4-5-2013

Encoding binary neural codes in networks of threshold-linear neurons

Carina Curto
University of Nebraska-Lincoln, ccurto2@math.unl.edu

Anda Degeratu
Albert-Ludwig-Universitat, a.degeratu@math.uni-freiburg.de

Vladimir Itskov
University of Nebraska-Lincoln, vladimir.itskov@psu.edu

Encoding binary neural codes in networks of threshold-linear neurons

Carina Curto[1] · Anda Degeratu[2] · Vladimir Itskov[1]

April 5, 2013

[1] Department of Mathematics, University of Nebraska-Lincoln
[2] Department of Mathematics, Albert-Ludwig-Universität, Freiburg, Germany
ccurto2@math.unl.edu, anda.degeratu@math.uni-freiburg.de, vladimir.itskov@math.unl.edu

Abstract

Networks of neurons in the brain encode preferred patterns of neural activity via their synaptic connections. Despite receiving considerable attention, the precise relationship between network connectivity and encoded patterns is still poorly understood. Here we consider this problem for networks of threshold-linear neurons whose computational function is to learn and store a set of binary patterns (e.g., a *neural code*) as "permitted sets" of the network. We introduce a simple Encoding Rule that selectively turns "on" synapses between neurons that co-appear in one or more patterns. The rule uses synapses that are *binary*, in the sense of having only two states ("on" or "off"), but also *heterogeneous*, with weights drawn from an underlying synaptic strength matrix S. Our main results precisely describe the stored patterns that result from the Encoding Rule – including unintended "spurious" states – and give an explicit characterization of the dependence on S. In particular, we find that binary patterns are successfully stored in these networks when the excitatory connections between neurons are *geometrically balanced* – i.e., they satisfy a set of geometric constraints. Furthermore, we find that certain types of neural codes are *natural* in the context of these networks, meaning that the full code can be accurately learned from a highly undersampled set of patterns. Interestingly, many commonly observed neural codes in cortical and hippocampal areas are natural in this sense. As an application, we construct networks that encode hippocampal place field codes nearly exactly, following presentation of only a small fraction of patterns. To obtain our results, we prove new theorems using classical ideas from convex and distance geometry, such as Cayley-Menger determinants, revealing a novel connection between these areas of mathematics and coding properties of neural networks.

Contents

1 Introduction

Recurrent networks in cortex and hippocampus exhibit highly constrained patterns of neural activity, even in the absence of sensory inputs [32, 48, 33, 11]. These patterns are strikingly similar in both stimulus-evoked and spontaneous activity [32, 33], suggesting that cortical networks store neural codes consisting of a relatively small number of allowed activity patterns [48, 11]. What is the relationship between the stored patterns of a network and its underlying connectivity? More specifically, given a prescribed set of binary patterns (e.g., a binary neural code), how can one arrange the connectivity of a network such that precisely those patterns are encoded as fixed point attractors of the dynamics, while minimizing the emergence of unwanted "spurious" states? This problem, which we refer to as the *Network Encoding (NE) Problem*, dates back at least to 1982 and has been most commonly studied in the context of the Hopfield model [26, 27, 5, 24]. A major challenge in this line of work has been to characterize the spurious states [3, 4, 2, 24, 43].

In this article we take a new look at the NE problem for networks of threshold-linear neurons whose computational function is to learn and store binary neural codes. Following [47, 23], we regard stored patterns of a threshold-linear network as "permitted sets" (a.k.a. "stable sets" [16]), corresponding to subsets of neurons that may be co-active at stable fixed points of the dynamics in the presence of one or more external inputs. Although our main results do not make any special assumptions about the prescribed sets of patterns to be stored, many commonly observed neural codes are sparse and have a rich internal structure, with correlated patterns reflecting similarities between represented stimuli. Our perspective thus differs somewhat from the traditional Hopfield model [26, 27], where binary patterns are typically assumed to be uncorrelated and dense [5, 24].

To tackle the NE problem we introduce a simple learning rule, called the Encoding Rule, that constructs a network W from a set of prescribed binary patterns $\mathcal{C}$. The rule selectively turns "on" connections between neurons that co-appear in one or more of the presented patterns, and uses synapses that are *binary* (in the sense of having only two states – "on" or "off"), but also *heterogeneous*, with weights drawn from an underlying synaptic strength matrix S. Our main result, Theorem 2, precisely characterizes the full set of permitted sets $\mathcal{P}(W)$ for any network constructed using the Encoding Rule, and shows explicitly the dependence on S. In particular, we find that binary patterns can be successfully stored in these networks if and only if the strengths of excitatory connections among co-active neurons in a pattern are *geometrically balanced* – i.e., they satisfy a set of geometric constraints. Theorem 3 shows that any set of binary patterns that can be exactly encoded as $\mathcal{C} = \mathcal{P}(W)$ for symmetric W can in fact be exactly encoded using our Encoding Rule. Furthermore, when a set of binary patterns $\mathcal{C}$ is *not* encoded exactly, we are able to completely describe the spurious states, and find that they correspond to cliques in the "co-firing" graph $G(\mathcal{C})$.

An important consequence of these findings is that certain neural codes are *natural* in the context of symmetric threshold-linear networks; i.e., the structure of the code closely matches the structure of emerging spurious states via the Encoding Rule, allowing the full code to be accurately learned from a highly undersampled set of patterns. Interestingly, using Helly's theorem [8] we can show that many commonly observed neural codes in cortical and hippocampal areas are natural in this sense. As an application, we construct networks that encode hippocampal place field codes nearly exactly, following presentation of only a small and randomly-sampled fraction of patterns in the code.

The organization of this paper is as follows. In Section 2 we introduce some necessary background on binary neural codes, threshold-linear networks, and permitted sets. In Section 3 we introduce the Encoding Rule and present our Results. The proofs of our main results are given in Section 4 and use ideas from classical distance and convex geometry, such as Cayley-Menger determinants [12], establishing a novel connection between these areas of mathematics and neural network theory. Finally, Sections 5 and 6 contain the Discussion and Appendices, respectively.

2 Background

2.1 Binary neural codes

A binary pattern on n neurons is simply a string of 0s and 1s, with a 1 for each active neuron and a 0 denoting silence; equivalently, it is a subset of (active) neurons

$$\sigma \subset \{1, \ldots, n\} \overset{\text{def}}{=} [n].$$

A *binary neural code* (a.k.a. a combinatorial neural code [18, 40]) is a collection of binary patterns $\mathcal{C} \subset 2^{[n]}$, where $2^{[n]}$ denotes the set of all subsets of $[n]$.

Experimentally observed neural activity in cortical and hippocampal areas suggests that neural codes are *sparse* [29, 7], meaning that relatively few neurons are co-active in response to any given stimulus. Correspondingly, we say that a binary neural code $\mathcal{C} \subset 2^{[n]}$ is *k-sparse*, for $k < n$, if all patterns $\sigma \in \mathcal{C}$ satisfy $|\sigma| \leq k$. Note that in order for a code $\mathcal{C}$ to have good error-correcting capability, the total number of codewords $|\mathcal{C}|$ must be considerably smaller than 2^n [34, 30, 18], a fact that may account for the limited repertoire of observed neural activity.

Important examples of binary neural codes are classical population codes, such as *receptive field codes* (RF codes) [18]. A simple yet paradigmatic example is the hippocampal *place field code* (PF code), where single neuron activity is characterized by place fields [38, 39]. We will consider general RF codes in Section 3.6, and specialize to sparse PF codes in Section 3.7.

2.2 Threshold-linear networks

A *threshold-linear network* [23, 16] is a firing rate model for a recurrent network [20, 22] where the neurons all have threshold nonlinearity, $\phi(z) = [z]_+ = \max\{z, 0\}$. The dynamics are given by

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i + \phi \left(\sum_{j=1}^{n} W_{ij} x_j + e_i - \theta_i \right), \quad i = 1, ..., n,$$

where n is the number of neurons, $x_i(t)$ is the firing rate of the ith neuron at time t, e_i is the external input to the ith neuron, and $\theta_i > 0$ is its threshold. The matrix entry W_{ij} denotes the effective strength of the connection from the jth to the ith neuron, and the timescale $\tau_i > 0$ gives the rate at which a neuron's activity decays to zero in the absence of any inputs (see Figure 1).

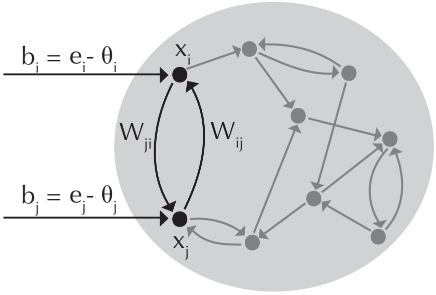

Figure 1: A recurrent network receiving an input vector $b = (b_1, \ldots, b_n)$. The firing rate of each neuron is given by $x_i = x_i(t)$, and evolves in time according to equation (1). The strengths of recurrent connection are captured by the matrix W.

Although sigmoids more closely match experimentally measured input-output curves for neurons, the above threshold nonlinearity is often a good approximation when neurons are far from saturation [20, 45]. Assuming that encoded patterns of a network are in fact realized by neurons that are firing far from saturation, it is reasonable to approximate them as stable fixed points of the threshold-linear dynamics.

These dynamics can be expressed more compactly as

$$\dot{x} = -Dx + [Wx + b]_+, \tag{1}$$

where $D \overset{\text{def}}{=} \text{diag}(1/\tau_1, ..., 1/\tau_n)$ is the diagonal matrix of inverse time constants, W is the synaptic connectivity matrix, $b = (b_1, ..., b_n) \in \mathbb{R}^n$ with $b_i = e_i - \theta_i$, and $[\cdot]_+$ is applied elementwise. Note that, unlike in the Hopfield model, the "input" to the network comes in the form of a constant (in time) external drive b, rather than an initial condition $x(0)$. We think of equation (1) as describing the fast-timescale dynamics of the network, and b as representing the effect of an external stimulus. So long as b changes slowly as compared to the fast network dynamics, the neural responses to individual stimuli are captured by the steady states of (1) in the presence of a constant input vector b.

In the Encoding Rule (Section 3.1), we assume homogeneous timescales and use $D = I$ (the identity matrix). Nevertheless, all results apply equally well to heterogeneous timescales – i.e., for any diagonal D having strictly positive diagonal. We also assume that $-D + W$ has strictly negative diagonal, so that the activity of an individual neuron always decays to zero in the absence of external or recurrent inputs. Although we consider responses to the full range of inputs $b \in \mathbb{R}^n$, the possible steady states of (1) are sharply constrained by the connectivity matrix W. Assuming fixed D, we refer to a particular threshold-linear network simply as W.

2.3 Permitted sets of threshold-linear networks

We consider threshold-linear networks whose computational function is to encode a set of binary patterns. These patterns are stored as "permitted sets" of the network. The theory of permitted (and forbidden) sets was first introduced in [47, 23], and many interesting results were obtained in the case of symmetric threshold-linear networks. Here we review some definitions and results that apply more generally, though later we will also restrict ourselves to the symmetric case.

notation	meaning
$[n]$	$\{1,...,n\}$, $n =$ # of neurons
$2^{[n]}$	the set of all subsets of $[n]$
$\sigma \subset [n]$	a subset of neurons; a binary pattern; a codeword; a permitted set
$\lvert\sigma\rvert$	# of elements (neurons) in the set σ
$\mathcal{C} \subset 2^{[n]}$	a prescribed set of binary patterns – e.g., a binary neural code
$G(\mathcal{C})$	the co-firing graph of $\mathcal{C}$; $(ij) \in G(\mathcal{C}) \Leftrightarrow \{i,j\} \subset \sigma$ for some $\sigma \in \mathcal{C}$
$X(G), X(G(\mathcal{C}))$	the clique complex of the graph G or $G(\mathcal{C})$, respectively
$\mathrm{supp}(x)$	$\{i \in [n] \mid x_i > 0\}$, for $x \in \mathbb{R}^n$
W	an $n \times n$ connectivity matrix; the network with dynamics (1)
D	fixed diagonal matrix of inverse time constants
$\mathcal{P}(W)$	$\{\sigma \subset [n] \mid \sigma$ a permitted set of $W\}$; set of all permitted sets of W
A	an $n \times n$ matrix
A_σ, for $\sigma \subset [n]$	the principal submatrix of A with index set σ
$\mathrm{stab}(A)$	$\{\sigma \subset [n] \mid A_\sigma$ is a stable matrix$\}$
$\mathrm{cm}(A)$	Cayley-Menger determinant of A
$1 \in \mathbb{R}^n$	the column vector with all entries equal to 1
-11^T	$n \times n$ rank 1 matrix with all entries equal to -1

Table 1: Frequently used notation.

Informally, a *permitted set* of a recurrent network is a binary pattern $\sigma \subset [n]$ that can be *activated*. This means there exists an external input to the network such that the neural activity $x(t) = (x_1(t), \ldots, x_n(t))$ converges to a steady state x^* (a stable fixed point) with support σ:

$$\sigma = \mathrm{supp}(x^*) \stackrel{\text{def}}{=} \{i \in [n] \mid x_i^* > 0\}.$$

Definition. A *permitted set* of the network (1) is a subset of neurons $\sigma \subset [n]$ with the property that for at least one external input $b \in \mathbb{R}^n$, there exists an asymptotically stable fixed point x^* such that $\sigma = \mathrm{supp}(x^*)$ [23]. For a given choice of network dynamics, the connectivity matrix W determines the set of all permitted sets of the network, denoted $\mathcal{P}(W)$.

For threshold-linear networks of the form (1), it has been previously shown that permitted sets of W correspond to stable principal submatrices of $-D + W$ [23, 16]. Recall that a *stable* matrix is one whose eigenvalues all have strictly negative real part. For any $n \times n$ matrix A, the notation A_σ denotes the *principal submatrix* obtained by restricting to the index set σ; if $\sigma = \{s_1, ..., s_k\}$, then A_σ is the $k \times k$ matrix with $(A_\sigma)_{ij} = A_{s_i s_j}$. We denote the set of all stable principal submatrices of A as

$$\mathrm{stab}(A) \stackrel{\text{def}}{=} \{\sigma \subset [n] \mid A_\sigma \text{ is a stable matrix}\}.$$

With this notation we can now restate our prior result, which generalizes an earlier result of [23] to non-symmetric networks.

Theorem 1 ([16, Theorem 1.2][1])**.** *Let W be a threshold-linear network on n neurons (not necessarily symmetric) with dynamics given by equation* (1)*, and let $\mathcal{P}(W)$ be the set of all permitted sets of W. Then*

$$\mathcal{P}(W) = \mathrm{stab}(-D + W).$$

[1] Note that in [16, Theorem 1.2], permitted sets were called "stable sets." See also [23] for an earlier proof specific to the symmetric case.

Theorem 1 implies that a binary neural code $\mathcal{C}$ can be exactly encoded as the set of permitted sets in a threshold-linear network if and only if there exists a pair of $n \times n$ matrices (D, W) such that $\mathcal{C} = \text{stab}(-D + W)$. From this observation, it is not difficult to see that not all codes are realizable by threshold-linear networks. This follows from a simple lemma.

Lemma 1. *Let A be an $n \times n$ real-valued matrix (not necessarily symmetric) with strictly negative diagonal and $n \geq 2$. If A is stable, then there exists a 2×2 principal submatrix of A that is also stable.*

Proof. We use the formula for the characteristic polynomial in terms of sums of principal minors:

$$p_A(X) \quad = \quad (-1)^n X^n + (-1)^{n-1} m_1(A) X^{n-1} + (-1)^{n-2} m_2(A) X^{n-2} + ... + m_n(A),$$

where $m_k(A)$ is the sum of the $k \times k$ principal minors of A. Writing the characteristic polynomial in terms of symmetric polynomials in the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$, and assuming A is stable, we have $m_2(A) = \sum_{i<j} \lambda_i \lambda_j > 0$. This implies that at least one 2×2 principal minor is positive. Since the corresponding 2×2 principal submatrix has negative trace, it must be stable. □

Combining Lemma 1 with Theorem 1 then gives:

Corollary 1. *Let $\mathcal{C} \subset 2^{[n]}$. If there exists a pattern $\sigma \in \mathcal{C}$ such that no order 2 subset of σ belongs to $\mathcal{C}$, then it is not realizable as $\mathcal{C} = \mathcal{P}(W)$ for any threshold-linear network W.*

Here we will not pay attention to the relationship between the input to the network b and the corresponding permitted sets that may be activated, as it is beyond the scope of this paper. In prior work, however, we were able to understand with significant detail the relationship between a given b and the set of resulting fixed points of the dynamics [16, Proposition 2.1]. For completeness, we summarize these findings in Appendix D.

Permitted sets of symmetric threshold-linear networks

In the remainder of this work, we will restrict attention to the case of symmetric networks. With this assumption, we can immediately say more about the structure of permitted sets $\mathcal{P}(W)$. Namely, if W is symmetric then the permitted sets $\mathcal{P}(W)$ have the combinatorial structure of a simplicial complex.

Definition. An (abstract) *simplicial complex* $\Delta \subset 2^{[n]}$ is a set of subsets of $[n] = \{1, \ldots, n\}$ such that the following two properties hold: (1) for each $i \in [n], \{i\} \in \Delta$, and (2) if $\sigma \in \Delta$ and $\tau \subset \sigma$, then $\tau \in \Delta$.

Lemma 2. *If W is a symmetric threshold-linear network, then $\mathcal{P}(W)$ is a simplicial complex.*

In other words, if W is symmetric then every subset of a permitted set is permitted, and every superset of a set that is not permitted is also not permitted. This was first observed in [23], using an earlier version of Theorem 1 for symmetric W. It follows from the fact that $\mathcal{P}(W) = \text{stab}(-D+W)$, by Theorem 1, and stab$(A)$ is a simplicial complex for any symmetric $n \times n$ matrix A having strictly negative diagonal (see Corollary 7 in Appendix A). The proof of this fact is a straightforward application of Cauchy's interlacing theorem (Appendix A), which applies only to symmetric matrices.

We are not currently aware of any simplicial complex Δ that is not realizable as $\Delta = \mathcal{P}(W)$ for a symmetric threshold-linear network, although we believe such examples are likely to exist.

3 Results

Theorem 1 allows one to find all permitted sets $\mathcal{P}(W)$ of a given network W. Our primary interest, however, is in the inverse problem.

> ***NE problem**: Given a set of binary patterns $\mathcal{C} \subset 2^{[n]}$, how can one construct a network W such that $\mathcal{C} \subseteq \mathcal{P}(W)$, while minimizing the emergence of unwanted spurious states?*

Note that *spurious states* are elements of $\mathcal{P}(W)$ that were *not* in the prescribed set of patterns to be stored; these are precisely the elements of $\mathcal{P}(W) \setminus \mathcal{C}$. If $\mathcal{C} \subset \mathcal{P}(W)$, so that all patterns in $\mathcal{C}$ are stored as permitted sets of W but $\mathcal{P}(W)$ may contain additional spurious states, then we say that $\mathcal{C}$ has been *encoded* by the network W. If $\mathcal{C} = \mathcal{P}(W)$, so that there are no spurious states, then we say that $\mathcal{C}$ has been *exactly encoded* by W.

We tackle the NE problem by analyzing a novel learning rule, called the Encoding Rule. In what follows, the problem is broken into four motivating questions that address (1) the learning rule, (2) the resulting structure of permitted sets, (3) binary codes that are exactly encodable, and (4) the structure of spurious states when codes are not encoded exactly. In Section 3.6 we use our results to uncover "natural" codes for symmetric threshold-linear networks, and illustrate this phenomenon in the case of hippocampal PF codes in Section 3.7.

3.1 The Encoding Rule

Question 1: Is there a biologically plausible learning rule that allows arbitrary neural codes to be stored as permitted sets in threshold-linear networks?

In this section we introduce a novel encoding rule that constructs a network W from a prescribed set of binary patterns $\mathcal{C}$. The rule is similar to the classical Hopfield learning rule [26] in that it updates the weights of the connectivity matrix W following sequential presentation of binary patterns, and strengthens excitatory synapses between co-active neurons in the patterns. In particular, the rule is Hebbian and *local*; i.e., each synapse is updated only in response to the co-activation of the two adjacent neurons, and the updates can be implemented by presenting only one pattern at a time [26, 20]. A key difference from the Hopfield rule is that the synapses are *binary*: once a synapse (ij) has been turned "on," the value of W_{ij} stays the same irrespective of the remaining patterns.[2] A new ingredient is that synapses are allowed to be *heterogeneous*: in other words, the actual weights of connections are varied among "on" synapses. These weights are assigned according to a predetermined *synaptic strength matrix* S, which is considered fixed and reflects the underlying architecture of the network. For example, if no physical connection exists between neurons i and j, then $S_{ij} = 0$ indicating that no amount of co-firing can cause a direct excitatory connection betwen those neurons. On the other hand, if two neurons have multiple points of physical contact, then S_{ij} will be greater than if there are only a few anatomical contacts. There is, in fact, experimental evidence in hippocampus for synapses that appear binary and heterogeneous in this sense [42], with individual synapses exhibiting potentiation in an all-or-nothing fashion, but having different "thresholds" for potentiation and heterogeneous synaptic strengths.

Here we describe the Encoding Rule in general, with minimal assumptions on S. Later, in Sections 3.4 and 3.5, we will investigate the consequences of various choices of S on the network's ability to encode different types of binary neural codes.

[2] The learning rule in [47] also had binary synapses in this sense.

Encoding Rule. This is a prescription for constructing (i.e., "learning") a network W from a set of binary patterns on n neurons, $\mathcal{C} \subset 2^{[n]}$ (e.g., $\mathcal{C}$ is a binary neural code). It consists of three steps: two initialization steps (Steps 1-2), followed by an update step (Step 3).

> **Step 1**: Fix an $n \times n$ synaptic strength matrix S and an $\varepsilon > 0$. We think of S and ε as *intrinsic* properties of the underlying network architecture, established prior to learning. Because S contains synaptic strengths for symmetric excitatory connections, we require that $S_{ij} = S_{ji} \geq 0$ and $S_{ii} = 0$.
>
> **Step 2**: The network W is initialized to be symmetric with effective connection strengths $W_{ij} = W_{ji} < -1$ for $i \neq j$, and $W_{ii} = 0$. (Beyond this requirement, the initial values of W do not affect the results.)
>
> **Step 3**: Following presentation of each pattern $\sigma \in \mathcal{C}$, we turn "on" all excitatory synapses between neurons that co-appear in σ. This means we update the relevant entries of W as follows:
>
> $$W_{ij} := -1 + \varepsilon S_{ij} \;\text{ if }\; i, j \in \sigma \text{ and } i \neq j.$$
>
> Note that the order of presentation does not matter; once an excitatory connection has been turned "on," the value of W_{ij} stays the same irrespective of remaining patterns.

To better understand what kinds of networks result from the Encoding Rule, observe that any initial W in Step 2 can be written as $W_{ij} = -1 - \varepsilon R_{ij}$, where $R_{ij} = R_{ji} > 0$ for $i \neq j$ and $R_{ii} = -1/\varepsilon$, so that $W_{ii} = 0$. Assuming a threshold-linear network with homogeneous timescales, i.e. fixing $D = I$, the final network W obtained from $\mathcal{C}$ after Step 3 satisfies,

$$(-D + W)_{ij} = \begin{cases} -1 + \varepsilon S_{ij}, & \text{if} \quad (ij) \in G(\mathcal{C}) \\ \quad -1, & \text{if} \qquad i = j \\ -1 - \varepsilon R_{ij} & \text{if} \quad (ij) \notin G(\mathcal{C}), \end{cases} \tag{2}$$

where $G(\mathcal{C})$ is the graph on n vertices (neurons) having an edge for each pair of neurons that co-appear in one or more patterns of $\mathcal{C}$. We call this graph the *co-firing graph* of $\mathcal{C}$. In essence, the rule allows the network to "learn" $G(\mathcal{C})$, selecting which excitatory synapses are turned "on" and assigned to their predetermined weights.

Consequently, any matrix $-D + W$ obtained via the Encoding Rule thus has the form

$$-11^T + \varepsilon A,$$

where -11^T denotes the $n \times n$ matrix of all -1s and A is a symmetric matrix with zero diagonal and off-diagonal entries $A_{ij} = S_{ij} \geq 0$ or $A_{ij} = -R_{ij} < 0$, depending on $\mathcal{C}$. It then follows from Theorem 1 that the permitted sets of this network are

$$\mathcal{P}(W) = \text{stab}(-11^T + \varepsilon A).$$

Furthermore, it turns out that $\mathcal{P}(W)$ for *any* symmetric W is of this form, even if $-D + W$ is not of the form $-11^T + \varepsilon A$.

Lemma 3. *If W is a symmetric threshold-linear network (with D not necessarily equal to the identity matrix I), then there exists a symmetric $n \times n$ matrix A with zero diagonal such that $\mathcal{P}(W) = \text{stab}(-11^T + A)$.*

The proof is given in Appendix B (Lemma 14).

In addition to being symmetric, the Encoding Rule (for small enough ε) generates "lateral inhibition" networks where the matrix $-D+W$ has strictly negative entries. In particular, this means that the matrix $D-W$ is *copositive* – i.e., $x^T(D-W)x > 0$ for all nonnegative x except $x=0$. It follows from [23, Theorem 1] that for all input vectors $b \in \mathbb{R}^n$ and for all initial conditions, the network dynamics (1) converge to an equilibrium point. This was proven by constructing a Lyapunov-like function, similar to the strategy in [15].[3]

3.2 Main Result

Question 2: What is the full set of permitted sets $\mathcal{P}(W)$ stored in a network constructed using the Encoding Rule?

Our main result, Theorem 2, characterizes the full set of permitted sets $\mathcal{P}(W)$ obtained using the Encoding Rule, revealing a detailed understanding of the structure of spurious states. Recall from Lemma 3 that the set of permitted sets of any symmetric network on n neurons has the form $\mathcal{P}(W) = \text{stab}(-11^T + \varepsilon A)$, for $\varepsilon > 0$ and A a symmetric $n \times n$ matrix with zero diagonal.[4] Describing $\mathcal{P}(W)$ thus requires understanding the stability of the principal submatrices $(-11^T + \varepsilon A)_\sigma$ for each $\sigma \subset [n]$. Note that these submatrices all have the same form: $-11^T + \varepsilon A_\sigma$, where -11^T is the all -1s matrix of size $|\sigma| \times |\sigma|$. Proposition 1 (below) provides an unexpected connection between the stability of these matrices and classical *distance geometry*.[5] We will first present Proposition 1, and then show how it leads us to Theorem 2.

For symmetric 2×2 matrices of the form $-11^T + \varepsilon A = \begin{bmatrix} -1 & -1+\varepsilon A_{12} \\ -1+\varepsilon A_{12} & -1 \end{bmatrix}$, with $\varepsilon > 0$, it is easy to identify the conditions for the matrix to be stable. Namely, one needs the determinant to be positive, so $A_{12} > 0$ and $\varepsilon < 2/A_{12}$. For 3×3 matrices, the conditions are more interesting, and the connection to distance geometry emerges.

Lemma 4. *Consider the 3×3 matrix $-11^T + \varepsilon A$, for a fixed symmetric A with zero diagonal:*

$$\begin{bmatrix} -1 & -1+\varepsilon A_{12} & -1+\varepsilon A_{13} \\ -1+\varepsilon A_{12} & -1 & -1+\varepsilon A_{23} \\ -1+\varepsilon A_{13} & -1+\varepsilon A_{23} & -1 \end{bmatrix}.$$

There exists an $\varepsilon > 0$ such that this matrix is stable if and only if $\sqrt{A_{12}}$, $\sqrt{A_{13}}$, and $\sqrt{A_{23}}$ are valid edge lengths for a nondegenerate triangle in $\mathbb{R}^2$.

In other words, the numbers $\sqrt{A_{ij}}$ must satisfy the triangle inequalities $\sqrt{A_{ij}} < \sqrt{A_{ik}} + \sqrt{A_{jk}}$ for distinct i, j, k. This can be proven by straightforward computation, using Heron's formula and the characteristic polynomial of the matrix. The upper bound on ε, however, is not so easy to identify.

Remarkably, the above observations completely generalize to $n \times n$ matrices of the form $-11^T + \varepsilon A$, and the precise limits on ε can also be computed for general n. This is the content of Proposition 1, below. To state it, however, we first need a few notions from distance geometry.

Definition. An $n \times n$ matrix A is a (Euclidean) *square distance matrix* if there exists a configuration of points $p_1, ..., p_n \in \mathbb{R}^{n-1}$ (not necessarily distinct) such that $A_{ij} = \|p_i - p_j\|^2$. A is a *nondegenerate*

[3] Note that threshold-linear networks do not directly fall into the very general class of networks discussed in [15].

[4] In fact, any $\mathcal{P}(W)$ of this form can be obtained by perturbing around any rank 1 matrix – not necessarily symmetric – having strictly negative diagonal (Proposition 3, in Appendix B).

[5] Distance geometry is a field of mathematics that was developed in the early 20th century, motivated by the following problem: find necessary and sufficient conditions such that a finite set of distances can be realized from a configuration of points in Euclidean space. The classical text on this subject is Blumenthal's 1953 book [12].

square distance matrix if the corresponding points are affinely independent; i.e., if the convex hull of $p_1, ..., p_n$ is a simplex with nonzero volume in $\mathbb{R}^{n-1}$.

Clearly, all square distance matrices are symmetric and have zero diagonal. Furthermore, a 2×2 matrix A is a *nondegenerate* square distance matrix if and only if the off-diagonal entry satisfies the additional condition $A_{12} > 0$. For a 3×3 matrix A, the necessary and sufficient condition to be a nondegenerate square distance matrix is that the entries $\sqrt{A_{12}}, \sqrt{A_{13}}$, and $\sqrt{A_{23}}$ are valid edge lengths for a nondegenerate triangle in $\mathbb{R}^2$ (this was precisely the condition in Lemma 4). For larger matrices, however, the conditions are less intuitive. A key object for determining whether or not and $n \times n$ matrix A is a nondegenerate square distance matrix is the *Cayley-Menger determinant*:

$$\mathrm{cm}(A) \overset{\text{def}}{=} \det \begin{bmatrix} 0 & 1^T \\ 1 & A \end{bmatrix},$$

where $1 \in \mathbb{R}^n$ is the column vector of all ones. If A is a square distance matrix, then $\mathrm{cm}(A)$ is proportional to the square volume of the simplex obtained as the convex hull of the points $\{p_i\}$ (see Lemma 11 in Appendix A). In particular, $\mathrm{cm}(A) \neq 0$ (and hence $|\mathrm{cm}(A)| > 0$) if A is a *nondegenerate* square distance matrix, while $\mathrm{cm}(A) = 0$ for any other (degenerate) square distance matrix.

Proposition 1. *Let A be a symmetric $n \times n$ matrix with zero diagonal, and $\varepsilon > 0$. Then the matrix*

$$-11^T + \varepsilon A$$

is stable if and only if the following two conditions hold:

(a) A is a nondegenerate square distance matrix, and

(b) $0 < \varepsilon < |\mathrm{cm}(A)/\det(A)|$.

Proposition 1 is essentially a special case of Theorem 4 – our core technical result – whose statement and proof are given in Section 4.1. The proof of Proposition 1 is then given in Section 4.2. To our knowledge, Theorem 4 is novel and connections to distance geometry have not previously been used in the study of neural networks or, more generally, the stability of fixed points in systems of ODEs.

The ratio $|\mathrm{cm}(A)/\det(A)|$ has a simple geometric interpretation in cases where condition (a) of Proposition 1 holds. Namely, if A is an $n \times n$ nondegenerate square distance matrix (with $n > 1$), then $|\mathrm{cm}(A)/\det(A)| = \dfrac{1}{2\rho^2}$, where ρ is the radius of the unique sphere circumscribed on any set of points in Euclidean space that can be used to generate A (see Remark 1 in Appendix C). Moreover, since $|\mathrm{cm}(A)| > 0$ whenever A is a nondegenerate square distance matrix, there always exists an ε small enough to satisfy the second condition, provided the first condition holds. Combining Proposition 1 together with Cauchy's interlacing theorem yields:

Lemma 5. *If A is an $n \times n$ nondegenerate square distance matrix, then*

$$0 < \left|\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)}\right| \leq \left|\frac{\mathrm{cm}(A_\tau)}{\det(A_\tau)}\right| \quad \textit{if } \tau \subseteq \sigma \subseteq [n].$$

Given any symmetric $n \times n$ matrix A with zero diagonal, and $\varepsilon > 0$, it is now natural to define the following simplicial complexes:

$$\begin{aligned} \mathrm{geom}_\varepsilon(A) &\overset{\text{def}}{=} \{\sigma \subseteq [n] \mid A_\sigma \text{ a nondeg. sq. dist. matrix and } \left|\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)}\right| > \varepsilon\}, \text{ and} \\ \mathrm{geom}(A) &\overset{\text{def}}{=} \lim_{\varepsilon \to 0} \mathrm{geom}_\varepsilon(A) = \{\sigma \subseteq [n] \mid A_\sigma \text{ a nondeg. sq. dist. matrix}\}. \end{aligned}$$

Lemma 5 implies that $\mathrm{geom}_\varepsilon(A)$ and $\mathrm{geom}(A)$ are simplicial complexes. Note that if $\sigma = \{i\}$, we have $A_\sigma = [0]$. In this case, $\{i\} \in \mathrm{geom}(A)$ and $\{i\} \in \mathrm{geom}_\varepsilon(A)$ for all $\varepsilon > 0$, by our convention. Also, $\mathrm{geom}_\varepsilon(A) = \mathrm{geom}(A)$ if and only if $0 < \varepsilon < \delta(A)$, where

$$\delta(A) \overset{\text{def}}{=} \min \left\{ \left| \frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)} \right| \right\}_{\sigma \in \mathrm{geom}(A)}.$$

If A is a nondegenerate square distance matrix, then $\delta(A) = |\mathrm{cm}(A)/\det(A)|$.

To state our main result, Theorem 2, we also need a few standard notions from graph theory. A *clique* in a graph G is a subset of vertices that is all-to-all connected.[6] The *clique complex* of G, denoted $X(G)$, is the set of all cliques in G; this is a simplicial complex for any G. Here we are primarily interested in the graph $G(\mathcal{C})$, the co-firing graph of a set of binary patterns $\mathcal{C} \subset 2^{[n]}$.

Theorem 2. *Let S be an $n \times n$ synaptic strength matrix satisfying $S_{ij} = S_{ji} \geq 0$ and $S_{ii} = 0$ for all $i, j \in [n]$, and fix $\varepsilon > 0$. Given a set set of prescribed patterns $\mathcal{C} \subset 2^{[n]}$, let W be the threshold-linear network (equation* (2)*) obtained from $\mathcal{C}$ using S and ε in the Encoding Rule. Then,*

$$\mathcal{P}(W) = \mathrm{geom}_\varepsilon(S) \cap X(G(\mathcal{C})).$$

If we further assume that $\varepsilon < \delta(S)$, then $\mathcal{P}(W) = \mathrm{geom}(S) \cap X(G(\mathcal{C}))$.

In other words, a binary pattern $\sigma \subset [n]$ is a permitted set of W if and only if S_σ is a nondegenerate square distance matrix, $\varepsilon < |\mathrm{cm}(S_\sigma)/\det(S_\sigma)|$, and σ is a clique in the graph $G(\mathcal{C})$.

The proof is given in Section 4.2. Theorem 2 answers Question 2, and makes explicit how $\mathcal{P}(W)$ depends on S, ε, and $\mathcal{C}$. One way of interpreting this result is to observe that a binary pattern $\sigma \in \mathcal{C}$ is successfully stored as a permitted set of W if and only if the excitatory connections between the neurons in σ, given by $\tilde{S}_\sigma = \varepsilon S_\sigma$, are *geometrically balanced* – i.e.,

- $\tilde{S}_\sigma$ is a nondegenerate square distance matrix, and
- $|\det(\tilde{S}_\sigma)| < |\mathrm{cm}(\tilde{S}_\sigma)|$.

The first condition ensures a certain balance among the relative strengths of excitatory connections in the clique σ, while the second condition bounds the overall excitation strengths relative to inhibition (which has been normalized to -1 in the Encoding Rule).

We next turn to an example that illustrates how this theorem can be used to solve the NE problem explicitly for a small binary neural code. In the following section, Section 3.4, we address more generally the question of what neural codes can be encoded exactly, and what is the structure of spurious states when a code is encoded inexactly.

3.3 An example

Suppose $\mathcal{C}$ is a binary neural code on $n = 6$ neurons, consisting of maximal patterns

$$\{110100, 101010, 011001, 000111\},$$

corresponding to subsets $\{124\}, \{135\}, \{236\}$, and $\{456\}$, together with all subpatterns (smaller subsets) of the maximal ones, thus ensuring that $\mathcal{C}$ is a simplicial complex. This is depicted in Figure 2A, using a standard method of illustrating simplicial complexes geometrically. The four maximal patterns correspond to the shaded triangles, while patterns with only one or two co-active neurons comprise the vertices and edges of the co-firing graph $G(\mathcal{C})$.[7]

[6] For recent work encoding cliques in Hopfield networks, see [25].

[7] In this example, there are no patterns having four or more neurons, but these would be illustrated by tetrahedra and other higher-order simplices.

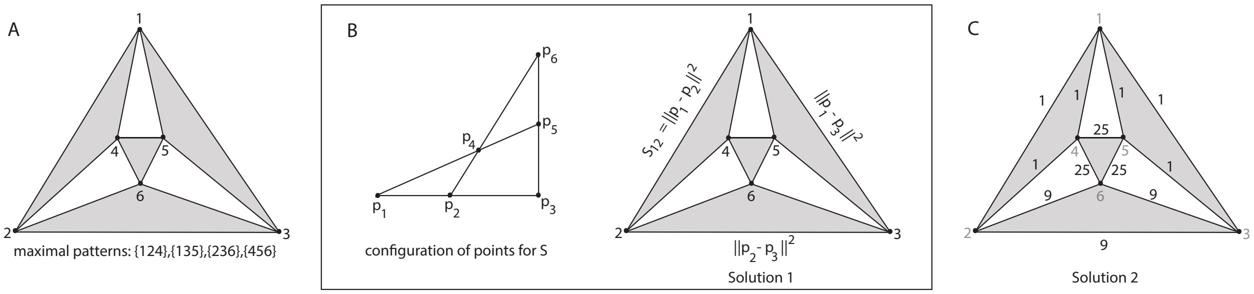

Figure 2: An example on $n = 6$ neurons. (A) The simplicial complex $\mathcal{C}$ consists of 4 two-dimensional facets (shaded triangles). The graph $G(\mathcal{C})$ contains the 6 vertices and 12 depicted edges; these are also included in $\mathcal{C}$, so the size of the code is $|\mathcal{C}| = 22$. (B) A configuration of points $p_1, ..., p_6 \in \mathbb{R}^2$ that can be used to exactly encode $\mathcal{C}$. Lines indicate triples of points that are collinear. From this configuration we construct a 6×6 synaptic strength matrix S, with $S_{ij} = \|p_i - p_j\|^2$, and choose $0 < \varepsilon < \delta(S)$. The geometry of the configuration implies that $\text{geom}(S)$ does not contain any patterns of size greater than 3, nor does it contain the triples $\{123\}, \{145\}, \{246\}$, or $\{356\}$. It is straightforward to check that $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$. (C) Another solution for exactly encoding $\mathcal{C}$ is provided by choosing the matrix S with S_{ij} given by the labeled edges in the figure. The square distances in S_{ij} were chosen to satisfy the triangle inequalities for shaded triangles, but to violoate them for empty triangles.

Without Theorem 2, it is difficult to find a network W that encodes $\mathcal{C}$ exactly – i.e., such that $\mathcal{P}(W) = \mathcal{C}$. This is in part because each connection strength W_{ij} belongs to two 3×3 matrices that must satisfy opposite stability properties. For example, the subset $\{124\}$ must be a permitted set of $\mathcal{P}(W)$ while $\{123\}$ is not permitted, imposing competing conditions on the entry W_{12}. In general, it may be difficult to "patch" together local ad-hoc solutions to obtain a single matrix W having all the desired stability properties.

Using Theorem 2, on the other hand, we can easily construct many exact solutions for encoding $\mathcal{C}$ as a set of permitted sets $\mathcal{P}(W)$. The main idea is as follows. Consider the Encoding Rule with synaptic strength matrix S and $0 < \varepsilon < \delta(S)$. Applying the rule to $\mathcal{C}$ yields a network with permitted sets

$$\mathcal{P}(W) = \text{geom}(S) \cap X(G(\mathcal{C})).$$

The goal is thus to find S so that $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$. From the co-firing graph $G(\mathcal{C})$, we see that the clique complex $X(G(\mathcal{C}))$ contains all triangles depicted in Figure 2A, including the "empty" (non-shaded) triangles: $\{123\}, \{145\}, \{246\}$, and $\{356\}$. The matrix S must therefore be chosen so that these triples are not in $\text{geom}(S)$, while ensuring that $\{124\}, \{135\}, \{236\}$, and $\{456\}$ are included. In other words, to obtain an exact solution we must find S such that S_σ is a nondegenerate square distance matrix for each $\sigma \in \{\{124\}, \{135\}, \{236\}, \{456\}\}$, but *not* for σ corresponding to an empty triangle.

Solution 1. Consider the configuration of points $p_1, ..., p_6 \in \mathbb{R}^2$ in Figure 2B, and let S be the 6×6 square distance matrix with entries $S_{ij} = \|p_i - p_j\|^2$. Because the points lie in the plane, the largest principal submatrices of S that can possibly be nondegenerate square distance matrices are 3×3. This means $\text{geom}(S)$ has no elements of size greater than 3. Because no two points have the same position, $\text{geom}(S)$ contains the complete graph with all edges (ij). It remains only to determine which triples are in $\text{geom}(S)$. The only 3×3 principal submatrices of S that are nondegenerate square distance matrices correspond to triples of points in general position. From Figure 2B (left) we see that $\text{geom}(S)$ includes all triples *except* $\{123\}, \{145\}, \{246\}$, and $\{356\}$, since these correspond to triples of points that are collinear (and thus yield *degenerate* square distance matrices). Although $\mathcal{C} \neq X(G(\mathcal{C}))$ and $\mathcal{C} \neq \text{geom}(S)$, it is now easy to check that $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$. Using

Theorem 2, we can conclude that $\mathcal{C} = \mathcal{P}(W)$ exactly, where W is the network obtained using the Encoding Rule with this S and any $0 < \varepsilon < \delta(S)$.

Solution 2. Let S be the symmetric matrix defined by the following equations for $i < j$: $S_{ij} = 1$ if $i = 1$; $S_{24} = S_{35} = 1$; $S_{23} = S_{26} = S_{36} = 3^2$; and $S_{ij} = 5^2$ if $i = 4$ or 5. Here we've only assigned values corresponding to each edge in $G(\mathcal{C})$ (see Figure 2C); remaining entries may be chosen arbitrarily, as they play no role after we intersect $\text{geom}(S) \cap X(G(\mathcal{C}))$. Note that S is *not* a square distance matrix at all, not even a degenerate one. Nevertheless, S_σ *is* a nondegenerate square distance matrix for $\sigma \in \{\{124\}, \{135\}, \{236\}, \{456\}\}$, because the distances correspond to nondegenerate triangles. For example, the triple $\{124\}$ has pairwise distances $(1, 1, 1)$, which satisfy the triangle inequality. In contrast, the triple $\{123\}$ has pairwise distances $(1, 1, 3)$, which violate the triangle inequality; hence, $S_{\{123\}}$ is *not* a square distance matrix. Similarly, the triangle inequality is violated for each of $\{145\}$, $\{246\}$, and $\{356\}$. It is straightforward to check that, among all cliques of $X(G(\mathcal{C}))$, only the desired patterns in $\mathcal{C}$ are also elements of $\text{geom}(S)$, so $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$.

3.4 Binary neural codes that can be encoded exactly

Question 3: What binary neural codes can be encoded exactly as $\mathcal{C} = \mathcal{P}(W)$ for a symmetric threshold-linear network W?

Question 4: If encoding is not exact, what is the structure of spurious states?

From Theorem 2, it is clear that if the set of patterns to be encoded happens to be of the form $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$, then $\mathcal{C}$ can be exactly encoded as $\mathcal{P}(W)$ for small enough ε and the same choice of S. Similarly, if the set of patterns has the form $\mathcal{C} = \text{geom}_\varepsilon(S) \cap X(G(\mathcal{C}))$, then $\mathcal{C}$ can be exactly encoded as $\mathcal{P}(W)$ using our Encoding Rule (Section 3.1) with the same S and ε. Can any other sets of binary patterns be encoded exactly via symmetric threshold-linear networks? The next theorem assures us that the answer is "No." This means that by focusing attention on networks constructed using our Encoding Rule, we are not missing any binary neural codes that could arise as $\mathcal{P}(W)$ for other symmetric networks.

Theorem 3. *Let $\mathcal{C} \subset 2^{[n]}$ be a binary neural code. There exists a symmetric threshold-linear network W such that $\mathcal{C} = \mathcal{P}(W)$ if and only if $\mathcal{C}$ is a simplicial complex of the form*

$$\mathcal{C} = \text{geom}_\varepsilon(S) \cap X(G(\mathcal{C})), \tag{3}$$

for some $\varepsilon > 0$ and S an $n \times n$ matrix satisfying $S_{ij} = S_{ji} \geq 0$ and $S_{ii} = 0$ for all $i, j \in [n]$. Moreover, W can be constructed using the Encoding Rule on $\mathcal{C}$, using this choice of S and ε.

The proof is given in Section 4.2. Theorem 3 allows us to make a preliminary classification of binary neural codes that can be encoded exactly, giving a partial answer to Question 3. To do this, it is useful to distinguish three different types of S matrices that can be used in the Encoding Rule.

- **universal** S. We say that a matrix S is "universal" if it is an $n \times n$ nondegenerate square distance matrix. In particular, any principal submatrix S_σ is also a nondegenerate square distance matrix, so if we let $0 < \varepsilon < \delta(S) = |\text{cm}(S)/\det(S)|$ then any $\sigma \in \mathcal{C}$ has corresponding excitatory connections εS_σ that are geometrically balanced (see Section 3.2). Furthermore, $\text{geom}_\varepsilon(S) = \text{geom}(S) = 2^{[n]}$, and hence $\text{geom}_\varepsilon(S) \cap X(G(\mathcal{C})) = X(G(\mathcal{C}))$, irrespective of S. It follows that if $\mathcal{C} = X(G)$ for any graph G, then $\mathcal{C}$ can be *exactly* encoded using any universal S and any $0 < \varepsilon < \delta(S)$ in the Encoding Rule.[8] Moreover, since $\mathcal{C} \subset X(G(\mathcal{C}))$ for any code

[8] Note that if $\mathcal{C} = X(G)$ is any clique complex with underlying graph G, then we automatically know that $G(\mathcal{C}) = G$, and hence $X(G(\mathcal{C})) = X(G) = \mathcal{C}$.

type of S matrix	$\mathcal{C}$ that can be *exactly* encoded: $\mathcal{C} = \mathcal{P}(W)$	$\mathcal{C}$ that can be encoded: $\mathcal{C} \subset \mathcal{P}(W)$	spurious states $\mathcal{P}(W) \setminus \mathcal{C}$
universal S	any clique complex $X(G)$	all codes	cliques of $G(\mathcal{C})$ that are not in $\mathcal{C}$
k-sparse universal S	any k-skeleton $X_k(G)$ of a clique complex	all k-sparse codes ($\lvert\sigma\rvert \leq k$ for all $\sigma \in \mathcal{C}$)	cliques of $G(\mathcal{C})$ of size $\leq k$, that are not in $\mathcal{C}$
specially- tuned S	$\mathcal{C}$ is of the form $\text{geom}(S) \cap X(G)$	depends on S	cliques of $G(\mathcal{C})$ that are in $\text{geom}(S)$ but not in $\mathcal{C}$

Table 2: Classification of S matrices, together with encodable codes and spurious states. Note: the above assumes using the Encoding Rule on the code $\mathcal{C}$ with synaptic strength matrix S and $0 < \varepsilon < \delta(S)$. Additional codes may be exactly encodable for other choices of ε.

$\mathcal{C}$, it follows that any code can be encoded – albeit inexactly – using a universal S in the Encoding Rule. Finally, the spurious states $\mathcal{P}(W) \setminus \mathcal{C}$ can be completely understood: they consist of all cliques in the graph $G(\mathcal{C})$ that are not elements of $\mathcal{C}$.

- **k-sparse universal** S. We say that a matrix S is "k-sparse universal" if it is a (degenerate) $n \times n$ square distance matrix for a configuration of n points that are in *general position*[9] in $\mathbb{R}^k$, for $k < n-1$ (otherwise S is universal). Let $0 < \varepsilon < \delta(S)$. Then, $\text{geom}_\varepsilon(S) = \text{geom}(S) = \{\sigma \subset [n] \mid |\sigma| \leq k+1\}$; this is the *k-skeleton*[10] of the complete simplicial complex $2^{[n]}$. This implies that $\text{geom}_\varepsilon(S) \cap X(G(\mathcal{C})) = X_k(G(\mathcal{C}))$, where X_k denotes the k-skeleton of the clique complex X:
$$X_k(G(\mathcal{C})) \overset{\text{def}}{=} \{\sigma \in X(G(\mathcal{C})) \mid |\sigma| \leq k+1\}.$$
It follows that any k-skeleton of a clique complex, $\mathcal{C} = X_k(G)$ for any graph G, can be encoded exactly. Furthermore, since any k-sparse code $\mathcal{C}$ satisfies $\mathcal{C} \subseteq X_k(G(\mathcal{C}))$, any k-sparse code can be encoded using this type of S matrix in the Encoding Rule. The spurious states in this case are cliques of $G(\mathcal{C})$ that have size no greater than k.

- **specially-tuned** S. We will refer to all S matrices that do not fall into the universal or k-sparse universal categories as "specially-tuned." In this case, we cannot say anything general about the codes that are exactly encodable without further knowledge about S. If we let $0 < \varepsilon < \delta(S)$, as above, Theorem 3 tells us that the binary codes $\mathcal{C}$ can be encoded exactly (via the Encoding Rule) are of the form $\mathcal{C} = \text{geom}(S) \cap X(G(\mathcal{C}))$. Unlike in the universal and k-sparse universal cases, the encodable codes depend on the precise form of S. Note that the example code $\mathcal{C}$ discussed in Section 3.3 was not a clique complex nor the k-skeleton of a clique complex. Nevertheless, it could be encoded exactly for the "specially-tuned" choices of S exhibited in Solution 1 and Solution 2 (see Figure 2B,C).

[9] This guarantees that all $k \times k$ principal submatrices of S are nondegenerate square distance matrices.

[10] The k-skeleton Δ_k of a simplicial complex Δ is obtained by restricting to faces of dimension $\leq k$, which corresponds to keeping only elements $\sigma \subset \Delta$ of size $|\sigma| \leq k+1$. Note that Δ_k is also a simplicial complex.

A summary of what codes are encodable and exactly encodable for each type of S matrix is shown in Table 2, under the assumption that $0 < \varepsilon < \delta(S)$ in the Encoding Rule.

We end this section with several technical Remarks, along with some open questions for further mathematical investigation.

Remarks

1. Fine-tuning? It is worth noting here that solutions obtained by choosing S to be a *degenerate* square distance matrix, as in the k-sparse universal S or the specially-tuned S of Figure 2B, are not as fine-tuned as they might first appear. This is because the ratio $|\mathrm{cm}(S_\sigma)/\det(S_\sigma)|$ approaches zero as subsets of points $\{p_i\}_{i\in\sigma}$ used to generate S become *approximately* degenerate, allowing elements to be eliminated from $\mathrm{geom}_\varepsilon(S)$ because of violations to condition (b) in Proposition 1, even if condition (a) is not quite violated. This means the appropriate matrices do not have to be exactly degenerate, but only approximately degenerate (see Remark 2 in Appendix C). In particular, the collinear points in Figure 2B need not be exactly collinear for Solution 1 to hold.

2. Controlling spurious cliques in sparse codes. If the set of patterns $\mathcal{C} \subset 2^{[n]}$ to be encoded is a k-*sparse* code, i.e. if $|\sigma| \leq k < n$ for all $\sigma \in \mathcal{C}$, then any clique of size $k+1$ or greater in $G(\mathcal{C})$ is potentially a spurious clique. We can eliminate these spurious states, however, by choosing a k-sparse universal S in the Encoding Rule. This guarantees that $\mathrm{geom}_\varepsilon(S)$ does not include any element of size greater than k, and hence $\mathcal{P}(W) \subseteq X_{k-1}(G(\mathcal{C}))$.

3. Uniform S. To use truly binary synapses, we can choose S in the Encoding Rule to be the uniform synaptic strength matrix having $S_{ij} = 1$ for $i \neq j$ and $S_{ii} = 0$ for all $i \in [n]$. In fact, S is a nondegenerate square distance matrix, so this is a special case of a "universal" S. Here $\delta(S)$ turns out to have a very simple form:

$$\delta(S) = |\mathrm{cm}(S)/\det(S)| = \frac{n}{n-1}.$$

Similarly, any $k \times k$ principal submatrix S_σ, with $|\sigma| = k$, satisfies $\delta(S_\sigma) = \dfrac{k}{k-1}$. This implies that $\mathrm{geom}_\varepsilon(S)$ is the k-skeleton of the complete simplicial complex on n vertices if and only if

$$\frac{k+2}{k+1} < \varepsilon < \frac{k+1}{k}.$$

It follows that for this choice of S and ε (note that $\varepsilon > \delta(S)$), the Encoding Rule yields $\mathcal{P}(W) = X_k(G(\mathcal{C}))$, just as in the case of k-sparse universal S. If, on the other hand, we choose $0 < \varepsilon \leq 1 < \delta(S)$, then $\mathrm{geom}_\varepsilon(S) = \mathrm{geom}(S) = 2^{[n]}$, then we have the usual properties for universal S (c.f. [47]).

4. Matroid complexes. In the special case where S is a square distance matrix, $\mathrm{geom}(S)$ is a *representable matroid complex* – i.e., it is the independent set complex of a real-representable matroid [41]. Moreover, all representable matroid complexes are of this form, and can thus be encoded exactly. To see this, take any code $\mathcal{C}$ having $G(\mathcal{C}) = K_n$, the complete graph on n vertices. Then $X(G(\mathcal{C})) = 2^{[n]}$, and the Encoding Rule (for $\varepsilon < \delta(S)$) yields

$$\mathcal{P}(W) = \mathrm{geom}(S).$$

Note that although the example code $\mathcal{C}$ of Section 3.3 is not a matroid complex (in particular, it violates the independent set exchange property [41]), $\mathrm{geom}(S)$ for the matrix S given in Solution 1

(Figure 2B) *is* a representable matroid complex, showing that $\mathcal{C}$ is the intersection of a representable matroid complex and the clique complex $X(G(\mathcal{C}))$.

5. Open questions. Can a combinatorial description be found for all simplicial complexes that are of the form $\text{geom}_\varepsilon(S)$ or $\text{geom}(S)$, where S and ε satisfy the conditions in Theorem 3? For such complexes, can the appropriate S and ε be obtained constructively? Does every simplicial complex $\mathcal{C}$ admit an exact solution to the NE problem via a *symmetric* network W? I.e., is every simplicial complex of the form $\text{geom}_\varepsilon(S) \cap X(G(\mathcal{C}))$, as in equation (3)? If not, what are the obstructions? More generally, does every simplicial complex admit an exact solution (not necessarily symmetric) to the NE problem? We have seen that all matroid complexes for representable matroids can be exactly encoded as $\text{geom}(S)$. Can non-representable matroids also be exactly encoded?

3.5 Spurious states and "natural" codes

Although it may be possible, as in the example of Section 3.3, to precisely tune the synaptic strength matrix S to exactly encode a particular neural code, this is somewhat contrary to the spirit of the Encoding Rule, which assumes S to be an intrinsic property of the underlying network. Fortunately, as seen in Section 3.4, Theorem 2 implies that certain "universal" choices of S enable any $\mathcal{C} \subset 2^{[n]}$ to be encoded. The price to pay, however, is the emergence of spurious states.

Recall that spurious states are permitted sets that arise in $\mathcal{P}(W)$ that were *not* in the prescribed list $\mathcal{C}$ of binary patterns to be encoded. Theorem 2 immediately implies that all spurious states lie in $X(G(\mathcal{C}))$ – i.e., every spurious state is a clique of the co-firing graph $G(\mathcal{C})$. We can divide them into two types:

- **Type 1: spurious subsets.** These are permitted sets $\sigma \in \mathcal{P}(W) \setminus \mathcal{C}$ that are subsets of patterns in $\mathcal{C}$. Note that if $\mathcal{C}$ is a simplicial complex, there will not be any spurious states of this type. On the other hand, if $\mathcal{C}$ is *not* a simplicial complex, then type 1 spurious states are guaranteed to be present for any symmetric encoding rule, because $\mathcal{P}(W) = \text{stab}(-D + W)$ is a simplicial complex for symmetric W (Lemma 2).

- **Type 2: spurious cliques.** These are permitted sets $\sigma \in \mathcal{P}(W) \setminus \mathcal{C}$ that are *not* of the first type. Note that, technically, the type 1 spurious states are also cliques in $G(\mathcal{C})$, but we will use the term "spurious clique" to refer only to spurious states that are not spurious subsets.

Perhaps surprisingly, some common neural codes have the property that the full set of patterns to be encoded naturally contains a large fraction of the cliques in the code's co-firing graph. In such cases, $\mathcal{C} \approx X(G(\mathcal{C}))$, or $\mathcal{C} \approx X_k(G(\mathcal{C}))$. These neural codes therefore have very few spurious states when encoded using a universal or k-sparse universal S, even though S has not been specially-tuned for the given code. We will refer to these as *natural* codes for symmetric threshold-linear networks, because they have two important properties that make them particularly fitting for these networks:

P1. Natural codes can be encoded exactly or nearly-exactly, using any universal or k-sparse universal matrix S in the Encoding Rule, and

P2. Natural codes can be fully encoded following presentation of only a small (randomly-sampled) fraction of the patterns in the code.

In other words, not only can natural codes be generically encoded with very few spurious states, but they can also be encoded from a highly undersampled set of codewords. This is because the network naturally "fills in" the missing elements via spurious states that emerge after encoding only part of the code. In the next two sections, we will explain why RF codes are "natural" in this sense, and illustrate the above two properties with a concrete application of encoding two-dimensional PF codes, an important example of RF codes.

3.6 Receptive field codes are natural codes

RF codes are binary neural codes consisting of activity patterns of populations of neurons that fire according to *receptive fields*.[11] Abstractly, a receptive field is a map $f_i : \mathcal{S} \to \mathbb{R}_{\geq 0}$ from a space of stimuli $\mathcal{S}$ to the average firing rate $f_i(s)$ of a single neuron i in response to each stimulus $s \in \mathcal{S}$. Receptive fields are computed from experimental data by correlating neural responses to external stimuli. We follow a common abuse of language, where both the map and its support (i.e., the subset $U_i \subset \mathcal{S}$ where f_i takes on strictly positive values) are referred to as "receptive fields." If the stimulus space is d-dimensional, i.e. $\mathcal{S} \subset \mathbb{R}^d$, we say that the receptive fields have *dimension* d. The paradigmatic examples of neurons with receptive fields are orientation-selective neurons in visual cortex [9] and hippocampal place cells [35]. Orientation-selective neurons have *tuning curves* that reflect a neuron's preference for a particular angle. Place cells are neurons that have *place fields* [38, 39]; i.e., each neuron has a preferred (convex) region of the animal's physical environment where it has a high firing rate. Both tuning curves and place fields are examples of low-dimensional receptive fields, having typical dimension $d = 1$ or $d = 2$.

The elements of a RF code $\mathcal{C}$ correspond to subsets of neurons that may be co-activated in response to a stimulus $s \in \mathbb{R}^d$ (see Figure 3). Here we define two variations of this notion, which we refer to as RF codes and *coarse* RF codes.

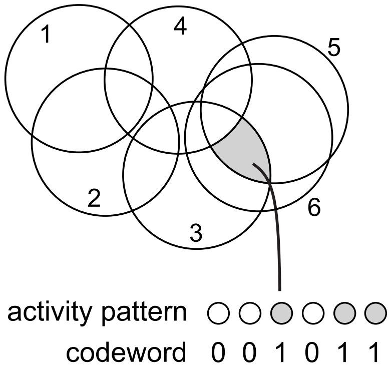

Figure 3: Two-dimensional receptive fields for 6 neurons. The RF code $\mathcal{C}$ has a codeword for each overlap region. For example, the shaded region corresponds to the binary pattern 001011; equivalently, we denote it as $\sigma = \{3, 5, 6\} \in \mathcal{C}$. The corresponding *coarse* RF code also includes all subsets, such as $\tau = \{3, 5\}$, even if they are not part of the original RF code.

Definition. Let $\{U_1, ..., U_n\}$ be a collection of convex open sets in $\mathbb{R}^d$, where each U_i is the receptive field corresponding to the ith neuron. To such a set of receptive fields, we associate a d-dimensional *RF code* $\mathcal{C}$, defined as follows: for each $\sigma \in 2^{[n]}$,

$$\sigma \in \mathcal{C} \quad \text{if and only if} \quad \bigcap_{i \in \sigma} U_i \setminus \bigcup_{j \notin \sigma} U_j \neq \emptyset.$$

This definition was previously introduced in [18, 19]. A *coarse* RF code is obtained from a RF code by including all subsets of codewords, so that for each $\sigma \in 2^{[n]}$,

$$\sigma \in \mathcal{C} \quad \text{if and only if} \quad \bigcap_{i \in \sigma} U_i \neq \emptyset.$$

Note that the codeword $\sigma = \{3, 5, 6\}$ in Figure 3 corresponds to stimuli in the shaded region, and *not* to the full intersection $U_3 \cap U_5 \cap U_6$. Moreover, the subset $\tau = \{3, 5\} \subset \sigma$ is not an element of the RF code, since $U_3 \cap U_5 \subset U_6$. Nevertheless, it often makes sense to also consider such subsets as codewords; for example, the cofiring of neurons 3 and 5 may still be observed, as neuron 6 may fail to fire even if the stimulus is in its receptive field. This is captured by the corresponding coarse RF code.

[11] In the vision literature, the term "receptive field" is reserved for subsets of the visual field; we use the term in a more general sense, applicable to any modality.

Coarse RF codes carry less detailed information about the underlying stimulus space [17, 19], but turn out to be more "natural" in the context of symmetric threshold-linear networks because they have the structure of a simplicial complex.[12] This implies that coarse RF codes do not yield any Type 1 spurious states – the spurious subsets – when encoded in a network using the Encoding Rule. Furthermore, both RF codes and coarse RF codes with *low-dimensional* receptive fields contain surprisingly few Type 2 spurious states – the spurious cliques. This follows from Helly's theorem, a classical theorem in convex geometry:

Helly's theorem [8]. Suppose that $U_1, ..., U_k$ is a finite collection of convex subsets of $\mathbb{R}^d$, for $d < k$. If the intersection of any $d + 1$ of these sets is nonempty, then the full intersection $\bigcap_{i=1}^k U_i$ is also nonempty.

To see the implications of Helly's theorem for RF codes, we define the notion of *Helly completion.*

Definition. Let Δ be a simplicial complex on n vertices, and let $\Delta_d = \{\sigma \in \Delta \mid |\sigma| \leq d+1\}$ denote its d-skeleton. The *Helly completion* is the largest simplicial complex, $\bar{\Delta}_d$, on n vertices that has Δ_d as its d-skeleton.

In other words, the Helly completion of a d-dimensional simplicial complex Δ_d is obtained by adding in all higher-dimensional faces in a way that is consistent with the existing lower-dimesional faces. In particular, the Helly completion of any graph G is the clique complex $X(G)$. For a two-dimensional simplicial complex, Δ_2, the Helly completion includes only cliques of the underlying graph $G(\Delta_2)$ that are consistent with Δ_2. For example, the Helly completion of the code in Section 3.3 does *not* include the 3-cliques corresponding to empty (non-shaded) triangles in Figure 3A. With this notion, Helly's theorem can now be reformulated as:

Lemma 6. *Let $\mathcal{C}$ be a coarse d-dimensional RF code, corresponding to a set of place fields $\{U_1, ..., U_n\}$ where each U_i is a convex open set in $\mathbb{R}^d$. Then $\mathcal{C}$ is the Helly completion of its own d-skeleton: $\mathcal{C} = \bar{\mathcal{C}}_d$.*

This lemma indicates that low-dimensional RF codes – whether coarse or not – have a relatively small number of spurious cliques, since most cliques in $X(G(\mathcal{C}))$ are also in the Helly completion $\bar{\mathcal{C}}_d$ for small d. In particular, it implies that coarse RF codes of dimensions $d = 1$ and $d = 2$ are very natural codes for symmetric threshold-linear networks.

Corollary 2. *If $\mathcal{C}$ is a coarse one-dimensional RF code, then it is a clique complex: $\mathcal{C} = \bar{\mathcal{C}}_1 = X(G(\mathcal{C}))$. Therefore, $\mathcal{C}$ can be exactly encoded using any universal S in the Encoding Rule.*

Corollary 3. *If $\mathcal{C}$ is a coarse two-dimensional RF code, then it is the Helly completion of its own 2-skeleton, $\mathcal{C} = \bar{\mathcal{C}}_2$, which can be obtained from knowledge of all pairwise and triple intersections of receptive fields.*

For coarse two-dimensional RF codes, the only possible spurious cliques are therefore spurious triples and the larger cliques of $G(\mathcal{C})$ that contain them. The spurious triples emerge when three receptive fields U_i, U_j and U_k have the property that each pair intersect, but $U_i \cap U_j \cap U_k = \emptyset$. For generic arrangements of receptive fields, this is relatively rare, allowing these codes to be encoded nearly exactly using any universal S in the Encoding Rule. In the next section, we illustrate this phenomenon in the case of two-dimensional place field codes.

[12] In topology, this simplicial complex is called the *nerve* of the cover $\{U_1, \ldots, U_n\}$ (see [13, 17]).

3.7 Encoding sparse place field codes in threshold-linear networks

As seen in the last section, Helly's theorem sharply limits the number of spurious cliques that result from encoding low-dimensional RF codes. Here we illustrate this phenomenon explicitly in the case of sparse place field codes (PF codes). In particular, we find that PF codes can be encoded nearly exactly from a very small, randomly selected sample of patterns. The near-exact encoding of PF codes from highly undersampled data shows that they are "natural" codes for symmetric threshold-linear networks, as defined in Section 3.5.

PF codes. Let $\{U_1, ..., U_n\}$ be a collection of convex open sets in $\mathbb{R}^d$, where each U_i is the *place field* corresponding to the ith neuron [38, 39]. To such a set of place fields we associate a *d-dimensional PF code*, $\mathcal{C}$, defined as follows: for each $\sigma \in 2^{[n]}$, $\sigma \in \mathcal{C}$ if and only if the intersection $\bigcap_{i\in\sigma} U_i$ is nonempty.

Note that in this definition, PF codes are coarse RF codes. PF codes are experimentally observed in recordings of neural activity in rodent hippocampus [35]. The elements of $\mathcal{C}$ correspond to subsets of neurons that may be co-activated as the animal's trajectory passes through a corresponding set of overlapping place fields. Typically $d = 1$ or $d = 2$, corresponding to the standard "linear track" and "open field" environments [36]; it has also been hypothesized that some animals possess $d = 3$ place fields [46].

It is clear from Corollary 2 above that one-dimensional PF codes can be encoded exactly – i.e., without any spurious states – using any universal S matrix in the Encoding Rule. Two-dimensional PF codes have no Type 1 spurious states, but may have Type 2 spurious cliques. For *sparse* PF codes, however, the spurious cliques can be further restricted (beyond what is expected from Helly's theorem) by choosing a k-sparse universal S.

Near-exact encoding of sparse PF codes. Consider a two-dimensional PF code $\mathcal{C}$ that is *k-sparse*, so that no more than k neurons can co-fire in a single pattern – even if there are higher-order overlaps of place fields. Experimental evidence suggests that the fraction of active neurons is typically on the order of $5 - 10\%$ [6], so we make the conservative choice of $k = n/10$ (our results improve with smaller k). In what follows, S was chosen to be k-sparse universal and ε so that $0 < \varepsilon < \delta(S)$, in order to control spurious cliques of size greater than k. We also assume the worst-case-scenario of $\mathcal{P}(W) = X_{k-1}(G(\mathcal{C}))$, providing an upper bound on the number of spurious cliques resulting from our Encoding Rule. What fraction of the stored patterns are spurious? This can be quantified by the following *error probability*,

$$P_{\text{error}} \overset{\text{def}}{=} \frac{|\mathcal{P}(W) \setminus \mathcal{C}|}{|\mathcal{P}(W)|} = \frac{|X_{k-1}(G(\mathcal{C}))| - |\mathcal{C}|}{|X_{k-1}(G(\mathcal{C}))|},$$

which assumes all permitted sets are equally likely to be retrieved from among the stored patterns in $\mathcal{P}(W)$. For exact encoding, $P_{\text{error}} = 0$, while large numbers of spurious states will push P_{error} close to 1.

To investigate how "exactly" two-dimensional PF codes are encoded, we generated random k-sparse PF codes with circular place fields, n = 80-100 neurons, and $k = n/10$ (see Appendix D). Because experimentally observed place fields do not have precise boundaries, we also generated "jittered" codes, where spurious triples were eliminated from the 2-skeleton of the code if they did not survive after enlarging the place field radii from r_0 to r_1 by a *jitter ratio*, $(r_1 - r_0)/r_0$. This has the effect of eliminating spurious cliques that are unlikely to be observed in neural activity, as they correspond to very small regions in the underlying environment. For each code and each jitter ratio (up to ~ 0.1), we computed P_{error} using the formula above. Even without jitter, the error probability was small, and P_{error} decreased quickly to values near zero for 10% jitter (Figure 4A).

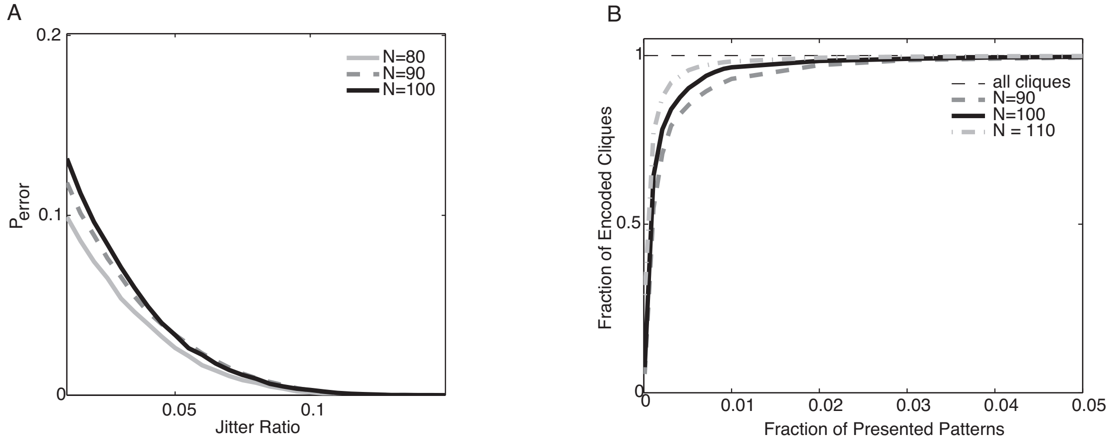

Figure 4: PF encoding is near-exact, and can be achieved by presenting a small fraction of patterns. (A) P_{error} was computed for randomly generated k-sparse PF codes having $n = 80, 90$ and 100 neurons and $k = n/10$. For each jitter ratio, the average value of P_{error} over 100 codes is shown. (B) For $n = 90, 100$ and 110 neurons, k-sparse PF codes with jitter ratio 0.1 were randomly generated and then randomly subsampled to contain a small fraction ($\leq 5\%$) of the total number of patterns. After applying the Encoding Rule to the subsampled code, the number of encoded cliques was computed. In each case, the fraction of encoded cliques for the subsampled code (as compared to the full PF code) was averaged over 10 codes. Cliques were counted using Cliquer [37], together with custom-made Matlab software.

Encoding full PF codes from highly undersampled sets of patterns. To investigate what fraction of patterns is needed to encode a two-dimensional PF code using the Encoding Rule, we generated randomly subsampled codes from k-sparse PF codes. We then computed the number of patterns that would be encoded by a network if a subsampled code was presented. Perhaps surprisingly, network codes obtained from highly subsampled PF codes (having only 1-5% of the patterns) are nearly identical to those obtained from full PF codes (Figure 4B). This is because large numbers of "spurious" states emerge when encoding subsampled codes, but most correspond to patterns in the full code. The spurious states of subsampled PF codes can therefore be *advantageous*, allowing networks to quickly encode full PF codes from only a small fraction of the patterns.

The results summarized in Figure 4 confirm the fact that sparse PF codes are natural codes, as they satisfy both properties P1 and P2 outlined in Section 3.5. These codes can be encoded nearly exactly because they have very few spurious states. The spurious cliques are limited by two factors: the implications of Helly's theorem (Section 3.6) and their sparsity, enabling the choice of a k-sparse universal S that automatically eliminates spurious cliques of size greater than k.

4 Proofs

To the best of our knowledge, all proofs in this section are original, as are the results presented in Theorems 2, 3, and 4. Theorem 4 is our core technical result, which we state and prove in Section 4.1. It appears to be closely related to some relatively recent results in convex geometry, involving correlation matrices and the geometry of the "elliptope" [21]. Our proof, however, relies only on classical distance geometry and well-known facts about stable symmetric matrices; these are summarized in Appendix A. The key new insight that allows us to connect stability of matrices of the form $-11^T + \varepsilon A$ to Cayley-Menger determinants is Lemma 7. In Section 4.2 we give the proofs of Proposition 1, Theorem 2 and Theorem 3, which all rely on Theorem 4.

4.1 Statement of Theorem 4 and its proof

The statement of Theorem 4 uses the following definition and some new notation.

Definition. A *Hebbian* matrix A is an $n \times n$ matrix satisfying $A_{ij} = A_{ji} \geq 0$ and $A_{ii} = 0$ for all $i, j \in [n]$.

The name reflects the fact that these are precisely the types of matrices that arise when synaptic weights are modified by a Hebbian learning rule. We also need the notation,

$$\mathbb{R}^n_\times \overset{\text{def}}{=} \{v \in \mathbb{R}^n \mid v_i \neq 0 \text{ for all } i \in [n]\}$$

for the set of vectors with all nonzero entries. Note that for $v \in \mathbb{R}^n_\times$, $-vv^T$ is a symmetric $n \times n$ rank 1 matrix with strictly negative diagonal. Next, given any $v \in \mathbb{R}^n$ and any $n \times n$ matrix A,

$$A^v \overset{\text{def}}{=} \operatorname{diag}(v) A \operatorname{diag}(v)$$

denotes the matrix with entries $A^v_{ij} = v_i v_j A_{ij}$. We are now ready to state Theorem 4.

Theorem 4. *Let A be a Hebbian matrix and $\varepsilon > 0$. For $v \in \mathbb{R}^n_\times$, consider the perturbed matrix*

$$M = -vv^T + \varepsilon A^v.$$

The following are equivalent:

1. *A is a nondegenerate square distance matrix.*
2. *There exists an $\varepsilon > 0$ such that M is stable.*
3. *There exists a $\delta > 0$ such that M is stable for all $0 < \varepsilon < \delta$.*
4. $0 < -\dfrac{\operatorname{cm}(A)}{\det A} < \infty$*; and M is stable if and only if* $0 < \varepsilon < -\dfrac{\operatorname{cm}(A)}{\det A}$.

The rest of this section is devoted to proving Theorem 4. A cornerstone of the proof is the following lemma, which allows us to connect perturbations of rank 1 matrices to Cayley-Menger determinants.

Lemma 7 (determinant lemma)**.** *Let $u, v \in \mathbb{R}^n$. For any real-valued $n \times n$ matrix A and any $t \in \mathbb{R}$,*

$$\det(-uv^T + t \operatorname{diag}(u) A \operatorname{diag}(v)) = \det(\operatorname{diag}(u)\operatorname{diag}(v)) \left(t^n \det A + t^{n-1}\operatorname{cm}(A)\right).$$

In particular, if $u = v \in \mathbb{R}^n_\times$ and $t > 0$, then

$$\operatorname{sgn}(\det(-vv^T + tA^v)) = \operatorname{sgn}\left(t \det A + \operatorname{cm}(A)\right),$$

where $\operatorname{sgn} : \mathbb{R} \to \{\pm 1, 0\}$ *is the sign function. Moreover, taking $u = v = 1 \in \mathbb{R}^n$ yields:*

$$\det(-11^T + tA) = t^n \det A + t^{n-1}\operatorname{cm}(A).$$

Proof of Lemma 7. Note that for any $n \times n$ matrix A, $t \in \mathbb{R}$, and $u, v \in \mathbb{R}^n$, we have

$$\det(-uv^T + t\,\mathrm{diag}(u) A\,\mathrm{diag}(v)) = \det(\mathrm{diag}(u)\,\mathrm{diag}(v)) \det(-11^T + tA),$$

where -11^T is as usual the rank 1 matrix of all -1s. It thus suffices to show that

$$\det(-11^T + tA) = t^n \det A + t^{n-1}\mathrm{cm}(A),$$

where $\mathrm{cm}(A)$ is the Cayley-Menger determinant of A.

Let $w, z \in \mathbb{R}^n$, and let Q be any $n \times n$ matrix. We have

$$\det \begin{bmatrix} 1 & z^T \\ w & Q \end{bmatrix} = \det(Q - wz^T),$$

where we have used the well-known formula for computing the determinant of a 2×2 block matrix.[13] On the other hand, the usual cofactor expansion along the first row gives

$$\det \begin{bmatrix} 1 & z^T \\ w & Q \end{bmatrix} = \det(Q) + \det \begin{bmatrix} 0 & z^T \\ w & Q \end{bmatrix}.$$

Therefore,

$$\det(-wz^T + Q) = \det(Q) + \det \begin{bmatrix} 0 & z^T \\ w & Q \end{bmatrix}.$$

In particular, taking $w = z = 1 \in \mathbb{R}^n$ (the column vector of all ones) and $Q = tA$, we have $\det(-11^T + tA) = \det(tA) + \mathrm{cm}(tA) = t^n \det A + t^{n-1}\mathrm{cm}(A)$. □

Finally, to prove Theorem 4 we will need the following technical lemma.

Lemma 8. *Fix $v \in \mathbb{R}^n_\times$, and let A be an $n \times n$ Hebbian matrix. If $(-1)^n \mathrm{cm}(A) \le 0$, then $-vv^T + tA^v$ is not stable for any $t > 0$. In particular, if there exists a $t > 0$ such that $-vv^T + tA^v$ is stable, then $(-1)^n\mathrm{cm}(A) > 0$.*

For its proof, we will need a simple convexity result.

Lemma 9. *Let M, N be real symmetric $n \times n$ matrices so that M is negative semidefinite (i.e., all eigenvalues are ≤ 0) and N is strictly negative definite (i.e., stable, with all eigenvalues < 0). Then $tM + (1-t)N$ is strictly negative definite (i.e., stable) for all $0 \le t < 1$.*

Proof. M and N satisfy $x^T M x \le 0$ and $x^T N x < 0$ for all $x \in \mathbb{R}^n$, so we have $x^T(tM + (1-t)N)x < 0$ for all nonzero $x \in \mathbb{R}^n$ if $0 \le t < 1$. □

The proof of Lemma 8 relies on Lemmas 7 and 9, which we have just proven, and also on some well-known results from classical distance geometry that are collected in Appendix A. These include facts about stable symmetric matrices (Cauchy's interlacing theorem, Corollary 6, and Lemma 10) as well as facts about square distance matrices (Lemma 12, Proposition 2, and Corollary 8). These facts are also used in the proof of Theorem 4.

[13] The formula $\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A)\det(D - CA^{-1}B)$ applies so long as A is invertible. It follows from observing that $\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & -CA^{-1}B + D \end{bmatrix}$.

Proof of Lemma 8. Since A is symmetric, so are A^v and $-vv^T + tA^v$ for any t. Hence, if any principal submatrix of $-vv^T + tA^v$ is unstable then $-vv^T + tA^v$ is also unstable, by Corollary 6. Therefore, without loss of generality we can assume $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) > 0$ for all proper principal submatrices A_σ, with $|\sigma| < n$ (otherwise, we use this argument on a smallest principal submatrix such that $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) \leq 0$). By Lemma 12, this implies that A_σ is a nondegenerate square distance matrix for all σ such that $|\sigma| < n$, and so we know by Proposition 2 that $(-1)^{|\sigma|}\det A_\sigma < 0$ and that each A_σ such that $1 < |\sigma| < n$ has one positive eigenvalue and all other eigenvalues negative.

We prove the lemma by contradiction. Suppose there exists a $t_0 > 0$ such that $-vv^T + t_0A^v$ is stable. Applying Lemma 9 with $M = -vv^T$ and $N = -vv^T + t_0A^v$, we have that $-vv^T + (1-t)t_0A^v$ is stable for all $0 \leq t < 1$. It follows that $-vv^T + tA^v$ is stable for all $0 < t \leq t_0$. Now Lemma 10 implies that $(-1)^n \det(-vv^T + tA^v) > 0$ for all $0 < t \leq t_0$. By Lemma 7, this is equivalent to having $(-1)^n(t\det A + \mathrm{cm}(A)) > 0$ for all $0 < t \leq t_0$. By assumption, $(-1)^n\mathrm{cm}(A) \leq 0$. But, if $(-1)^n\mathrm{cm}(A) < 0$, then there would exist a small enough $t > 0$ such that $(-1)^n(t\det A + \mathrm{cm}(A)) < 0$. Therefore we conclude that $\mathrm{cm}(A) = 0$ and hence $(-1)^n \det A > 0$.

Next, let $\lambda_1 \leq ... \leq \lambda_n \leq \lambda_{n+1}$ denote the eigenvalues of the Cayley-Menger matrix $CM(A) = \begin{bmatrix} 0 & 1^T \\ 1 & A \end{bmatrix}$, and observe that A, $A_{[n-1]}$, and $CM(A_{[n-1]})$ are all principal submatrices of $CM(A)$. Since everything is symmetric, Cauchy's interlacing theorem applies. We have seen above that $A_{[n-1]}$ has one positive eigenvalue and all others negative, so by Cauchy interlacing $\lambda_{n+1} > 0$ and $\lambda_{n-2} < 0$. Because $\mathrm{cm}(A) = \det CM(A) = 0$, then $CM(A)$ must have a zero eigenvalue, while $\det A \neq 0$ implies that it is unique. We thus have two cases.

Case 1: Suppose $\lambda_{n-1} = 0$ and thus $\lambda_n > 0$. Since we assume $(-1)^{n-1}\mathrm{cm}(A_{[n-1]}) > 0$, the $n \times n$ matrix $CM(A_{[n-1]})$ must have an odd number of positive eigenvalues, but by Cauchy interlacing the top two eigenvalues must be positive, so we have a contradiction.

Case 2: Suppose $\lambda_n = 0$ and thus $\lambda_{n-1} < 0$. Then by Cauchy interlacing A has exactly one positive eigenvalue. On the other hand, the fact that $(-1)^n \det A > 0$ implies that A has an even number of positive eigenvalues, which is a contradiction. □

We can now prove Theorem 4.

Proof of Theorem 4. We prove (4) ⇒ (3) ⇒ (2) ⇒ (1) ⇒ (4).

(4) ⇒ (3) ⇒ (2) is obvious.

(2) ⇒ (1): Suppose there exists a $t > 0$ such that $-vv^T + tA^v$ is stable. Then, by Corollary 6 and Lemma 8, $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) > 0$ for all principal submatrices A_σ. By Lemma 12 it follows that A is a nondegenerate square distance matrix.

(1) ⇒ (4): Suppose A is a nondegenerate square distance matrix. By Lemma 12 we have $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) > 0$ for all A_σ, while Proposition 2 implies $(-1)^{|\sigma|}\det(A_\sigma) < 0$ for all A_σ with $|\sigma| > 1$. This implies that for $|\sigma| > 1$ we have $-\dfrac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)} > 0$ (by Corollary 8), and that if $\varepsilon > 0$,

$$(-1)^{|\sigma|}\left(\varepsilon \det(A_\sigma) + \mathrm{cm}(A_\sigma)\right) > 0 \quad \Leftrightarrow \quad \varepsilon < -\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)}.$$

Applying now Lemma 7,

$$(-1)^{|\sigma|}\det(-vv^T + \varepsilon A^v)_\sigma > 0 \quad \Leftrightarrow \quad \varepsilon < -\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)}.$$

For $|\sigma| = 1$, we have diagonal entries $A_\sigma = A^v_\sigma = 0$ and $(-vv^T)_\sigma < 0$, so $(-1)\det(-vv^T + \varepsilon A^v)_\sigma > 0$ for all ε. Using Lemma 10, we conclude (assuming $\varepsilon > 0$):

$$-vv^T + \varepsilon A^v \text{ is stable} \quad \Leftrightarrow \quad \varepsilon < \delta,$$

where

$$\delta = \min \left\{ -\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)} \right\}_{\sigma \subseteq [n]} > 0.$$

It remains only to show that $\delta = -\mathrm{cm}(A)/\det(A)$. Note that we can not use Lemma 5 from the Main Text because that lemma follows from Proposition 1, and is hence a consequence of Theorem 4.

On the other hand, because the matrix $-vv^T + \varepsilon A^v$ changes from stable to unstable at $\varepsilon = \delta$, by continuity of the eigenvalues as functions of ε it must be that

$$\det(-vv^T + \delta A^v) = 0.$$

Using Lemma 7 it follows that $\delta \det(A) + \mathrm{cm}(A) = 0$, which implies $\delta = -\mathrm{cm}(A)/\det(A)$. □

4.2 Proofs of Proposition 1, Theorem 2 and Theorem 3

Here we prove our main results from Sections 3.2 and 3.4. We begin with the proof of Proposition 1.

Proof of Proposition 1. Setting $v = 1 \in \mathbb{R}^n_\times$ (the column vector of all ones) in Theorem 4 yields a slightly weaker version of Proposition 1, as the hypothesis in Theorem 4 is that A is *Hebbian*, which is more constrained than the Proposition 1 hypothesis that A is symmetric with zero diagonal. To see why Proposition 1 holds more generally, suppose A is symmetric with zero diagonal but *not* Hebbian. Then there exists an off-diagonal pair of negative entries, $A_{ij} = A_{ji} < 0$, and the 2×2 principal submatrix

$$(-11^T + \varepsilon A)_{\{ij\}} = \begin{pmatrix} -1 & -1 + \varepsilon A_{ij} \\ -1 + \varepsilon A_{ij} & -1 \end{pmatrix}$$

is unstable as it has negative trace and negative determinant. It follows from Cauchy's interlacing theorem (see Corollary 6 in Appendix A) that $-11^T + \varepsilon A$ is unstable for any $\varepsilon > 0$. Correspondingly, condition (a) in Proposition 1 is violated, as the existence of negative entries guarantees that A cannot be a nondegenerate square distance matrix. □

To prove Theorems 2 and 3, we will need the following two corollaries of Proposition 1. First, recall the definitions for $\mathrm{geom}(A), \mathrm{geom}_\varepsilon(A)$, $\delta(A)$ from Section 3.2. Applying Proposition 1 to each of the principal submatrices of the perturbed matrix $-11^T + \varepsilon A$ we obtain:

Corollary 4. *If A is a symmetric matrix with zero diagonal, and $\varepsilon > 0$, then*

$$\mathrm{stab}(-11^T + \varepsilon A) = \mathrm{geom}_\varepsilon(A).$$

For $0 < \varepsilon < \delta(A)$, $\mathrm{stab}(-11^T + \varepsilon A) = \mathrm{geom}(A)$.

Next, recall that $X(G)$ is the clique complex of the graph G.

Corollary 5. *Let A be a symmetric $n \times n$ matrix with zero diagonal, and $\varepsilon > 0$. Let G be the graph on n vertices having $(ij) \in G$ if and only if $A_{ij} \geq 0$. For any $n \times n$ matrix S with $S_{ij} = S_{ji} \geq 0$ and $S_{ii} = 0$, if S "matches" A on G (i.e., if $S_{ij} = A_{ij}$ for all $(ij) \in G$), then*

$$\mathrm{geom}_\varepsilon(A) = \mathrm{geom}_\varepsilon(S) \cap X(G).$$

In particular, $\mathrm{geom}(A) = \mathrm{geom}(S) \cap X(G)$.

We can now prove Theorems 2 and 3.

Proof of Theorem 2. Any network W obtained via the Encoding Rule (equation (2)) has the form $-D+W=-11^T+\varepsilon A$, where A is symmetric with zero diagonal and "matches" the (nonnegative) synaptic strength matrix S precisely on the entries A_{ij} such that $(ij)\in G(\mathcal{C})$. All other off-diagonal entries of A are negative. It follows that

$$\begin{aligned}\mathcal{P}(W) &= \text{stab}(-11^T+\varepsilon A)=\text{geom}_\varepsilon(A)\\ &= \text{geom}_\varepsilon(S)\cap X(G(\mathcal{C})),\end{aligned}$$

where the last two equalities are due to Corollaries 4 and 5, respectively. □

Proof of Theorem 3. ($\Leftarrow$) This is an immediate consequence of Theorem 2.

($\Rightarrow$) Suppose there exists a symmetric network W with $\mathcal{P}(W)=\mathcal{C}$, and observe by Theorem 1 that $\mathcal{P}(W)=\text{stab}(-11^T+A)$, for some symmetric $n\times n$ matrix A with zero diagonal. By Corollaries 4 and 5,

$$\mathcal{C}=\mathcal{P}(W)=\text{geom}_\varepsilon(A)=\text{geom}_\varepsilon(S)\cap X(G),$$

where $\varepsilon=1$, G is the graph associated to A (as in Corollary 5) and S is an $n\times n$ matrix with $S_{ij}=S_{ji}\geq 0$ and zero diagonal that "matches" A on G. It remains only to show that $\text{geom}_\varepsilon(S)\cap X(G)=\text{geom}_\varepsilon(S)\cap X(G(\mathcal{C}))$. Since $\mathcal{C}=\text{geom}_\varepsilon(A)$, any element $\{ij\}\in\mathcal{C}$ must have corresponding $A_{ij}>0$, so $G(\mathcal{C})\subseteq G$ and hence $X(G(\mathcal{C}))\subseteq X(G)$. On the other hand, $\mathcal{C}=\mathcal{C}\cap X(G(\mathcal{C}))$, so we conclude that $\mathcal{C}=\text{geom}_\varepsilon(S)\cap X(G(\mathcal{C}))$. □

5 Discussion

Understanding the relationship between the connectivity matrix and the activity patterns of a neural network is one of the central challenges in theoretical neuroscience. We have found that in the context of symmetric threshold-linear networks, one can obtain an unexpectedly precise understanding of the binary activity patterns stored by network steady states. In particular, we have arrived at a complete and precise combinatorial characterization of spurious states, something that has not yet been achieved in the context of the Hopfield model [3, 4, 2, 24, 43]. Moreover, we have shown that network solutions to the NE problem can be obtained constructively, using a simple Encoding Rule. A new concept that emerges from our results is that of *geometric balance*, whereby the excitatory synapses between neurons in a stored pattern must satisfy a set of geometric constraints, ensuring they are appropriately bounded and balanced in their strengths.

As a consequence of our main results, we have discovered that threshold-linear networks naturally encode neural codes arising from low-dimensional receptive fields (such as place fields) while introducing very few spurious states. Remarkably, these codes can be "learned" by the network from a highly undersampled set of patterns. Neural codes representing (continuous) parametric stimuli, such as place field codes, have typically been modeled as arising from continuous attractor networks whose synaptic matrices have symmetric "Mexican hat"-type connectivity [9, 35]. This is in large part due to the fact that there is a well-developed mathematical handle on these networks [1, 14, 31]. Our work shows that one can have fine mathematical control over a much wider class of networks, encompassing all symmetric connectivity matrices. It may thus provide a novel foundation for understanding – and engineering – neural networks with prescribed steady state properties.

Acknowledgments

The authors would like to thank Christopher Hillar, Caroline Klivans, Katie Morrison and Bernd Sturmfels for valuable comments, as well as Nora Youngs and Zachary Roth for assistance with figures. CC was supported by NSF DMS 0920845, NSF DMS 1225666, a Woodrow Wilson Fellowship, and the Sloan Research Fellowship. AD was supported by the Max Planck Society and the DFG via SFB/Transregio 71 "Geometric Partial Differential Equations." VI was supported by NSF DMS 0967377 and NSF DMS 1122519.

6 Appendices

6.1 Appendix A: Stable symmetric matrices and square-distance matrices

In this appendix we review some classical facts about stable symmetric matrices and square-distance matrices that are critical to many of our proofs. Everything in this section is well-known.

6.1.1 Stable symmetric matrices

Here we summarize some well-known facts about the stability of symmetric matrices that we use in various proofs. The first is Cauchy's interlacing theorem, which relates eigenvalues of a symmetric matrix to those of its principal submatrices. Recall that the eigenvalues of a symmetric matrix are always real.

Theorem 5 (Cauchy's interlacing theorem [28])**.** *Let A be a symmetric $n \times n$ matrix, and let B be an $m \times m$ principal submatrix of A. If the eigenvalues of A are $\alpha_1 \leq ...\alpha_j... \leq \alpha_n$ and those of B are $\beta_1 \leq ...\beta_j... \leq \beta_m$, then $\alpha_j \leq \beta_j \leq \alpha_{n-m+j}$ for all j.*

Some immediate consequences of this theorem are:

Corollary 6. *Any principal submatrix of a stable symmetric matrix is stable. Any symmetric matrix containing an unstable principal submatrix is unstable.*

Corollary 7. *Let A be a symmetric $n \times n$ matrix with strictly negative diagonal. Then* $\text{stab}(A)$ *is a simplicial complex.*

Proof. First, recall the definitions of $\text{stab}(A)$ and simplicial complex from Section 2.3. We need to check the two properties in the definition of a simplicial complex. Property (1) holds for $\text{stab}(A)$, because A has strictly negative diagonal. Property (2) follows from Corollary 6. □

Another well-known consequence of Cauchy's interlacing theorem is the following lemma. Here $A_{[k]}$ refers to the principal submatrix obtained by taking the upper left $k \times k$ entries of A.

Lemma 10. *Let A be a real symmetric $n \times n$ matrix. Then the following are equivalent:*

1. *A is a stable matrix.*
2. $(-1)^k \det(A_{[k]}) > 0$ *for all* $1 \leq k \leq n$.
3. $(-1)^{|\sigma|} \det(A_\sigma) > 0$ *for every* $\sigma \subseteq [n]$.

Proof. We prove (1) ⇔ (2). The equivalence between (1) and (3) follows using a very similar argument.

(⇒) Assume A is stable. Then $\lambda_1(A) \leq ... \leq \lambda_n(A) < 0$. By Cauchy's interlacing theorem, $\lambda_1(A) \leq \lambda_i(A_{[k]}) \leq \lambda_n(A)$ for all $i = 1, ..., k$ and $k = 1, ..., n$. Therefore, all eigenvalues of the matrices $A_{[k]}$ are strictly negative, and hence $(-1)^k \det(A_{[k]}) > 0$ for all k.

(⇐) We prove this by induction. The base case is $n = 1$: indeed, a 1×1 matrix $A = [a]$ is stable if $-\det(A_{[1]}) > 0$, i.e. if $a < 0$. Now suppose (⇐) of the theorem is true for $(n-1) \times (n-1)$

matrices, and also that $(-1)^k \det(A_{[k]}) > 0$, $k = 1, ..., n$, for an $n \times n$ matrix A. This implies $A_{[n-1]}$ is stable. By Cauchy's interlacing theorem, the highest eigenvalue $\lambda_{n-1}(A_{[n-1]})$ lies between the top two eigenvalues of A:

$$\lambda_{n-1}(A) \leq \lambda_{n-1}(A_{[n-1]}) \leq \lambda_n(A).$$

The stability of $A_{[n-1]}$ thus implies $\lambda_1(A) \leq ... \leq \lambda_{n-1}(A) < 0$. It remains only to check that $\lambda_n(A) < 0$. For n even, $(-1)^n \det(A_{[n]}) > 0$ implies $\lambda_1(A)\lambda_2(A) \cdots \lambda_n(A) > 0$, hence $\lambda_n(A) < 0$. For n odd, $(-1)^n \det(A_{[n]}) > 0$ implies $\lambda_1(A)\lambda_2(A) \cdots \lambda_n(A) < 0$, hence $\lambda_n(A) < 0$. It follows that A is stable. □

6.1.2 Square distance matrices

Recall from Section 3.2 the definitions of square distance matrix, nondegenerate square distance matrix, and Cayley-Menger determinant. Our convention is that the 1×1 zero matrix $[0]$ is a nondegenerate square distance matrix, as $|\mathrm{cm}([0])| = 1 > 0$. As an example, a 3×3 symmetric matrix A with zero diagonal is a nondegenerate square distance matrix if and only if the off-diagonal entries A_{ij} are all positive, and their square roots ($\sqrt{A_{12}}$, $\sqrt{A_{13}}$, and $\sqrt{A_{23}}$) satisfy all three triangle inequalities.

There are two classical characterizations of square distance matrices. The first, due to Menger [12], relies on Cayley-Menger determinants. The second, due to Schoenberg [44], uses eigenvalues of principal submatrices. Both are needed for our proof of Theorem 4.

The relationship between Cayley-Menger determinants and simplex volumes is well-known:

Lemma 11. *Let $p_1, .., p_k$ be k points in a Euclidean space. Assume that $A_{ij} = \|p_i - p_j\|^2$ is the matrix of square distances between these points. Then the $(k-1)$-dim volume V of the convex hull of the points $\{p_i\}_{i=1}^k$ can be computed as*

$$V^2 = \frac{(-1)^k}{2^{(k-1)}\left((k-1)!\right)^2}\mathrm{cm}(A). \tag{4}$$

In particular, if A is a degenerate square distance matrix then $\mathrm{cm}(A) = 0$.

This leads to Menger's characterization of square distance matrices. Recall that A_σ is the principal submatrix obtained by restricting A to the index set σ.

Lemma 12. *Let A be an $n \times n$ matrix satisfying $A_{ij} = A_{ji} \geq 0$ and $A_{ii} = 0$ for all $i, j \in [n]$ (i.e., A is a Hebbian matrix). Then,*

1. *A is a square distance matrix if and only if $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) \geq 0$ for every A_σ.*
2. *A is a nondegenerate square distance matrix if and only if $(-1)^{|\sigma|}\mathrm{cm}(A_\sigma) > 0$ for every A_σ.*

Proof. (1) is equivalent to the Corollary of Theorem 42.2 in [12]. (2) is equivalent to Theorem 41.1 in [12]. □

Schoenberg's characterization implies that if a matrix is a square distance matrix, then the determinant of any principal submatrix has opposite sign to that of its Cayley-Menger determinant.

Proposition 2. *Let A be an $n \times n$ square distance matrix that is not the zero matrix. Then:*

1. *A has one strictly positive eigenvalue and $n-1$ eigenvalues that are less than or equal to zero. In particular, $(-1)^{|\sigma|}\det(A_\sigma) \leq 0$ for every principal submatrix A_σ.*

2. *If A is a nondegenerate square distance matrix, A has no zero eigenvalues and $(-1)^{|\sigma|}\det(A_\sigma) < 0$ for every principal submatrix A_σ with $|\sigma| > 1$.*

Proof. This Proposition is contained in [21, Theorem 6.2.16]. It can also be proven directly from Theorem 1 of Schoenberg's 1935 paper [44]. □

Corollary 8. *If A is an $n \times n$ nondegenerate square distance matrix with $n > 1$, then*

$$-\frac{\mathrm{cm}(A)}{\det A} > 0.$$

6.2 Appendix B: Some more facts about permitted sets of symmetric threshold-linear networks

This section proves a few additional (and novel) facts about permitted sets in symmetric threshold-linear networks. These were not included in the main text in order not to disrupt the flow of the exposition.

Recall from Theorem 1 that the permitted sets $\mathcal{P}(W)$, where W is a threshold-linear network with dynamics given by equation (1), always have the form

$$\mathcal{P}(W) = \mathrm{stab}(-D + W).$$

Here we show that when W is *symmetric* (like the networks obtained using the Encoding Rule (2)), $\mathcal{P}(W)$ can always be expressed as $\mathrm{stab}(-11^T + A)$ or $\mathrm{stab}(-xy^T + B)$, where $-xy^T$ is any rank 1 matrix having strictly negative diagonal, and A, B are square matrices with zero diagonal. In what follows, we use the notation $\mathbb{R}^n_\times$ and A^v defined at the beginning of Section 4.1.

Lemma 13. *Let M be a symmetric $n \times n$ matrix, and $v \in \mathbb{R}^n_\times$. Then,*

$$\mathrm{stab}(M^v) = \mathrm{stab}(M).$$

In other words, a principal submatrix M^v_σ is stable if and only if M_σ is stable.

Proof. By Lemma 10, $\tau \in \mathrm{stab}(M)$ if and only if $(-1)^{|\sigma|}\det(M_\sigma) > 0$ for every $\sigma \subseteq \tau$. Observe that, since $M^v = \mathrm{diag}(v) M \,\mathrm{diag}(v)$, we have $\mathrm{sgn}(\det(M^v_\sigma)) = \mathrm{sgn}(\det(M_\sigma))$ for all $\sigma \subseteq [n]$. It follows that $\tau \in \mathrm{stab}(M^v)$ if and only if $\tau \in \mathrm{stab}(M)$. □

Lemma 14. *For any symmetric threshold-linear network W on n neurons, there exists a symmetric $n \times n$ matrix A with zero diagonal such that*

$$\mathcal{P}(W) = \mathrm{stab}(-11^T + A).$$

Proof. Let $x \in \mathbb{R}^n_\times$ be the vector such that $\mathrm{diag}(-xx^T) = \mathrm{diag}(-D + W)$, and write

$$-D + W = -xx^T + (-D + W + xx^T),$$

where the term in parentheses is symmetric and has zero diagonal. This can be rewritten as

$$-D + W = \mathrm{diag}(x)(-11^T + A)\,\mathrm{diag}(x) = (-11^T + A)^x,$$

where

$$A = \mathrm{diag}(x)^{-1}(-D + W + xx^T)\,\mathrm{diag}(x)^{-1}$$

is a symmetric $n \times n$ matrix with zero diagonal. It follows from Lemma 13 that $\mathcal{P}(W) = \mathrm{stab}(-D + W) = \mathrm{stab}(-11^T + A)$. □

Lemma 14 implies that all sets of permitted sets $\mathcal{P}(W)$ for symmetric networks W have the form $\mathcal{P}(W) = \text{stab}(-11^T + A)$, where A is a symmetric matrix having zero diagonal. The following Proposition implies that all such codes can also be obtained by perturbing around *any* rank 1 matrix with negative diagonal, not necessarily symmetric. Note that if $x, y \in \mathbb{R}^n_\times$, the rank 1 matrix $-xy^T$ has strictly negative diagonal if and only if $x_i y_i > 0$ for all $i \in [n]$.

Proposition 3. *Fix $x, y \in \mathbb{R}^n_\times$ with $x_i y_i > 0$ for all $i \in [n]$. For any symmetric threshold-linear network W on n neurons, there exists an $n \times n$ matrix B with zero diagonal such that*

$$\mathcal{P}(W) = \text{stab}(-xy^T + B).$$

The proof of this Proposition constructs the matrix B explicitly, and relies on the following Lemma.

Lemma 15. *Let M be any $n \times n$ matrix, and T an $n \times n$ invertible diagonal matrix. Then*

$$\text{stab}(TMT^{-1}) = \text{stab}(M).$$

Proof. We have $(TMT^{-1})_\sigma = T_\sigma M_\sigma T_\sigma^{-1}$. Since conjugation preserves the eigenvalue spectrum, the statement follows. □

Proof of Proposition 3. Let W be a symmetric threshold-linear network on n neurons. By Lemma 14, there exists a symmetric $n \times n$ matrix A with zero diagonal such that $\mathcal{P}(W) = \text{stab}(-11^T + A)$. It thus remains only to construct an $n \times n$ matrix B with zero diagonal such that

$$\text{stab}(-xy^T + B) = \text{stab}(-11^T + A).$$

We prove that this can always be done in two steps: first, we prove that it can be done in the special case $x = y$, and then we show that B can be constructed in general.

Step 1: Fix $x = y \in \mathbb{R}^n_\times$, and observe that $-xx^T + A^x = (-11^T + A)^x$, so by Lemma 13 we have $\text{stab}(-xx^T + A^x) = \text{stab}(-11^T + A)$. Letting $B = A^x$ we obtain the desired statement.

Step 2: Fix $x, y \in \mathbb{R}^n_\times$ so that $x_i y_i > 0$ for all $i \in [n]$, and let T be the diagonal matrix with entries $T_{ii} = \sqrt{y_i/x_i}$. Then

$$(T(-xy^T)T^{-1})_{ij} = \sqrt{\frac{y_i}{x_i}}(-x_i y_j)\sqrt{\frac{x_j}{y_j}} = -\sqrt{x_i y_i}\sqrt{x_j y_j},$$

so $T(-xy^T)T^{-1} = -zz^T$ for $z \in \mathbb{R}^n_\times$ having entries $z_i = \sqrt{x_i y_i}$. It follows from Step 1 that $\text{stab}(-11^T + A) = \text{stab}(-zz^T + A^z) = \text{stab}(T(-xy^T)T^{-1} + A^z)$. Let

$$B = T^{-1}A^z T.$$

Then, using Lemma 15, $\text{stab}(-xy^T + B) = \text{stab}(T(-xy^T + B)T^{-1}) = \text{stab}(-11^T + A)$. Since A has zero diagonal, so do A^z and B. Note that B can be obtained explicitly, using the expression for A in the proof of Lemma 14. □

6.3 Appendix C: Remarks on the ratio $-\frac{\text{cm}(A)}{\det(A)}$

Remark 1. If A is an $n \times n$ nondegenerate square distance matrix for $n > 1$, then the ratio $-\dfrac{\text{cm}(A)}{\det(A)}$ has a very nice geometric interpretation:

$$-\frac{\text{cm}(A)}{\det(A)} = \left|\frac{\text{cm}(A)}{\det(A)}\right| = \frac{1}{2\rho^2},$$

where ρ is the radius of the unique sphere circumscribed on the points used to generate A. This is proven in [10, Proposition 9.7.3.7], where it is also shown that $\det(A) \neq 0$ not only if A is a nondegenerate square distance matrix, but also if A is a *degenerate* square distance matrix corresponding to n points in general position in $\mathbb{R}^{n-2}$. Since $\mathrm{cm}(A)$ vanishes in this case, we see that the ratio $-\dfrac{\mathrm{cm}(A)}{\det(A)}$ goes to zero as n points that are initially in general position in $\mathbb{R}^{n-1}$ approach general position on a hyperplane of dimension $n-2$.

Remark 2. The previous remark has important implications for the apparent "fine-tuning" that is involved in eliminating spurious cliques by arranging points to be collinear, or coplanar, so that the corresponding principal submatrix A_σ is degenerate (as in Figure 2B). Since $-11^T + \varepsilon A_\sigma$ is only stable for

$$0 < \varepsilon < -\frac{\mathrm{cm}(A_\sigma)}{\det(A_\sigma)} = \frac{1}{2\rho^2},$$

where ρ is the radius of the circumscribed sphere, then by making the points $\{p_i\}_{i\in\sigma}$ corresponding to A_σ *approximately* degenerate, ρ can be made large enough so that $-11^T + \varepsilon A_\sigma$ is unstable – without the fine-tuning required to make A_σ exactly degenerate.

Similarly, exact solutions for k-skeleta of clique complexes, obtained using a k-sparse universal S which is a *degenerate* square distance matrix, are also not as fine-tuned as they might first appear. If, in fact, S is a nondegenerate square distance matrix, corresponding to a configuration of n points in $\mathbb{R}^{n-1}$ that *approximately* lies on a k-dimensional plane, the value of $\delta(S_\sigma)$ will be very small for any pattern of size $|\sigma| > k+1$; one can thus choose ε large enough to ensure that $\mathrm{geom}_\varepsilon(S) = \{\sigma \subset [n] \mid |\sigma| \leq k+1\}$, as in the case where S is truly degenerate.

Remark 3. It is quite simple to understand the scaling properties of $-\mathrm{cm}(A)/\det(A)$. If A is any $n \times n$ matrix, then $\mathrm{cm}(tA) = t^{n-1}\mathrm{cm}(A)$, while $\det(tA) = t^n \det(A)$, so

$$-\frac{\mathrm{cm}(tA)}{\det(tA)} = \frac{1}{t}\left(-\frac{\mathrm{cm}(A)}{\det(A)}\right),$$

independent of n. If $A_{ij} = \|p_i - p_j\|^2$, for $p_1, ..., p_n \in \mathbb{R}^{n-1}$, and we scale the position vectors so that $p_i \mapsto tp_i$ for each $i \in [n]$, then $A \mapsto t^2 A$ and we have

$$-\frac{\mathrm{cm}(A)}{\det(A)} \mapsto \frac{1}{t^2}\left(-\frac{\mathrm{cm}(A)}{\det(A)}\right).$$

This is consistent with the fact that the radius ρ of the circumscribed sphere scales as $\rho \mapsto t\rho$ in this case (see Remark 1).

Remark 4. Consider an $n \times n$ matrix A satisfying the Hebbian conditions $A_{ij} = A_{ji} \geq 0$ and $A_{ii} = 0$. If n is large, it is computationally intensive to test whether or not A is a nondegenerate square distance matrix using the criteria of Lemma 12, which potentially require computing $\mathrm{cm}(A_\sigma)$ for all $\sigma \subset [n]$.

On the other hand, our results imply that in order to test whether or not a Hebbian matrix A is a nondegenerate square distance matrix it is enough to check the stability of the matrix

$$-11^T + \varepsilon A \quad \text{for} \quad \varepsilon = \frac{1}{2}\left|\frac{\mathrm{cm}(A)}{\det(A)}\right|.$$

Here the factor of 1/2 was chosen arbitrarily, and can be replaced with any number $0 < c < 1$. For large n, this is a computationally efficient strategy, as it requires checking the eigenvalues of just one matrix.

6.4 Appendix D: Additional supporting text

Relationship between the input b and permitted sets $\sigma \in \mathcal{P}(W)$ that can be activated by the network

In the following proposition we use the notation that $x < y$, for vectors $x, y \in \mathbb{R}^n$, if and only if $x_i < y_i$ for each $i \in [n]$. The symbols $>$ and $\leq$ are interpreted analogously.

Proposition 4 ([16, Proposition 2.1])**.** *Consider the threshold-linear network W (not necessarily symmetric) with dynamics given by equation* (1)*, in the presence of a particular fixed input b. Let $\sigma \subset [n]$ be a subset of neurons, and $\bar{\sigma}$ its complement. Then, a point $x^* \in \mathbb{R}^n$ with $x^*_\sigma > 0$ and $x^*_{\bar{\sigma}} = 0$ is a fixed point if and only if*

(i) $b_\sigma = (D - W)_\sigma x^*_\sigma$, and

(ii) $b_{\bar{\sigma}} \leq -W_{\bar{\sigma}\sigma} x^*_\sigma$,

where $W_{\bar{\sigma}\sigma}$ is the submatrix with rows and columns restricted to $\bar{\sigma}$ and σ, respectively. In particular, if $\det(D - W)_\sigma \neq 0$, then there exists at most one fixed point with support σ and it is given by

$$x^*_\sigma = (D - W)^{-1}_\sigma b_\sigma \quad \textit{and} \quad x^*_{\bar{\sigma}} = 0,$$

provided that $(D - W)^{-1}_\sigma b_\sigma > 0$ and properties (i) and (ii) hold. Moreover, if x^* is a fixed point and $b_{\bar{\sigma}} < -W_{\bar{\sigma}\sigma} x^*_\sigma$, then x^* is asymptotically stable if and only if $(-D + W)_\sigma$ is stable.

This proposition allows one to find candidate inputs b that may activate a given permitted set $\sigma \in \mathcal{P}(W)$. Suppose $-W$ and $D - W$ are nonnegative and invertible matrices (as in the typical output of the Encoding Rule), and let $\sigma \in \mathcal{P}(W)$ be a permitted set. Then if $b_\sigma > 0$ and $b_{\bar{\sigma}} < 0$, using the above proposition we know that there will be a unique stable fixed point with support σ, given by $x^*_\sigma = (D - W)^{-1}_\sigma b_\sigma$, provided $(D - W)^{-1}_\sigma b_\sigma > 0$ (otherwise there is no stable fixed point with support σ). Note, however, that stable fixed points with other supports are also possible for the same input b.

Details related to generation of PF codes for Figure 4

To produce Figure 4, we generated random k-sparse PF codes with circular place fields, $n = 80$-100 neurons, and $k = .1n$. For each code, n place field centers were selected uniformly at random from a square box environment of side length 1, and n place field radii were drawn independently from an experimentally observed gamma distribution (Figure 5). We then computed the 2-skeleton for each PF code, with pairwise and triple overlaps of place fields determined from simple geometric considerations. The full PF code was obtained as the Helly completion of the 2-skeleton (see Lemma 6). Finally, to obtain the k-sparse PF code, we restricted the full code to its $(k - 1)$-skeleton, thereby eliminating patterns of size larger than k.

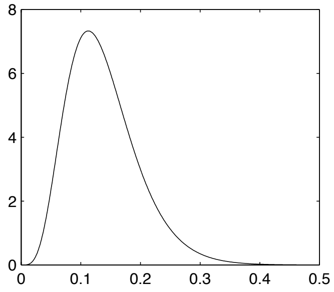

Figure 5: Gamma distribution used for generating random place field radii; this fits the experimentally-observed mean and variability (see [17, Figure 4B]).

References

[1] Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biol Cybern **27**(2), 77–87 (1977)

[2] Amit, D.: Modeling Brain Function: the World of Attractor Neural Networks. Cambridge University press. (1989)

[3] Amit, D., Gutfreund, H., Sompolinsky, H.: Spin-glass models of neural networks. Physical Review. A **32**(2), 1007–1018 (1985)

[4] Amit, D., Gutfreund, H., Sompolinsky, H.: Statistical mechanics of neural networks near saturation. Annals of Physics NY 173, 30 (1987)

[5] Amit, D.J.: Modeling brain function. Cambridge University Press, Cambridge (1989). The world of attractor neural networks

[6] Andersen, P., Morris, R., Amaral, D., Bliss, T., O'Keefe, J.: The Hippocampus Book. Oxford University Press (2006)

[7] Barth, A.L., Poulet, J.F.: Experimental evidence for sparse firing in the neocortex. Trends Neurosci **35**(6), 345–355 (2012)

[8] Barvinok, A.: A course in convexity, *Graduate Studies in Mathematics*, vol. 54 (2002)

[9] Ben-Yishai, R., Bar-Or, R.L., Sompolinsky, H.: Theory of orientation tuning in visual cortex. Proc Natl Acad Sci U S A **92**(9), 3844–8 (1995)

[10] Berger, M.: Geometry. I. Universitext. Springer-Verlag, Berlin (1994). Translated from the 1977 French original by M. Cole and S. Levy, Corrected reprint of the 1987 translation

[11] Berkes, P., Orban, G., Lengyel, M., Fiser, J.: Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science **331**, 83–87 (2011)

[12] Blumenthal, L.M.: Theory and applications of distance geometry. Oxford, at the Clarendon Press (1953)

[13] Bott, R., Tu, L.W.: Differential forms in algebraic topology. Springer-Verlag, New York (1982)

[14] Bressloff, P.C.: Spatiotemporal dynamics of continuum neural fields. J. Phys. A **45**(3) (2012)

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

[16] Curto, C., Degeratu, A., Itskov, V.: Flexible memory networks. Bulletin of Mathematical Biology **74**(3), 590–614 (2012)

[17] Curto, C., Itskov, V.: Cell groups reveal structure of stimulus space. PLoS Comput Biol **4**(10) (2008)

[18] Curto, C., Itskov, V., Morrison, K., Roth, Z., Walker, J.: Combinatorial neural codes from a mathematical coding theory perspective. Neural Computation (In press)

[19] Curto, C., Itskov, V., Veliz-Cuba, A., Youngs, N.: The neural ring: an algebraic tool for analyzing the intrinsic structure of neural codes. arXiv:1212.4201 [q-bio.NC] (2012)

[20] Dayan, P., Abbott, L.F.: Theoretical neuroscience. MIT Press, Cambridge, MA (2001)

[21] Deza, M.M., Laurent, M.: Geometry of cuts and metrics, *Algorithms and Combinatorics*, vol. 15. Springer-Verlag, Berlin (1997)

[22] Ermentrout, G., Terman, D.: Mathematical Foundations of Neuroscience. Springer (2010)

[23] Hahnloser, R.H., Seung, H.S., Slotine, J.J.: Permitted and forbidden sets in symmetric threshold-linear networks. Neural Comput **15**(3), 621–638 (2003)

[24] Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the theory of neural computation. Addison-Wesley, Redwood City, CA (1991)

[25] Hillar, C., Tran, N., Koepsell, K.: Robust exponential binary pattern storage in little-hopfield networks. arXiv:1206.2081 [q-bio.NC] (2012)

[26] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79**(8), 2554–2558 (1982)

[27] Hopfield, J.J.: Neurons with graded response have collective computational properties like those of two-sate neurons. Proc. Natl. Acad. Sci. **81**, 3088–3092 (1984)

[28] Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press (1985)

[29] Hromádka, T., Deweese, M.R., Zador, A.M.: Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol **6**(1) (2008)

[30] Huffman, W., Pless, V.: Fundamentals of Error-Correcting Codes. Cambridge University Press, Cambridge, MA (2003)

[31] Itskov, V., Hansel, D., Tsodyks, M.: Short-term facilitation may stabilize parametric working memory trace. Frontiers in Computational Neuroscience **5**, 1–19 (2011)

[32] Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., Arieli, A.: Spontaneously emerging cortical representations of visual attributes. Nature **425**, 954–956 (2003)

[33] Luczak, A., Bartho, P., Harris, K.: Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron **62**, 413–425 (2009)

[34] MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes. North-Holland, Amsterdam (1983)

[35] McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.B.: Path integration and the neural basis of the 'cognitive map'. Nat Rev Neurosci **7**(8), 663–78 (2006)

[36] Muller, R.: A quarter of a century of place cells. Neuron **17**(5), 813–822 (1996)

[37] Niskanen, S., Ostergard, P.: Cliquer - routines for clique searching (2010). Available at http://users.tkk.fi/pat/cliquer.html

[38] O'Keefe, J.: Place units in the hippocampus of the freely moving rat. Exp. Neurol. **51**, 78–109 (1976)

[39] O'Keefe, J., Nadel, L.: The Hippocampus as a Cognitive Map. Clarendon Press, Oxford, UK (1978)

[40] Osborne, L., Palmer, S., Lisberger, S., Bialek, W.: The neural basis for combinatorial coding in a cortical population response. Journal of Neuroscience **50**(28), 13,522–13,531 (2008)

[41] Oxley, J.: Matroid theory, *Oxford Graduate Texts in Mathematics*, vol. 21, second edn. Oxford University Press, Oxford (2011)

[42] Petersen, C.C.H., Malenka, R.C., Nicoll, R.A., Hopfield, J.J.: All-or-none potentiation at CA3-CA1 synapses. PNAS **95**, 4732–4737 (1998)

[43] Roudi, Y., Treves, A.: Disappearance of spurious states in analog associative memories. Phys. Rev. E **67**, 041,906 (2003)

[44] Schoenberg, I.J.: Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". Ann. of Math. (2) **36**(3), 724–732 (1935). DOI 10.2307/1968654. URL `http://dx.doi.org/10.2307/1968654`

[45] Shriki, O., Hansel, D., Sompolinsky, H.: Rate models for conductance-based cortical neuronal networks. Neural Comput **15**(8), 1809–1841 (2003)

[46] Ulanovsky, N.: Neuroscience: how is three-dimensional space encoded in the brain? Curr Biol **21**, 886–888 (2011)

[47] Xie, X., Hahnloser, R.H., Seung, H.S.: Selectively grouping neurons in recurrent networks of lateral inhibition. Neural Comput **14**, 2627–46 (2002)

[48] Yuste, R., MacLean, J., Smith, J., Lansner, A.: The cortex as a central pattern generator. Nat Rev Neurosci **6**, 477–83 (2005)