

Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks

Richard H. R. Hahnloser

rhahnloser@mit.edu

H. Sebastian Seung

seung@mit.edu

Howard Hughes Medical Institute, Department of Brain and Cognitive Sciences, MIT E25-210, Cambridge, MA 02139, U.S.A.

Jean-Jacques Slotine

jjs@mit.edu

Department of Mechanical Engineering and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

The richness and complexity of recurrent cortical circuits is an inexhaustible source of inspiration for thinking about high-level biological computation. In past theoretical studies, constraints on the synaptic connection patterns of threshold-linear networks were found that guaranteed bounded network dynamics, convergence to attractive fixed points, and multistability, all fundamental aspects of cortical information processing. However, these conditions were only sufficient, and it remained unclear which were the minimal (necessary) conditions for convergence and multistability.

We show that symmetric threshold-linear networks converge to a set of attractive fixed points if and only if the network matrix is copositive. Furthermore, the set of attractive fixed points is nonconnected (the network is multiattractive) if and only if the network matrix is not positive semidefinite. There are permitted sets of neurons that can be coactive at a stable steady state and forbidden sets that cannot. Permitted sets are clustered in the sense that subsets of permitted sets are permitted and supersets of forbidden sets are forbidden. By viewing permitted sets as memories stored in the synaptic connections, we provide a formulation of long-term memory that is more general than the traditional perspective of fixed-point attractor networks. There is a close correspondence between threshold-linear networks and networks defined by the generalized Lotka-Volterra equations.

1 Introduction

Recurrent networks of neurons are believed to play an important role in the perceptual interpretation and memorization of sensory events. The dynamics of simplified networks are often studied in a regime where as a response to constant sensory stimulation, neural activities converge to a steady state. Such networks can efficiently implement complex nonlinear mappings between input stimulation and response activity. In a very general framework, we characterize the geometry of potentially multiple stable, steady responses and show that in addition to the dependence on synaptic connections, steady responses also depend on the inherent ambiguity of the sensory input.

A Lyapunov function can be used to prove that a fixed point of a set of differential equations is stable. By extension, for some differential equations that have many distinct fixed points, Lyapunov-like functions can be constructed with multiple local minima, each corresponding to a stable fixed point. These local minima were previously interpreted as memorized patterns that can be recalled by suitable initialization. Lyapunov theory applies mainly to symmetric networks in which neurons have monotonic activation functions (Hopfield, 1984; Cohen & Grossberg, 1983). Here we show that the restriction of activation functions to threshold linear ones can yield new insights into the computational behavior of recurrent networks. The results we present elaborate on previous work (Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000; Hahnloser & Seung, 2001).

We present three main theorems about the neural responses to constant inputs. The first theorem provides necessary and sufficient conditions on the synaptic weight matrix for the existence of a set of globally attractive fixed points. These conditions can be expressed in terms of copositivity, a concept from quadratic programming and linear complementarity theory. Alternatively, they can be expressed in terms of certain eigenvalues and eigenvectors of submatrices of the synaptic weight matrix, making a connection to linear systems theory. The theorem guarantees that the network will produce a steady-state response to any constant input. We regard this response as the computational output of the network, and its characterization is the topic of the second and third theorems.

In the second theorem, we introduce the idea of permitted and forbidden sets. Under certain conditions on the synaptic weight matrix, we show that there exist sets of neurons that are "forbidden" by the recurrent synaptic connections from being coactivated at a stable steady state, no matter what input is applied. Other sets are "permitted," in the sense that they can be stably coactivated for some input. Our main result is that if and only if a network possesses a forbidden set is it conditionally multiattractive, that is, there exists an input for which there is a nonconnected set of attractive fixed points, each of which can be retrieved by suitable initial conditions.

Hence, we find that forbidden sets and conditional multiattractiveness are inseparable concepts.

The conceptual importance of permitted and forbidden sets suggests a new way of thinking about memory in neural networks. When an input is applied, the network must select a set of active neurons, and this selection is constrained to be one of the permitted sets. Therefore, the permitted sets can be regarded as memories stored in the synaptic connections. The memory storage in threshold-linear networks is limited by the topological properties of the memories. Our third theorem states that there are constraints on the groups of permitted and forbidden sets that can be stored by a network. No matter which learning algorithm is used to store memories, active neurons cannot arbitrarily be divided into permitted and forbidden sets, because subsets of permitted sets have to be permitted and supersets of forbidden sets have to be forbidden.

In the appendix, we show that symmetric threshold-linear networks have the same set of attractive fixed points as do networks defined by the generalized Lotka-Volterra equations.

2 Threshold-Linear Network Equations

Our theory is applicable to the network dynamics,

$$\frac{dx_i}{dt} + x_i = \left[b_i + \sum_j W_{ij}x_j \right]^+, \quad (2.1)$$

where $[u]^+ = \max\{u, 0\}$ is a (half-wave) rectification nonlinearity and the synaptic weight matrix is symmetric, $W_{ij} = W_{ji}$. The dynamics can also be written in a more compact matrix vector form as $\dot{x} + x = [b + Wx]^+$. The state of the network is x . An input to the network is an arbitrary vector b . An output of the network is a steady state $\underline{x}(\dot{x} = 0)$ in response to b . The existence of outputs and their relationship to the input are determined by the synaptic weight matrix W . Threshold-linear units were introduced to neural modeling by Hartline and Ratliff (1958). Under certain circumstances, networks of conductance-based model neurons can be reduced to nonspiking dynamics like equation 2.1 (Ermentrout, 1994). In this context, the term in square brackets is interpreted as a spike rate, whereas x_i is interpreted as a synaptic activation variable.

Definition 1. *A vector v is said to be nonnegative, $v \geq 0$, if all of its components are nonnegative. Similarly, a vector is said to be positive if all of its components are positive. The nonnegative orthant $\{v: v \geq 0\}$ is the set of all nonnegative vectors.*

It can be shown that any trajectory of equation 2.1 starting in the nonnegative orthant remains in the nonnegative orthant. Therefore, for simplicity, we will consider initial conditions that are confined to the nonnegative orthant.

3 Convergence to an Equilibrium

The following theorem establishes necessary and sufficient conditions for a symmetric network of the form in equation 2.1 to converge always to a steady state.

Definition 2. Let A be an $N \times N$ matrix, and let $\Sigma = (i_1, \dots, i_k)$ be a set of k ordered indices, $1 \leq i_1 < i_2 < \dots < i_k \leq N$. The matrix A^Σ is said to be a $k \times k$ submatrix of A if $A_{uv}^\Sigma = A_{i_u i_v}$ for $\forall u, v \in \Sigma$. The matrix A^Σ can be constructed from A simply by removing from A all rows and columns not indexed by Σ .

Theorem 1. Assuming that W is symmetric, the following statements are equivalent:

1. All positive eigenvectors of all submatrices of $I - W$ have positive eigenvalues.
2. The matrix $I - W$ is copositive; that is, $x^T(I - W)x > 0$ for all nonnegative x , except $x = 0$.
3. For all constant b and all initial conditions, the dynamics converge to an equilibrium point.

Proof.

- *Statement 2 follows from statement 1.* We will assume that $I - W$ is not copositive and show that there exists a positive eigenvector of a submatrix of W with nonpositive eigenvalue. Let v^* be the minimum of $Q = v^T(I - W)v$ over nonnegative v on the unit sphere, $v^* = \operatorname{argmin}_{v \geq 0} Q$, where $G = Q - \alpha v^T v$ (α is a Lagrange multiplier). Define by $\Sigma = (i_1, \dots, i_k)$ the k positive components of v^* , $v_{i \in \Sigma}^* > 0$. The gradient ∇G has to vanish in these components, and so for all $i \in \Sigma$, we have the Kuhn-Tucker conditions $\alpha v_i^* = v_i^* - \sum_j W_{ij} v_j^* = v_i^* - \sum_{j \in \Sigma} W_{ij} v_j^*$. Thus, the vector $v_{i \in \Sigma}^*$ is a positive eigenvector of the matrix $(I - W)^\Sigma$ with eigenvalue α . And because the minimum $Q(v^*) = \alpha$ is nonpositive by assumption, we find that the eigenvalue of this positive eigenvector is nonpositive.
- *Statement 3 follows from statement 2.* We show that given the copositivity of $I - W$, we can find a Lyapunov-like function that is strictly decreasing under the dynamics in equation 2.1 and is zero only at steady states. By the copositivity of $I - W$, the function $L = \frac{1}{2} x^T(I - W)x - b^T x$ is lower bounded and radially unbounded in the nonnegative orthant, $x \geq 0$. For all fixed b , L is nonincreasing under the network dynamics:

$\dot{L} = \dot{x}^T(x - Wx - b) = -(x - [Wx + b]^+)^T(x - Wx - b) = -(x - y^+)^T(x - y) \leq 0$, where $y = Wx + b$. This last inequality follows from the fact that all components of x and y contribute to the decrease of L : (1) if a component i satisfies $y_i \geq 0$, we have that $y_i^+ = y_i$ and so $-(x_i - y_i^+)(x_i - y_i) = -(x_i - y_i)^2 \leq 0$; (2) if a component j satisfies $y_j < 0$, we have that $-(x_j - y_j^+)^T(x_j - y_j) = -x_j(x_j - y_j) \leq 0$ (because $x_j \geq 0$).

All equilibria of L correspond to steady states of the dynamics of equation 2.1. From $\dot{L} = -(x - y^+)^T(x - y) = 0$, it follows that for every component i , we have that either $(x_i - y_i^+) = 0$ or $(x_i - y_i) = 0$. If $x_i - y_i^+ = 0$, then by definition $\dot{x}_i = 0$; if $x_i - y_i = 0$, then we must have that $y_i = y_i^+$, which by positivity of x_i implies $\dot{x}_i = 0$.

Hence, for all b and all initial conditions, the network converges to an equilibrium.

- *Statement 1 follows from statement 2.* By contradiction, assume there is a positive eigenvector v^Σ of a submatrix $(I - W)^\Sigma$ with eigenvalue $\lambda \leq 0$. Define the vector v by $v_i = v_i^\Sigma$ if $i \in \Sigma$ and $v_i = 0$ otherwise. Then $v^T(I - W)v = \lambda v^T v \leq 0$, and therefore $I - W$ is not copositive.
- *Statement 2 follows from statement 3.* By contradiction, assume that $I - W$ is not copositive. From the equivalence of statement 2 with statement 1, we know that the positive components Σ of the minimum $v^* = \operatorname{argmin}_{v \geq 0} G$ correspond to a positive eigenvector of the submatrix $(I - W)^\Sigma$ with nonpositive eigenvalue $\alpha \leq 0$. According to the Kuhn-Tucker conditions, minimization also implies that the gradient ∇G restricted to the vanishing components Z of v^* ($v_{i \in Z}^* = 0$) has to be nonnegative: $-\alpha v_i^* + v_i^* - \sum_j W_{ij} v_j^* = -\sum_j W_{ij} v_j^* \geq 0$ for $i \in Z$. Thus, by choosing inputs $b_\Sigma = v_\Sigma^*$ and $b_Z = 0$ as well as initial conditions $x_\Sigma(0) = v_\Sigma^*$ and $x_Z(0) = 0$, we have constructed a nonconverging trajectory: if $\alpha < 0$, this trajectory is given by $x_\Sigma(t) = (e^{-\alpha t} + \frac{1}{\alpha})v_\Sigma^*$ and $x_Z(t) = 0$; if $\alpha = 0$, it is given by $x_\Sigma(t) = tv_\Sigma^*$ and $x_Z(t) = 0$.

The function L that was constructed for the proof of theorem 1 is locally a Lyapunov function in the sense that the stable fixed points of the dynamics correspond to local minima of L . Conversely, we also have that all local minima of L correspond to stable fixed points. We will again make use of L in the next section, where we characterize the geometry of attractive fixed points. The fact that L is a quadratic function will be particularly useful.

Theorem 1 states sufficient and necessary conditions that guarantee bounded dynamics, the absence of limit cycles, and the absence of chaos. The first two conditions for convergence in theorem 1 are best appreciated by comparison with the conditions for convergence of a purely linear network constructed from equation 2.1 by dropping the rectification nonlinearity. Let us first consider statement 1: in a linear network, all eigenvalues of $I - W$ have to be positive to ensure convergence. In theorem 1, only eigenvalues of nonnegative eigenvectors have to be positive to ensure convergence. Eigen-

values of nonpositive eigenvectors may well be nonpositive, a fact that will turn out to be important for multistability, as we shall see. But all submatrices of $I - W$ must be considered in theorem 1, because different sets of feedback connections are active for different sets of neurons that are above threshold. Let us now consider statement 2: in a linear network, $I - W$ would have to be positive definite to ensure asymptotic stability, but because of the rectification, in theorem 1 positive definiteness is replaced by the weaker condition of copositivity.

Although theorem 1 implies convergence to a fixed point, it does not imply that there is only a single fixed point. As we shall see, for example, there can be an attractive line of fixed points (a line attractor) toward which the dynamics must converge. Nevertheless, theorem 1 excludes line attractor networks of the integrator type (Seung, 1996), that is, where there exists a positive eigenvector with zero eigenvalue, such as in $\dot{x} + x = [x + b]^+$. In such networks, a positive input is indefinitely integrated, resulting in unbounded behavior.

Lemma 1. *For any nonnegative vector $v \geq 0$, there exists an input b , such that v is a steady state of equation 2.1 with input b .*

Proof. Choose $b = (I - W)v$.

Lemma 1 states that any nonnegative vector can be realized as a fixed point. Sometimes this fixed point is stable, such as in symmetric and copositive networks in which only a single neuron is active. Indeed, the submatrix of $I - W$ corresponding to a single active neuron corresponds to a diagonal element, which according to statement 1 of theorem 1 must be positive. Hence, it is always possible to activate only a single neuron at an asymptotically stable fixed point. However, as will become clear from theorem 2, not all nonnegative vectors can be realized as an asymptotically stable fixed point.

4 Forbidden and Permitted Sets

In this section, we present a geometrical description of the set of attractive fixed points by introducing the notion of forbidden and permitted sets of neurons. Recall first the classical definition of stability in the sense of Lyapunov.

Definition 3. *A steady state \underline{x} is stable if for all balls B with $\underline{x} \in B$, there exists a ball A with $\underline{x} \in A$, such that for any initial condition $x(0) \in A$, we have that $x(t) \in B$ for all times $t \geq 0$. A steady state is asymptotically stable if it is stable and there is a ball of initial conditions that converge to it.*

Definition 4. *A neuron is said to be activated if its activity is nonzero.*

Definition 5. *A set of neurons is permitted if the neurons can be coactivated at a stable steady state for some input b . A set of neurons is forbidden if it is not permitted, that is, if the neurons cannot be coactivated at a stable steady state no matter what the input b .*

According to these definitions, permitted and forbidden sets are network characteristics that are independent of the input. Their significance can best be appreciated by comparison to linear systems. In a linear network, any set of neurons is permitted if and only if all eigenvalues of $I - W$ are nonnegative (the linear system is stable). This means that either all sets of neurons are permitted or none. For a threshold-linear network, this is not the case. It is possible that some sets are permitted, whereas others are forbidden. Whether a particular set is permitted depends on the eigenvalues of $(I - W)^\Sigma$. For example, a given set Σ is forbidden means that $(I - W)^\Sigma$ must have at least one negative eigenvalue. And by copositivity of $I - W$, we know that the corresponding eigenvector cannot be nonnegative and so must have both positive and negative components.

Definition 6. *A threshold-linear network is said to be conditionally multi-attractive if there exists an input b such that the set of stable equilibrium points is disconnected.*

Multiattractiveness is a statement about the topology of stable fixed points. The word *conditional* refers to the fact that multiattractiveness exists only for certain inputs, but not for others. For example, consider a network and a particular input for which all stable fixed points lie on a line. The network is not multiattractive for this input, because from any point of the line, any other point of the line can be reached by staying on the line only (i.e., a line is connected). However, it may well be that for another input, the network has two stable fixed points that are separate from each other, which would characterize the network as conditionally multiattractive.

To prove our theorems about conditional multiattractiveness and forbidden sets, we will need the following lemma about the minimization of convex functions over convex sets (Bertsekas, 1995).

Definition 7. *A set C is convex if $\alpha x + (1 - \alpha)y \in C$ for all $x, y \in C$, and $0 \leq \alpha \leq 1$ (if any line between two points remains in C).*

Definition 8. *A scalar function f is convex if $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$ for all $x, y \in C$, and $0 \leq \alpha \leq 1$ (if any line between two points does not pass below the function f).*

Lemma 2. (a) If C is a convex set (such as the nonnegative orthant) and f a convex function on C , then a local minimum of f is also a global minimum. (b) The set of global minima of f is convex (and thus connected).

Proof. (a) Let x be a local minimum of f but not a global minimum. Then there exists a y such that $f(y) < f(x)$. From the convexity of f , we have that $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \leq f(x)$ for every $\alpha \in [0, 1]$. This contradicts the assumption that x is a local minimum. (b) Assume there exist two global minima $x, y \in C$ and $\alpha \in [0, 1]$ such that $z = \alpha x + (1 - \alpha)y$ is not a global minimum, $f(z) > f(x) = f(y)$. By convexity of f , we have that $f(z) \leq \alpha f(x) + (1 - \alpha)f(y) = f(x)$, contradicting the assumption that z is not a global minimum.

Definition 9. A symmetric $N \times N$ matrix A is positive semidefinite if for all vectors x we have that $x^T A x \geq 0$.

If a matrix is positive semidefinite, then all its eigenvalues are nonnegative. We will use that if $I - W$ is a positive semidefinite matrix, then for all inputs b , the Lyapunov-like energy function L is convex (this follows from the fact that L is quadratic).

We will also need the following interlacing theorem for eigenvalues of symmetric matrices.

Lemma 3. Let A be a symmetric $N \times N$ matrix with real eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$, and let A' be a $(N - 1) \times (N - 1)$ submatrix of A . Then the eigenvalues $\eta_1 \leq \dots \leq \eta_{N-1}$ of A' interlace the eigenvalues of A : $\lambda_1 \leq \eta_1 \leq \lambda_2 \leq \dots \leq \eta_{N-1} \leq \lambda_N$.

Proof. The proof is given in Horn and Johnson (1985).

As a corollary of lemma 3, if A has k negative eigenvalues, then A' must have at least $k - 1$ negative eigenvalues.

A symmetric threshold-linear network has a unique stable, steady state if and only if $I - W$ is positive definite (all eigenvalues are positive; Feng & Haderler, 1996). As an extension, the following theorem contains necessary and sufficient conditions for symmetric threshold-linear networks to be conditionally multiattractive.

Theorem 2. Assume that the matrix $I - W$ is symmetric and copositive. The following statements are equivalent:

1. The matrix $I - W$ is not positive semidefinite.
2. There exists a forbidden set.
3. The network is conditionally multiattractive.

Proof.

- *Statement 2 follows from statement 1.* $I - W$ is not positive semidefinite, and so $I - W$ has a negative eigenvalue. Consider a steady state with all neurons active. Because of this negative eigenvalue, the steady state cannot be stable, and so the set of all neurons is forbidden.
- *Statement 1 follows from statement 2.* Let Σ be the forbidden set. The smallest eigenvalue of $(I - W)^\Sigma$ must be negative. By the interlacing theorem, the smallest eigenvalue of $I - W$ must be negative as well and so $I - W$ is not positive semidefinite.
- *Statement 1 follows from statement 3.* Suppose that statement 1 is false. Choose an input b , and let X be a nonconnected set of attractive fixed points of the network. By Lyapunov stability, X corresponds to the set of local minima of L . Because $I - W$ is positive semidefinite, L is a convex function. By lemma 2, the local minima of L in the nonnegative orthant are connected, contradicting our initial assumption that X is nonconnected.
- *Statement 3 follows from statement 1.* The plan is to construct an input for which the network is multiattractive. $I - W$ is not positive semidefinite, and so must have at least one negative eigenvalue. If $I - W$ has more than one negative eigenvalue, then define a sequence of subsets $\Sigma_N \supset \Sigma_{N-1} \supset \dots \supset \Sigma_k = \Sigma$ by removing neurons one by one until only one of the k eigenvalues of $(I - W)^\Sigma$ is negative. By the corollary of the interlacing theorem, such a set Σ can always be found. The forbidden set Σ contains at least $k \geq 2$ neurons because by copositivity, any single-neuron set is permitted. Denote the negative eigenvalue of $(I - W)^\Sigma$ by λ , and let v be the corresponding eigenvector. Without loss of generality, assume that $v = (v^p, v^n)$ has q nonnegative and $k - q$ negative components, $v_i^p = v_i \geq 0$ for $i = 1, \dots, q$ and $v_i^n = v_{q+i} < 0$ for $i = 1, \dots, k - q$. Define a nonnegative vector $y = \left(\frac{v_p}{\|v_p\|^2}, -\frac{v_n}{\|v_n\|^2} \right)$ that is orthogonal to v , and define the input to the network as $b_i = y_i$ for $i \in \Sigma$ and $b_i = -K \ll 0$ for $i \notin \Sigma$. The value K has to be sufficiently large to prevent neurons not belonging to Σ from being activated during the dynamics (how to construct K is shown after the next paragraph).

Define a hyperplane H passing through b , spanned by the $k - 1$ stable eigenvectors orthogonal to v . The hyperplane divides N_Σ —the nonnegative orthant restricted to Σ —into two disjoint regions. The function L restricted to this hyperplane is convex, so by lemma 2, it has a connected set of local minima that are also global minima. Choose an arbitrary minimum $z = \operatorname{argmin}_{x \in H} L(x)$, and define two initial conditions of the dynamics, both contained in N_Σ but on opposite sides of the hyperplane, $x_i^1 = z_i + \epsilon v_i$, $x_i^2 = z_i - \epsilon v_i$ for $i \in \Sigma$, and $x_i^1 = 0$, $x_i^2 = 0$ for $i \notin \Sigma$, with $\epsilon \ll 1$. Because $L(x^1) = L(x^2) = L(z) + \frac{\lambda}{2} \epsilon^2 < L(z)$ and because L is nonincreasing along

trajectories, the two trajectories will not cross the hyperplane, but will converge toward two sets of attractive fixed points X^1 and X^2 on either side of the hyperplane.

How to choose K : Define a region $C \subset N_\Sigma$ by $C = \{x \geq 0 \mid x \in N_\Sigma, L(x) < L(z)\}$. Because L is radially unbounded, C is contained in a ball of radius R centered at the origin. By choosing $K = R\|w\| + 1$ (where $\|\cdot\|$ denotes an arbitrary matrix norm), we guarantee that $\sum_j w_{ij}x_j < b_i$ for all $i \notin \Sigma$ and all $x \in C$, and thus neurons that are initially inactive remain inactive for all times.

It remains to be shown that X^1 and X^2 are sets of fixed points that are not connected, neither by a curve confined within N_Σ nor by a curve extending beyond N_Σ . First, X^1 and X^2 are not connected by a curve Γ_1 within N_Σ because this curve would have to cross the hyperplane, implying that there exists a point $u \in H$ for which $L(u) < L(x^1) < L(z)$, contradicting the assumption that z is a global minimum of L restricted to H . Second, X^1 and X^2 are not connected by a curve Γ_2 extending beyond N_Σ , because the strongly inhibitory inputs prevent the existence of fixed points in a small neighborhood around N_Σ . By contradiction, assume there exists a fixed point $q \notin N_\Sigma$ on a curve c connecting X^1 and X^2 and satisfying $\|q - p\| < \frac{1}{\|w\|}$, where $p \in c$ and $p \in N_\Sigma$. Because $q \notin N_\Sigma$, at least one component $i \notin \Sigma$ of the steady state q must be positive, $q_i > 0$. Thus, the total input to neuron i must be positive: $0 < \sum_j w_{ij}q_j - K = \sum_j w_{ij}(q_j - p_j + p_j) - K \leq \|w\|\|q - p\| + \sum_j w_{ij}p_j - K \leq 1 + \|w\|R - K = 0$, a contradiction. In conclusion, we have constructed an input b for which there exist two nonconnected sets X^1 and X^2 of attractive fixed points (see Figure 1).

How does our new concept of multiattractiveness compare to the older concept of multistability? Given our definition of multiattractiveness, multistability is a subtly more general concept: whereas a single line attractor is multistable (it has multiple stable fixed points), it is not multiattractive. On the other hand, any multiattractive system is necessarily also multistable, implying that the difference between multistability and multiattractiveness is that the latter excludes single line (and surface) attractors.

Previous results about stability of rectified linear networks were concerned with either necessary and sufficient conditions for monostability (Feng & Hadel, 1996; Hadel & Kuhn, 1987) or sufficient conditions for multistability (Hahnloser, 1998; Wersing, Beyn, & Ritter, 2001). As an extension to these results, theorem 2 contains a sufficient and necessary condition for multiattractiveness: the existence of a forbidden set. To give an intuitive understanding of the equivalence of conditional multiattractiveness and forbidden sets, let us consider a one-dimensional system of the form $\dot{x} = -f(x)$. If such a system has to have two stable fixed points, then the function $F(x) = \int^x dy f(y)$ must have two separate local minima. And by the smoothness of f , it must be that F has a local maximum between the two local minima, that is, the system must have an instability. A similar

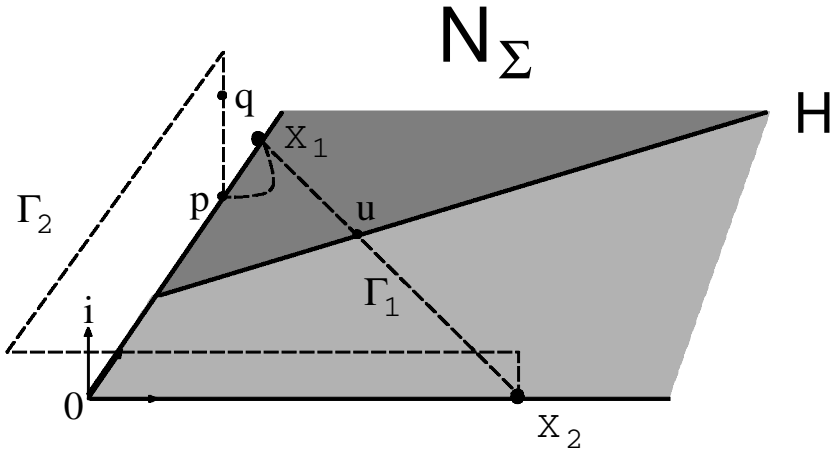


Figure 1: Proof of theorem 2. The nonnegative orthant N_Σ is two-dimensional and is drawn in the horizontal plane. The hyperplane H is a line dividing N_Σ into two disjoint regions (shaded areas). Because Σ is a forbidden set, the two sets of stable fixed points X_1 and X_2 must lie on the boundary of N_Σ . They are not connected, neither by a trajectory Γ_1 (straight line) passing inside N_Σ and intersecting H in u , nor by a trajectory Γ_2 passing outside N_Σ (along axis i), passing by points $p \in N_\Sigma$ and $q \notin N_\Sigma$.

principle also exists for two-dimensional systems, in which case f is a two-dimensional vector field. The index theorem says that the contour integral of f along a closed curve C equals the sum of indices of the isolated fixed points inside C (Strogatz, 1994). Maxima and minima have an index of 1, and saddles have an index of -1 . Hence, if f is radially unbounded, then the index of a contour integral along C equals 1. And if C is to contain two minima, then it must also contain a saddle: $1 = 2 - 1$. Hence, two-dimensional systems are multistable only if they contain an instability. Unfortunately, there is no similarly general result for multistability in dynamical systems of more than two dimensions. But we show here that for rectified linear networks, the intuitive generalization of the index theorem—the equivalence of multistability and instability—is true in arbitrary dimensions. Using theorem 2 and previous results, we can summarize all possible geometries of stable steady states in symmetric networks:

- $I - W$ is positive definite: There exists a single globally attractive fixed point.
- $I - W$ is copositive and positive semidefinite but not positive definite: Multiple steady states can exist in the form of a line (or surface) attractor. For example, the two-neuron network given by $W = (0, -1; -1, 0)$

results in eigenvalues 0 and 2 of $I - W$, and so the two-neuron set is permitted. Let $b = (1; 1)$; it follows that there is a line of attractive fixed points given by $x_1 + x_2 = 1$ and $x_1 \geq 0, x_2 \geq 0$. The fixed points are connected, and so the network is not multiattractive for this input. By theorem 2, the network is not conditionally multiattractive. Recall that no surface attractor of the integrator type is possible for copositive networks.

- $I - W$ is copositive but not positive semidefinite. There may exist multiple, nonconnected sets of attractive fixed points dependent on the input. These might either be separate points or nonintersecting line or surface attractors.

The following theorem characterizes the relationship between forbidden and permitted sets.

Theorem 3. *Assume that the matrix $I - W$ is symmetric and copositive. Any subset of a permitted set is permitted. Any superset of a forbidden set is forbidden.*

Proof. Use the interlacing theorem. If the smallest eigenvalue of a symmetric matrix is nonnegative, then so are the smallest eigenvalues of all its submatrices. And if the smallest eigenvalue of a symmetric submatrix is negative, then so is the smallest eigenvalue of the original matrix.

The hierarchical grouping of permitted sets holds only for symmetric networks and may be violated in nonsymmetric networks. Consider the two-neuron network defined by $w = (2, -1; 4, -2)$. The matrix $I - W$ has the unique eigenvalue 1, and so the two neurons form a permitted set. However, the first diagonal element is -1 , and so the first neuron, as a subset of a permitted set, forms a forbidden set, in disagreement with theorem 3. What happens in this network is that due to the strong self-excitation, the activity of the first neuron diverges until it necessarily activates the second neuron, the inhibition of which has a stabilizing effect.

5 An Example: The Ring Network

Symmetric threshold-linear networks with local excitation and more global inhibition have been studied in the past as models for cortical processing (Ben-Yishai, Lev Bar-Or, & Sompolinsky, 1995; Douglas, Koch, Mahowald, Martin, & Suarez, 1995). Here we argue that invariances of neural responses to sensory stimulation in these models arise from the existence of forbidden sets. We also show that many of these networks exhibit spurious permitted sets, consisting of sets of noncontiguous neurons.

Let the synaptic matrix of a 10-neuron ring network be translationally invariant. The connection between neurons i and j is given by $W_{ij} = -\beta + \alpha_0 \delta_{ij} + \alpha_1 (\delta_{i,j+1} + \delta_{i+1,j}) + \alpha_2 (\delta_{i,j+2} + \delta_{i+2,j})$, where β quantifies global inhi-

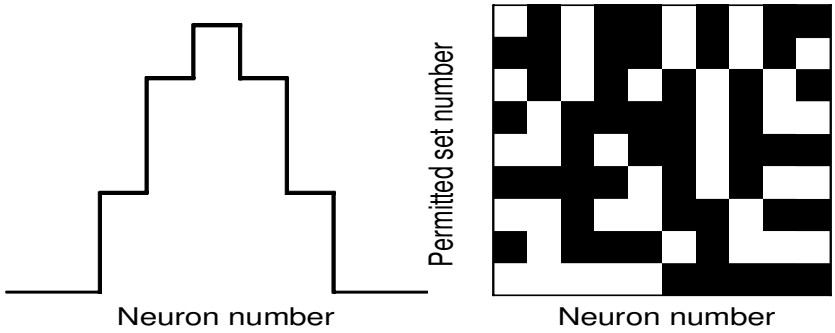


Figure 2: (Left) Output of a ring network of 10 neurons to uniform input (random initial condition). (Right) The 9-parent permitted sets (x -axis: neuron number; y -axis: set number). White means that a neuron belongs to a set, and black means that it does not. Left-right and translation-symmetric parent-permitted sets of the ones shown have been excluded (they can be obtained by translation of the sets shown). The first parent permitted set (first row from the bottom) corresponds to the output on the left.

bition, α_0 self-excitation, α_1 first-neighbor lateral excitation, and α_2 second-neighbor lateral excitation. In Figure 1, we numerically computed the permitted sets of this network, with the parameters taken from Hahnloser et al. (2000), that is, $\alpha_0 = 0$, $\alpha_1 = 1.1$, $\alpha_2 = 1$, and $\beta = 0.55$. The permitted sets were determined by diagonalizing the 2^{10} submatrices of $I - W$ and classifying the eigenvalues. In agreement with theorem 1, we find that all positive eigenvectors have positive eigenvalues, and so the network has a set of globally attractive fixed points. (For a sufficient but not necessary criterion for stability of a similar ring network, see Wersing et al., 2001.) We classified sets of neurons Σ into permitted and forbidden sets according to the eigenvalues of $(I - W)^\Sigma$. In Figure 2, we depict the parent permitted sets (sets that have no permitted supersets). Consistent with the finding that such ring networks can explain contrast invariant tuning of V1 cells (Ben-Yishai et al., 1995) and multiplicative response modulation of parietal cells (Salinas & Abbott, 1996), we find that there are no permitted sets that consist of more than five contiguous active neurons. For example, as a model for parietal cells, any combination of positive visual and positive eye-position inputs retrieves a permitted set consisting of a group of contiguous neurons. Because the resulting activity pattern in the network is localized, the tuning widths of visual responses are limited, making the combined eye position and visual responses look invariant in shape. It can be shown that in order to retrieve one of the spurious permitted sets in Figure 2, some inputs must be negative, a rather irrelevant case, because cortical inputs are usually excitatory.

6 Discussion

In earlier network models of associative memory (Hopfield, 1982; Hinton & Sejnowski, 1986), computational inputs were encoded as initial conditions to the neural dynamics and memories as attractive fixed points. After applying an input in these models, the dynamics converge to a steady state, which represents the retrieved memory in response to that input. At the steady state, all information about the input is lost. In threshold-linear networks, the situation is different; the attractive fixed points are input dependent, meaning that a small change in the input leads to a small change in the location of the retrieved steady state. Thus, in these networks, memories cannot be represented by steady states.

We have shown that memories in threshold-linear networks can be defined in terms of permitted sets of neurons, for example, sets of neurons that can be stably coactivated at a steady state. This offers a new concept of pattern memorization that fulfills the criterion of input independence also for the case of threshold-linear networks. Although memories are input independent, the retrieval of a memory is strongly constrained by the input. A typical input will not allow for the retrieval of arbitrary stored permitted sets. This comes from the fact that the existence of nonconnected sets of stable fixed points is dependent not just on whether there exists a forbidden set but also on the input (see theorem 2). Generally, multiattractiveness in the ring network is possible only when more than a single neuron is excited.

Notice that threshold-linear networks can behave as traditional attractor networks when the inputs are represented as initial conditions of the dynamics. For example, by fixing $b = 1$ and initializing a copositive network with some input, the permitted sets unequivocally determine the stable fixed points. Thus, in this case, the notion of permitted sets is no different from fixed-point attractors. However, the hierarchical grouping of permitted sets (see theorem 3) becomes irrelevant, since there can be only one attractive fixed point per hierarchical group defined by a parent permitted set.

We would like to discuss the limits within which we believe the network dynamics studied here are approximations of real biological neural networks. The fact that threshold-linear neurons are nonsaturating is a simplification of real neurons that loses its validity when interspike intervals are close to the refractory time of action potentials or when postsynaptic receptors become saturated. Unfortunately, for saturating nonlinearities, we are unable to derive any formal and input-independent results about the computations performed by a network. The reason is that the effect of a saturating nonlinearity is to induce an activity-dependent change in the slopes of transfer functions. As a result, depending on the network architecture, some eigenvalues may decrease, whereas others may increase as firing rates change, making stability criteria subtly activity dependent and hard to characterize. The symmetry of connections is a simplification that

is valid on average when excitatory neurons are recurrently connected (as is common, e.g., in cortex; Douglas et al., 1995) and when inhibitory neurons are nonspecific and fast acting (in which case, inhibitory firing rates can be expressed in terms of excitatory firing rates, then inhibitory neurons can be eliminated from the network, and one ends up with symmetric connections). Last but not least, the notion of forbidden sets does not have to be incompatible with the nonzero spontaneous (background) firing rates of cortical neurons. We would like to argue that the spontaneous background firing of cortical neurons may involve the rapid switching of permitted sets so that on a larger timescale, all neurons seem to be simultaneously active, whereas in reality they are not. The switching of permitted sets may be driven, for example, by the dynamics of long-range cortico-cortical inputs, which reflect the behavioral state of an animal and its expectations about the sensory environment; switching may also involve the intrinsic noisiness of synapses (such as their release probability).

The fact that permitted sets never have forbidden subsets implies a hierarchical organization of symmetric networks. As a model for perception, if neurons are feature detectors, then different parent permitted sets may represent different objects. Certain combinations of features will not be perceived as a recognizable object because the features are encoded by neurons forming a forbidden set. If not all but just a subset of the features of an object is present, such as may occur during occlusion, for example, then a subset of the neurons encoding the object may still be coactivated because they still form a permitted set. But how would a network encode, for example, the letter *P* as a separate object from the letter *R* and not as a subset thereof? Are symmetric networks not a suitable medium for encoding such object pairs? We do not believe so, because cortical feature detectors have receptive fields comprising both excitatory and inhibitory subfields. In this way, they can signal the absence of a feature as well as they can signal its presence. For example, a neuron that responds to a dark vertical bar surrounded by white flanks will respond to the letter *P* but not to the letter *R*, because the slanted bar in *R* excites the inhibitory subfield, causing the neuron to be silent. It is thus highly plausible that the encoding of the letter *P* is based on the activation of sensory neurons that will not be activated for the letter *R* and so the parent permitted sets of *P* and *R* will not be subsets of each other. The representational strategy of excitatory and inhibitory receptive subfields makes it possible in principle to store just about any collection of objects as an antichain of subsets, that is, as a collection of mutually exclusive sets. For a simple procedure on how to store a given family of possibly overlapping patterns as permitted sets, see Xiaohui, Hahnloser, and Seung (2002).

Appendix: The Generalized Lotka-Volterra Equations

In many previous studies, the following generalized Lotka-Volterra equations were used to model neuron dynamics, instead of the rectified linear

equation 2.1 (Fukai & Tanaka, 1997; Rabinovich et al., 2001):

$$\frac{dx_i}{dt} = x_i \left(-x_i + b_i + \sum_j W_{ij}x_j \right). \quad (\text{A.1})$$

The generalized Lotka-Volterra equation A.1 has a similar property as the rectified linear equation, which is that once initialized to nonnegative values, neural activities x_i remain nonnegative for all future times. The reason for nonnegativity is that the nonnegative orthant cannot be left, because for all inputs b , we have that $\dot{x}_i = 0$ whenever $x_i = 0$.

Here we show that given the same connection matrix W and input b , the rectified linear dynamics of equation 2.1 and the generalized Lotka-Volterra dynamics of equation A.1 have the same set of stable fixed points. This correspondence of attractors will follow from theorem 4, establishing that the quadratic function L that was a Lyapunov-like function for equation 2.1 is also a Lyapunov-like function for equation A.1.

Theorem 4. *If W is symmetric and $I - W$ copositive, then for all constant b and initial conditions, the generalized Lotka-Volterra equation A.1 converges to an equilibrium. Furthermore, for all b , the set of attractive fixed points is identical to that of the threshold-linear equation 2.1.*

Proof. Because $I - W$ is copositive, $L = \frac{1}{2}x^T(I - W)x - b^T x$ is bounded below. The dynamics of equation A.1 can be written as $\dot{x}_i = -x_i \frac{\partial L}{\partial x_i}$. From this, it follows that L is nonincreasing along trajectories: $\dot{L} = \frac{\partial L}{\partial x_i} \dot{x}_i = -\left(\frac{\partial L}{\partial x_i}\right)^2 x_i \leq 0$. Fixed points of 2 correspond to steady states of L and vice versa: from $\dot{x}_i = 0$, it follows that either $x_i = 0$ or $\frac{\partial L}{\partial x_i} = 0$, and so $\dot{L} = 0$. Similarly, from $\dot{L} = 0$, it follows that $\dot{x}_i = 0$. Hence, for all b , the dynamics of equation A.2 converge to a steady state. Furthermore, because L is a Lyapunov-like function for both equations 2.1 and A.1, it follows that their sets of attractive fixed points must be identical.

The two networks in equations 2.1 and A.1 have the same Lyapunov-like function. It follows that theorems 2 and 3 about conditional multiattractiveness and the grouping of permitted sets also apply to the generalized Lotka-Volterra equations, and so make these networks practically identical. The only differences between them are their basins of attraction, their speed of convergence, and the pathological unstable fixed points $\underline{x}_i = 0$ of the generalized Lotka-Volterra equation (recall that the origin is always a fixed point of equation A.1 because it is not possible to leave this fixed point by the action of b , changing b).

In practical applications, however, the origin is not a worrisome fixed point. An approximation to a threshold-linear network was recently im-

plemented in terms of a silicon circuit using analog very large-scale integration (aVLSI) technology (Hahnloser et al., 2000). There, neural activities are represented by electrical currents. The equations describing the current dynamics were shown to be the generalized Lotka-Volterra equation A.1. The origin is not a pathological fixed point in this physical system because thermal noise causes the dynamics to leave the repulsive origin for positive b .

Acknowledgments

We thank Misha Tsodyks for his critical reading of the manuscript.

References

- Ben-Yishai, R., Lev Bar-Or, R., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA*, *92*, 3844–3848.
- Bertsekas, D. (1995). *Nonlinear programming*. Belmont, MA: Athena Scientific.
- Cohen, M., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 288–307.
- Douglas, R., Koch, C., Mahowald, M., Martin, K., & Suarez, H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981–985.
- Ermentrout, B. (1994). Reduction of conductance-based models with slow synapses to neural nets. *Neural Computation*, *6*, 679–695.
- Feng, J., & Hadel, K. (1996). Qualitative behaviour of some simple networks. *J. Phys. A*, *29*, 5019–5033.
- Fukui, T., & Tanaka, S. (1997). A simple neural network exhibiting selective activation of neuronal ensembles: From winner-take-all to winners-share-all. *Neural Computation*, *9*(1), 77–97.
- Hadel, K., & Kuhn, D. (1987). Stationary states of the Hartline-Ratliff model. *Biological Cybernetics*, *56*, 411–417.
- Hahnloser, R. H. (1998). About the piecewise analysis of networks of linear threshold neurons. *Neural Networks*, *11*, 691–697.
- Hahnloser, R., Sarpeshkar, R., Mahowald, M., Douglas, R., & Seung, H. (2000). Digital selection and analog amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*, 947–951.
- Hahnloser, R. H., & Seung, H. (2001). Permitted and forbidden sets in symmetric threshold linear networks. In *Proceedings of NIPS2001—Neural Information Processing Systems: Natural and Synthetic* (Vol. 13, pp. 217–223). Cambridge, MA: MIT Press.
- Hartline, H. K., & Ratliff, F. (1958). Spatial summation of inhibitory influence in the eye of limulus and the mutual interaction of receptor units. *J. Gen. Physiol.*, *41*, 1049–1066.
- Hinton, G., & Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing* (pp. 448–453). Cambridge, MA: MIT Press.

- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, *79*(8), 2554–2558.
- Hopfield, J. J. (1984). Neurons with graded response have collective properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, *81*, 3088–3092.
- Horn, R., & Johnson, C. (1985). *Matrix analysis*. Cambridge: Cambridge University Press.
- Rabinovich, M., Volkovskii, A., Lecanda, P., Huerta, R., Abarbanel, H., & Laurent, G. (2001). Dynamical encoding by networks of competing neuron groups: Winnerless competition. *Physical Review Letters*, *87*(6), 068102 (1–4).
- Salinas, E., & Abbott, L. (1996). A model of multiplicative neural responses in parietal cortex. *Proc. Natl. Acad. Sci. USA*, *93*, 11956–11961.
- Seung, H. S. (1996). How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. USA*, *93*, 13339–13344.
- Strogatz, S. (1994). *Nonlinear dynamics and chaos*. Reading, MA: Addison-Wesley.
- Wersing, H., Beyn, W.-J., & Ritter, H. (2001). Dynamical stability conditions for recurrent neural networks with unsaturating piecewise linear transfer functions. *Neural Computation*, *13*(8), 1811–1825.
- Xiaohui, X., Hahnloser, R. H., & Seung, H. S. (2002). Selectively grouping neurons in recurrent networks of lateral inhibition. *Neural Computation*, *14*(11), 2627–2646.

Received April 29, 2002; accepted August 2, 2002.