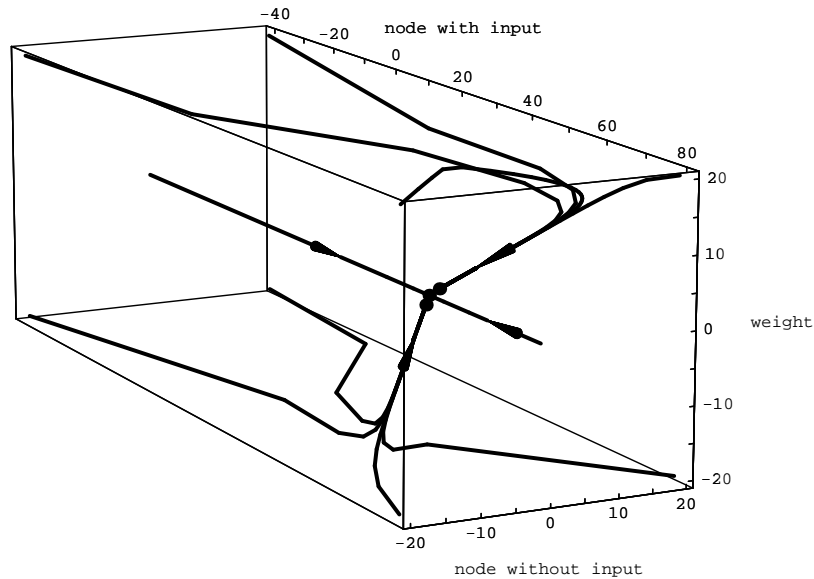


Figure 4.3: 3-dimensional section of the phase space

features. Such a 3-dimensional section is shown in Figure 4.3. In this figure, the three black dots clustered together in a triangular pattern mark the locations of the three equilibrium points. The two points that are approached by the trajectories running along the top and bottom surfaces of the cube are equilibrium points #1 and #2 respectively. The point in between them which is approached by the trajectories running through the middle of the cube is equilibrium point #3. Notice that the three equilibrium points are coplanar.

It appears from Figure 4.3 that the trajectories change directions rather abruptly at some points, and that all of the trajectories leading to equilibria #1 and #2 merge before going to these equilibria. These phenomena can be seen more clearly in the 2-dimensional sections of the phase space shown in Figure 4.4. This figure shows the phase space projected onto the two dimensions representing the node activation values. The top part of the figure shows the trajectories which converge to equilibrium #1, which was calculated with the weight values c_{12} and c_{21} fixed at 20. Similarly the bottom part, which shows trajectories that converge to equilibrium #2, was calculated with the weights fixed at -20. The black dots in the upper and lower part of the figure, mark the locations of equilibrium #1 and #2, respectively. Notice that the directions of the trajectories change when either of the axes are crossed. This is due to the sign change of the derivative at this point. Although this change looks discontinuous in the figure, in fact the trajectories are still differentiable at these points. Decreasing the value of \mathcal{G}_1 and \mathcal{G}_2 makes these transitions more gradual.

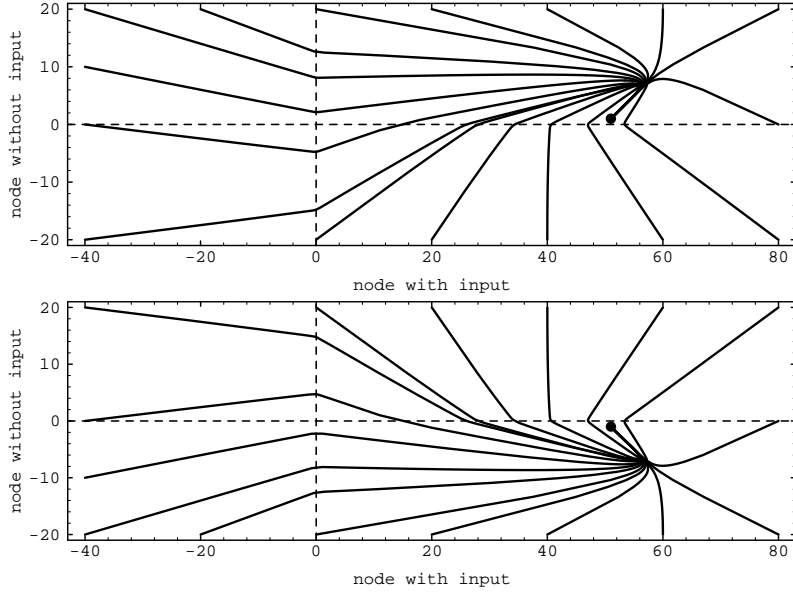


Figure 4.4: 2-dimensional section of the node activation phase space

If the input K_1 is negative (e.g. -50) then the phase space diagram is identical to Figure 4.3, except the signs of the numbers on the “node with input” and “node without input” axes are reversed. The equilibrium locations for this system are given in Table 4.2. Note that the three

	state variables			
	x_1	x_2	c_{12}	c_{21}
Equilibrium #1	-50.9899	-0.994902	0.994902	0.994902
Equilibrium #2	-50.9899	0.994902	-0.994902	-0.994902
Equilibrium #3	-50	0	0	0

Table 4.2: Equilibrium solutions for $K_1 = -50$

equilibria for $K_1 = -50$ do not lie on the same plane as those for $K_1 = 50$; in fact the two planes are perpendicular. It was observed that smaller values of \mathcal{G}_1 and \mathcal{G}_2 caused equilibria #1 and #2 to move closer to equilibrium #3. In the limit as $\mathcal{G}_1 = \mathcal{G}_2 \rightarrow 0$, only equilibrium #3 exists. Making the values larger has the opposite effect, however the equilibrium values of x_2 , c_{12} , and c_{21} at equilibria #1 and #2 are limited to $|x_2| = |c_{12}| = |c_{21}| \sim 1$ as $\mathcal{G}_1 = \mathcal{G}_2 \rightarrow \infty$.

As discussed in [24], the type of equilibrium point may be determined by finding the eigenvalues of the Jacobian of the system evaluated at each equilibrium point. Because the system is gradient-like, Theorem 2.12 shows that the eigenvalues of the Jacobian at every equilibrium point must be real valued. Further, the Jacobian must be diagonalizable at each equilibrium point. For this

system the Jacobian is given by

$$\mathbf{J}_S = \begin{pmatrix} -10 & 30c_{12} \operatorname{sech}^2(3x_2) & 10 \tanh(3x_2) & 0 \\ 30c_{21} \operatorname{sech}^2(3x_1) & -10 & 0 & 10 \tanh(3x_1) \\ 3 \operatorname{sech}^2(3x_1) \tanh(3x_2) & 3 \operatorname{sech}^2(3x_2) \tanh(3x_1) & -1 & 0 \\ 3 \operatorname{sech}^2(3x_1) \tanh(3x_2) & 3 \operatorname{sech}^2(3x_2) \tanh(3x_1) & 0 & -1 \end{pmatrix}. \quad (4.3)$$

The eigenvalues of the Jacobian at each of the three equilibrium points are given in Table 4.3. Since all of the eigenvalues of the Jacobian are negative at equilibrium points #1 and #2, both of

	eigenvalues			
	λ_1	λ_2	λ_3	λ_4
Equilibrium #1	-10	-0.966224	-10.0338	-1
Equilibrium #2	-10	-0.966224	-10.0338	-1
Equilibrium #3	-1	-10	1.58872	-12.5887

Table 4.3: Eigenvalues of the Jacobian at the equilibrium points

these points are stable equilibria. Since the eigenvalues of the Jacobian at equilibrium #3 are both positive and negative, this equilibria is a saddle point. Notice that the eigenvalues of the Jacobian are in fact real valued. Further, since the eigenvalues are distinct, the Jacobian is diagonalizable at each equilibrium point. Note that the eigenvalues of the Jacobian at each equilibrium point are all nonzero, which implies that all three equilibrium points are isolated [25].

The gradient potential function $V(\mathbf{u})$, and associated matrix $\mathbf{P}(\mathbf{u})$, which lead to these dynamics are given by Equations (3.27) and (3.28) respectively. In this example the gradient potential $V(\mathbf{u})$ is

$$V(\mathbf{u}) = -\frac{1}{2} [\tanh(3x_1) c_{12} \tanh(3x_2) + \tanh(3x_2) c_{21} \tanh(3x_1)] \\ + \int_0^{x_1} 3(\zeta_1 - K_1) \operatorname{sech}^2(3\zeta_1) d\zeta_1 + \int_0^{x_2} 3\zeta_2 \operatorname{sech}^2(3\zeta_2) d\zeta_2 + \frac{1}{4} [c_{12}^2 + c_{21}^2], \quad (4.4)$$

and the matrix $\mathbf{P}(\mathbf{u})$ is given by

$$\mathbf{P}(\mathbf{u}) = \begin{pmatrix} \frac{10}{3 \operatorname{sech}^2(3x_1)} & 0 & 0 & 0 \\ 0 & \frac{10}{3 \operatorname{sech}^2(3x_2)} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \quad (4.5)$$

with the state vector $\mathbf{u} = [x_1, x_2, c_{12}, c_{21}]^T$. Graphing $V(\mathbf{u})$ with respect to the state variables

gives the surface along which the trajectories are constrained to move. Unfortunately in this example that surface is 5-dimensional. For a gradient-like system, Theorem 2.8 proves that the value of $V(\mathbf{u})$ can only decrease or remain constant with time for any trajectory. Figure 4.5 shows the variation of $V(\mathbf{u})$ over time for one of the trajectories which approaches equilibrium #1. Different trajectories, all ending at equilibrium #1, have different initial values of $V(\mathbf{u})$, but

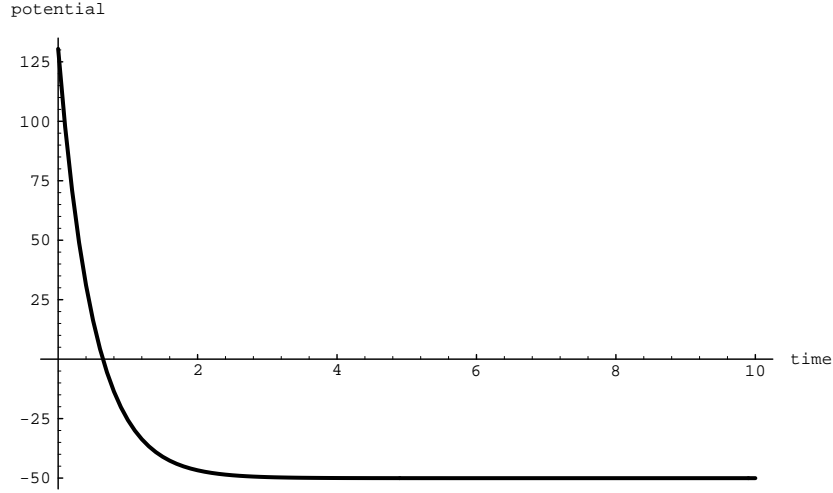


Figure 4.5: Change of the gradient potential over time

obviously must have the same final value of $V(\mathbf{u})$. The same is true of trajectories ending at equilibrium #2. Interestingly, the final values of $V(\mathbf{u})$ at both equilibrium #1 and #2 are identical. In a gradient-like system, Corollary 2.2 shows that all local minima of $V(\mathbf{u})$ are asymptotically stable equilibrium points. From calculus, $\tilde{\mathbf{u}}$ is a minimum of $V(\mathbf{u})$ if $\nabla_{\mathbf{u}}V(\tilde{\mathbf{u}}) = \mathbf{0}$ and $\nabla_{\mathbf{u}}^2V(\tilde{\mathbf{u}})$ is positive definite. Evidently, for all three equilibrium points in this example $\nabla_{\mathbf{u}}V(\tilde{\mathbf{u}}) = \mathbf{0}$. For this system the Hessian of $V(\mathbf{u})$ is

$$\nabla_{\mathbf{u}}^2 V(\mathbf{u}) = \begin{pmatrix} 3 \operatorname{sech}^2(3x_1) & & & & \\ -18(-50 + x_1) & -4.5(c_{12} + c_{21}) & & & \\ \operatorname{sech}^2(3x_1) \tanh(3x_1) & \operatorname{sech}^2(3x_1) & -1.5 \operatorname{sech}^2(3x_1) & -1.5 \operatorname{sech}^2(3x_1) & \\ +9(c_{12} + c_{21}) \operatorname{sech}^2(3x_1) & \operatorname{sech}^2(3x_2) & \tanh(3x_2) & \tanh(3x_2) & \\ \tanh(3x_1) \tanh(3x_2) & & & & \\ & -4.5(c_{12} + c_{21}) & 3 \operatorname{sech}^2(3x_2) & & \\ & \operatorname{sech}^2(3x_1) & -18x_2 \operatorname{sech}^2(3x_2) \tanh(3x_2) & -1.5 \operatorname{sech}^2(3x_2) & -1.5 \operatorname{sech}^2(3x_2) \\ & \operatorname{sech}^2(3x_2) & +9(c_{12} + c_{21}) \operatorname{sech}^2(3x_2) & \tanh(3x_1) & \tanh(3x_1) \\ & & \tanh(3x_1) \tanh(3x_2) & & \\ & & & 0.5 & 0 \\ & & & & \\ & & & & \\ & -1.5 \operatorname{sech}^2(3x_1) & -1.5 \operatorname{sech}^2(3x_2) & & \\ & \tanh(3x_2) & \tanh(3x_1) & & \\ & & & & \\ & -1.5 \operatorname{sech}^2(3x_1) & -1.5 \operatorname{sech}^2(3x_2) & & \\ & \tanh(3x_2) & \tanh(3x_1) & & \\ & & & 0 & 0.5 \end{pmatrix}. \quad (4.6)$$

The value of $V(\mathbf{u})$ and the eigenvalues of $\nabla_{\mathbf{u}}^2V(\mathbf{u})$, at each equilibrium point, are shown in

Table 4.4. Since $\nabla_{\mathbf{u}}^2 V(\mathbf{u})$ is symmetric and since its eigenvalues at equilibria #1 and #2 are

	potential	eigenvalues			
	$V(\bar{\mathbf{u}})$	λ_1	λ_2	λ_3	λ_4
Equilibrium #1	-50.0387	$0 < \lambda_1 \ll 10^{-6}$	0.500989	0.0295233	0.5
Equilibrium #2	-50.0387	$0 < \lambda_1 \ll 10^{-6}$	0.500989	0.0295233	0.5
Equilibrium #3	-49.769	$0 < \lambda_1 \ll 10^{-6}$	4.21221	0.5	-0.712214

Table 4.4: Potential value and eigenvalues of the Hessian, at the equilibrium points

all positive, these two points are minima of $V(\mathbf{u})$. This means that equilibria #1 and #2 are approached asymptotically by any trajectory started in their respective regions of attraction. The extremely small size of one of the eigenvalues at both of these points can be explained as follows. Notice that each of the elements in the first row and column of Equation (4.6) contains the term $\text{sech}^2(3x_1)$. The large value of x_1 at equilibria #1 and #2 makes this term quite small, however it is never zero or negative. This in turn causes every entry in row and column one to be very small but positive. So the size of the eigenvalue is not due to numerical error. This explanation is supported by the fact that the difference between the values of $V(\mathbf{u})$ at equilibrium #1 or #2 and equilibrium #3 is very small. Conceptually all of this implies that $V(\mathbf{u})$ is very flat along certain directions. This phenomena is illustrated in Figure 4.6. This figure shows four cross sections of

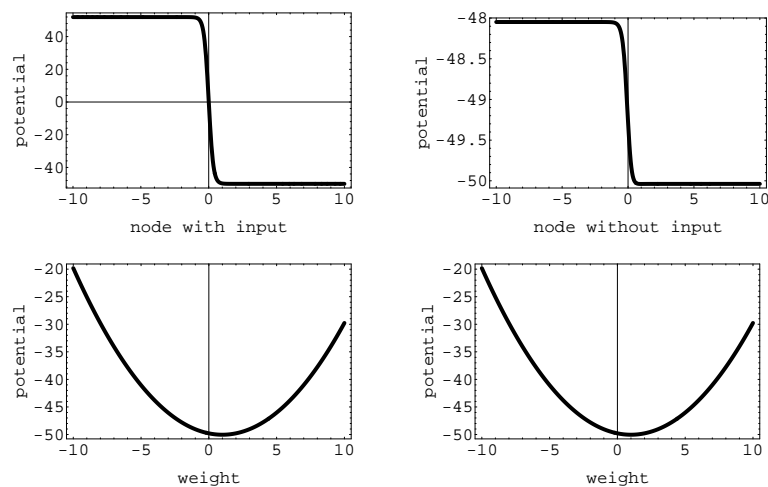


Figure 4.6: Change in the gradient potential with the state variables at equilibrium #1

the graph of the potential $V(\mathbf{u})$ with respect to the states. For each of the sections, the state values that are not displayed are set to their values at equilibrium #1. The cross sections with respect to the two node activation values, x_1 and x_2 , are clearly quite flat. Since the equilibrium of x_1 is approximately equal to the value of the input, making the input smaller will make $V(\mathbf{u})$

more curved in the neighborhood of the equilibria. In fact, this was observed to be the case in simulation. It is clear from the top two panels of Figure 4.6 that the set

$$\mathcal{N}_c = \{\mathbf{u} \in \mathbb{R}^n : V(\mathbf{u}) \leq c\} \tag{4.7}$$

is neither bounded nor closed in this example. Therefore the conditions of Theorem 2.11 are violated, and the system is not guaranteed to converge to one of the equilibrium for every set of initial conditions. In spite of this, no set of initial conditions has yet been found that does not converge to one of the three equilibria.

This analysis can be used to study more complex networks. The number of state variables in a fully connected network with p nodes is $p^2 + p$. This means that the number of dimensions in the state space increases rapidly with increasing network size. This increase causes the phase space and potential cross sections to have a much more complex structure. However this formalism is still an extremely useful one for analyzing high dimensional networks.

Chapter 5

Conclusion

The behavior of a neural network that can be written in the form of equation (3.23) is described by the theorems in Section 2.2. Conceptually the node activation and connection weight values must, if possible, move downhill along the surface of the function $V(\mathbf{u})$ defined in equation (3.20). If they can no longer move downhill in one or more directions, then they must remain in place. This is shown by Theorem 2.8 and Corollary 2.2. The activations and weights of the neural network cannot cycle periodically, their values must go asymptotically either to an equilibrium point or to infinity. This is shown by Theorems 2.9 and 2.10. A restriction on the gradient potential function $V(\mathbf{u})$ is provided in Theorem 2.11 which guarantees that the activations and weights must go to one of their equilibrium values no matter where the system is started. An example of a gradient potential function which satisfies this restriction is one which is bounded below and radially unbounded. Close to any equilibrium point the trajectories will be in the shape of either a hyperbola, parabola or a line. Theorem 2.12 shows this. The most important direction of future research is extending this treatment to systems which have asymmetric connection matrices. This can be accomplished by decomposing the network dynamics into the sum of a gradient-like system and a non-gradient-like system. A less restrictive way to incorporate multiplicative dynamics into this formalism will also be investigated.

Bibliography

- [1] S.-I. Amari. Neural theory of association and concept-formation. *Biological Cybernetics*, 26(2):175–185, 1977.
- [2] G.A. Carpenter. Neural network models for pattern recognition and associative memory. *Neural Networks*, 2(4):243–257, 1989.
- [3] M.A. Cohen and S. Grossberg. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man and Cybernetics*, 13(5):815–825, 1983.
- [4] P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2):165–170, 1990.
- [5] F.R. Gantmacher. *The theory of matrices*, volume 1. Chelsea Publishing Co., New York, NY, 1977.
- [6] C.L. Giles and T. Maxwell. Learning, invariance, and generalization in high-order neural networks. *Applied Optics*, 26(23):4972–4978, 1987.
- [7] S. Grossberg. Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt, editor, *Delay and Functional Differential Equations and their Applications*, pages 121–160. Academic Press, Inc., San Diego, CA, 1972.
- [8] S. Grossberg. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52(3):213–257, 1973.
- [9] S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1(1):17–61, 1988.
- [10] J. Guckenheimer and P. Holmes. *Nonlinear oscillations, dynamical systems and bifurcations of vector fields*, volume 42 of *Applied Mathematical Sciences*. Springer-Verlag, Inc., New York, NY, 1983.

- [11] M.W. Hirsch and S. Smale. *Differential equations, dynamical systems, and linear algebra*, volume 60 of *Pure and Applied Mathematics*. Academic Press, Inc., San Diego, CA, 1974.
- [12] H.K. Khalil. *Nonlinear systems*. Macmillan Publishing Co., New York, NY, 1992.
- [13] B. Kosko. Adaptive bidirectional associative memories. *Applied Optics*, 26(23):4947–4960, 1987.
- [14] B. Kosko. *Neural networks and fuzzy systems: A dynamical systems approach to machine intelligence*. Prentice Hall, Inc., Englewood Cliffs, NJ, 1992.
- [15] J.P. LaSalle and S. Lefschetz. *Stability by Lyapunov's direct method, with applications*, volume 4 of *Mathematics in Science and Engineering*. Academic Press, Inc., New York, NY, 1st edition, 1961.
- [16] L. Markus. Structurally stable differential systems. *Annals of Mathematics*, 73(1):1–19, 1961.
- [17] S.E. Newhouse. Nondensity of axiom A(a) on S^2 . In S.-S. Chern and S. Smale, editors, *Global Analysis*, volume XIV of *Proceedings of Symposia in Pure Mathematics*, pages 191–202. American Mathematical Society, 1970.
- [18] J. Palis and S. Smale. Structural stability theorems. In S.-S. Chern and S. Smale, editors, *Global Analysis*, volume XIV of *Proceedings of Symposia in Pure Mathematics*, pages 223–231. American Mathematical Society, 1970.
- [19] M.M. Peixoto. Structural stability on 2-dimensional manifolds. *Topology*, 1(2):101–120, 1962.
- [20] D. Psaltis, C.H. Park, and J. Hong. Higher order associative memories and their optical implementations. *Neural Networks*, 1(2):149–163, 1988.
- [21] F.M.A. Salam, Y. Wang, and M.R. Choi. On the analysis of dynamic feedback neural nets. *IEEE Transactions on Circuits and Systems*, 38(2):196–201, 1991.
- [22] S. Smale. Structurally stable systems are not dense. *American Journal of Mathematics*, 88(2):491–496, 1966.
- [23] G. Strang. *Linear algebra and its applications*. Harcourt Brace Jovanovich, Inc., San Diego, CA, 3rd edition, 1988.
- [24] F. Verhulst. *Nonlinear differential equations and dynamical systems*. Springer-Verlag, Inc., Berlin, Germany, 1990.

- [25] M. Vidyasagar. *Nonlinear systems analysis*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1st edition, 1978.