9-2022

Permitted Sets and Convex Coding in Nonthreshold Linear Networks

Steven Collazos

Duane Nykamp

LETTER Communicated by Richard Hahnloser

Permitted Sets and Convex Coding in Nonthreshold Linear Networks

Steven Collazos
colla054@umn.edu
Science and Math Division, University of Minnesota Morris, Morris, MN 56267, U.S.A.

Duane Nykamp
nykamp@umn.edu
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Hebbian theory proposes that ensembles of neurons form a basis for neural processing. It is possible to gain insight into the activity patterns of these neural ensembles through a binary analysis, regarding neurons as either active or inactive. The framework of permitted and forbidden sets, introduced by Hahnloser, Seung, and Slotine (2003), is a mathematical model of such a binary analysis: groups of coactive neurons can be permitted or forbidden depending on the network's structure.

In order to widen the applicability of the framework of permitted sets, we extend the permitted set analysis from the original threshold-linear regime. Specifically, we generalize permitted sets to firing rate models in which Φ is a nonnegative continuous piecewise C^1 activation function. In our framework, the focus is shifted from a neuron's firing rate to its responsiveness to inputs; if a neuron's firing rate is sufficiently sensitive to changes in its input, we say that the neuron is responsive. The algorithm for categorizing a neuron as responsive depends on thresholds that a user can select arbitrarily and that are independent of the dynamics.

Given a synaptic weight matrix W, we say that a set of neurons is permitted if it is possible to find a stimulus where those neurons, and no others, remain responsive. The main coding property we establish about $P_\Phi(W)$, the collection of all permitted sets of the network, is that $P_\Phi(W)$ is a convex code when W is almost rank one. This means that $P_\Phi(W)$ in the low-rank regime can be realized as a neural code resulting from the pattern of overlaps of receptive fields that are convex.

1 Introduction

A central unsolved question in neuroscience is how neural activity and network connectivity influence each other. As a result, groups of coactive neurons are phenomena of interest in neuroscience (Hebb, 2005; Harris,

Neural Computation **34, 1978–2008** (2022)
https://doi.org/10.1162/neco_a_01523

2005; Thompson & Scott, 2016; Wilson & McNaughton, 1994). A framework in which the activity patterns of coactive neurons—and how they are shaped by the network's structure—can be analyzed is the firing rate model known as threshold-linear network (TLN). The neurons' response to overall synaptic input in a TLN is given by a rectifier, $\Phi(x) = \max(x, 0)$, where $x \in \mathbb{R}$.

In previous work, Hahnloser et al. (Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000; Hahnloser, Seung, & Slotine, 2003) introduced TLNs with symmetric weight matrices and studied their fixed points. In their framework, the input is a stimulus and the output is a collection of neurons that can become stably coactive, called a permitted set of the network. (Here "stably coactive" refers to an asymptotically stable fixed point of the dynamics whose nonzero coordinates correspond to the coactive neurons.) Similarly, there are groups of neurons that cannot be stably coactive regardless of the choice of stimulus; such a group of neurons is called a forbidden set. Thus, their framework combines the interplay between digital and analog coding in neurons. On the other hand, since a pattern of active and silent neurons in a population can be identified with a binary list of neuron indices, the set of all such activity patterns can be considered a combinatorial neural code (Osborne, Palmer, Lisberger, & Bialek, 2008). Therefore, the collection of permitted sets of a network is a combinatorial neural code (Curto, Degeratu, & Itskov, 2013). It turns out that permitted sets for TLNs can also be considered when the synaptic weight matrix is not symmetric (Curto, Degeratu, & Itskov, 2012).

The main goal of this letter is to generalize the notion of permitted and forbidden sets to networks with activation functions Φ that can be any nonnegative continuous piecewise C^1 function. (A function Φ is C^1 when Φ is differentiable and its derivative is a continuous function.) A key step for this generalization is recasting the definition of these sets to be based on the concept of a neuron being responsive to input rather than active. In the TLN framework, the active neurons that define permitted sets are neurons with a positive firing rate. In our generalization, the responsive neurons that define permitted sets are neurons whose gain, or derivative of Φ, is sufficiently high. In the case of TLNs, the gain is binary, either 0 or 1, and the gain of 1 corresponds exactly to the original definition of active neurons with a positive firing rate. In general, the gain can vary continuously with input, so to categorize responsive neurons, we introduce a threshold gain that can be chosen arbitrarily.

Unlike TLN networks, the threshold for responsive neurons is strictly a tool to transform neuronal activity to a combinatorial code for analysis. The threshold does not have an impact on the dynamics: both unresponsive and responsive neurons can have graded effects on a postsynaptic neuron. Nonetheless, the gain itself is a natural quantity to characterize network dynamics, such as stability, or the encoding properties, such as in the score function used to define Fisher information.

By establishing thresholds to categorize neurons into responsive and unresponsive sets, we create the combinatorial neural code of permitted sets based on patterns of coresponsive neurons. We prove results regarding patterns of coresponsive neurons that can be supported by networks with low-rank synaptic weight matrix, paving a way for further binary analysis of activity patterns of firing rate models.

This letter is organized into six sections. In section 2, we give an overview of threshold-linear networks. In section 3, we introduce our definition of permitted set for a firing network model in the general setting. We also show related results and an example of a network where we find its permitted sets. In section 4, we prove that if the network's synaptic weight matrix is close to being rank one; then the collection of permitted sets forms a convex code. In section 5, we discuss our results. Finally, we present proofs of several results in the appendix.

2 Firing-Rate Network Models

One way of modeling the dynamics of recurrent neural networks is via firing-rate models, where the network consists of neuron-like units whose outputs are firing rates. Two advantages of firing-rate models are that they avoid the short-timescale dynamics required to simulate action potentials, and they allow us to perform analytic calculations of some aspects of network dynamics (Dayan & Abbott, 2001).

A *threshold-linear network* (Hahnloser, Seung, & Slotine, 2003) is a neural network model where the dynamics of each neuron is given by

$$\tau \dot{x}_i + x_i = \left[\sum_{j=1}^{N} w_{ij} x_j + b_i \right]_+ ,$$

where we use the following notation:

$[\cdot]_+$:	Rectification nonlinearity, $[x]_+ = \max(x, 0)$, where $x \in \mathbb{R}$
$x_i(t)$:	Firing rate of neuron i at time t
b_i:	Input current to neuron i
$\tau > 0$:	Neuron's timescale
w_{ij}:	Effective strength of the synapse of neuron j onto neuron i

In general, a firing-rate model of N neurons has the form

$$\tau \dot{x}_i + x_i = \Phi \left(\sum_{j=1}^{N} w_{ij} x_j + b_i \right),$$

where Φ, the activation function, describes the steady-state firing rate of the neurons as a function of the total synaptic input.

The systems of N differential equations we discuss in this section can be expressed more compactly in terms of the firing-rate vector $x = (x_1, x_2, \ldots, x_N)$ as

$$D\dot{x} + x = [Wx + b]_+ \tag{2.1}$$

in the TLN regime and as

$$D\dot{x} + x = \Phi\,(Wx + b) \tag{2.2}$$

in general. Here W is an $N \times N$ synaptic weight matrix; $D = \mathrm{diag}(\tau, \tau, \ldots, \tau)$ is an $N \times N$ diagonal matrix of time constants; and $b \in \mathbb{R}^N$ is interpreted as an external stimulus to the network. Furthermore, $\Phi(x)$ denotes $(\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N))$. Throughout this letter, Φ is nonnegative continuous piecewise C^1, that is, continuous everywhere and differentiable everywhere except possibly at finitely many points. Finally, $\Phi'(s) \to 0$ as $s \to -\infty$.

3 Permitted and Forbidden Sets

A permitted set is a set of neurons that can be made stably coresponsive (where, implicitly, we mean all neurons outside the set are unresponsive). In general, if a neuronal network is weakly coupled, then external input can drive any activity pattern and any set of neurons can be made stably coresponsive with a suitable stimulus. In this case, the set of permitted sets is trivial: it consists of all subsets of neurons. Of more interest is the case where the feedback of synaptic connections is strong enough to amplify perturbations and destabilize certain activity patterns. If the network feedback prevents a set of neurons from being stably coresponsive, no matter the external input, then that set corresponds to a forbidden set. The activity might stabilize to a different pattern of coresponsive neurons, as the gains on some neurons change. The resulting stably coresponsive set would correspond to a permitted set.

3.1 Effective Gain Matrix. We need a few preliminary results before formalizing the concept of a permitted set when the network is not a TLN. Here, we state a lemma showing we can create any steady-state activity pattern and introduce the concept of the effective gain of the network.

Lemma 1. *Let W be an $N \times N$ synaptic weight matrix, D an $N \times N$ diagonal matrix of time constants, and Φ be an activation function (which need not be differentiable). Let $x^* = (\Phi(I_1), \ldots, \Phi(I_N))$ be given, where $I_k \in \mathbb{R}$. Then there is $b \in \mathbb{R}^N$ such that x^* is a fixed point of $D\dot{x} + x = \Phi(Wx + b)$.*

Proof. See the appendix. □

Lemma 1 states that provided we have freedom over how to stimulate a neuronal network, we can create a steady state for any pattern of activity across the network. Since the lemma does not say anything about the stability of the steady states, it does not distinguish between activity patterns that could correspond to permitted or forbidden sets. As a first step toward addressing stability, we define a quantity that will play a fundamental role in our stability calculations.

Definition 1 (Effective gain of a network). *For a neuronal network modeled by equation 2.2, we define the* effective gain of the network *at* $\boldsymbol{x}$ *to be*

$$\Lambda(x) = \begin{pmatrix} \Phi'(W^{(1)} \cdot x + b_1) & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Phi'(W^{(N)} \cdot x + b_N) \end{pmatrix}, \tag{3.1}$$

where $W^{(k)}$ *denotes the kth row of* W.

The effective gain $\Lambda(x)$ captures the influence of the network on the responsiveness of each neuron. Given some stimulus b in $\mathbb{R}^N$, let x^* be a fixed point of equation 2.2, which we can rewrite as

$$\dot{x} = D^{-1} \left(\Phi(Wx + b) - x \right).$$

The stability of the fixed point x^* is determined by the Jacobian matrix of the right-hand side. However, since we assume all time constants are the same, the matrix D^{-1} is proportional to the identity and simply rescales the eigenvalues of the Jacobian matrix. Hence, the stability depends only on the Jacobian matrix of $\Phi(Wx + b) - x$ at x^*, which is $\Lambda(x^*)W - I$. In terms of neurobiology, $\Lambda(x^*)W$ is the effective recurrence matrix.

3.2 Permitted, Marginally Permitted, and Forbidden Sets.

To our knowledge, there is no definition of permitted sets for firing-rate models outside of TLNs. One main issue in the generalization is that the gain is no longer binary as in the TLN regime. We can recover the binary distinction for our analysis by dichotomizing using a gain threshold. For additional flexibility, we introduce two thresholds: r_{off} and r_{on}, with $0 \leq r_{\text{off}} \leq r_{\text{on}}$. We regard a neuron with gain at or below r_{off} as unresponsive; we regard a neuron with gain above r_{on} as responsive. A user can optionally choose $r_{\text{on}} > r_{\text{off}}$ to create stricter definitions of forbidden and permitted sets that exclude neurons with intermediate gains. Alternatively, one could set $r_{\text{on}} = r_{\text{off}}$ so that all neurons can be classified as responsive or unresponsive. These thresholds allow analysis of the network activity in terms of combinatorial codes even when the dynamics are governed by a Φ that is continuous piecewise C^1 with a slope that spans a continuum.

The two thresholds introduce three possible categorizations of a neuron's activity. The magnitude of the gain of neuron i is the absolute value of the derivative, that is, $|\Phi'(W^{(i)} \cdot x + b_i)|$, where $W^{(i)}$ denotes the ith column of W. We categorize neuron i as unresponsive, marginally responsive, or responsive by comparing the magnitude of the gain with the gain thresholds. (Note that the category marginally responsive is possible only if $r_{\text{on}} > r_{\text{off}}$.)

Definition 2 (Unresponsive and responsive neurons). *Let x^* be a vector of firing rates of a network with N neurons. Suppose that equation 2.2 describes the network's dynamics. Let Φ'_i denote $\Phi'\left(W^{(i)} \cdot x + b_i\right)$. A neuron i is* unresponsive *when $|\Phi'_i| \leq r_{off}$ and* responsive *when $|\Phi'_i| > r_{on}$. Additionally, we say that i is* marginally responsive *if $r_{off} < |\Phi'_i| \leq r_{on}$.*

We next define the concept of coresponsive neurons as a set of neurons that are responsive with all other neurons being unresponsive. (We insist that no neurons are marginally responsive in order to use the coresponsive designation.)

Definition 3 (Coresponsive neurons). *We say that a nonempty subset $\sigma \subseteq \{1, 2, \ldots, N\}$ of neurons is* coresponsive *when $|\Phi'_i| > r_{on}$ for all $i \in \sigma$ and $|\Phi'_j| \leq r_{off}$ for all $j \notin \sigma$.*

Equipped with a way of classifying groups of neurons based on whether they are responsive, marginally responsive, or uresponsive, we now introduce a combinatorial neural code for networks whose dynamics are described by equation 2.2. The combinatorial code is determined by the stable patterns of neurons that can be elicited by any stimulus b. If a set of neurons σ can be made stably coresponsive by some stimulus, σ is a permitted set. For the case with $r_{\text{on}} > r_{\text{off}}$, we can also have marginally permitted sets by allowing some neurons in σ to be only marginally responsive. A forbidden set σ is a set of neurons that cannot be made stably coresponsive (even if we allow marginally responsive neurons) no matter the value of the stimulus b.

Definition 4 (Permitted and forbidden sets). *Suppose the network's dynamics are described by equation 2.2. Let $0 \leq r_{off} \leq r_{on}$ be constants such that there is $s_{off} \in \mathbb{R}$ for which $|\Phi'(s)| \leq r_{off}$ for all $s \leq s_{off}$. When σ is a subset of $\{1, 2, \ldots, N\}$, then:*

1. *We call σ a* permitted set *if there exists a b in $\mathbb{R}^N$ such that there is an asymptotically stable fixed point $x^* = (\Phi(I_1), \ldots, \Phi(I_N))$ for which the neurons in σ are coresponsive.*
2. *We call σ a* marginally permitted set *if σ is not permitted and there exists a b in $\mathbb{R}^N$ such that there is an asymptotically stable fixed point $x^* = (\Phi(I_1), \ldots, \Phi(I_N))$ for which the neurons in σ are marginally responsive or responsive, whereas neurons that are not in σ are unresponsive.*

3. *If σ is neither permitted nor marginally permitted, then we call it a* forbidden set.

We denote the network's collection of all permitted sets by $\mathcal{P}_\Phi(W)$. (Note that $\mathcal{P}_\Phi(W)$ depends on r_{off} and r_{on}.)

Note that definition 4 ensures that neurons become unresponsive for sufficiently small input, which is also the input range where neurons can be given an arbitrary low gain. (Recall that $\lim_{s\to-\infty} \Phi'(s) = 0$.) An important consequence is that the empty set is permitted. By giving each neuron a sufficiently negative input, one can obtain an asymptotically stable fixed point supporting a group of neurons that are all unresponsive.

3.3 Permitted Sets in a Rank-One Network. The next useful lemma gives a formula for the one eigenvalue we will usually be concerned with when dealing with rank-one W synaptic weight matrices. Recall that rank-one matrices satisfy the property that $W = uv^T$, where u and v are nonzero and v^T denotes the transpose of v. Further, the eigenvalues of W are tr(W) and 0, where tr(W) denotes the trace of W (Horn & Johnson, 2012).

Lemma 2. *Suppose that the network's dynamics are described by equation 2.2. Further suppose that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero. Let $x^* = (\Phi(I_1), \Phi(I_2), \dots, \Phi(I_N))$. Then x^* is an asymptotically stable fixed point if and only if there is some $I^* = (I_1^*, I_2^*, \dots, I_N^*)$ in $\Phi^{-1}(x_1^*) \times \Phi^{-1}(x_2^*) \times \cdots \times \Phi^{-1}(x_N^*)$ such that*

$$tr(\Lambda(x^*)W) = \sum_{i=1}^{N} u_i v_i \, \Phi'(I_i^*) < 1.$$

(Here $\Phi^{-1}(x_1^*) \times \Phi^{-1}(x_2^*) \times \cdots \times \Phi^{-1}(x_N^*)$ denotes the Cartesian product of the sets $\Phi^{-1}(x_1^*)$, $\Phi^{-1}(x_2^*)$, ..., and $\Phi^{-1}(x_N^*)$. In general, $\Phi^{-1}(x_i^*)$ is the preimage of x_i^*, so it is a set of values.)

Proof. See the appendix. □

Next, we show a simple example of finding a network's permitted and forbidden sets. Recall that following definition 4, sets of neurons σ are subsets of $\{1, 2, \dots, N\}$, so numbers in the set $\{1, 2, \dots, N\}$ stand for neurons' indices.

Example 1. Let W be the synaptic weight matrix

$$W = uv^T = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Suppose D is the 2×2 identity matrix and $\Phi(x) = \exp(x)$. Then equation 2.2 will take the form

$$\dot{x}_i + x_i = \exp\left(b_i + \frac{1}{2}x_1 + \frac{1}{2}x_2\right),$$

where i can be 1 or 2. (Here $\tau = 1$.)

Next, we pick gain thresholds: let $r_{\text{off}} = \dfrac{1}{4}$ and $r_{\text{on}} = \dfrac{3}{2}$.

When W is rank one, $|||W|||_2 = ||u||_2\,||v||_2$, where $||\cdot||_2$ denotes the Euclidean norm. Hence, $||W||_2 = 1$ because $||u||_2 = ||v||_2 = 1$. In particular, $1/|||W|||_2 = 1$. Now let us find the permitted sets of the network.

First, we claim that $\{1, 2\}$ is not a permitted set, which means that neurons 1 and 2 cannot be stably coresponsive: let $b \in \mathbb{R}^2$ be such that $x^* = (\exp(I_1), \exp(I_2))$ is a fixed point of the dynamics for which $|\Phi'(I_1)| = e^{I_1} > r_{\text{on}}$ and $|\Phi'(I_2)| = e^{I_2} > r_{\text{on}}$. Since Φ is monotonic, I_1 and I_2 are the only net inputs possible in lemma 2. We calculate

$$\sum_{k=1}^{2} u_k v_k \Phi'(I_k) = \frac{1}{2}e^{I_1^*} + \frac{1}{2}e^{I_2^*} > \frac{1}{2}r_{\text{on}} + \frac{1}{2}r_{\text{on}} = \frac{3}{2} > 1.$$

By lemma 2, the fixed point is not asymptotically stable and $\{1, 2\}$ is not permitted.

On the other hand, $\{1\}$ and $\{2\}$ are permitted; it suffices to show that $\{1\}$ is a permitted set (since the argument for $\{2\}$ will be identical in this example).

Let

$$b_1 = \log\left(\frac{5}{3}\right) - \left(\frac{1}{2}e^{\log(5/3)} + \frac{1}{2}e^{\log(1/5)}\right)$$

and

$$b_2 = \log\left(\frac{1}{5}\right) - \left(\frac{1}{2}e^{\log(5/3)} + \frac{1}{2}e^{\log(1/5)}\right).$$

Observe that $x^* = (e^{\log(5/3)}, e^{\log(1/5)}) = (5/3, 1/5)$ is a fixed point of the dynamics. Further, $\Phi'(\log(5/3)) = 5/3 > r_{\text{on}}$ and $\Phi'(\log(1/5)) = 1/5 < r_{\text{off}}$. Now apply lemma 2 to verify that x^* is indeed asymptotically stable:

$$\begin{aligned}\sum_{k=1}^{2} u_k v_k \Phi'(I_k^*) &= \frac{1}{2}e^{\log(5/3)} + \frac{1}{2}e^{\log(1/5)} \\ &= \frac{5}{6} + \frac{1}{10} < 1.\end{aligned}$$

Therefore, $\{1\}$ is permitted.

3.4 Relationship between Weakly Coupled Networks and Responsiveness Thresholds. The framework we introduce allows one to freely choose responsiveness thresholds r_{on} and r_{off}. Although the results we present are valid for any such values (aside from the technical restriction involving s_{off} in definition 4), the results are not interesting if one chooses r_{on} so small that neurons can become responsive without affecting the stability of fixed points. For such a small r_{on}, all sets would be permitted. We show that the threshold for interesting behavior varies inversely with coupling strength so that weakly coupled networks would require a large r_{on} to allow for the possibility of forbidden sets.

Before presenting the result, recall that (Horn & Johnson, 2012) when $||\cdot||$ is a norm on a vector space, the matrix norm induced by $||\cdot||$ of an $N \times N$ matrix A is defined by

$$|||A||| = \max_{||x||=1} ||Ax||.$$

Later in this letter $||\cdot||$ will be the usual Euclidean norm on $\mathbb{R}^N$. We also remind readers that (Horn & Johnson, 2012) the spectral radius $\rho(A)$ of A is defined as the largest absolute value of the eigenvalues of A:

$$\rho(A) = \max(|\lambda_1|, |\lambda_2|, \ldots, |\lambda_N|),$$

where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A. It turns out that $\rho(A) \leq |||A|||$. Finally, recall that the matrix norm is submultiplicative (Horn & Johnson, 2012):

$$|||AB||| \leq |||A|||\,|||B|||$$

for any $N \times N$ matrices A and B.

Proposition 1. *Suppose that the network's dynamics are described by equation 2.2. Let $0 \leq r_{off} \leq r_{on}$ be constants such that there is $s_{off} \in \mathbb{R}$ for which $|\Phi'(s)| \leq r_{off}$ for all $s \leq s_{off}$. Suppose that*

$$r_{on} < \frac{1}{|||W|||}$$

and that there is $J \in \mathbb{R}$ such that $r_{on} < |\Phi'(J)| < 1/|||W|||$.

Then any subset of $\{1, 2, \ldots, N\}$ is a permitted set, that is, for any subset σ of $\{1, 2, \ldots, N\}$, there is $b \in \mathbb{R}^N$ so that $D\dot{x} + x = \Phi(Wx + b)$ has an asymptotically stable fixed point $x^* = (\Phi(I_1), \Phi(I_2), \ldots, \Phi(I_N))$ satisfying

$$|\Phi'(I_i)| > r_{on} \quad \text{for all } i \in \sigma \qquad \text{and} \qquad |\Phi'(I_j)| \leq r_{off} \quad \text{for all } j \notin \sigma.$$

Proof. Let $J \in \mathbb{R}$ be such that $1/|||W||| > |\Phi'(J)| > r_{\text{on}}$. Continuity of Φ' at J implies that there is an open interval $\mathcal{N}_{\text{on}} \subset \mathbb{R}$ containing J such that for all $x \in \mathcal{N}_{\text{on}}$, we have $r_{\text{on}} < |\Phi'(x)| < 1/|||W|||$. Furthermore, $|\Phi'(x)| \leq r_{\text{off}}$ for all $x \in \mathcal{N}_{\text{off}}$ for some open interval $\mathcal{N}_{\text{off}} \subset \mathbb{R}$ because $\Phi'(s) \to 0$ as $s \to -\infty$, $|\Phi'(s)| \leq r_{\text{off}}$ for all $s \leq s_{\text{off}}$, and Φ' is continuous except possibly at finitely many values. Let x^* be such that for all $i \in \sigma$, we have $x_i^* = \Phi(I_i)$, where $I_i \in \mathcal{N}_{\text{on}}$, while for all $j \notin \sigma$, we have $x_j^* = \Phi(I_j)$, where $I_j \in \mathcal{N}_{\text{off}}$.

Next, observe that if $i \in \sigma$, then $\Phi^{-1}(x_i^*) \cap \mathcal{N}_{\text{on}} \neq \emptyset$ because $I_i \in \Phi^{-1}(x_i^*)$. Similarly, $\Phi^{-1}(x_j^*) \cap \mathcal{N}_{\text{off}} \neq \emptyset$. Hence, by lemma 1, if $b = L^* - Wx^*$, where $L_i^* \in \Phi^{-1}(x_i^*) \cap \mathcal{N}_{\text{on}}$ for all $i \in \sigma$ and $L_j^* \in \Phi^{-1}(x_j^*) \cap \mathcal{N}_{\text{off}}$ for all $j \notin \sigma$, then x^* is a fixed point.

To prove that x^* is in fact an asymptotically stable fixed point, we show that the eigenvalues of $JF(x^*) = \Lambda(x^*)W - I$ have negative real part. Since $\Lambda(x^*)$ is a diagonal matrix whose entries are strictly bounded by $1/|||W|||$, we get $|||\Lambda(x^*)||| < 1/|||W|||$. Then by the submultiplicativity property of matrix norms, we observe

$$\begin{aligned} |||\Lambda(x^*)W||| &\leq |||\Lambda(x^*)|||\,|||W||| \\ &< \frac{1}{|||W|||}|||W||| \\ &= 1. \end{aligned}$$

Therefore, $\rho(\Lambda(x^*)W) < 1$ because $\rho(\Lambda(x^*)W) \leq |||\Lambda(x^*)W|||$, which implies that the eigenvalues of $\Lambda(x^*)W - I$ are negative. Hence, x^* is asymptotically stable. □

Roughly speaking, proposition 1 shows that given a suitable activation function, one should be careful not to select a responsiveness threshold that is too small. Otherwise, every coresponsive group of neurons will be a codeword in our neural code (i.e., the collection of permitted sets). The resulting trivial combinatorial neural code would be uninteresting. The framework of permitted sets can provide insight into network behavior when some groups of neurons can be made coresponsive given a suitable stimulus, whereas there are other groups that cannot be made stably coresponsive regardless of the stimulus that is applied to them.

4 Convex Coding and Rank-One Networks

In the previous section, we introduced permitted sets $\mathcal{P}_\Phi(W)$ of networks whose dynamics are described by $D\dot{x} + x = \Phi(Wx + b)$. We said that $\mathcal{P}_\Phi(W)$ is the collection of all groups of neurons that can be made stably coresponsive by a stimulus. If we think of $\mathcal{P}_\Phi(W)$ as a combinatorial neural code, it is natural to inquire about the properties of such a code. Recall that a codeword in a combinatorial neural code is a pattern of neural activity where neurons are either responsive or unresponsive. Equivalently, codewords in a combinatorial neural code are subsets of $\{1, 2, \ldots, N\}$; for example, any neuron in an ensemble $\sigma \subseteq \{1, 2, \ldots, N\}$ is responsive and neurons that are not in σ are unresponsive.

4.1 Convex Codes. One property of combinatorial neural codes is convexity. Recall that a set C of $\mathbb{R}^d$ is convex: any two points in C can be connected by a straight line contained entirely within C. The notion of a convex code is based on endowing each neuron with a receptive field in a stimulus space: an activation pattern across all neurons is created by choosing a stimulus s, which is a point in $\mathbb{R}^d$, and for a given s, a neuron is active if s is in its receptive field and silent s is outside its receptive field. (As a reminder, we use the term *active* in the usual sense, i.e., to say that a neuron's magnitude of firing is large enough.) The set of all possible activation patterns generated by all stimuli is the *receptive field code*.

We next consider an example of a convex code.

Example 2. The code

$$\mathcal{C}_1 = \{\emptyset, \{1\}, \{4\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$$

is an open convex code, a combinatorial neural code that is generated by the pattern of intersections of a collection of open and convex subsets in some $\mathbb{R}^d$. One can demonstrate this by using a one-dimensional stimuli space consisting of four open intervals corresponding to the four neurons whose overlaps generate $\mathcal{C}_1$. For this example, the stimuli space is $\mathbb{R}$, and consider the receptive fields

$$U_1 = (-1, 5),$$
$$U_2 = (0, 3),$$
$$U_3 = (2, 5),$$
$$U_4 = (6, 7).$$

Here each U_k is an open interval denoting neuron k's receptive field. To illustrate why $\mathcal{S} = \{U_1, U_2, U_3, U_4\}$ generates $\mathcal{C}_1$, we select some stimuli and groups of neurons and then discuss whether the group of neurons can be a codeword:

- If $s = 1$, then s falls in the receptive fields of neurons 1 and 2. As a result, $\{1, 2\}$ is a codeword in the receptive field code determined by the four neurons in this example.
- If $s = 5.5$, then s is not in the receptive fields of any of the four neurons. Hence, $\emptyset$ is a codeword because none of the neurons become responsive when s is presented.

- The set of neurons {2, 3} does not appear as an activation pattern, so {2, 3} is not a codeword. If s is a stimulus presented to neurons 2 and 3 that falls in their receptive fields, then s must necessarily fall in the receptive field of neuron 1 as $U_2 \cap U_3 \subseteq U_1$.

The combinatorial neural code in example 2 is an instance of an open convex code. The dimension d of the stimulus space corresponds to the number of parameters needed to describe the stimulus; for example, a simple neuron in the primary visual cortex tuned to the orientation of edges in two dimensions would have $d = 1$ because orientation can be parameterized by an angle.

To make the definition of a receptive field (RF) code precise, we define the RF code generated by $\mathcal{S} = \{U_1, U_2, \ldots, U_N\}$, which is a collection of open subsets of $\mathbb{R}^d$, as (Curto et al., 2017):

$$\mathcal{C}(\mathcal{S}) = \left\{ \sigma \subseteq \{1, 2, \ldots, N\} : \bigcap_{i \in \sigma} U_i \backslash \bigcup_{j \notin \sigma} U_j \neq \emptyset \right\}.$$

(Here the symbol "\" is set subtraction; for example, $\{2, 3, 5, 9\} \backslash \{1, 2, 5\} = \{3, 9\}$.) If, in addition, every open subset in $\mathcal{S}$ is convex, then $\mathcal{C}(\mathcal{S})$ is referred to as a *convex RF code*. From this point of view, then, codewords correspond to ensembles of neurons σ that can be coactivated when a suitable stimulus falls in a part of the stimulus space that is covered by the receptive fields of the neurons in σ.

Example 3. For a concrete example of a nonconvex RF code,

$$\mathcal{C}_2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{2, 3\}, \{2, 4\}, \{1, 2, 3\}, \{1, 2, 4\}\}$$

is a combinatorial code that is not convex (due to Vladimir Itskov, personal communication). We provide an (uninstructive) argument in the appendix that shows that the receptive field of either neuron 1 or neuron 2 (i.e., U_1 or U_2) must be nonconvex.

There is a similar notion (Cruz, Giusti, Itskov, & Kronholm, 2019) involving closed subsets of $\mathbb{R}^N$. Observe that if $\mathcal{S}$ is a collection of receptive fields, then $\mathcal{C}(\mathcal{S})$ is a combinatorial neural code.

Although the above notion of an RF code was based on the usual sense of a neuron being active, we reformulate the code back in terms of a neuron being responsive so that we can investigate arbitrary activation functions Φ. We will regard a neuron as responsive when a stimulus falls within that neuron's receptive field, which will allow us to explore the question of how neural activity is shaped by the network structure. We investigate whether the collection of permitted sets generated by the network W, that is, $\mathcal{P}_\Phi(W)$, could be realized as an (open) convex RF code. In other words, given $\mathcal{P}_\Phi(W)$, we would like to know if $\mathcal{P}_\Phi(W)$ could be the result of the pattern of overlaps among the neurons' convex receptive fields.

4.2 Permitted Sets of Low-Rank Networks Form a Convex Code. We investigate conditions of the network connectivities W under which the collection of permitted sets $\mathcal{P}_\Phi(W)$ form a convex code. Given a network with dynamics described by equation 2.2, our goal is to uncover links between the network structure and the structure of receptive fields observed in response to a stimulus. Convex neural codes are of particular interest as receptive fields tend to be convex. Our main result in this letter, theorem 1, which is at the end of this section, is that if W is rank one and P is a suitable perturbation, then $\mathcal{P}_\Phi(W+P)$ is a convex code.

Our route for obtaining convex codes is through maximal-intersection complete codes. A combinatorial code is maximal-intersection complete if intersections of maximal codewords are also codewords (Curto et al., 2017), where a maximal codeword is one that is not a proper subset of any other codeword. Since Cruz et al. (2019) proved that maximal-intersection complete codes are open and closed convex, our goal is to show conditions under which a permitted set code is maximal-intersection complete.

We first prove in proposition 2 that the collection of permitted sets is maximal-intersection complete (and therefore convex) if the synaptic weight matrix is rank one. We then generalize this result in theorem 1 to matrices that are close to rank one. For the remainder of the letter, we denote the set $\{1, 2, \ldots, N\}$ by $[N]$.

4.2.1 Rank-One Synaptic Weight Matrices.

Proposition 2. *Assume that the network's dynamics are described by equation 2.2 and that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero vectors.*

Suppose that σ and μ are permitted sets. Further suppose that μ is a maximal codeword.

Then $\sigma \cap \mu$ is a permitted set. In particular, since $\mathcal{P}_\Phi(W)$ is maximum-intersection complete, it is a convex code.

Proof. Let $\mu, \sigma \in \mathcal{P}_\Phi(W)$, where μ is maximal. Let $x^* = (\Phi(I_1), \ldots, \Phi(I_N))$ and $y^* = (\Phi(L_1), \ldots, \Phi(L_N))$ be asymptotically stable fixed points associated with μ and σ, respectively.

We need an observation pertaining to neurons that are not part of μ. Since μ is a maximal codeword, it follows that for any $k \notin \mu$, $\widetilde{\mu} = \mu \cup \{k\}$ is not permitted. In other words, if $z^* = (\Phi(J_1), \ldots, \Phi(J_N))$ is a fixed point of the dynamics such that $|\Phi'(J_i)| > r_{\text{on}}$ for all $i \in \widetilde{\mu}$ and $|\Phi'(J_j)| \leq r_{\text{off}}$ for all $j \notin \widetilde{\mu}$, then $\text{tr}(\Lambda_{\widetilde{\mu}} W) \geq 1$, where $\Lambda_{\widetilde{\mu}} = \Lambda(z^*)$. We make a particular choice for J_i, setting $J_i = I_i$ for all $i \neq k$. However, the important point is that we let J_k be any value satisfying $|\Phi'(J_k)| > r_{\text{on}}$. Observing that $\text{tr}(\Lambda_\mu W) < 1$, where $\Lambda_\mu = \Lambda(x^*)$, and putting together the inequalities

$$0 > -1 + \mathrm{tr}(\Lambda_{\mu} W) = -1 + u_k v_k \Phi'(I_k) + \sum_{i \in [N]\setminus\{k\}} u_i v_i \Phi'(I_i), \text{ and}$$

$$-1 + \mathrm{tr}(\Lambda_{\widetilde{\mu}} W) = -1 + u_k v_k \Phi'(J_k) + \sum_{i \in [N]\setminus\{k\}} u_i v_i \Phi'(I_i) \geq 0,$$

we see that

$$u_k v_k \left(\Phi'(J_k) - \Phi'(I_k)\right) > 0. \tag{4.1}$$

We reiterate that the inequality 4.1 must hold for any $k \notin \mu$ and any $J_k \in \mathbb{R}$ satisfying $|\Phi'(J_k)| > r_{\text{on}}$.

Next, set $\tau = \mu \cap \sigma$. Assume that τ is a strict subset of σ. Further, assume that $\tau \neq \emptyset$ (since $\emptyset$ is always permitted).

We define $\widetilde{x}^*$ to be a fixed point associated with τ by starting with the fixed point y^* (the fixed point associated with σ), then making all neurons on $\sigma\setminus\tau$ be unresponsive by setting them equal to the values from x^* (the fixed point associated with μ). In order words, define $\widetilde{x}^* = (\Phi(\widetilde{I_1}), \ldots, \Phi(\widetilde{I_N}))$ to be such that $\widetilde{x}^*_i = \Phi(L_i)$ for any $i \in \tau \cup ([N]\setminus\sigma)$, and $\widetilde{x}^*_j = \Phi(I_j)$ for $j \in \sigma\setminus\tau$. By lemma 1, there is $b \in \mathbb{R}^N$ such that $\widetilde{x}^*$ is a fixed point associated with τ.

We will show that $\widetilde{x}^*$ is asymptotically stable:

$$\begin{aligned}
-1 + \mathrm{tr}(\Lambda_\tau W) &= -1 + \sum_{i\in\tau} u_i v_i \Phi'(L_i) + \sum_{j \in [N]\setminus\sigma} u_j v_j \Phi'(L_j) + \sum_{j\in\sigma\setminus\tau} u_j v_j \Phi'(I_j) \\
&= \left(-1 + \sum_{i\in\tau} u_i v_i \Phi'(L_i) + \sum_{j \in [N]\setminus\sigma} u_j v_j \Phi'(L_j) + \sum_{l\in\sigma\setminus\tau} u_l v_l \Phi'(L_l)\right) \\
&\quad + \sum_{l\in\sigma\setminus\tau} u_l v_l \left(\Phi'(I_l) - \Phi'(L_l)\right).
\end{aligned}$$

Recalling that $y^* = (\Phi(L_1), \ldots, \Phi(L_N))$ is an asymptotically stable fixed point associated with σ, observe that

$$\sum_{l\in\sigma\setminus\tau} u_l v_l \left(\Phi'(I_l) - \Phi'(L_l)\right) < 0$$

by inequality 4.1, given that $|\Phi'(L_l)| > r_{\text{on}}$. Moreover,

$$-1 + \sum_{i\in\tau} u_i v_i \Phi'(L_i) + \sum_{j \in [N]\setminus\sigma} u_j v_j \Phi'(L_j) + \sum_{l\in\sigma\setminus\tau} u_l v_l \Phi'(L_l) = \mathrm{tr}(\Lambda(y^*)W) - 1 < 0$$

by the assumption that σ is permitted. Therefore, $-1 + \mathrm{tr}(W_\tau W) < 0$ so that τ is in $\mathcal{P}_\Phi(W)$. We have shown that $\mu \cap \sigma$ is a permitted set and $P_\Phi(W)$ is maximum-intersection complete. □

4.2.2 Perturbation of Rank-One Synaptic Weight Matrices. Next, we tackle a perturbed version of proposition 2. The notation $|| \cdot ||_{\max}$ is the max matrix norm (Horn & Johnson, 2012); that is, if $A = (a_{ij})$ is an $N \times N$ matrix, then

$$||A||_{\max} = \max_{1 \le i, j \le N} |a_{ij}|.$$

In neurobiological terms, theorem 1 has almost the same interpretation as proposition 2, except that we relax the requirement that the synaptic weight matrix be low-dimensional. It turns out that the conclusion of proposition 1 still holds if the synaptic weight matrix is required to be sufficiently close to a rank-one matrix.

Theorem 1. *Assume that the network's dynamics are described by equation 2.2 and that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero vectors. Suppose that $u_j v_j \neq 0$ for all $j \in \{1, 2, \ldots, N\}$. Let P be an $N \times N$ matrix.*

There exists $d_P > 0$ depending on P such that if it turns out that $||P||_{\max} < d_P$, then the following it true: for any $\mu, \sigma \in \mathcal{P}_\Phi(W + P)$, where μ is maximal, we have that $\mu \cap \sigma \in \mathcal{P}_\Phi(W + P)$. In particular, since $\mathcal{P}_\Phi(W)$ is maximum-intersection complete, it is a convex code.

Proof. See the appendix. □

5 Discussion

In Hahnloser et al. (2003), the notion of permitted sets of $D\dot{x} + x = [Wx + b]_+$ was introduced and analyzed. In that setting, the nonzero entries of an asymptotically stable fixed point correspond to active neurons (and zero entries correspond to neurons that are not active). If the rectifier $[\cdot]_+$ is replaced by an activation function Φ that is nonnegative continuous and piecewise C^1 such that $\Phi'(s) \to 0$ as $s \to -\infty$, then it is not generally possible to dichotomize the coordinates of asymptotically stable fixed points of $D\dot{x} + x = \Phi(Wx + b)$ into active and not active neurons.

In order to address the above gap, we define permitted sets in a more general setting by changing the focus from neural activity to a notion of responsiveness that is based on the gain Φ' of each neuron. By introducing user-chosen responsiveness thresholds r_{off} and r_{on}, we categorize neurons as responsive, marginally responsive, or unresponsive. In our framework, permitted sets are groups of neurons that can be made stably coresponsive (while all other neurons are unresponsive). A user can choose the responsiveness thresholds to create a definition of responsiveness that reflects the nature of the neurons under consideration. A stricter definition for a set being considered coresponsive can be made by creating a gap between r_{off} and r_{on}. Alternatively, one can set $r_{\text{on}} = r_{\text{off}}$ so that every neuron is categorized as responsive or unresponsive (eliminating the marginally responsive category). Although the framework will yield results for any choice of r_{off} and r_{on}, the resulting combinatorial code will be nontrivial only if r_{on} is chosen large enough so that some groups of neurons cannot be made stably coresponsive, that is, so that some sets are forbidden and are not in the combinatorial code.

In the special case that $r_{\text{on}} = r_{\text{off}} = 0$, $\Phi(x) = \max(x, 0)$ and W is symmetric, our framework for permitted sets recovers the TLN version of permitted sets introduced in Hahnloser et al. (2003). With this choice of parameters, our notion of a responsive neuron matches their definition of an active neuron having $\Phi(x) > 0$. In their framework, a nonempty subset σ of $\{1, 2, \ldots, N\}$ is permitted depending on the eigenvalues of the principal submatrix of $W - I$ built from removing the rows and columns not indexed by σ. When applying our framework to the TLN, we obtain the eigenvalues of that principal submatrix (plus additional eigenvalues of -1) from the Jacobian matrix $\Lambda W - I$, where for the TLN, the effective gain matrix Λ (defined in equation 3.1) becomes a diagonal matrix with $\Lambda_{ii} = 1$ if $i \in \sigma$ and $\Lambda_{ii} = 0$ otherwise.

Finally, when W is almost a rank-one synaptic weight matrix, we proved in theorem 1 that $\mathcal{P}_\Phi(W)$ is a convex code. Low-rank synaptic weight matrices appear in multiple other models and lead to models that display desirable computational properties. For example, in the theory of flexible memory networks (Curto et al., 2012), rank-one matrices appear in the decomposition of synaptic weight matrices of networks that are maximally flexible. In another context where again the synaptic weight matrix is thought of as the sum of a low-rank matrix and a perturbation, it is shown that a low-rank connectivity matrix leads to low-dimensional dynamics in a class of models the authors call low-rank recurrent networks (Mastrogiuseppe & Ostojic, 2018).

We highlight a few limits of the framework we presented. First, finding the permitted sets $\mathcal{P}_\Phi(W)$ of $D\dot{x} + x = \Phi(Wx + b)$ is challenging: assessing the stability of fixed points is activity-dependent, a shortcoming that permitted sets in the TLN regime do not suffer. Although one can prove general structural results about $\mathcal{P}_\Phi(W)$, as in theorem 1, explicitly calculating the members of $\mathcal{P}_\Phi(W)$ can be computationally expensive. Second, although proposition 1 gives some guidance about choosing responsiveness thresholds, the converse of the proposition is not true. Even if one chooses $r_{\text{on}} > 1/|||W|||$, there is no guarantee that $\mathcal{P}_\Phi(W)$ will be nontrivial. As we have not found general criteria for choosing the thresholds, it is likely that the user will need to make use of additional information about the network in order to determine reasonable responsiveness thresholds.

The theory of permitted sets was introduced in the early 2000s in the context of threshold-linear networks. Researchers have found necessary and sufficient conditions under which symmetric TLNs converge to a set of attractive fixed points and under which the network is multiattractive (Hahnloser et al., 2003). Furthermore, encoding neural codes in TLNs has been studied (Curto et al., 2013). Changing the basis of permitted sets from active to responsive neurons allowed us to generalize this framework to a large class of firing-rate neuron models. More work is needed to determine how, for example, multiattractiveness or neural encoding rules can be generalized to our new framework involving responsiveness thresholds in general firing rate models.

Appendix

A.1 Proof of Lemmas 1 and 2. We prove the two lemmas presented in section 3.

Proof of Lemma 1. Let $b = I^* - Wx^*$, where

$$I^* \in \Phi^{-1}(x_1^*) \times \Phi^{-1}(x_2^*) \times \cdots \times \Phi^{-1}(x_N^*).$$

Then

$$\begin{aligned}\Phi(Wx^* + b) &= \Phi(Wx^* + I^* - Wx^*)\\ &= (\Phi(I_1^*), \Phi(I_2^*), \ldots, \Phi(I_N^*))\\ &= (x_1^*, x_2^*, \ldots, x_N^*)\\ &= x^*.\end{aligned}$$

Hence, if $b = I^* - Wx^*$, then $\Phi(Wx^* + b) = x^*$, which shows that x^* is a fixed point of $D\dot{x} + x = \Phi(Wx + b)$. □

Proof of Lemma 2. Recall that x^* is an asymptotically stable fixed point of the dynamics if and only if $JF(x^*) = \Lambda(x^*)W - I$ is stable, where $\Lambda(x^*)$ is the diagonal matrix defined in equation 3.1. The eigenvalues of $JF(x^*)$ are of the form $\lambda - 1$, where λ is an eigenvalue of $\Lambda(x^*)W$. Since $\Lambda(x^*)W$ is rank one, the eigenvalues of $\Lambda(x^*)W$ will be 0 and $\mathrm{tr}(\Lambda(x^*)W)$; therefore, the eigenvalues of $\Lambda(x^*)W - I$ will be -1 and $\mathrm{tr}(\Lambda(x^*)W) - 1$. Asymptotic stability of x^* corresponds to $\mathrm{tr}(\Lambda(x^*)W) - 1 < 0$.

By lemma 1, there exists an I^* in $\Phi^{-1}(x_1^*) \times \cdots \times \Phi^{-1}(x_N^*)$ such that $x^* = \Phi(Wx^* + b)$, where $b = I^* - Wx^*$. In particular, for all $i \in \{1, 2, \ldots, N\}$,

$$\begin{aligned}\Lambda(x^*)_{ii} &= \Phi'(W^{(i)} \cdot x^* + b_i)\\ &= \Phi'(W^{(i)} \cdot x^* + I_i^* - W^{(i)} \cdot x^*)\\ &= \Phi'(I_i^*).\end{aligned}$$

(Recall that $W^{(i)} \cdot x^*$ denotes the dot product between the ith column of W and x^*.) Therefore, we can rewrite $\mathrm{tr}(\Lambda(x^*)W) - 1 < 0$ as

$$\mathrm{tr}(\Lambda(x^*)W) = \sum_{i=1}^{N} u_i v_i \Phi'(I_i^*) < 1. \qquad \square$$

A.2 Nonconvexity of RF Code in Example 3. We will show that

$$\mathcal{C}_2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{2,3\}, \{2,4\}, \{1,2,3\}, \{1,2,4\}\}$$

is a not a convex code: since $\{1,2\} \notin \mathcal{C}_2$, it follows that $(U_1 \cap U_2) \backslash (U_3 \cup U_4) = \emptyset$. (See the discussion in section 4.1 preceding example 3 for the definition of RF code.) This implies that $U_1 \cap U_2 \subseteq U_3 \cup U_4$. Hence,

$$U_1 \cap U_2 = (U_1 \cap U_2 \cap U_3) \cup (U_1 \cap U_2 \cap U_4).$$

Notice that $U_1 \cap U_2 \neq \emptyset$ because $U_1 \cap U_2 \cap U_3 \neq \emptyset$ (since $\{1,2,3\} \in \mathcal{C}_2$). Furthermore, $U_1 \cap U_2 \cap U_3$ and $U_1 \cap U_2 \cap U_4$ are disjoint because $\{1,2,3,4\} \notin \mathcal{C}_2$. Hence, $U_1 \cap U_2$ is nonconvex as we have shown that $U_1 \cap U_2$ is disconnected, that is, it is the union of two disjoint nonempty open sets. This implies that either U_1 or U_2 is nonconvex because the intersection of two convex sets is a convex set.

A.3 Proof of Theorem 1. Here we provide a full proof of theorem 1, which is the perturbed version of proposition 2. First we review what it means for the spectra of two matrices to be close given that the corresponding matrices are close. The space of $N \times N$ matrices with complex entries, denoted $\mathbb{C}^{N\times N}$, is naturally topologized by the maximum norm (Horn & Johnson, 2012), or "max-norm" for short, on $\mathbb{C}^{N^2}$, that is, for any $A = (a_{ij}) \in \mathbb{C}^{N\times N}$,

$$||A||_{\max} = \max_{i,j} |a_{ij}|.$$

Let $\mathfrak{S}_N$ denote the symmetric group on N elements and $\mathcal{A}_N$ the quotient space $\mathbb{C}^N/\mathfrak{S}_N$. That is, if $z, w \in \mathcal{A}_N$, then z and w are equivalent if and only if there exists a permutation π in $\mathfrak{S}_N$ such that

$$(z_1, z_2, \ldots, z_N) = (w_{\pi(1)}, w_{\pi(2)}, \ldots, w_{\pi(N)}).$$

With such a notion of equivalence for any two elements in $\mathcal{A}_N$, it turns out that $\mathcal{A}_N$ can be topologized (Serre, 2020) via the metric

$$d(w, z) = \min_{\pi \in \mathfrak{S}_N} \max_{1 \le j \le N} |w_j - z_{\pi(j)}|.$$

This metric turns $\mathcal{A}_N$ into a complete metric space. Further, if we denote the set of eigenvalues of an $N \times N$ matrix A by Spec(A), the map Spec : $\mathbb{C}^{N\times N} \to \mathcal{A}_N$ defined by $M \mapsto$ Spec(M) is continuous (Serre, 2020).

Now we have the notions for lemma 3 that we will use to show theorem 1. As usual, we denote $\{1, 2, \dots, N\}$ by $[N]$.

Lemma 3. *Let $A \in \mathbb{C}^{N\times N}$ and $\varepsilon > 0$ be given. Suppose $d > 0$ is such that for any $X \in \mathbb{C}^{N\times N}$ satisfying $||A - X||_{\max} < d$, we have*

$$\min_{\pi\in\mathfrak{S}_N} \max_{1\le j\le N} \left|\lambda_j(A) - \lambda_{\pi(j)}(X)\right| < \varepsilon$$

(where $\lambda_n(A)$ and $\lambda_m(X)$ denote an nth eigenvalue of A and an mth eigenvalue of X, respectively). If $X_0 \in \mathbb{C}^{N\times N}$ is such that $||A - X_0||_{\max} < d$, then

$$\mathit{Spec}(X_0) = \{\lambda_j + \Delta_j : j \in [N]\,; \lambda_j \in \mathit{Spec}(A)\,; |\Delta_j| < \varepsilon\}.$$

Proof. By continuity of Spec : $\mathbb{C}^{N\times N} \to \mathcal{A}_N$ at A, there is $d > 0$ such that for any $X \in \mathbb{C}^{N\times N}$ satisfying $||A - X||_{\max} < d$,

$$\min_{\pi\in\mathfrak{S}_N} \max_{1\le j\le N} \left|\lambda_j(A) - \lambda_{\pi(j)}(X)\right| < \varepsilon.$$

Let $X_0 \in \mathbb{C}^{N\times N}$ be such that $||A - X_0||_{\max} < d$.

We show next that Spec(X_0) has the form in the conclusion of the lemma. Let $\pi_0 \in \mathfrak{S}_N$ be such that

$$\min_{\pi\in\mathfrak{S}_N} \max_{1\le j\le N} \left|\lambda_j(A) - \lambda_{\pi(j)}(X_0)\right| = \max_{1\le j\le N} \left|\lambda_j(A) - \lambda_{\pi_0(j)}(X_0)\right|.$$

Define $\Delta_j = \lambda_{\pi_0(j)}(X_0) - \lambda_j$, where $\lambda_j \in$ Spec(A), for all $j \in [N]$, so we have

$$\lambda_{\pi_0(j)}(X_0) = \lambda_j + \Delta_j.$$

Note that by definition $|\Delta_j| < \varepsilon$ for all $j \in [N]$. Thus,

$$\text{Spec}(X_0) = \{\lambda_j + \Delta_j : j \in [N]\,; \lambda_j \in \text{Spec}(A)\,; |\Delta_j| < \varepsilon\}. \qquad \square$$

Over the course of the proof of the next lemma, we use the following notation: if ω is a subset of $\{1, 2, \dots, N\}$ that is a permitted set, then

$$x^*_{\text{stable}}(\omega) = (\Phi(I_1^\omega), \Phi(I_2^\omega), \dots, \Phi(I_N^\omega))$$

will be used to denote an asymptotically stable fixed point supporting ω, that is, $I_k^\omega \in \mathbb{R}$, $|\Phi'(I_k^\omega)| > r_{\text{on}}$ when $k \in \omega$, and $|\Phi'(I_j^\omega)| \leq r_{\text{off}}$ when $j \notin \omega$. We will use the notation

$$x^*(\omega) = (\Phi(I_1^\omega), \Phi(I_2^\omega), \ldots, \Phi(I_N^\omega))$$

to refer to a fixed point supporting ω that may not be asymptotically stable. We will also write $[N]_{<0}$ to denote the set of neurons for which there is an input that makes them responsive and to have negative effective self-coupling strength–specifically,

$$[N]_{<0} = \left\{k \in \{1, 2, \ldots, N\} : \text{ there is } \alpha \in \mathbb{R} \text{ such that } u_k v_k \Phi'(\alpha) < -r_{\text{on}}\right\}.$$

Remember that we use the term "effective self-coupling strength" for the expression $u_k v_k \Phi'(\alpha)$. Finally, we use the notation $d(\mathcal{S}, c)$, where $\mathcal{S}$ is a subset of $[N]$ and c is a positive constant, to denote a positive number satisfying the following property: for any $X \in \mathbb{C}^{N\times N}$ satisfying

$$||\Lambda(x^*(\mathcal{S}))W - I - X||_{\max} < d(\mathcal{S}, c), \tag{A.1}$$

we have

$$\min_{\sigma\in\mathfrak{S}_N} \max_{1\leq j\leq N} |\lambda_j(\Lambda(x^*(\mathcal{S}))W - I) - \lambda_{\pi(j)}(X)| < c.$$

(If it is known that $\mathcal{S}$ is a permitted set, then we would denote the associated asymptotically stable fixed point as $x^*_{\text{stable}}(\mathcal{S})$ instead of $x^*(\mathcal{S})$.) Here $d(\mathcal{S}, c)$ is associated with invoking the continuity of Spec : $\mathbb{C}^{N\times N} \to \mathcal{A}_N$.

The next lemma, lemma 4, shows that if $[N]_{<0} \neq \emptyset$ and there is a permitted set σ missing a neuron n in $[N]_{<0}$, then there is a permitted set consisting of σ and n.

Lemma 4. *Assume that the network's dynamics are described by equation 2.2 and that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero. Let P be an $N \times N$ matrix.*

There exists a $d_P > 0$ depending on P such that if it turns out that $||P||_{\max} < d_P$, then the following is true: if $\sigma \in \mathcal{P}_\Phi(W + P)$ and $n \in [N]_{<0}$ with $n \notin \sigma$, then $\sigma \cup \{n\}$ will be supported by a $x^*_{stable}(\sigma \cup \{n\})$ such that $u_n v_n \Phi'(I_n^{\sigma\cup\{n\}}) < 0$. (In particular, $\sigma \cup \{n\} \in \mathcal{P}_\Phi(W + P)$.)

Proof. First, we introduce two numbers, ε_1 and d_P, to quantify how much we can perturb W. In order to define ε_1, we consider the smallest negative effective response gap: Given a neuron that (1) is unresponsive in some permitted set and (2) can have an effective self-coupling strength that is negative when it is responsive, we take the difference between the strengths under those conditions. We use ε_1 to put a bound on the magnitude of the contributions of the perturbation to the spectrum of the rank-one network.

For every $n \in [N]_{<0}$, let $\alpha_n \in \mathcal{A}$ be such that $u_n v_n \Phi'(\alpha_n) < 0$. Let $\varepsilon_1 > 0$ be such that

$$\varepsilon_1 < \min\left(1, \min_{\sigma \in \mathcal{P}_\Phi(W+P)} \min_{\substack{n \notin \sigma \\ n \in [N]_{<0}}} -\frac{u_n v_n}{2}\left(\Phi'(\alpha_n) - \Phi'(I_n^\sigma)\right)\right), \tag{A.2}$$

where, as outlined above, we denoted the stable fixed point supporting the permitted set σ as $x^*_{\text{stable}}(\sigma) = (\Phi(I_1^\sigma), \ldots, \Phi(I_N^\sigma))$. Next, we define d_P. The role of d_P is to place a bound on how large the entries of P are allowed to be. For all $n \in [N]_{<0}$ such that $n \notin \sigma$, define

$$x^*(\omega_n) = (\Phi(I_1^{\omega_n}), \ldots, \Phi(I_N^{\omega_n})),$$

where $\omega_n = \sigma \cup \{n\}$, $I_k^{\omega_n} = I_k^\sigma$ for $k \neq n$, and $I_n^{\omega_n} = \alpha_n$.

Let

$$\delta_1 = \min\left(\min_{\sigma \in \mathcal{P}_\Phi(W+P)} d\left(\sigma, \varepsilon_1\right), \min_{\sigma \in \mathcal{P}_\Phi(W+P)} \min_{\substack{n \notin \sigma \\ n \in [N]_{<0}}} d\left(\sigma \cup \{n\}, \varepsilon_1\right)\right) \tag{A.3}$$

(see the definition of $d(\mathcal{S}, c)$ in equation A.1) and

$$\widetilde{I} = \underset{\substack{\sigma \in \mathcal{P}_\Phi(W+P) \\ 1 \leq k \leq N}}{\arg\max} |\Phi'(I_k^\sigma)|.$$

Define $I_{\text{on}} = \underset{x \in \{\widetilde{I}, \alpha_{i_1}, \ldots, \alpha_{i_m}\}}{\arg\max} |\Phi'(x)|$ and

$$M = |\Phi'(I_{\text{on}})|. \tag{A.4}$$

Let $\Lambda(x^*_{\text{stable}}(\sigma)) = \Lambda_\sigma$ and $\Lambda(x^*(\omega_n)) = \Lambda_{\omega_n}$. Observe that $||\Lambda_\sigma||_{\max}$, $||\Lambda_{\omega_n}||_{\max} \leq M$: by the definition of effective gain of a network at x (see equation 3.1), if we are given

$$x^* = (\Phi(I_1), \ldots, \Phi(I_N)),$$

then $b = \mathcal{I} - Wx^*$, where $\mathcal{I} = (I_1, \ldots, I_N)$, is such that x^* is a fixed point. Hence, for bounding $||\Lambda_\sigma||_{\max}$ and $||\Lambda_{\omega_n}||_{\max}$, we have

$$\begin{aligned} |(\Lambda_\sigma)_{ii}| &= \left|\Phi'\left(W^{(i)} \cdot x^*_{\text{stable}}(\sigma) + b_i\right)\right| \\ &= |\Phi'(I_i^\sigma)| \leq M \end{aligned}$$

and similarly for $|(\Lambda_{\omega_n})_{ii}|$.

Let

$$d_P = \frac{\delta_1}{M},$$

and suppose that $||P||_{\max} < d_P$.

Letting ω_n and $x^*(\omega_n)$ be as defined before, we know that $x^*(\omega_n)$ is a fixed point of the dynamics supporting ω_n. As a result, it remains to prove that $x^*(\omega_n)$ is asymptotically stable.

Although the matrix max norm is not submultiplicative, it is easy to see that if A is an $N \times N$ matrix and D is a diagonal $N \times N$ matrix, then

$$||DA||_{\max} \leq ||D||_{\max}||A||_{\max};$$

hence,

$$\begin{aligned} ||\Lambda_\sigma W - I - (\Lambda_\sigma(W+P) - I)\,||_{\max} &= ||\Lambda_\sigma P||_{\max} \\ &\leq ||\Lambda_\sigma||_{\max}||P||_{\max} \\ &\leq M||P||_{\max} \\ &< \delta_1 \end{aligned}$$

and

$$\begin{aligned} ||\Lambda_{\omega_n} W - I - (\Lambda_{\omega_n}(W+P) - I)\,||_{\max} &= ||\Lambda_{\omega_n} P||_{\max} \\ &\leq ||\Lambda_{\omega_n}||_{\max}||P||_{\max} \\ &\leq M||P||_{\max} \\ &< \delta_1, \end{aligned}$$

so it follows by lemma 3 that

$$\text{Spec}(\Lambda_\sigma(W+P) - I) = \{\text{tr}(\Lambda_\sigma W) - 1 + \Delta_1^\sigma, \Delta_2^\sigma - 1, \ldots, \Delta_N^\sigma - 1\} \text{ and}$$
$$\text{Spec}(\Lambda_{\omega_n}(W+P) - I) = \{\text{tr}(\Lambda_{\omega_n} W) - 1 + \Delta_1^{\omega_n}, \Delta_2^{\omega_n} - 1, \ldots, \Delta_N^{\omega_n} - 1\},$$

where

$$|\Delta_k^\sigma|, |\Delta_k^{\omega_n}| < \varepsilon_1 \tag{A.5}$$

for every $k \in \{1, 2, \ldots, N\}$.

Note that $\text{Re}(\Delta_k^\sigma) - 1 < 0$ and $\text{Re}(\Delta_k^{\omega_n}) - 1 < 0$ for all $k \in \{2, 3, \ldots, N\}$ because $\varepsilon_1 < 1$ by equations A.2 and A.5. Thus, showing $\text{tr}(\Lambda_{\omega_n} W) - 1 +$

$\mathrm{Re}(\Delta_1^{\omega_n}) < 0$ will suffice to prove that $x^*(\omega_n)$ is asymptotically stable:

$$\begin{aligned}
\mathrm{tr}(\Lambda_{\omega_n} W) - 1 + \mathrm{Re}(\Delta_1^{\omega_n}) &= -1 + \mathrm{Re}(\Delta_1^{\omega_n}) + \sum_{i\in\omega_n} u_i v_i \Phi'(I_i^{\omega_n}) + \sum_{j\notin\omega_n} u_j v_j \Phi'(I_j^{\omega_n}) \\
&= -1 + \mathrm{Re}(\Delta_1^{\omega_n} - \Delta_1^{\sigma}) + \mathrm{Re}(\Delta_1^{\sigma}) \\
&\quad + u_n v_n \Phi'(\alpha_n) - u_n v_n \Phi'(I_n^{\sigma}) + \sum_{i\in\sigma} u_i v_i \Phi'(I_i^{\sigma}) \\
&\quad + \sum_{j\notin\sigma} u_j v_j \Phi'(I_j^{\sigma}).
\end{aligned}$$

Let $\eta = \Delta_1^{\omega_n} - \Delta_1^{\sigma}$. Since $x^*_{\text{stable}}(\sigma)$ is an asymptotically stable fixed point supporting σ, it follows that

$$\mathrm{tr}(\Lambda_{\sigma} W) - 1 + \mathrm{Re}(\Delta_1^{\sigma}) = -1 + \mathrm{Re}(\Delta_1^{\sigma}) + \sum_{i\in\sigma} u_i v_i \Phi'(I_i^{\sigma}) + \sum_{j\notin\sigma} u_j v_j \Phi'(I_j^{\sigma}) < 0.$$

As for the remaining terms, first recall that

$$-\varepsilon_1 < \mathrm{Re}(\Delta_1^{\omega_n}), \mathrm{Re}(\Delta_1^{\sigma}) < \varepsilon_1$$

by equation A.5, so

$$\mathrm{Re}(\eta) = \mathrm{Re}(\Delta_1^{\omega_n}) - \mathrm{Re}(\Delta_1^{\sigma}) < \varepsilon_1 - (-\varepsilon_1) = 2\varepsilon_1.$$

Thus,

$$\begin{aligned}
\mathrm{Re}(\eta) + u_n v_n \Phi'(\alpha_n) - u_n v_n \Phi'(I_n^{\sigma}) &< 2\varepsilon_1 + u_n v_n \left(\Phi'(\alpha_n) - \Phi'(I_n^{\sigma})\right) \\
&< -u_m v_m \left(\Phi'(\alpha_m) - \Phi'(I_m^{\mu})\right) \\
&\quad + u_n v_n \left(\Phi'(\alpha_n) - \Phi'(I_n^{\sigma})\right) \\
&\le 0,
\end{aligned}$$

where $\mu \in \mathcal{P}_\Phi(W+P)$ and $m \notin \mu$ and $m \in [N]_{<0}$ are such that

$$\varepsilon_1 < -\frac{u_m v_m}{2}\left(\Phi'(\alpha_m) - \Phi'(I_m^{\mu})\right) \le -\frac{u_p v_p}{2}\left(\Phi'(\alpha_p) - \Phi'(I_p^{\tau})\right)$$

for every $\tau \in \mathcal{P}_\Phi(W+P)$, $p \notin \tau$ and $p \in [N]_{<0}$.

Hence, $\mathrm{tr}(\Lambda_{\omega_n} W) - 1 + \mathrm{Re}(\Delta_1^{\omega_n}) < 0$, so $\omega_n = \sigma \cup \{n\}$ is permitted. □

Now we show that if σ is a permitted set and $k \in \sigma \cap [N]_{<0}$, then there exists an asymptotically stable fixed point supporting σ such that neuron k has a negative effective self-coupling strength. As a consequence of lemma 5, we will be able to say that if $\sigma \in \mathcal{P}_\Phi(W+P)$, where W is a rank-one matrix and P is a suitable perturbation, then there is an asymptotically stable fixed point $x^*_{\text{stable}}(\sigma)$ supporting σ such that $u_k v_k \Phi'(I_k^\sigma) < 0$ for every $k \in \sigma \cap [N]_{<0}$:

Lemma 5. *Assume that the network's dynamics are described by equation 2.2 and that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero. Let P be an $N \times N$ matrix.*

There exists a $d_P > 0$ depending on P such that if it turns out that $||P||_{\max} < d_P$, then the following is true: if $\sigma \in \mathcal{P}_\Phi(W+P)$, $p \in \sigma \cap [N]_{<0}$, and

$$x^*_{stable}(\sigma) = (\Phi(I_1^\sigma), \ldots, \Phi(I_N^\sigma))$$

is an asymptotically stable fixed point supporting σ such that $u_p v_p \Phi'(I_p^\sigma) > 0$, then there exists an asymptotically stable fixed point

$$\widetilde{x}^*_{stable}(\sigma) = (\Phi(\widetilde{I}_1^\sigma), \ldots, \Phi(\widetilde{I}_N^\sigma))$$

supporting σ such that $u_p v_p \Phi'(\widetilde{I}_p^\sigma) < 0$.

Proof. Let $\alpha_k \in \mathcal{A}$ be such that $u_k v_k \Phi'(\alpha_k) < 0$ for every $k \in [N]_{<0}$. Let $\varepsilon_2 > 0$ be such that

$$\varepsilon_2 < \min\left(1, \min_{\sigma \in \mathcal{P}_\Phi(W+P)} \min_{\substack{k \in \sigma \cap [N]_{<0} \\ u_k v_k \Phi'(I_k^\sigma) > 0}} -\frac{u_k v_k}{2}\left(\Phi'(\alpha_k) - \Phi'(I_k^\sigma)\right)\right). \tag{A.6}$$

For every $n \in \sigma$, where $\sigma \in \mathcal{P}_\Phi(W+P)$, satisfying $u_n v_n \Phi'(I_n^\sigma) > 0$, define

$$\widetilde{x}^*(\sigma) = (\Phi(\widetilde{I}_1^\sigma), \ldots, \Phi(\widetilde{I}_N^\sigma)),$$

where $\widetilde{I}_k^\sigma = I_k^\sigma$ for $k \neq n$ and $\widetilde{I}_n^\sigma = \alpha_n$. Now define $\widetilde{d}_\sigma > 0$ such that for any $X \in \mathbb{C}^{N \times N}$ satisfying $||\Lambda(\widetilde{x}^*(\sigma))W - I - X||_{\max} < \widetilde{d}_\sigma$,

$$\min_{\sigma \in \mathfrak{S}_N} \max_{1 \le j \le N} |\lambda_j(\Lambda(\widetilde{x}^*(\sigma))W - I) - \lambda_{\pi(j)}(X)| < \varepsilon_2.$$

Let

$$\delta_2 = \min\left(\delta_1, \min_{\sigma \in \mathcal{P}_\Phi(W+P)} \min_{\substack{n \in \sigma \cap [N]_{<0} \\ u_n v_n \Phi'(I_n^\sigma) > 0}} \widetilde{d}_\sigma\right). \tag{A.7}$$

Let $d_P = \delta_2/M$, where M is as defined in equation A.4. Suppose that $||P||_{\max} < d_P$.

Suppose $\sigma \in \mathcal{P}_\Phi(W+P)$ and $n \in \sigma \cap [N]_{<0}$. We know that $\widetilde{x}^*(\sigma)$ is a fixed point of the dynamics supporting σ; it remains to prove that $\widetilde{x}^*(\sigma)$ is asymptotically stable.

Letting $\Lambda_\sigma = \Lambda(x^*_{\text{stable}}(\sigma))$ and $\widetilde{\Lambda}_\sigma = \Lambda(\widetilde{x}^*(\sigma))$, we observe that

$$\operatorname{Spec}(\Lambda_\sigma(W+P) - I) = \{\operatorname{tr}(\Lambda_\sigma W) - 1 + \Delta_1^\sigma, \Delta_2^\sigma - 1, \ldots, \Delta_N^\sigma - 1\}$$

and

$$\operatorname{Spec}(\widetilde{\Lambda}_\sigma(W+P) - I) = \{\operatorname{tr}(\widetilde{\Lambda}_\sigma W) - 1 + \widetilde{\Delta}_1^\sigma, \widetilde{\Delta}_2^\sigma - 1, \ldots, \widetilde{\Delta}_N^\sigma - 1\},$$

where $|\Delta_k^\sigma|, |\widetilde{\Delta}_k^\sigma| < \varepsilon_2$ for every $k \in \{1, 2, \ldots, N\}$, by lemma 3.

Next, we show that $\operatorname{tr}(\widetilde{\Lambda}_\sigma W) - 1 + \operatorname{Re}(\widetilde{\Delta}_1^\sigma) < 0$:

$$\begin{aligned}\operatorname{tr}(\widetilde{\Lambda}_\sigma W) - 1 + \operatorname{Re}(\widetilde{\Delta}_1^\sigma) = -1 &+ \operatorname{Re}(\widetilde{\Delta}_1^\sigma - \Delta_1^\sigma) + \operatorname{Re}(\Delta_1^\sigma) \\ &+ u_n v_n \Phi'(\widetilde{I}_n^\sigma) - u_n v_n \Phi'(I_n^\sigma) + \sum_{i \in \sigma} u_i v_i \Phi'(I_i^\sigma) \\ &+ \sum_{j \notin \sigma} u_j v_j \Phi'(I_j^\sigma).\end{aligned}$$

Since $x^*_{\text{stable}}(\sigma)$ is asymptotically stable, $\operatorname{tr}(\Lambda_\sigma W) - 1 + \operatorname{Re}(\Delta_1^\sigma) < 0$. Further, $n \in \sigma$ is such that $\widetilde{I}_n^\sigma = \alpha_n$ and $u_n v_n \Phi'(\alpha_n) < 0$, so

$$\operatorname{Re}(\eta) + u_n v_n \Phi'(\widetilde{I}_n^\sigma) - u_n v_n \Phi'(I_n^\sigma) < 2\varepsilon_2 + u_n v_n \left(\Phi'(\alpha_n) - \Phi'(I_n^\sigma)\right) \leq 0,$$

where $\eta = \widetilde{\Delta}_1^\sigma - \Delta_1^\sigma$. □

Next, we need an auxiliary result that tells us which neurons are recruited by maximal permitted sets of almost rank-one networks. Lemma 6 shows that if an ensemble of neurons is permitted and it is maximal (with respect to set containment), then every neuron that can be made responsive by using an input that makes its effective self-coupling strength negative must be part of such a maximal permitted set. In the unperturbed, rank-one synaptic weight matrices, such a statement is immediate: stability of the dynamics is essentially described by the trace of the effective gain of the network, so if an ensemble of neurons is permitted and maximal, then any neuron in the network that could have a negative effective self-coupling strength will preserve the stability of the ensemble.

Lemma 6. *Assume that the network's dynamics are described by equation 2.2 and that $W = uv^T$, where $u, v \in \mathbb{R}^N$ are nonzero. Let P be an $N \times N$ matrix.*

There is $d_P > 0$ depending on P such that if it turns out that $||P||_{\max} < d_P$, then the following is true: each maximal permitted set contains every neuron whose effective self-coupling strength can be negative—that is, $[N]_{<0} \subseteq \mu$ for every maximal $\mu \in \mathcal{P}_\Phi(W + P)$.

Proof. Define $\varepsilon > 0$ to be such that

$$\varepsilon < \min\left(1, C_1, C_2, \varepsilon_1, \varepsilon_2\right), \tag{A.8}$$

where

$$C_1 = \min_{\sigma,\omega\in\mathcal{P}_\Phi(W+P)} \min_{\substack{k\in\sigma\setminus\omega\\ k\in[N]_{<0}}} -\frac{u_k v_k}{2}\left(\Phi'(I_k^\sigma) - \Phi'(I_k^\omega)\right) \tag{A.9}$$

and

$$C_2 = \min_{\sigma,\omega\in\mathcal{P}_\Phi(W+P)} \min_{\substack{l\in\sigma\setminus\omega\\ l\notin[N]_{<0}\\ u_l v_l\neq 0}} \frac{u_l v_l}{2}\left(\Phi'(I_l^\sigma) - \Phi'(I_l^\omega)\right). \tag{A.10}$$

Here ε_1 and ε_2 are as defined in equations A.2 and A.6, respectively. We remark that by lemma 5, we choose $x^*_{\text{stable}}(\sigma)$, where σ in $\mathcal{P}_\Phi(W+P)$, such that $u_k v_k \Phi'(I_k^\sigma) < 0$ for all $k \in \sigma \cap [N]_{<0}$.

If $s \in [N]_{<0}$ such that $s \notin \sigma$, where $\sigma \in \mathcal{P}_\Phi(W+P)$, define

$$x^*(\omega) = \left(\Phi\left(I_1^\omega\right), \Phi\left(I_2^\omega\right), \ldots, \Phi\left(I_N^\omega\right)\right),$$

where $\omega = \sigma \cup \{s\}$, be such that

$$I_k^\omega = I_k^\sigma \text{ for all } k \neq s$$

and

$$I_s^\omega = I_s^\rho \text{ satisfying } u_s v_s \Phi'\left(I_s^\omega\right) < 0 \text{ for some } \rho \in \mathcal{P}_\Phi(W+P). \tag{A.11}$$

(For defining I_s^ω in equation A.11, we know that I_s^ρ will exist by lemma 4: if $[N]_{<0} \neq \emptyset$, σ is permitted, and $s \in [N]_{<0}$ is such that $s \notin \sigma$, then $\rho = \sigma \cup \{s\}$ is permitted and $x^*_{\text{stable}}(\rho)$ is such that $u_s v_s \Phi'(I_s^\rho) < 0$.) Define

$$\delta_3 = \min\left(\delta_2, \quad \min_{\sigma\in\mathcal{P}_\Phi(W+P)} \min_{\substack{s\notin\sigma\\ s\in[N]_{<0}}} d\left(\sigma \cup \{s\}, \varepsilon\right)\right) \tag{A.12}$$

(see the definition of $d(\mathcal{S}, c)$ in equation A.1) where δ_2 is as defined in equation A.7. Let $d_P = \delta_3/M$. Assume that $||P||_{\max} < d_P$.

Now that we have constructed ε and d_P, we prove the conclusion of lemma 6 by way of contradiction. The impossibility is reached when we show the difference between the contributions of the perturbation to the spectrum to a maximal permitted set and a carefully selected augmented ensemble is a negative number.

By way of contradiction, suppose μ is a maximal permitted set such that $[N]_{<0} \not\subseteq \mu$ and let

$$x^*_{\text{stable}}(\mu) = (\Phi(I_1^\mu), \Phi(I_2^\mu), \ldots, \Phi(I_N^\mu))$$

be an asymptotically stable fixed point supporting μ. Let $s \in [N]_{<0}$ be such that $s \notin \mu$ and set

$$\widetilde{\mu} = \mu \cup \{s\};$$

we will demonstrate that $\widetilde{\mu} \in \mathcal{P}_\Phi(W + P)$ (which will contradict μ being a maximal codeword).

Let

$$x^*(\widetilde{\mu}) = (\Phi(I_1^{\widetilde{\mu}}), \Phi(I_2^{\widetilde{\mu}}), \ldots, \Phi(I_N^{\widetilde{\mu}}))$$

be a fixed point (so $I_s^{\widetilde{\mu}}$ is as defined in equation A.11). We focus next on showing that $x^*(\widetilde{\mu})$ is in fact asymptotically stable.

If we let $\Lambda_\mu = \Lambda(x^*_{\text{stable}}(\mu))$ and $\Lambda_{\widetilde{\mu}} = \Lambda(x^*(\widetilde{\mu}))$,

$$\operatorname{Spec}(\Lambda_\mu(W + P) - I) = \left\{\operatorname{tr}(\Lambda_\mu W) - 1 + \Delta_1^\mu, \Delta_2^\mu - 1, \ldots, \Delta_N^\mu - 1\right\}$$

and

$$\operatorname{Spec}(\Lambda_{\widetilde{\mu}}(W + P) - I) = \left\{\operatorname{tr}(\Lambda_{\widetilde{\mu}} W) - 1 + \Delta_1^{\widetilde{\mu}}, \Delta_2^{\widetilde{\mu}} - 1, \ldots, \Delta_N^{\widetilde{\mu}} - 1\right\},$$

where $|\Delta_i^\mu|, |\Delta_i^{\widetilde{\mu}}| < \varepsilon$.

Finally, we prove that $\Lambda_{\widetilde{\mu}}(W + P) - I$ is stable:

$$\begin{aligned}
\operatorname{tr}(\Lambda_{\widetilde{\mu}} W) - 1 + \operatorname{Re}(\Delta_1^{\widetilde{\mu}}) = -1 &+ \left(\sum_{i \in \mu} u_i v_i \Phi'(I_i^\mu) + u_s v_s \Phi'(I_s^{\widetilde{\mu}})\right) \\
&+ \left(\sum_{j \notin \mu} u_j v_j \Phi'(I_j^\mu) - u_s v_s \Phi'(I_s^\mu)\right) \\
&+ \left(\operatorname{Re}(\Delta_1^\mu) + \operatorname{Re}(\eta)\right),
\end{aligned}$$

where $\eta = \Delta_1^{\widetilde{\mu}} - \Delta_1^\mu$ and we noted that

$$\sum_{i \in \widetilde{\mu}} u_i v_i \Phi'(I_i^{\widetilde{\mu}}) = u_s v_s \Phi'(I_s^{\widetilde{\mu}}) + \sum_{i \in \mu} u_i v_i \Phi'(I_i^\mu)$$

(because $I_i^{\widetilde{\mu}} = I_i^{\mu}$ for $i \neq s$ and $\widetilde{\mu} = \mu \cup \{s\}$ by definition) and

$$\sum_{j \notin \widetilde{\mu}} u_j v_j \Phi'(I_j^{\widetilde{\mu}}) = -u_s v_s \Phi'(I_s^{\mu}) + \sum_{j \notin \mu} u_j v_j \Phi'(I_j^{\mu}).$$

Since $x^*_{\text{stable}}(\mu)$ is asymptotically stable, $\text{tr}(\Lambda_\mu W) - 1 + \text{Re}(\Delta_1^{\mu}) < 0$. Finally, after rewriting $I_s^{\widetilde{\mu}}$ as I_s^{ρ} for some $\rho \in \mathcal{P}_\Phi(W+P)$ such that $u_s v_s \Phi'(I_s^{\rho}) < 0$, we see

$$\begin{aligned}\text{Re}(\eta) + u_s v_s \Phi'(I_s^{\widetilde{\mu}}) - u_s v_s \Phi'(I_s^{\mu}) &= \text{Re}(\eta) + u_s v_s \Phi'(I_s^{\rho}) - u_s v_s \Phi'(I_s^{\mu}) \\ &< 2\varepsilon + u_s v_s \left(\Phi'(I_s^{\rho}) - \Phi'(I_s^{\mu})\right) \\ &< 0,\end{aligned}$$

where in the last step we used the definition of ε (see equations A.8 and A.9). We conclude that $[N]_{<0} \subseteq \mu$. □

Next, we prove that $\mathcal{P}_\Phi(W+P)$ is a convex code when W is a rank-one synaptic weight matrix. Recall that we assume that W is a rank-one synaptic weight matrix and Φ is a continuous nonnegative piecewise C^1 activation function (see the statement of theorem 1 in section 4). In the following proof, define $\min\limits_{x \in \emptyset} f(x) = +\infty$.

Proof of Theorem 1. If μ and σ are permitted sets such that μ is maximal, let

$$x^*(\mu \cap \sigma) = \left(\Phi\left(I_1^{\mu\cap\sigma}\right), \Phi\left(I_2^{\mu\cap\sigma}\right), \ldots, \Phi\left(I_N^{\mu\cap\sigma}\right)\right), \qquad \text{(A.13)}$$

where $I_i^{\mu\cap\sigma} = I_i^{\sigma}$ for every $i \in \mu \cap \sigma$ or $i \in [N]\backslash\sigma$, and $I_j^{\mu\cap\sigma} = I_j^{\mu}$ for every $j \in \sigma\backslash(\mu \cap \sigma)$. Let

$$\delta = \min\left(\delta_3, \quad \min_{\substack{\mu \in \mathcal{P}_\Phi(W+P) \\ \mu \text{ maximal}}} \; \min_{\sigma \in \mathcal{P}_\Phi(W+P)} d\left(\mu \cap \sigma, \varepsilon\right)\right)$$

(see the definition of $d(\mathcal{S}, c)$ in equation A.1), where δ_3 is as defined in equation A.12. Define $d_P = \delta/M$, where M is as defined in equation A.4, and assume that $||P||_{\max} < d_P$.

Next, let $\mu, \sigma \in \mathcal{P}_\Phi(W+P)$, where μ is maximal. Let $\tau = \mu \cap \sigma$, where $\tau \subset \sigma$ is such that $\tau \neq \sigma$, and $x^*(\tau) = (\Phi(I_1^{\tau}), \ldots, \Phi(I_N^{\tau}))$. Note that $x^*(\tau)$ supports τ by construction. Our next step in the argument is showing that $x^*(\tau)$ is in fact an asymptotically stable fixed point supporting τ.

As usual, for readibility's sake, let $\Lambda_\sigma = \Lambda(x^*_{\text{stable}}(\sigma))$ and $\Lambda_\tau = \Lambda(x^*(\tau))$. Then

$$\mathrm{Spec}(\Lambda_\sigma(W+P)-I)=\left\{\mathrm{tr}(\Lambda_\sigma W)-1+\Delta_1^\sigma, \Delta_2^\sigma-1,\ldots,\Delta_N^\sigma-1\right\}$$

and

$$\mathrm{Spec}(\Lambda_\tau(W+P)-I)=\left\{\mathrm{tr}(\Lambda_\tau W)-1+\Delta_1^\tau, \Delta_2^\tau-1,\ldots,\Delta_N^\tau-1\right\},$$

where $|\Delta_k^\sigma|, |\Delta_k^\tau|<\varepsilon$ for every $k\in[N]$.

Now, define $\eta=\Delta_1^\tau-\Delta_1^\sigma$. We next show that $\mathrm{tr}(\Lambda_\tau W)-1+\Delta_1^\tau$ has a negative real part:

$$\begin{aligned}-1+\mathrm{tr}(\Lambda_\tau W)+\mathrm{Re}(\Delta_1^\tau)&=-1+\sum_{i\in\tau}u_iv_i\Phi'(I_i^\tau)+\sum_{j\in[N]\backslash\sigma}u_jv_j\Phi'(I_j^\tau)\\&\quad+\sum_{j\in\sigma\backslash\tau}u_jv_j\Phi'(I_j^\tau)+\mathrm{Re}(\eta)+\mathrm{Re}(\Delta_1^\sigma)\\&=\sum_{l\in\sigma\backslash\tau}u_lv_l(\Phi'(I_l^\tau)-\Phi'(I_l^\sigma))+\mathrm{Re}(\eta)\\&\quad+\left(-1+\sum_{i\in\tau}u_iv_i\Phi'(I_i^\tau)+\sum_{l\in\sigma\backslash\tau}u_lv_l\Phi'(I_l^\sigma)\right.\\&\quad\left.+\sum_{j\in[N]\backslash\sigma}u_jv_j\Phi'(I_j^\tau)+\mathrm{Re}(\Delta_1^\sigma)\right).\end{aligned}$$

First, we show that

$$-1+\sum_{i\in\tau}u_iv_i\Phi'(I_i^\tau)+\sum_{l\in\sigma\backslash\tau}u_lv_l\Phi'(I_l^\sigma)+\sum_{j\in[N]\backslash\sigma}u_jv_j\Phi'(I_j^\tau)+\mathrm{Re}(\Delta_1^\sigma)\qquad\text{(A.14)}$$

is negative by showing that it is equal to $\mathrm{tr}(\Lambda_\sigma W)-1+\mathrm{Re}(\Delta_1^\sigma)$. Observe that since $I_i^\tau=I_i^\sigma$ for every $i\in\tau$ or $i\in[N]\backslash\sigma$,

$$\sum_{i\in\tau}u_iv_i\Phi'(I_i^\tau)=\sum_{i\in\tau}u_iv_i\Phi'(I_i^\sigma)\ \text{ and }\ \sum_{j\in[N]\backslash\sigma}u_jv_j\Phi'(I_j^\tau)=\sum_{j\in[N]\backslash\sigma}u_jv_j\Phi'(I_j^\sigma),$$

so we see that equation A.14 is equal to $\mathrm{tr}(\Lambda_\sigma W)-1+\mathrm{Re}(\Delta_1^\sigma)$, which is a negative number by the assumption that $\Lambda_\sigma(W+P)-I$ is stable.

Second, we show that what remains of $-1+\mathrm{tr}(\Lambda_\tau W)+\mathrm{Re}(\Delta_1^\tau)$ is also negative:

$$\sum_{l\in\sigma\backslash\tau}u_lv_l\left(\Phi'(I_l^\tau)-\Phi'(I_l^\sigma)\right)+\mathrm{Re}(\eta).\qquad\text{(A.15)}$$

Observe that since $\tau = \sigma \cap \mu$ by definition, $l \in \sigma \backslash \tau$ implies $l \notin \mu$. It is worth remembering at this stage that if a neuron l is in a permitted set σ, then its effective self-coupling strength satisfies $|u_l v_l \Phi'(I_l^\sigma)| > r_{\text{on}}$. Otherwise, if l is not in σ, then l must be unresponsive: $|u_l v_l \Phi'(I_l^\sigma)| \leq r_{\text{off}}$. Next, we know that μ is maximal by assumption, so it follows that $l \notin [N]_{<0}$ by lemma 6. Therefore, $x^*_{\text{stable}}(\sigma)$ is such that

$$u_l v_l \Phi'(I_l^\sigma) > 0,$$

which implies that

$$u_l v_l \left(\Phi'(I_l^\tau) - \Phi'(I_l^\sigma)\right) < 0 \tag{A.16}$$

for any $l \in \sigma \backslash \tau$. By the definition of ε in equation A.8 and C_2 in equation A.10,

$$2\varepsilon < -u_k v_k \left(\Phi'(I_k^\theta) - \Phi'(I_k^\rho)\right) \tag{A.17}$$

for all $k \in \rho \backslash \theta$, where $\rho, \theta \in \mathcal{P}_\Phi(W + P)$, such that $u_k v_k \Phi'(I_k^\rho) > 0$. By equations A.16 and A.17, we have for any $l \in \sigma \backslash \tau$ that

$$\begin{aligned} 2\varepsilon + u_l v_l \left(\Phi'(I_l^\tau) - \Phi'(I_l^\sigma)\right) &< 2\varepsilon + u_l v_l \left(\Phi'(I_l^\mu) - \Phi'(I_l^\sigma)\right) \\ &< 0, \end{aligned}$$

where we used the fact that $I_l^\tau = I_l^\mu$ for every $l \in \sigma \backslash \tau$ (see the definition of $x^*(\mu \cap \sigma)$ in equation A.13). Hence, we have exhibited a fixed-point $x^*(\tau)$ that is asymptotically stable and that supports τ, so we conclude that $\tau \in \mathcal{P}_\Phi(W + P)$. □

Acknowledgments

We thank Carina Curto and Vladimir Itskov for helpful conversations.

References

Cruz, J., Giusti, C., Itskov, V., & Kronholm, B. (2019). On open and closed convex codes. *Discrete and Computational Geometry*, *61*(2), 247–270. 10.1007/s00454-018-00050-1, PubMed: 31571705

Curto, C., Degeratu, A., & Itskov, V. (2012). Flexible memory networks. *Bulletin of Mathematical Biology*, *74*(3), 590–614. 10.1007/s11538-011-9678-9, PubMed: 21826564

Curto, C., Degeratu, A., & Itskov, V. (2013). Encoding binary neural codes in networks of threshold-linear neurons. *Neural Computation*, *25*(11), 2858–2903. 10.1162/NECO_a_00504, PubMed: 23895048

Curto, C., Gross, E., Jeffries, J., Morrison, K., Omar, M., Rosen, Z., . . . Youngs, N. (2017). What makes a neural code convex? *SIAM Journal on Applied Algebra and Geometry*, *1*(1), 222–238. 10.1137/16M1073170

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*(6789), 947–951. 10.1038/35016072, PubMed: 10879535

Hahnloser, R. H., Seung, H. S., & Slotine, J.-J. (2003). Permitted and forbidden sets in symmetric threshold-linear networks. *Neural Computation*, *15*(3), 621–638. 10.1162/089976603321192103, PubMed: 12620160

Harris, K. D. (2005). Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, *6*(5), 399–407. 10.1038/nrn1669, PubMed: 15861182

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. London: Psychology Press.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge: Cambridge University Press.

Mastrogiuseppe, F., & Ostojic, S. (2018). Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, *99*(3), 609–623. 10.1016/j.neuron.2018.07.003, PubMed: 30057201

Osborne, L. C., Palmer, S. E., Lisberger, S. G., & Bialek, W. (2008). The neural basis for combinatorial coding in a cortical population response. *Journal of Neuroscience*, *28*(50), 13522–13531. 10.1523/JNEUROSCI.4390-08.2008, PubMed: 19074026

Serre, D. (2020). *Matrices: Theory and applications* (2nd ed.). Berlin: Springer-Verlag.

Thompson, A. W., & Scott, E. K. (2016). Characterisation of sensitivity and orientation tuning for visually responsive ensembles in the zebrafish tectum. *Scientific Reports*, *6*.

Wilson, M., & McNaughton, B. (1994). Dynamics of the hippocampal ensemble code for space. *Science*, *264*(5155), 16–16. 10.1126/science.264.5155.16.c

Received December 2, 2021; accepted April 19, 2022.