Selectively Grouping Neurons in Recurrent Networks of Lateral Inhibition

Xiaohui Xie
xhx@ai.mit.edu
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Richard H. R. Hahnloser
rhahnloser@mit.edu
H. Sebastian Seung
seung@mit.edu
Department of Brain and Cognitive Sciences and Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Winner-take-all networks have been proposed to underlie many of the brain's fundamental computational abilities. However, not much is known about how to extend the grouping of potential winners in these networks beyond single neuron or uniformly arranged groups of neurons. We show that competition between arbitrary groups of neurons can be realized by organizing lateral inhibition in linear threshold networks. Given a collection of potentially overlapping groups (with the exception of some degenerate cases), the lateral inhibition results in network dynamics such that any permitted set of neurons that can be coactivated by some input at a stable steady state is contained in one of the groups. The information about the input is preserved in this operation. The activity level of a neuron in a permitted set corresponds to its stimulus strength, amplified by some constant. Sets of neurons that are not part of a group cannot be coactivated by any input at a stable steady state. We analyze the storage capacity of such a network for random groups—the number of random groups the network can store as permitted sets without creating too many spurious ones. In this framework, we calculate the optimal sparsity of the groups (maximizing group entropy). We find that for dense inputs, the optimal sparsity is unphysiologically small. However, when the inputs and the groups are equally sparse, we derive a more plausible optimal sparsity. We believe our results are the first steps toward attractor theories in hybrid analog-digital networks.

1 Introduction

It has long been known that lateral inhibition in neural networks can lead to winner-take-all competition, so that only a single neuron is active at a steady state (Amari & Arbib, 1977; Feng & Hadeler, 1996; Hahnloser, 1998; Sum & Tam, 1996; Coultrip, Granger, & Lynch, 1992; Maass, 2000). When used for unsupervised learning, such winner-take-all networks enforce grandmother-cell representations as in vector quantization (Kohonen, 1989). Recently, much research has focused on unsupervised learning algorithms for sparsely distributed representations (Olshausen & Field, 1996; Lee & Seung, 1999). These algorithms lead to representations where multiple neurons participate in the encoding of an object and so are more distributed than vector quantization. Therefore, it is of interest to find ways of using lateral inhibition to mediate winner-take-all competition between groups of neurons, enforcing the sparse representation at a network level.

Competing groups of neurons are the essence of attractor models of associative memory. Selectively grouped neurons correspond to patterns that are stored as attractors in the network, with only one of these patterns retrieved at a steady state (Hopfield, 1982; Willshaw, Buneman, & Longuet-Higgins, 1969; Willshaw & Longuet-Higgins, 1970). In this case, the input to the network is represented in the initial conditions of the dynamic system, and the winning group is the resulting steady state. However, the binary behavior of an individual neuron in associative memory models is much different and computationally less informative than a biophysical neuron, whose firing rate encodes information on the signal it is processing. Although there have been extensions of these discrete and digital attractor networks to networks with graded (Hopfield, 1984; Miller & Zucker, 1999) or stochastic neurons (Golomb, Rubin, & Sompolinsky, 1990), the behavior of the individual neuron tends to be inactive or saturate and thus remains binary in essence.

In this article, we show how winner-take-all competition between groups of neurons can be realized in networks of nonbinary, analog neurons. In a network model introduced later, neurons at a steady state can be either active or inactive and form a binary pattern representing a permitted grouping of the neurons. At the same time, the activated neurons carry analog values resulting from computations implemented by the network.

We present a natural way of wiring the network to group neurons selectively by adding strong lateral inhibition between them. Given a collection of potentially overlapping groups, the inhibitory connectivity is set by a simple formula that can be interpreted as arising from an on-line learning rule. To show that the resulting network functions as group winner-take-all, we perform a stability analysis. If the strength of inhibition is sufficiently great and the group organization satisfies certain conditions, one and only one group of neurons can be activated at a stable steady state. In general, the identity of the winning group depends on the network inputs and also the initial conditions of the dynamics.

We characterize the storage capacity, the maximum number of groups the network can mediate to produce winner-take-all competitions, for random sparse groups in which each neuron has the probability p to be included in each group. Let n be the total number of neurons in the network. We determine the optimal sparsities p that maximize group entropy in two cases: (1) when the input is dense, the optimal sparsity scales as $\ln(n)/n$, and (2) when the inputs are of equal sparsity as the groups themselves, the optimal sparsity scales as $\sqrt{\ln(n)/n}$. In the first case, the storage capacity roughly scales as n^2, and in the second case, the storage capacity scales as $n/\ln(n)$.

2 Basic Definitions

Let m groups of neurons be given, where the group membership of the ath group is specified by

$$\xi_i^a = \begin{cases} 1 & \text{if the } i\text{th neuron is in the } a\text{th group} \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

for $i = 1, \ldots, n$.

We will assume that every neuron belongs to at least one group[1] and that every group contains at least one neuron. A neuron is allowed to belong to more than one group, so that the groups are potentially overlapping. The inhibitory synaptic connectivity of the network is defined in terms of the group membership,

$$J_{ij} = \prod_{a=1}^{m} (1 - \xi_i^a \xi_j^a) = \begin{cases} 0 & \text{if } i \text{ and } j \text{ both belong to a group} \\ 1 & \text{otherwise.} \end{cases} \tag{2.2}$$

The matrix J basically states that a connection between neuron i and j is established only if they do not belong to any of the same groups. This pattern of connectivity could arise from a simple learning mechanism. Suppose that all elements of J are initialized to be unity, and the groups are presented sequentially as binary vectors $\xi^1, \ldots, \xi^m$. The ath pattern is learned through the update,

$$J_{ij} \leftarrow J_{ij}(1 - \xi_i^a \xi_j^a). \tag{2.3}$$

In other words, if both neurons i and j belong to pattern a, then the connection between them is removed. After presentation of all m patterns, this leads to equation 2.2. At the start of the learning process, the initial state of J corresponds to uniform inhibition, which is known to implement winner-take-all competition between individual neurons. It will be seen that as inhibitory connections are removed during learning, the competition evolves to mediate competition between groups of neurons rather than individual neurons.

[1] This condition can be relaxed but is kept for simplicity.

Let x_i be the activity of neuron i. The dynamics of the network is given by

$$\frac{dx_i}{dt} + x_i = \left[b_i + \alpha x_i - \beta \sum_{j=1}^{n} J_{ij} x_j \right]^+, \tag{2.4}$$

for all $i = 1, \ldots, n$, where $[z]^+ \equiv \max\{z, 0\}$ denotes rectification, $\alpha > 0$ is the strength of self-excitation, and $\beta > 0$ is the strength of lateral inhibition. b_i is the external input. Equivalently, the dynamics can be written in matrix vector form as

$$\dot{\boldsymbol{x}} + \boldsymbol{x} = [\boldsymbol{b} + W\boldsymbol{x}]^+, \tag{2.5}$$

where $W = \alpha I - \beta J$ includes both self-excitation and lateral inhibition. The state of the network is specified by the vector $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ and the external input by the vector $\boldsymbol{b} = [b_1, \ldots, b_n]^T$. Recurrent networks with linear threshold units have been used in a variety of neural modeling studies (Hansel & Sompolinsky, 1998; Salinas & Abbott, 1996; Wersing, Beyn, & Ritter, 2001; Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000).

A vector $\boldsymbol{v}$ is said to be nonnegative, $\boldsymbol{v} \geq 0$, if all of its components are nonnegative. The nonnegative orthant is the set of all nonnegative vectors. Notice that in equation 2.4, $\dot{x}_i \geq 0$ whenever $x_i = 0$. Moreover, the linear threshold function is obviously Lipschitz continuous. These two properties are sufficient to guarantee that the nonnegative orthant is a positive invariant set of the dynamics, that is, any trajectory of equation 2.4 starting in the nonnegative orthant remains there (Khalil, 1996). Furthermore, even if the initial state of $\boldsymbol{x}$ is negative, it will become nonnegative after some transient period. Therefore, for simplicity, we consider trajectories that are confined to the nonnegative orthant $\boldsymbol{x} \geq 0$. However, we consider input vectors $\boldsymbol{b}$ whose components are of arbitrary sign.

3 Network Performance

Next, we briefly state some of the properties of the network. The detailed analysis is deferred to later sections.

We start with a simple case with n different groups, each containing one of the n neurons, which is the traditional winner-take-all network. Suppose $k > 1$ neurons are active initially. After proper ordering, the interaction matrix between these k active neurons is $W = (\alpha + \beta)I - \beta \mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is the column vector consisting of all ones. One eigenvector of W is $\mathbf{1}$ with eigenvalue $\alpha - (k-1)\beta$. The other $k-1$ eigenvectors are differential modes whose components sum to zero, with eigenvalue $\alpha + \beta$. If the inhibition strength is strong enough, $\beta > 1 - \alpha$, the differential modes are unstable, therefore, the network cannot have more than one neuron active at a steady state. Moreover, the network is guaranteed to converge to a steady state provided that $\alpha < 1$. Under these conditions, we can conclude that for all $\boldsymbol{b}$ and initial conditions of $\boldsymbol{x}$, the network always converges to one of the given groups.

For the general case with arbitrary group membership matrix ξ, the above conclusion still holds true, except in some degenerate cases (which will be described in the next section). If the lateral inhibition is strong enough ($\beta > 1-\alpha$) as in the previous case, any steady state with two active neurons not contained in the same group is unstable. If $\alpha < 1$, the network is again guaranteed to converge to a steady state. Therefore, one and only one of the given groups can be active at each steady state.

Which groups could potentially be the winner is specified by the input $\boldsymbol{b}$. In the case of nonoverlapping groups, the potential winners are determined by the aggregate positive input $B^a = \sum_{i=1}^{n}[b_i]^+\xi_i^a$ that each group receives. Any group with $B^a \geq (1-\alpha)\beta^{-1}b_{\max}$ could end up as the winning group, where $b_{\max} \equiv \max_i\{b_i\}$. Which group wins in the end depends on the initial conditions. It is possible for a specific group to win for all initial conditions if its inputs are sufficiently large.

The synaptic connections between neurons within a group are restricted to self-excitation. This causes the activities of winning neurons to be equal to their rectified input, amplified by a gain factor $1/(1-\alpha)$. Thus, the network implements a form of hybrid analog-digital computation, selectively amplifying activities in only one group of neurons.

4 Analysis of the Network Dynamics

4.1 Convergence to a Steady State.

This section characterizes the steady-state responses of the network equation 2.4 to an input $\boldsymbol{b}$ that is constant in time. For this to be a sensible goal, we need some guarantee that the dynamics converges to a steady state and does not diverge. This is provided by the following theorem.

Theorem 1. *Consider the network equation 2.4. The following statements are equivalent:*

1. *For any input $\boldsymbol{b}$, the network state $\boldsymbol{x}$ converges to a steady state that is stable in the sense of Lyapunov, except for initial conditions in a set of measure zero consisting of unstable equilibria.*
2. *The strength α of self-excitation is less than one.*

Proof. To prove equation 2.2 ⇒ equation 2.1, if $\alpha < 1$, the function

$$E(\boldsymbol{x}) = \frac{1}{2}(1-\alpha)\boldsymbol{x}^T\boldsymbol{x} + \frac{\beta}{2}\boldsymbol{x}^T J\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \tag{4.1}$$

is bounded below and radially unbounded in the nonnegative orthant. Furthermore, E is nonincreasing following the dynamics

$$\begin{aligned} dE/dt &= -((I-W)\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{x}-[W\boldsymbol{x}+\boldsymbol{b}]^+) \\ &= -\sum_{i\in M}((I-W)\boldsymbol{x}-\boldsymbol{b})_i^2 - \sum_{i\notin M}(x_i^2-(W\boldsymbol{x}+\boldsymbol{b})_i\,x_i) \\ &\le -\sum_{i\in M}((I-W)\boldsymbol{x}-\boldsymbol{b})_i^2 - \sum_{i\notin M}x_i^2 \\ &\le 0, \end{aligned}$$

where $M \equiv \{i \mid (W\boldsymbol{x}+\boldsymbol{b})_i > 0, \forall i = 1, \ldots, n\}$. The notation $(\boldsymbol{z})_i$ denotes the ith component of the vector $\boldsymbol{z}$.

Equality above holds if and only if $\boldsymbol{x}$ is at the steady state. Therefore, $E(\boldsymbol{x})$ is a Lyapunov-like function ensuring convergence to a stable steady state, except for initial conditions in a set of measure zero.

To prove equation 2.1 ⇒ equation 2.2, let us suppose that equation 2.2 is false. If $\alpha \ge 1$, choose $\boldsymbol{b} = (1, 0, \ldots, 0)^T$ and initial conditions $\boldsymbol{x}(0) = (1, 0, \ldots, 0)^T$ so that the dynamics of the first neuron is reduced to $\dot{x}_1 + x_1 = [\alpha x_1 + 1]^+ \ge x_i + 1$, in which x_1 diverges. In addition, x_1 diverges for initial conditions in a set of nonzero measure, so equation 2.1 is contradicted. Therefore, $\alpha < 1$ is both the necessary and sufficient condition for convergence to a stable steady state.

In the following, we restrict the network to $\alpha < 1$.

4.2 Permitted and Forbidden Sets. In general, the network may have many fixed points. However, only those that are stable are typically observed at a steady state. We will call a set of neurons that can be coactivated by some input at a stable (in the sense of Lyapunov) steady state a *permitted set*. Otherwise, it is termed a *forbidden set*.

For a set of neurons to be a permitted set, two conditions have to be satisfied: its neurons have to be steadily coactivated by some input, and the steady coactivation must be stable. For the network we are considering, it is always possible to choose an input that realizes a steady coactivation of the given set of neurons. Hence, the first condition is readily satisfied. Consequently, whether a set is permitted or forbidden depends on only its stability, which is determined by the synaptic connection matrix between the coactivated neurons. If the largest eigenvalue of that matrix is less than unity, then the set is permitted. Otherwise, it is forbidden.

One special property of the permitted and forbidden sets is that any superset of a forbidden set is forbidden, and any subset of a permitted set is permitted (Hahnloser et al., 2000). An intuitive understanding of this property is that by inactivating a neuron, its feedback is removed. Because the connections in a symmetric network form effectively positive feedback loops, in the form of mutual excitation or disinhibition, removing feedbacks increases stability of the network. Similarly, adding positive feedbacks decreases stability, in agreement with the property that any superset of a forbidden set is forbidden.

The above property adds convenience for verifying whether a set is permitted or forbidden. For example, if we know a subset of a set is forbidden, then the set itself is forbidden. We will use this property in the following sections.

4.3 Relationship Between Groups and Permitted Sets. The network in equation 2.4 is constructed to make the groups and their subgroups the only permitted sets of the network. To determine whether this is the case, we must answer two questions. First, are all groups and their subgroups permitted? Second, are all permitted sets contained in the given groups? The first question is answered by the following lemma.

Lemma 1. *All groups and their subgroups are permitted.*

Proof. If a set is contained in a group, then there is no lateral inhibition between the neurons in the set. Provided that $\alpha < 1$, all eigenvalues of the interaction matrix between neurons in the group are less than unity, so the set is permitted.

The answer to the second question, whether all permitted sets are contained in the groups, is not necessarily affirmative. For example, consider the network defined by the group membership matrix $\xi = \{(1, 1, 0), (0, 1, 1), (1, 0, 1)\}$. Since every pair of neurons belongs to some group, there is no lateral inhibition ($J = 0$), which means that there are no forbidden sets. As a result, $(1, 1, 1)$ is a permitted set, but obviously it is not contained in any group.

Let us define a spurious permitted set to be a permitted set that is not contained in any group. For example, $(1, 1, 1)$ is a spurious permitted set in the above example. To eliminate all the spurious permitted sets in the network, certain conditions on the group membership matrix ξ have to be satisfied.

Definition 1. The membership matrix ξ is *degenerate* if there exists a set of $k \geq 3$ neurons that is not contained in any group, but all of its subsets with $k - 1$ neurons belong to some group. Otherwise, ξ is called *nondegenerate*.

For example, $\xi = \{(1, 1, 0), (0, 1, 1), (1, 0, 1)\}$ is degenerate. Using this definition, we can formulate the following theorem.

Theorem 2. *The neural dynamics equation 2.4 with $\alpha < 1$ and $\beta > 1 - \alpha$ has a spurious permitted set if and only if ξ is degenerate.*

To prove the theorem, we need the following lemma.

Lemma 2. *If $\beta > 1 - \alpha$, any set containing two neurons not in any same group is forbidden under the neural dynamics equation 2.4.*

Proof. We start by analyzing a very simple case where there are two neurons belonging to two different groups. Let the group membership be $\{(1, 0), (0, 1)\}$. In this case, $W = \{(\alpha, -\beta); (-\beta, \alpha)\}$. This matrix has eigenvectors $(1, 1)^T$ and $(1, -1)^T$ with eigenvalues being $\alpha - \beta$ and $\alpha + \beta$, respectively. Since $\alpha < 1$ for convergence to a steady state and $\beta > 0$ by definition, the $(1, 1)^T$ mode is always stable. But if $\beta > 1 - \alpha$, the $(1, -1)^T$ mode is unstable. This means that it is impossible for the two neurons to be coactivated at a stable steady state. Since any superset of a forbidden set is also forbidden, the result generalizes to more than two neurons.

Now we are ready to prove theorem 2 by using lemma 2.

Proof of Theorem 2. If ξ is degenerate, there must exist a set $k \geq 3$ neurons that is not contained in any group, but all of its subsets with $k - 1$ neurons belong to some group. There is no lateral inhibition between these k neurons, since every pair of neurons belongs to some group. Thus, the set containing all k neurons is permitted and spurious.

On the other hand, if there exists a spurious permitted set P, we need to prove that ξ must be degenerate. We will prove this by contradiction and induction. Let us assume ξ is nondegenerate.

P must contain at least two neurons since any one neuron subset is permitted and not spurious. By lemma 2, these two neurons must be contained in some group, or else it is forbidden. Thus, P must contain at least three neurons to be spurious, and any pair of neurons in P belongs to some group by lemma 2.

If P contains at least k neurons and all of its subsets with $k - 1$ neurons belong to some group, then the set with these k neurons must belong to some group; otherwise, ξ is degenerate. Thus, P must contain at least $k + 1$ neurons to be spurious, and all its k subsets must belong to some group.

By induction, this implies that P must contain all neurons in the network, in which case P is either forbidden or nonspurious. This contradicts the assumption that P is a spurious permitted set.

Remark. The group winner-take-all competition described above holds only for the case of strong inhibition $\beta > 1 - \alpha$. If β is small, the competition will be weak and may not result in group winner-take-all. In particular, if $\beta < (1 - \alpha)/\lambda_{\max}(-J)$, where $\lambda_{\max}(-J)$ is the largest eigenvalue of $-J$, then the set of all neurons is permitted. Since every subset of a permitted set is permitted, this means that there are no forbidden sets, and the network is monostable.

Hence, group winner-take-all does not hold. In the intermediate regime, $(1-\alpha)/\lambda_{\max}(-J) < \beta < 1-\alpha$, the network has forbidden sets, but the possibility of spurious permitted sets cannot be excluded.

5 The Potential Winners

We have seen that if ξ is nondegenerate, any stable coactive set of neurons must be contained in a group, provided that lateral inhibition is strong ($\beta > 1-\alpha$). The group that contains the coactive set is the "winner" of the competition between groups. The identity of the winner depends on input $\boldsymbol{b}$ and also on the initial conditions of the dynamics.

Suppose the ath group is the winner. For all neurons not in this group to be inactive, the self-consistent condition should read

$$\sum_i [b_i]^+ \xi_i^a J_{ij} \geq (1-\alpha)\beta^{-1}[b_j]^+, \tag{5.1}$$

for all $j \notin a$. If the group a contains the neuron with the largest input, this condition is always satisfied. Hence, any group containing the neuron with the largest input is always a potential winner.

In the case of nonoverlapping groups, the condition in equation 5.1 can be simplified as

$$\sum_i [b_i]^+ \xi_i^a \geq (1-\alpha)\beta^{-1} \max_{j \notin a}\{[b_j]^+\}, \tag{5.2}$$

and therefore potential winners are determined by the aggregate group inputs $B^a = \sum_i [b_i]^+ \xi_i^a$. Denote the largest input as $b_{\max} = \max_i\{b_i\}$, and assume $b_{\max} > 0$. Only those groups whose aggregate inputs are not smaller than $(1-\alpha)\beta^{-1} b_{\max}$ can win, with the exact winner identity determined by the initial conditions of the dynamics.

6 An Example: The Ring Network

In this section, we take the ring network as an example to illustrate several results we have obtained so far. Let n neurons be organized into a ring, and let every set of d contiguous neurons form a group. Thus, in total, there are n patterns to be stored. In the special case $d = 1$, this network becomes a traditional winner-take-all network.

In the case $d > 1$, the groups are overlapping and ξ could be degenerate. In fact, it can be shown that ξ becomes degenerate when $d \geq n/3 + 1$. This is illustrated in Figure 1, which shows the permitted sets of a ring network with 15 neurons. If the group width is $d = 5$ neurons, there are no spurious permitted sets (see Figures 1A–1C). However, when the group width is 6, the network contains 5 spurious permitted sets (see Figure 1F).

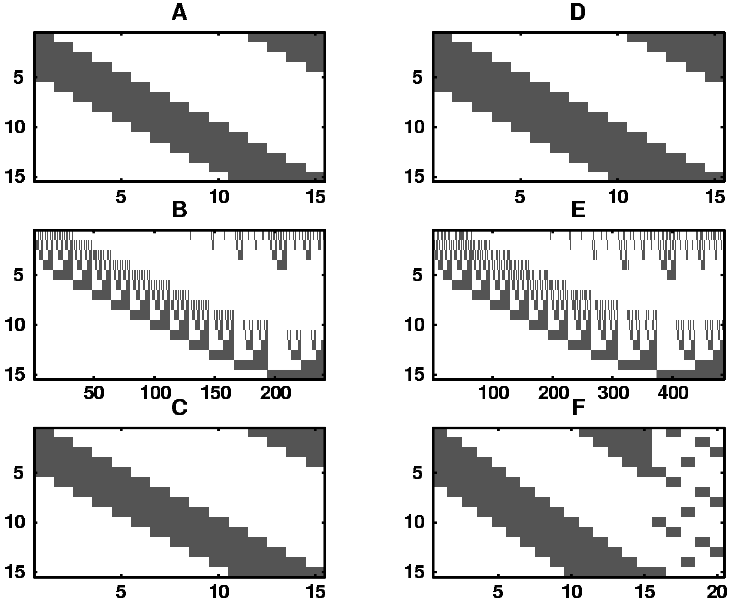

Figure 1: Permitted sets of the ring network. The ring network consists of 15 neurons with $\alpha = 0.4$ and $\beta = 1$. (A, D) The 15 groups are represented by columns. Black refers to active neurons, and white refers to inactive neurons. (A) Fifteen groups of width $d = 5$. (B) All permitted sets corresponding to the groups in *A*. (C) The 15 permitted sets in *B* that have no permitted supersets. They are the same as the groups in *A*. (D) Fifteen groups with width $d = 6$. (E) All permitted sets corresponding to groups in *D*. (F) There are 20 permitted sets in *E* that have no permitted supersets. Note that there are five spurious permitted sets.

Figure 2 shows the effect of changing the strength of lateral inhibition. When the strength of inhibition is strong ($\beta > 1 - \alpha$), there are no spurious permitted sets provided that ξ is nondegenerate (see Figure 2D). At the other extreme, when $\beta < (1 - \alpha)/\lambda_{\max}(-J)$, there is no unstable differential mode in the network. All neurons could potentially be active at a stable steady state, given a suitable input (see Figure 2A). Between these two critical values ($1 - \alpha < \beta < (1 - \alpha)/\lambda_{\max}(-J)$), there exist both unstable differential modes and spurious permitted sets (see Figure 2C).

7 Storage Capacity for Random Sparse Groups

An important characterization of any attractor network is its storage capacity for random patterns, that is, random groups (Amit, Gutfreund, & Som-

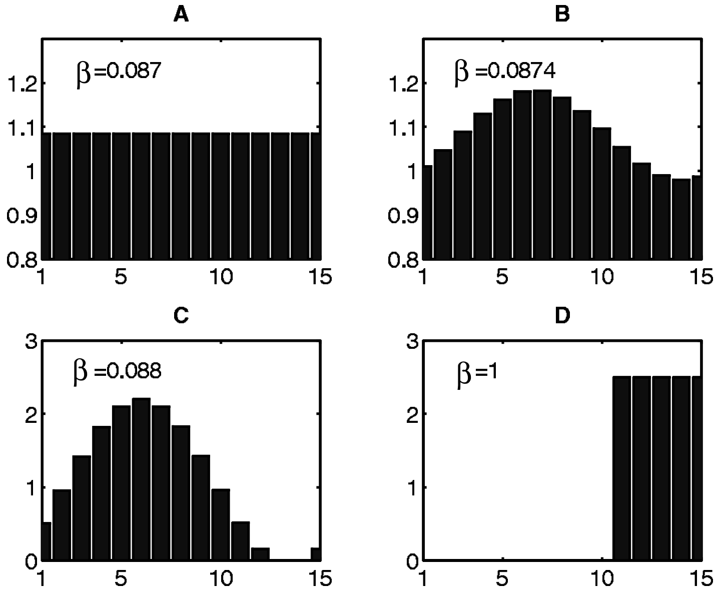

Figure 2: Lateral inhibition strength β determines the behavior of the network. The network is a ring network of 15 neurons with width $d = 5$ and where $\alpha = 0.6$, and input $b_i = 1$ for all i. The panels show the steady-state activities of the 15 neurons. (A) There are no forbidden sets. (B) The marginal state $\beta = (1 - \alpha)/\lambda_{\max}(-J) = 0.0874$, in which the network forms a continuous attractor. (C) Forbidden sets exist, and so do spurious permitted sets. (D) Group winner-take-all case; no spurious permitted sets.

polinsky, 1985; Miller & Zucker, 1999). In our case, as the number of groups gets larger, the probability of the groups' being degenerate increases. We call the probability of error the probability that a neuron outside a group is activated by mistake.

We choose random sparse groups; $p \ll 1$ is the probability that a particular neuron is part of a particular group. The storage capacity is defined to be the maximum number of groups the network can store, such that the error probability remains smaller than a given bound. After constructing the synaptic weight matrix, we present random inputs to the network. We assume that each component of the input $\boldsymbol{b}$ has the probability q of being positive. The expected number of neurons receiving positive inputs is $\bar{n} = nq$. Since a neuron receiving a nonpositive input can never become active in our network, the error probability is effectively determined by the network of the $\bar{n}$ neurons receiving positive inputs.

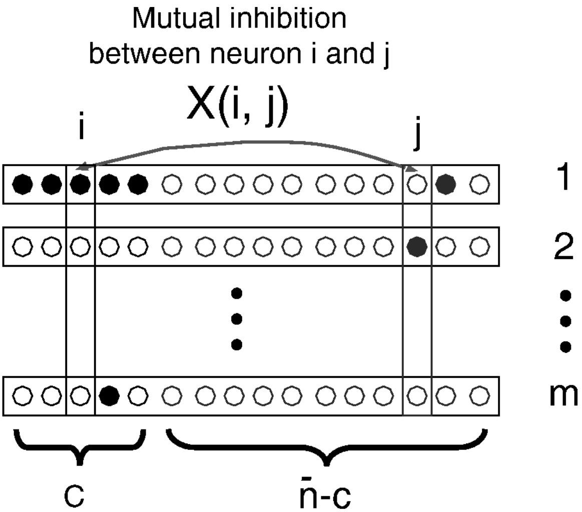

Figure 3: Diagram of m random groups. Filled circles represent active neurons. The first c neurons in group 1 are coactivated. For a perfect retrieval, all the other $\bar{n} - c$ neurons must be inactive; that is, all must be inhibited by at least one of the c active neurons. The error probability $\mathcal{P}_{\mathcal{E}}$ is the probability that at least one of the $\bar{n} - c$ neurons is active.

Under the randomness in both the groups and the inputs, the expected number of coactive neurons in a stable steady state is $c = \bar{n}p = npq$. Next, we assume that c neurons are coactivated and calculate the probability $\mathcal{P}_{\mathcal{E}}$ of mistakenly activating any of the other $\bar{n} - c$ neurons.

We use $X(i, j)$ to denote the existence of synaptic inhibition between neurons i and j, which in our network implies that neurons i and j are not contained in any same group (see Figure 3). According to lemma 2, $X(i, j)$ also represents mutual exclusion of neuron i and j at any stable steady state.

Without loss of generality, we index the c active neurons from 1 to c. For neuron j within the other $\bar{n} - c$ neurons to be inactive, it must make an inhibitory connection with at least one of the c neurons. The probability of this happening is $\Pr\left\{\bigvee_{i=1}^{c} X(i, j)\right\}$, where $\bigvee$ represents logical OR. Extending this to all the other $\bar{n} - c$ neurons, we derive the probability for all the $\bar{n} - c$ neurons being inactive as follows:

$$\mathcal{P}_{\mathcal{C}} = \Pr\left\{\bigwedge_{j=1}^{\bar{n}-c} \bigvee_{i=1}^{c} X(i, j)\right\}, \tag{7.1}$$

where $\bigwedge$ represents logical AND. The error probability $\mathcal{P}_{\mathcal{E}}$ for at least one neuron being mistakenly activated is then

$$\mathcal{P}_{\mathcal{E}} = 1 - \mathcal{P}_{\mathcal{C}} = \Pr\left\{\bigvee_{j=1}^{\bar{n}-c}\overline{\bigvee_{i=1}^{c} X(i, j)}\right\}, \tag{7.2}$$

where the overbar denotes logic complement. Next, we find an upper bound on $\mathcal{P}_{\mathcal{E}}$ and use it to estimate the capacity of the network.

7.1 Capacity. The error probability is upper bounded by

$$\mathcal{P}_{\mathcal{E}} \leq (\bar{n} - c)\Pr\left\{\overline{\bigvee_{i=1}^{c} X(i, j)}\right\} = (\bar{n} - c)\left(1 - \Pr\left\{\bigvee_{i=1}^{c} X(i, j)\right\}\right), \tag{7.3}$$

where $\Pr\{\bigvee_{i=1}^{c} X(i, j)\}$ can be exactly calculated using the inclusion-exclusion principle (Stanley, 1999) as follows:

$$\Pr\left\{\bigvee_{i=1}^{c} X(i, j)\right\} = \sum_{k=1}^{c}(-1)^{k+1}\binom{c}{k}\Pr\left\{\bigwedge_{i_1, i_2, \ldots, i_k} X(i_k, j)\right\} \tag{7.4}$$

$$= \sum_{k=1}^{c}(-1)^{k+1}\binom{c}{k}[1 - p + p(1-p)^k]^{m-1}. \tag{7.5}$$

In the above equation, the term $1 - p + p(1-p)^k$ represents the probability that neuron j does not coexist with other k neurons. This can happen in two cases: with neuron j being inactive (with probability $1-p$) or neuron j being active but all other k neurons being inactive (with probability $p(1-p)^k$).

Equation 7.5 can be further simplified by

$$\Pr\left\{\bigvee_{i=1}^{c} X(i, j)\right\} = 1 + \sum_{k=0}^{c}(-1)^{k+1}\binom{c}{k}[1 - kp^2 + O(p^3)]^{m-1} \tag{7.6}$$

$$\approx 1 - \sum_{k=0}^{c}(-1)^{k}\binom{c}{k}\exp(-kmp^2) \tag{7.7}$$

$$= 1 - [1 - \exp(-mp^2)]^c. \tag{7.8}$$

We have made two approximations in the above calculation. To derive equation 7.6, we have assumed that cp is sufficiently small, implying sparse groups. In the approximation made in equation 7.7, we have assumed $m(cp^2)^2 \to 0$ in the large n limit, that is, the number of groups m should

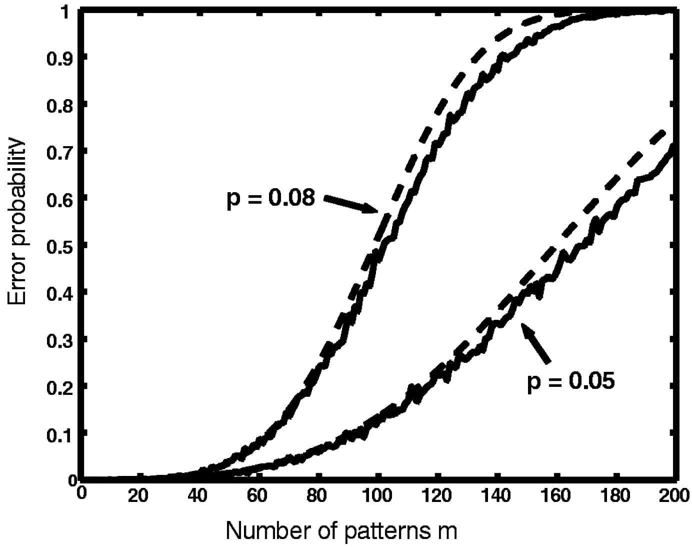

Figure 4: The error probability $\mathcal{P}_{\mathcal{E}}$ is plotted as a function of the number of groups m. Here, $q = 1$, and the number of neurons $n = 100$. The solid curves show the results from numerical simulations, and the dashed curves are the upper bounds calculated by equation 7.8. Two different sparsities p are used.

scale in order less than $1/(cp^2)^2$. We will see later, after deriving the capacity m, that these assumptions are indeed satisfied.

By substituting equation 7.8 into equation 7.3, we derive the upper bound on the error probability:

$$\mathcal{P}_{\mathcal{E}} \leq (\bar{n} - c)[1 - \exp(-mp^2)]^c. \tag{7.9}$$

We observed close tightness of this bound when compared to the true error probability from numerical simulations of random groups (see Figure 4).

Given some small number d, the error probability $\mathcal{P}_{\mathcal{E}}$ is guaranteed not to exceed this number, provided that $m < m^*(d)$, where

$$m^*(d) = -p^{-2} \ln\{1 - [d/(\bar{n} - c)]^{1/c}\} \tag{7.10}$$

$$\approx -p^{-2} \ln\{1 - [d/(nq)]^{1/(npq)}\}, \tag{7.11}$$

where equation 7.11 follows from $c \ll \bar{n}$.

Given n, p, and q, using equation 7.11, we can estimate the maximum number of random groups the network can store in such a way that the probability of incorrect retrieval remains smaller than d.

7.2 Optimal Sparsity. How sparse should the random groups optimally be? We define the optimal sparsity p^* as the sparsity p that maximizes the information capacity of the network. We measure the information capacity I by the normalized entropy of the m^* random groups,

$$I = -m^* n\, [p \log_2 p + (1-p) \log_2 (1-p)]/n^2. \tag{7.12}$$

In other words, I is the entropy of the m^* binary words with length n with probability p of 1, normalized by the number of synaptic connections.

The denominator n^2 corresponds to the total entropy the binary synaptic weight matrix J can hold. Thus, I is expressed as the fraction of the possible entropy of J, used to store groups. The optimal sparsity p^* is given by $p^* = \text{argmax}_p\{I\}$. To calculate p^*, we first have to choose some value for q, determining the probability that a neuron receives an excitatory input. We consider two different cases. First, q is independent of p, and without loss of generality we take $q = 1$, which corresponds to the case where the inputs are excitatory and nonsparse. Second, q depends on p, and for simplicity we choose $q = p$, which is the case where the inputs are of equal sparsity as the groups.

The optimal sparsity calculations for both cases are derived in the appendix. Here we state only the results:

$$p^* \approx \begin{cases} \log_2(n)/n & \text{when } q = 1 \\ \sqrt{k \ln(n)/n} & \text{when } q = p, \end{cases} \tag{7.13}$$

where $k = 2.86$ is a constant. The approximation becomes exact when the number of neurons goes to infinity. This result shows that to achieve the maximum information capacity, p^* should scale as $\ln(n)/n$ when $q = 1$ and as $\sqrt{\ln(n)/n}$ when $q = p$. Correspondingly, the average number of neurons in each pattern scales as $\ln(n)$ for $q = 1$ and $\sqrt{n \ln(n)}$ for $q = p$.

By substituting p^* into equation 7.11, we derive the storage capacity for these optimal sparsities,

$$m^* \approx \begin{cases} \alpha n^{2-1/c} & \text{when } q = 1 \\ k_m n/\ln(n) & \text{when } q = p, \end{cases} \tag{7.14}$$

where $c \approx \log_2(n)$, $\alpha = d^{1/c}/\log_2^2(n)$ and $k_m = -\ln[1 - \exp(-1/(2k^2))]/k^2 \approx 0.35$. Since $\ln(n)$ hardly increases for large n, the capacity in the $q = 1$ case roughly scales as n^2 and in the $q = p$ case it roughly scales as n.

Equation 7.7 is derived under the assumption that $m(cp^2)^2 \to 0$ in the large n limit. Now we check the validity of this assumption. Self-consistently, in the case $q = 1$, we find $m^*(cp^{*2})^2 \sim 1/n^2$, and in the case $q = p$, $m^*(cp^{*2})^2 \sim 1/n$. Both approach zero in the large n limit.

8 Discussion

We have presented a network that uses structured lateral inhibition to mediate winner-take-all competition between potentially overlapping groups of neurons. Our construction uses the distinction between permitted and forbidden sets of neurons and identifies the allowed groupings as permitted sets inherent in the network.

Our capacity calculation in the $q = 1$ case reveals similarity with the Willshaw model (Willshaw & Longuet-Higgins, 1970): we find that the optimal sparsity scales as $\ln(n)/n$, for example, for a network of 10^{10} neurons, an optimal group consists of fewer than 30 neurons and is thus unrealistically small. In the case where inputs are sparse, $q = p$, we find that the optimal sparsity scales roughly as $\sqrt{n}$ and is thus within the realm of real networks.

A distinct feature of our generalized winner-take-all network is the coexistence of discrete pattern selection and analog computation. We use strong lateral inhibitory interactions to constrain certain groupings of neurons, but leave the analog values of the active neurons unconstrained, except by the input. It might be interesting to apply our principle of how to constrain active groups to the problem of data reconstruction using a constrained set of basis vectors. The constraints on the linear combination of basis vectors could, for example, implement sparsity or nonnegativity constraints (Lee & Seung, 1999).

The coexistence of analog filtering with logical constraints on neural activation represents a form of hybrid analog-digital computation that may be especially appropriate for perceptual tasks. Using this network model for object recognition, the perception of an object could be represented by the set of active neurons, while activities of these neurons correspond to continuous instantiations of the object such as viewpoint, illumination, and scale (Seung & Lee, 2000). In addition, this type of network may constitute a neural mechanism for feature binding and sensory segmentation problems, as suggested by Wersing et al. (Wersing, Steil, & Ritter, 2001; Wersing, 2002). In the domain of olfactory perception, recent experimental data on odor-evoked population responses in the olfactory bulb also show some promising applications of our model (Hildebrand & Shepherd, 1997; Christensen, Pawlowski, Lei, & Hildebrand, 2000; Mori, Nagao, & Yoshihara, 1999).

As we have shown, there are some degenerate cases of overlapping groups to which our method does not apply. It is an interesting open question whether there exists a general way of translating arbitrary groups of coactive neurons into permitted sets without involving spurious permitted sets. There are several possible approaches. For example, we could use a more sophisticated interaction matrix, including both lateral inhibition and excitation. For instance, in the three-neuron degenerate example given earlier, if we choose the interaction matrix $W = avv^T$ with $v = [1, -1, 1]^T$ and $1/3 < a < 1/2$, then the spurious set $(1, 1, 1)$ is forbidden, whereas its subsets are still permitted. Another possible approach would be to use higher-order interactions. Take again the three-neuron degenerate case as an example. If we added quadratic interactions to the dynamics, $\dot{x}_i + x_i = [b_i + \alpha x_i - \beta \sum_j J_{ij} x_j - \gamma \sum_{j,k} x_j x_k]^+$, it would follow that for large enough inputs and suitable parameters, the set $(1, 1, 1)$ would not be permitted, but its subsets would. One more possible approach would be to use hierarchical networks with interlayer excitation and intralayer inhibition.

In the past, a great deal of research has been inspired by the idea of storing memories as fixed-point attractors in neural networks with a fixed input. Our model suggests an alternative viewpoint, which is to regard permitted sets as memories latent in the synaptic connections, while the fixed points corresponding to permitted sets can continuously change depending on the input. From this viewpoint, the contribution of this article is a method of storing and retrieving memories as permitted sets in neural networks.

Appendix: Calculation of the Optimal Sparsity for Random Patterns

Start from the information capacity of the network, given by

$$I = -m^* n[p \log_2 p + (1-p)\log_2(1-p)]/n^2$$
$$\approx \log_2(p)(pn)^{-1} \ln\{1 - [(nq)^{-1} d]^{1/(npq)}\}.$$

Here, m^* is from equation 7.11 and the approximation is made in the small p limit. Next we consider two cases for choosing the value of q and find the optimal $p^* = \operatorname{argmax}_p\{I\}$ for these two cases, respectively. The calculation is done under the condition that the number of neurons n is sufficiently large.

A.1 Dense inputs, $q = 1$.

The information capacity I can be written as $I = c^{-1} \ln(1 - (d/n)^{1/c}) \log_2(c/n)$, where $c = pn$. By setting the derivative of I with respect to c equal to zero, we find

$$\ln(d/n)^{1/c} \ln(c/n) + [(d/n)^{-1/c} - 1] \ln[1 - (d/n)^{1/c}]\, [1 - \ln(c/n)] = 0.$$

Let $z = (d/n)^{1/c}$. Then we have

$$\ln z \ln(c/n) + (z^{-1} - 1)\ln(1-z)\, [1 - \ln(c/n)] = 0.$$

Under sparsity assumption, $p = c/n \ll 1$, we have $|\ln(c/n)| \gg 1$. Hence, the above equation can be simplified to

$$(1-z)\ln(1-z) = z \ln(z). \tag{A.1}$$

The solutions of the above equation are $z = 0$, $1/2$, and 1. Given a fixed n, c can be only a finite number. Therefore, the solution $z = 1$ is impossible. The other two solutions lead to $c = 0$ or $c = \log_2(n)$. Correspondingly,

$p = 0$ or $p = \log_2(n)/n$. Substituting the value of $p = 0$ into I, we find that $p = 0$ corresponds to a local minimum. Furthermore, the boundary value $p = 1$ also corresponds to a local minimum. From these, we conclude that the optimal probability is given by $p^* = \log_2(n)/n$. Notice that it satisfies sparsity assumption ($p^* \ll 1$).

A.2 Sparse inputs, $q = p$. The derivative of I with respect to p is

$$\begin{aligned} I'(p) &= \{\ln(1-t)[1-\ln(p)] \\ &\qquad + t(1-t)^{-1}\ln(p)/(np^2)[1+2\ln(d/(pn))]\}/(np^2\ln 2) \\ &\approx -[1-2k\ln(pn)/(np^2)]\ln(p)\ln(1-t)/(np^2\ln 2), \end{aligned}$$

where $k \equiv -t[(1-t)\ln(1-t)]^{-1}$ and $t \equiv [d/(pn)]^{1/(p^2 n)}$. To derive the above equation, we have neglected small terms by assuming that n is sufficiently large.

By setting $I'(p^*) = 0$, we find that p^* obeys,

$$\frac{\ln(p^* n)}{np^{*2}} = \frac{1}{2k}.$$

Deriving the exact form of p^* as a function of n from the above equation is not easy. However, when n is sufficiently large, we can simplify the calculation by assuming that k is independent of n. Under this ansatz, we derive p^* to scale as

$$p^* = \sqrt{k\ln(n)/n}. \tag{A.2}$$

Next, we need to self-consistently verify that our ansatz still holds by replacing p^* in the definition of t,

$$\ln t = \frac{\ln(d) - \ln[kn\ln(n)]/2}{k\ln(n)} \approx -\frac{1}{2k}. \tag{A.3}$$

The approximation becomes exact as n goes to infinity. Thus, we have verified that t is approximately constant, equal to $\exp(-1/(2k))$. This completes our ansatz.

We still need to determine the value of k. Substituting equation A.3 into the definition of k, we derive that

$$\frac{1-t}{2t} = \frac{\ln(t)}{\ln(1-t)}. \tag{A.4}$$

The root of this algebra equation can be found numerically. The final result is $t = 0.8396$ and $k = 2.86$. We can further check that the boundary values p at 0 or 1 lead only to local minima of I. Therefore, we conclude that p^* in equation A.2 is the optimal sparsity.

Acknowledgments

We acknowledge comments on the manuscript for this article by M. Goldman and R. Tedrake.

References

Amari, S., & Arbib, M. A. (1977). Competition and cooperation in neural nets. In J. Metzler (Ed.), *Systems neuroscience* (pp. 119–165). San Diego, CA: Academic Press.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Statistical mechanics of neural networks near saturation. *Ann. Phys. (New York)*, *173*, 30.

Christensen, T. A., Pawlowski, V. M., Lei, H., & Hildebrand, J. G. (2000). Multi-unit recordings reveal context-dependent modulation of synchrony in odor-specific neural ensembles. *Nature Neuroscience*, *12*, 927–931.

Coultrip, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, *5*, 47–54.

Feng, J., & Hadeler, K. (1996). Qualitative behaviour of some simple networks. *J. Phys. A*, *29*, 5019–5033.

Golomb, D., Rubin, N., & Sompolinsky, H. (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Physical Review A*, *41*, 1843–1854.

Hahnloser, R. H. (1998). About the piecewise analysis of networks of linear threshold neurons. *Neural Networks*, *11*, 691–697.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analog amplification coexist in a silicon circuit inspired by cortex. *Nature*, *405*, 947–951.

Hansel, D., & Sompolinsky, H. (1998). Modeling feature selectivity in local cortical circuits. In C. Koch & I. Segev (Eds.), *Methods in neuronal modeling* (pp. 499–567). Cambridge, MA: MIT Press.

Hildebrand, J. G., & Shepherd, G. M. (1997). Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annu. Rev. Neurosci.*, *20*, 595–631.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, *79*, 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.

Khalil, H. K. (1996). *Nonlinear systems* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). Berlin: Springer-Verlag.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*, 788–791.

Maass, W. (2000). On the computational power of winner-take-all. *Neural Computation*, *12*, 2519–2535.

Miller, D. A., & Zucker, S. W. (1999). Computing with self-excitatory cliques: A model and an application to hyperacuity-scale computation in visual cortex. *Neural Computation*, *11*, 21–66.

Mori, K., Nagao, H., & Yoshihara, Y. (1999). The olfactory bulb: Coding and processing of odor molecule information. *Science*, *286*, 711–715.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Salinas, E., & Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proc. Natl. Acad. Sci. USA*, *93*, 11956–11961.

Seung, H. S., & Lee, D. D. (2000). Cognition the manifold ways of perception. *Science*, *290*, 2268–2269.

Stanley, R. P. (1999). *Enumerative combinatorics* (Vol. 1). Cambridge: Cambridge University Press.

Sum, J. P. F., & Tam, P. K. S. (1996). Note on the maxnet dynamics. *Neural Computation*, *8*(3), 491–499.

Wersing, H. (2002). Learning lateral interactions for feature binding and sensory segmentation. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *1*. Cambridge, MA: MIT Press.

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.

Wersing, H., Steil, J. J., & Ritter, H. (2001). A competitive-layer model for feature binding and sensory segmentation. *Neural Computation*, *13*, 357–387.

Willshaw, D. J., Buneman, & Longuet-Higgins, H. C. (1969). Nonholographic associative memory. *Nature*, *222*, 960–962.

Willshaw, D., & Longuet-Higgins, H. (1970). Associative memory models. In B. Meltzer & O. Michie (Eds.), *Machine Intelligence* (Vol. 5).

Received January 8, 2002; accepted May 7, 2002.