Minimax Dynamics of Optimally Balanced Spiking Networks of Excitatory and Inhibitory Neurons

Qianyi Li  
Biophysics Graduate Program  
Harvard University  
Cambridge, MA 02138  
qianyi_li@g.harvard.edu

Cengiz Pehlevan  
John A. Paulson School of Engineering and Applied Sciences  
Harvard University  
Cambridge, MA 02138  
cpehlevan@seas.harvard.edu

Abstract

Excitation-inhibition balance is ubiquitously observed in the cortex. Recent studies suggest an intriguing link between balance on fast timescales, tight balance, and efficient information coding with spikes. We further this connection by taking a principled approach to optimal balanced networks of excitatory (E) and inhibitory (I) neurons. By deriving E-I spiking neural networks from greedy spike-based optimizations of constrained minimax objectives, we show that tight balance arises from correcting for deviations from the minimax optimum. We predict specific neuron firing rates in the networks by solving the minimax problems, going beyond statistical theories of balanced networks. Finally, we design minimax objectives for reconstruction of an input signal, associative memory, and storage of manifold attractors, and derive from them E-I networks that perform the computation. Overall, we present a novel normative modeling approach for spiking E-I networks, going beyond the widely-used energy-minimizing networks that violate Dale’s law. Our networks can be used to model cortical circuits and computations.

1 Introduction

While spiking neural networks empower our brains, a thorough understanding of how spikes accurately represent and transmit information is lacking. Any explanation should address a striking and widely observed property of cortical spiking networks that excitatory (E) and inhibitory (I) currents for individual neurons are balanced (detailed balance), and spikes are only generated during the brief occasions when inhibition fails to track spontaneous fluctuations in excitation (tight balance) [1–8]. In this work, we show that tight and detailed balance could arise from correcting for deviations from a minimax optimum that governs the dynamics of E-I balanced networks. Having access to an objective allows us to design spiking E-I networks that perform various biologically relevant functions.

While the theory of balanced networks goes further back [9–11], the idea that tight and detailed balance is related to the efficiency of spike coding was proposed in recent influential theoretical work [6,12,13]. These results are typically limited to spiking dynamics performing a greedy-minimization of a reconstruction error loss function without considering more diverse computations carried out in the cortex. Moreover, minimizing an energy function leads to symmetric connection weights [14] that violate, among other biological constraints, Dale’s law: a neuron’s influence on its post-synaptic neurons are either all excitatory or all inhibitory. This violation has been addressed by introducing separate reconstruction error cost functions for E and I neurons [12,15]. While E neurons reconstruct an input signal, I neurons reconstruct the non-Dalian part of recurrent E interactions, assuming that I neuron dynamics equilibrates faster than E neurons. Our work extends these previous accounts in two ways.

Preprint. Under review.
First, we take a more principled approach and propose a common minimax dynamics that E and I spiking neurons collectively optimize, adopting an idea from previous work on networks of rate-coding (continuous outputs) E-I neurons [16]. The minimax approach, besides other benefits, provides an intriguing interpretation of the antagonism between E and I neural populations.

Second, we consider minimax objectives that perform functions beyond signal reconstruction. Energy minimizing networks with symmetric interactions have been a powerful modeling framework for neuroscience since Hopfield’s seminal contribution [14]. Phenomena such as associative memory [14], oculomotor integration [17,18], coding of periodic variables like head direction [19,20] and spatial navigation [21,22] were modeled with fixed-point or manifold attractors resulting from such energy minimizing neural dynamics. Here, by moving from energy minimization to minimax optimization, we extend the reach of normative modeling to E-I spiking networks and provide derivations of circuits for each of the functions listed above.

Our technical contributions in this paper are:

- Derivation of a greedy spiking algorithm from a minimax objective. We derive a detailed and tightly balanced E-I spiking network and its dynamics from the optimization of a minimax objective function. We discuss the optimality conditions leading to firing rate predictions of individual neurons and detailed balance. We argue that greedy spiking dynamics leads to tight balance, and provide necessary conditions for convergence of the dynamics.
- Applications. We design spiking networks that reconstruct signals, and exhibit fixed-point and manifold attractors, all biologically relevant computational problems, while obeying Dale’s law and remaining in tight and detailed balance. We verify our theory in simulations. These applications indicate that our approach offers a principled method for designing spiking neural network models for cortical function.

## 2 Minimax dynamics of optimally balanced E-I spiking networks

In this section, we consider the spiking dynamics of an integrate-and-fire E-I network, and show how such dynamics can be derived from a minimax objective function as a greedy optimization algorithm. By analyzing the optimum of the constrained minimax objective, we observe that the network is in detailed balance and derive conditions for stability of the optimum. Finally, we derive conditions for convergence of the dynamics and demonstrate tight balance. Our derivation extends the methods of [12,13] to minimax optima.

Our network is composed of separate E and I neuron populations of $N^E$ and $N^I$ neurons (Fig.1a). For simplicity, only the E neurons receive $N^0$ dimensional external inputs, although our results and analysis still hold when both populations receive external inputs. The spiking dynamics is given by

$$
\frac{dV^E_i}{dt} = -\frac{V^E_i}{\tau^E} + \sum_j W_{ij}^{EE} s^E_j - \sum_j W_{ij}^{EI} s^I_j + \sum k F_{jk} s^\text{input}_j,
$$

$$
\frac{dV^I_i}{dt} = -\frac{V^I_i}{\tau^I} + \sum_j W_{ij}^{IE} s^E_j - \sum_j W_{ij}^{II} s^I_j.
$$

(1)

Here, $V^E_i, i = 1, \ldots, N^E$ and $V^I_i, i = 1, \ldots, N^I$, denote the membrane potentials for E and I neurons respectively, $\tau^E$ and $\tau^I$ are the corresponding membrane time constants, and $s^E_j$ and $s^I_j$ denote the spike trains of E and I neurons: e.g. $s^E_j(t) = \sum_k \delta(t - t_{j,k})$, where $t_{j,k}$ is the time of the $k$-th spike of the $j$-th neuron. $s^\text{input}_j, j = 1, \ldots, N^0$ denotes the input signal, which is not required to be a spike train. $F \in \mathbb{R}^{N^E \times N^0}$ is the feedforward connectivity matrix and $W^{EE,EI,II} \in \mathbb{R}^{N^E \times N^E}$ are the connectivity matrices within and between the E-I populations. We require $W^{EE} \in \mathbb{R}^{N^E \times N^E}$ and $W^{II} \in \mathbb{R}^{N^I \times N^I}$ to be symmetric and $W^{EI} = W^{IE}$ for our minimax objective approach. All the off-diagonal elements of the weight matrices are nonnegative so that the network obeys Dale’s law. The spiking reset is incorporated into the diagonals of $W^{EE}$ and $W^{II}$, which define how much the membrane potentials decrease after the arrival of a spike. Therefore, the diagonal elements of $W^{EE}$ are negative and $W^{II}$ are positive. The spiking thresholds for the E and I neurons are given by $T^E_i = -\frac{1}{2} W^{EE}_{ii}$ and $T^I_i = \frac{1}{2} W^{II}_{ii}$ respectively.
We can obtain implicit expressions for $V^E_i$ and $V^I_i$ by integrating Eq. (1):

$$V^E_i(t) = \sum_j W_{ij}^E r^E_j(t) - \sum_j W_{ij}^I r^I_j(t) + \sum_j F_{ij} x_j(t),$$

$$V^I_i(t) = \sum_j W_{ij}^I r^E_j(t) - \sum_j W_{ij}^I r^I_j(t),$$

where $r^E_j$ and $x_j$ are defined by filtering the spike train or the input signal with an exponential kernel:

$$r^E_j(t) = \int_0^\infty e^{-(t-t')/\tau^E} S^E_j(t-t') dt', \quad x_j(t) = \int_0^\infty e^{-(t-t')/\tau^I} S^I_j(t-t') dt'.$$

Our intuition is that the network is in a balanced state when the dynamics reaches the optimum of an objective function. Detailed balance of E and I inputs require that $V^E_i$ should fluctuate around zero for each neuron $i$ [6]. Tight balance requires the imbalance between E and I to be short-lived [6]. As we show next, both of these goals are achieved when the dynamics (1) performs a greedy optimization of a minimax objective $S$, where the implicit expressions (2) correspond to the saddle-point condition.

**Spiking dynamics is a greedy algorithm optimizing a minimax objective:** Consider the minimax optimization problem:

$$\min_{r^E} \max_{r^I} S(r^E, r^I), \quad S = -\frac{1}{2} r^E^T W^E E r^E + r^E^T W^E I r^I - \frac{1}{2} r^I^T W^I I r^I - x^T F^T r^E. \quad (4)$$

We can derive from our objective function a greedy algorithm that performs the optimization, which corresponds to the spiking dynamics (1), by adopting a similar approach to [12][13]. The details of the derivation is provided in Supplementary Information (SI) A. Here we outline the approach. We track filtered spike trains, $r^E_j$ given in Eq. (3), instantaneously. At each time, each E/I neuron makes an individual decision to spike. An E neuron spikes if the spike decreases the instantaneous value of $S$: $S(r^E + e_i, r^I) < S(r^E, r^I)$, where $e_i$ is the $i$-th standard basis vector. An I neuron spikes if the spike increases the instantaneous value of $S$: $S(r^E, r^I + e_i) > S(r^E, r^I)$. This series of spiking decisions lead to the spiking dynamics defined in Eq. (1) when $\tau^E = \tau_I = \tau$, and the dynamics in Eq. (2) and Eq. (3) for the general case where the time constants can be different.

**KKT conditions imply detailed and tight balance of active neurons:** We can verify that the detailed balance is reached at the optimum of the constrained minimax problem by examining the KKT conditions [27] obtained by considering the optimizations with respect to $r^I$ and $r^E$ sequentially. These KKT conditions are satisfied for local optimal points of the minimax problem [28][29]. We obtain, $\forall i \in \{1, \ldots, N^I\}$

$$r^I_i = 0, \quad \text{and} \quad V^I_i = \sum_j W_{ij}^I r^E_j - \sum_j W_{ij}^I r^I_j \leq 0, \quad \text{and} \quad \lambda^I_i \geq 0,$$

or $r^I_i > 0, \quad \text{and} \quad V^I_i = \sum_j W_{ij}^I r^E_j - \sum_j W_{ij}^I r^I_j = 0, \quad \text{and} \quad \lambda^I_i = 0.$

and $\forall i \in \{1, \ldots, N^E\}$

$$r^E_i = 0, \quad \text{and} \quad V^E_i = \sum_j W_{ij}^E r^E_j - \sum_j W_{ij}^E r^I_j + \sum_j F_{ij} x_j \leq 0, \quad \text{and} \quad \lambda^E_i \geq 0,$$

or $r^E_i > 0, \quad \text{and} \quad V^E_i = \sum_j W_{ij}^E r^E_j - \sum_j W_{ij}^E r^I_j + \sum_j F_{ij} x_j = 0, \quad \text{and} \quad \lambda^E_i = 0.$

From the KKT conditions, we see that for active neurons with $r^E_i > 0$, we have $V^E_i = 0$ at the saddle point, which suggests that any neuron that is active is in detailed balance. Because of the greedy nature of optimization, when the network state deviates from the balanced state, the dynamics will automatically self-correct leading to tight balance. We verify this with simulations. The E and I input for active neurons closely track each other (Fig. 1), and their total
The error of firing rate prediction exhibits a power law dependence on the time constant. The network is more tightly balanced with increasing time constants. The prediction is obtained by directly solving the KKT conditions of the minimax optimization problem. Left: \( \tau = 0.1 \), Right: \( \tau = 1 \). Larger time constant results in more accurate prediction. (f) The average CV for active neurons increases as \( \tau \) becomes larger, but remains close to 1, and saturates around 1.7.

Further optimality conditions: KKT conditions are necessary but not sufficient. The second order sufficient conditions of the minimax problem (by considering maximization and minimization sequentially, details in Subsection 1.4) give us

\[
\tilde{W}^{II} \succeq 0, \quad \tilde{W}^{EI} \tilde{W}^{II-1} \tilde{W}^{IE} - \tilde{W}^{EE} \succeq 0,
\]

where \( \tilde{W}^{II} \) is the principal submatrix of \( W^{II} \) whose columns and rows correspond to nonzero elements of \( r^I \). \( \tilde{W}^{EE} \) is the principal submatrix of \( W^{EE} \) whose columns and rows correspond to nonzero elements of \( r^E \), and \( W^{EI} \) is a submatrix of \( W^{EI} \) with its rows corresponding to nonzero elements of \( r^E \) and its columns corresponding to nonzero elements of \( r^I \).
Convergence conditions: The conditions discussed above describe an optimum of the objective, however, they do not guarantee that the spiking dynamics converges to the optimum. The convergence of spiking dynamics is challenging to prove. Instead, we characterize the convergence conditions for a rate dynamics optimizing the minimax objective Eq. (4), hypothesizing that similar conditions would hold for the spiking dynamics. This thinking is also motivated by previous work that proved the convergence of a certain type of spiking dynamics to the fixed point of a rate dynamics, both minimizing the reconstruction error cost function \([30][31]\). By constructing an energy (or Lyapunov) function \([16][32]\) on the rate dynamics, we derive the sufficient convergence conditions below (see SI. B.2 for details):

1. The following eigenvalue condition guarantees the convergence of any bounded trajectory that does not exhibit a change in the set of active/inactive neurons through its evolution:
   \[
   \lambda_{\min}(W^{II}) > \lambda_{\max}(\tilde{W}^{EE})
   \]  
   (8)

2. Replacing the positive semidefiniteness in the second order sufficient conditions for the optimality of the objective (Eq. (7)) with strict positive definiteness guarantees the boundedness of trajectories.

We empirically observed that these conditions are also valid for the convergence of spiking dynamics.

3 Applications

Having access to an objective function allows us to design balanced E-I networks to perform specific functions. Such normative approach has been common in neuroscience using energy minimizing (not minimax optimizing) dynamics. Minimization of signal reconstruction have been widely used to derive efficient coding circuits \([12][15][33][37]\). Energy functions for various types of attractors have been thoroughly studied: fixed point attractor networks are commonly used as models for associative memory \([14]\), ring attractor networks are used to model head-direction systems \([19–23]\) and orientation selectivity \([24]\), and grid attractor models are used to model grid cell responses \([25,26]\). Here, we revisit all these systems and show how balanced E-I circuits performing the same tasks can be obtained by designing the minimax objectives and weights.

3.1 Input reconstruction

Consider the typical objective for a signal reconstruction problem: minimization of the mean squared error with an \(l_2\)-regularizer on the response,

\[
\arg\min_{r^E} \|x - F^T r^E\|_2^2 + \frac{\lambda}{2} \|r^E\|_2^2 = \arg\min_{r^E} -x^T F^T r^E + \frac{1}{2} r^E F^T F r^E + \frac{\lambda}{2} r^E r^E. 
\]  
(9)

Our goal is to transform this problem to the one in (4), allowing a mapping to an E-I network.

To achieve our goal, we first perform a matrix factorization \(FF^T = U \Sigma U^T\) where \(U \in \mathbb{R}^{N_E \times N_I}\), \(\Sigma \in \mathbb{R}_{\geq 0}^{N_I \times N_I}\). We emphasize that the elements of \(\Sigma\) are nonnegative, \(\Sigma\) is symmetric and can be, but does not have to be, diagonal. For the factorization to be plausible \(N_I \geq \text{rank}(F)\). We will show examples of two different ways of performing the factorization, 1) by simply choosing \(\Sigma = I\) and \(U = F\), and 2) performing a singular value decomposition (SVD) on \(FF^T\). We also separate out the positive and negative parts in \(U\), and denote them by \([U]_+ = U_+\), and \([U]_- = U_-\). With all this set up, the reconstruction error cost function can be written as:

\[
\min_{r^E} -x^T F^T r^E + r^E^T (U_+ \Sigma U_+^T + U_- \Sigma U_-^T + \frac{\lambda}{2} I) r^E + \frac{1}{2} r^E^T (U_+ - U_-) \Sigma (U_+ - U_-)^T r^E
\]  
(10)

Next, with a “variable substitution” trick commonly used in statistical mechanics and the similarity matching framework \([38]\), we transform the problem into a minimax problem and introduce the I neurons:

\[
\min_{r^E} \max_{r^I} -x^T F^T r^E + r^E^T \left( U_+ \Sigma U_+^T + U_- \Sigma U_-^T + \frac{\lambda}{2} I \right) r^E + r^I^T \Sigma (U_+ - U_-) r^E - \frac{1}{2} r^I^T \Sigma r^I.
\]  
(11)
We can store attractors in our network as discrete fixed points with properly designed weights, allowing different fixed points to have overlapping sets of active neurons. In Fig. 1f for a given set of fixed
points, the network weights are obtained by performing a nonlinear constrained optimization problem. We minimize the mean-squared-error of the membrane potentials $V_t^{E,I}$ for the active neurons at the fixed points, subject to the constraints given by the KKT conditions for the inactive neurons (Eq. (5) (6)), the second order sufficient (Eq. (7)) condition, and the condition for convergence of rate dynamics (Eq. (8)) (see SI C.1). We show in Fig. 3 that with the optimized weights, our network can store different attractors (Fig. 3a&b) while remaining tightly balanced (Fig. 3c).

Ring attractor. Ring attractors have been used to model the head direction system [19, 23] and orientation selectivity in primary visual cortex [24]. Its dynamics can be derived from energy functions. To store a ring attractor in our E-I network, we design the weight matrices such that the effective energy function for the E neurons matches the energy function in standard ring attractor network [20]:

$$L = -\frac{1}{2N_E} \sum_{i,j} r_i^E(w_0 + w_1 \cos(\theta_i - \theta_j))r_j^E - \sum_i (h_0 + h_1 \cos(\theta_0 - \theta_i))r_i^E + \sum_i \int_0^{r_i^E} f^{-1}(r_i^E),$$

(12)

where $w_0, w_1, h_0, h_1$ are scalar parameters that control the system’s behavior (see SI C.2), and $f(x) = [x]_+$ represents ReLU nonlinearity (see SI B.2). Each neuron $i$ in the E population is assigned a distinct preferred angle $\theta_i \in \{0, \frac{2\pi}{N_E}, \ldots, \frac{2\pi(N_E-1)}{N_E}\}$.

We assume that the inhibitory neurons are all active ($r_i^I(t) > 0$) at the optima for simplicity, which can be achieved by carefully designing the weights (eg. $W_{II}^{I-E} > 0$). We obtain an effective energy function for E neurons by minimizing over $r_i^I$ in (4) and plugging in the optimum value:

$$L_{eff} = -\frac{1}{2} r^E \top (W_{EE} - W_{EI} W_{II}^{-1} W_{IE}^I) r^E - x \top F \top r^E.$$  

(13)

To match this effective energy function with (12), we design weights such that $(W_{EE} - W_{EI} W_{II}^{-1} W_{IE}^I)_{ij} = \frac{1}{N_E} (w_0 + w_1 \cos(\theta_i - \theta_j) - \delta_{ij}) (\delta_{ij}$ term is due to the nonlinearity, similar to SI B.2). For simplicity, we choose $W_{II}$ to be an identity matrix and $W_{EI}$ to be a uniform matrix whose elements are all equivalent. The amplitude of the two matrices are tuned for the constraints to be satisfied. The input to the $i$-th neuron is given by $h_0 + h_1 \cos(\theta_0 - \theta_i)$. $\theta_0$ reflects the inhomogeneity of the input: if $h_1 > 0$, neurons with their preferred angle $\theta_i$ closer to $\theta_0$ receive larger inputs. As the standard ring attractor model, this network exhibits different behaviors in different parameter regimes [41] (see SI C.2). In a certain regime the network self-organizes into a bump attractor state even with homogeneous input (Fig 4). We observe that it also exhibits tight balance (Fig 4b).

We can introduce anisotropy to our weights to obtain further functions while loosing the objective function interpretation. By following [41][42], when we design the weights as $(W_{EE} - W_{EI} (W_{II})^{-1} W_{IE}^I)_{ij} = \frac{1}{N_E} (w_0 + w_1 \cos(\theta_i - \theta_j) - w_1 \gamma \sin(\theta_i - \theta_j) - \delta_{ij}) (W_{II}$ and $W_{EI}$ are chosen as above), the network produces traveling waves, where $\gamma$ controls the angular frequency (Fig 4d). In this example, the network stays on a limit cycle and does not reach an equilibrium, and therefore is not balanced (SI C.2).
We showed that spiking dynamics in an optimally balanced E-I network that obeys Dale’s law can be derived from a greedy optimization of a minimax objective. By inspecting the KKT conditions, we proved that the network is in detailed balance at the saddle point. Simulations confirmed tight balance. We also described conditions on the weights for the rate dynamics (gradient descent-ascent) to converge. We then showed several applications relevant to neuroscience by relating our effective objective to the well-studied energy functions for different types of computational goals. Our approach provides a theoretical understanding of how a network can maintain its tight and detailed balance while performing computational tasks.

While spiking neural networks were used to solve convex optimization problems before [12, 30, 31, 43, 44], we showed that they can also be used to optimize minimax problems. Other types of dynamics for optimizing constrained minimax objectives are given in [32, 45, 47].
Our work provides a novel normative framework for modeling of biologically realistic neural networks. While energy minimization has been used widely in computational neuroscience [12,14,15,19,26,33–37], our work extends normative modeling to minimax optimization and spiking networks obeying Dale’s law.

Broader Impact

This work introduces a principled approach to designing spiking neural networks of excitatory (E) and inhibitory (I) neurons to perform various computations. As any spiking neural network model, networks derived from our approach could be applied in the field of neuromorphic computing, where information is transmitted through spikes instead of rate, and is thus more energy efficient. Previous work have shown that E-I balanced networks can serve as fast responding modules [9,10], and our approach could be applied to designing E-I balanced modules that potentially speed up solving optimization problems in general neuromorphic computing systems.

Furthermore, we provided several conditions for the regular functioning of our spiking networks. These conditions could potentially have implications on understanding neural connectivity in the cortex, and how pathological activities in the brain may arise from disrupted synaptic interactions.

Acknowledgments

This work was supported by funds from the Intel Corporation. We thank Sreya Vemuri for help in the initial stages of this project.

References

[1] Michael Wehr and Anthony M Zador. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature, 426(6965):442–446, 2003.
[2] Yousheng Shu, Andrea Hasenstaub, and David A McCormick. Turning on and off recurrent balanced cortical activity. Nature, 423(6937):288–293, 2003.
[3] Bilal Haider, Alvaro Duque, Andrea R Hasenstaub, and David A McCormick. Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. Journal of Neuroscience, 26(17):4535–4545, 2006.
[4] Michael Okun and Ilan Lampl. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nature neuroscience, 11(5):535, 2008.
[5] Bassam V Atallah and Massimo Scanziani. Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. Neuron, 62(4):566–577, 2009.
[6] Sophie Denève and Christian K Machens. Efficient codes and balanced networks. Nature neuroscience, 19(3):375, 2016.
[7] Peter Rupprecht and Rainer W Friedrich. Precise synaptic balance in the zebrafish homolog of olfactory cortex. Neuron, 100(3):669–683, 2018.
[8] Cindy Poo and Jeffry S Isaacson. Odor representations in olfactory cortex:”sparse” coding, global inhibition, and oscillations. Neuron, 62(6):850–861, 2009.
[9] Carl van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. Neural computation, 10(6):1321–1371, 1998.
[10] Carl Van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science, 274(5293):1724–1726, 1996.
[11] Michael N Shadlen and William T Newsome. Noise, neural codes and cortical organization. Current opinion in neurobiology, 4(4):569–579, 1994.
[12] Martin Boerlin, Christian K Machens, and Sophie Denève. Predictive coding of dynamical variables in balanced spiking networks. PLoS computational biology, 9(11), 2013.
[13] David G Barrett, Sophie Denève, and Christian K Machens. Firing rate predictions in optimal balanced networks. In Advances in Neural Information Processing Systems, pages 1538–1546, 2013.
[14] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8):2554–2558, 1982.
[15] David GT Barrett, Sophie Deneve, and Christian K Machens. Optimal compensation for neuron loss. *Elife*, 5:e12454, 2016.

[16] H Sebastian Seung, Tom J Richardson, Jeffrey C Lagarias, and John J Hopfield. Minimax and hamiltonian dynamics of excitatory-inhibitory networks. In *Advances in neural information processing systems*, pages 329–335, 1998.

[17] H Sebastian Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, 1996.

[18] Mark S Goldman, A Compte, and Xiao-Jing Wang. Neural integrator models. In *Encyclopedia of neuroscience*, pages 165–178. Elsevier Ltd, 2010.

[19] William E Skaggs, James J Knierim, Hemant S Kudrimoti, and Bruce L McNaughton. A model of the neural basis of the rat’s sense of direction. In *Advances in neural information processing systems*, pages 173–180, 1995.

[20] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[21] Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853, 2017.

[22] Xiaohui Xie, Richard HR Hahnloser, and H Sebastian Seung. Double-ring network model of the head-direction system. *Physical Review E*, 66(4):041902, 2002.

[23] Rishidev Chaudhuri, Berk Gercek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature neuroscience*, 22(9):1512–1520, 2019.

[24] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[25] Himanshu Mhatre, Anatoli Gorchetchnikov, and Stephen Grossberg. Grid cell hexagonal patterns formed by fast self-organized learning within entorhinal cortex. *Hippocampus*, 22(2):320–334, 2012.

[26] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2), 2009.

[27] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Calif., 1951. University of California Press.

[28] Kyle L Luther, Runzhe Yang, and H Sebastian Seung. Unsupervised learning by a “softened” correlation game: duality and convergence. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 876–883. IEEE, 2019.

[29] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

[30] Ping Tak Peter Tang, Tsung-Han Lin, and Mike Davies. Sparse coding by spiking neural networks: Convergence theory and computational results. *arXiv preprint arXiv:1705.05475*, 2017.

[31] Chi-Ning Chou, Kai-Min Chung, and Chi-Jen Lu. On the algorithmic power of spiking neural networks. *arXiv preprint arXiv:1803.10375*, 2018.

[32] H Sebastian Seung. Convergence of gradient descent-ascent analyzed as a newtonian dynamical system with dissipation. *arXiv preprint arXiv:1903.02536*, 2019.

[33] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[34] Anthony J Bell and Terrence J Sejnowski. The “independent components” of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[35] Surya Ganguli and Haim Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, 35:485–508, 2012.

[36] Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neuronal circuits. *Neural computation*, 20(10):2526–2563, 2008.

[37] Wiebrend Brendel, Ralph Bourdoukan, Pietro Vertechi, Christian K Machens, and Sophie Denève. Learning to represent signals spike by spike. *PLoS computational biology*, 16(3):e1007692, 2020.

[38] Cengiz Pehlevan and Dmitri B Chklovskii. Neuroscience-inspired online unsupervised learning algorithms: Artificial neural networks. *IEEE Signal Processing Magazine*, 36(6):88–96, 2019.

[39] H Sebastian Seung, Daniel D Lee, Ben Y Reis, and David W Tank. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26(1):259–271, 2000.
[40] A David Redish, Adam N Elga, and David S Touretzky. A coupled attractor model of the rodent head direction system. *Network: Computation in Neural Systems*, 7(4):671–685, 1996.

[41] David Hansel and Haim Sompolinsky. Modeling feature selectivity in local cortical circuits, 1998.

[42] Chris Eliasmith. A unified approach to building and controlling spiking attractor networks. *Neural Computation*, 17(6):1276–1314, 2005.

[43] Rubén Moreno-Bote and Jan Drugowitsch. Causal inference and explaining away in a spiking network. *Scientific Reports*, 5:17531, 2015.

[44] Cengiz Pehlevan. A spiking neural network with local learning rules derived from nonnegative similarity matching. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7958–7962. IEEE, 2019.

[45] Guannan Qu and Na Li. On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 3(1):43–48, 2018.

[46] Michele Benzi, Gene H Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[47] MJ Kallio and Andrzej Ruszczynski. Perturbation methods for saddle point computation. 1994.

[48] Peter Dayan and Laurence F Abbott. Theoretical neuroscience: computational and mathematical modeling of neural systems. 2001.

[49] Garth P McCormick. Second order conditions for constrained minima. In *Traces and Emergence of Nonlinear Programming*, pages 259–270. Springer, 2014.

[50] MATLAB. version 9.6.0 (R2019a). The MathWorks Inc., Natick, Massachusetts, 2019.
Supplementary Information (SI)

A  Spiking dynamics as a greedy optimization algorithm on the minimax objective

We first show the derivation of the spiking dynamics for I neurons. Because we are maximizing over $r^I$, the neuron fires a spike when it increases the objective. The firing condition for neuron $k$ correspond to $S(r^E, r^I + e_k) > S(r^E, r^I)$, where $e_k$ denotes the standard basis vector. By plugging in Eq.3 we have

$$\sum_{ij} r_i^E W_{ij} W_{ij}^I r_j^I < \sum_{ij} (r_i^E W_{ij} W_{ij}^I (r_j^I + \delta_{jk}) - \frac{1}{2} (r_i^I + \delta_{ik}) W_{ij}^I (r_j^I + \delta_{jk}))$$

$$\Rightarrow \sum_i W_{ik}^I r_i^I - \frac{1}{2} \sum_j W_{ik}^I r_i^I - \frac{1}{2} \sum_j W_{ik}^I r_j^I > \frac{1}{2} W_{kk}^I.$$  \hfill (SL.1)

We define the membrane potential and the firing threshold of the I neurons.

$$V_i^I = \sum_i W_{ik}^I r_i^I - \sum_i W_{ik}^I r_i^I$$

$$T_i^I = \frac{1}{2} W_{kk}^I.$$  \hfill (SI.2)

Next, we derive the dynamics of the membrane potential. For $\tau_E = \tau_I$ and using Eq.3 for the definition of $r_{i/E}^I$, we arrive at:

$$\frac{dV_i^I}{dt} = -\frac{1}{\tau} V_i^I - \sum_i W_{ik}^I s_i + \sum_i W_{ik}^I s_i.$$  \hfill (SI.3)

We recognize this as the standard integrate-and-fire spiking dynamics with threshold $T_i^I = \frac{1}{2} W_{kk}^I$ [48].

For the E neurons, we proceed similarly. The firing condition for a neuron $k$ is $S(r^E + e_i, r^I) < S(r^E, r^I)$, where $e_i$ is the $i$-th standard basis vector. Eq.4 implies

$$\sum_{ij} \left( -\frac{1}{2} r_i^E W_{ij}^E r_j^E - \frac{1}{2} r_i^E W_{ij}^E r_j^E \right) + \sum_{ij} x_i F_{ij} r_j^E$$

$$> \sum_{ij} \left( -\frac{1}{2} (r_i^E + \delta_{ik}) W_{ij}^E (r_j^E + \delta_{jk}) + (r_i^E + \delta_{ik}) W_{ij}^E r_j^E - \sum_{ij} x_i F_{ij} (r_j^E + \delta_{jk})) \right)$$

$$\Rightarrow \sum_i \left( W_{ki}^E r_i^E - W_{ki}^E r_i + F_{ki} x_i \right) > -\frac{1}{2} W_{kk}^E$$

Defining the membrane potential and the firing threshold

$$V_k^E = \sum_i W_{ki}^E r_i^E - \sum_i W_{ki}^I r_i^I + \sum_i F_{ki} x_i$$

$$T_k^E = -\frac{1}{2} W_{kk}^E,$$  \hfill (SI.5)

we can obtain dynamics of the membrane potential for $\tau_E = \tau_I = \tau$:

$$\frac{dV_k^E}{dt} = -\frac{1}{\tau} V_k^E + \sum_i W_{ki}^E s_i - \sum_i W_{ki}^I s_i + \sum_i F_{ki} s_{input}$$  \hfill (SI.6)

with spiking threshold $T_k^E = -\frac{1}{2} W_{kk}^E$.

B  Convergence of the dynamics

B.1  Second order sufficient condition for optimality

We cite a theorem from [49].

**Theorem 1** The solution $x^*, \lambda^*$ obeying the KKT conditions is a constrained local minimum if for the Lagrangian

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x),$$  \hfill (SI.7)
we have
\[ s^T \nabla_x^2 L(x^*, \lambda^*) s \geq 0, \]  
(8.16)
where \( s \neq 0 \) is a vector satisfying
\[ \nabla_x g_i(x^*)^T s = 0, \]  
(8.17)
where only those active inequality constraints \( g_i(x) \) corresponding to strict complementarity (i.e. where \( \lambda_i > 0 \) are applied.

Apply Thm 1 to our case, for the maximization problem, the second order sufficient condition for optimality is \( W^{II} \neq 0 \), \( W^{II} \) denotes the submatrix corresponding to where strict complementarity is applied, namely when \( r^* > 0 \). Plugging in the optimal solution for \( r^* \) and deriving the condition for optimality for the minimization over \( r^E \), we arrive at \( W^{EI} W^{II}^{-1} W^{IE} - W^{EE} \neq 0 \).

### B.2 Rate dynamics and convergence

We can prove the convergence of a rate dynamics derived from the same minimax objective, for trajectories that do not include a switch between active and inactive neurons, i.e., active neurons at initialization remain active, and silent neurons remain silent throughout the trajectory. In practice we observe convergence even when a transition between active and inactive states occurs during the dynamics for some cases.

We use a slightly modified objective function of the form
\[ S = -\frac{1}{2} r^T W^{EE} r^E + r^T W^{EI} r^I - \frac{1}{2} r^T W^{II} r^I - x^T F r^E + \sum_i H(r^E_i) - \sum_i G(r^I_i). \]  
(10.16)

The last two terms are related to nonlinear neural activations to be shown shortly. For ReLU neurons with thresholds \( \theta_E \) and \( \theta_I \), \( H(r^E_i) = \int_{0}^{r^E_i} f^{-1}(x) dx = \frac{1}{2} (r^E_i + \theta_E)^2 \), \( G(r^I_i) = \int_{0}^{r^I_i} f^{-1}(x) dx = \frac{1}{2} (r^I_i + \theta_I)^2 \).

Function \( H(x) \) and \( G(x) \) are only defined for \( x \geq 0 \). If we choose \( W^{EE} = W^{EE} + I, W^{II} = W^{II} - I, \theta_E = \theta_I = 0 \), then this objective reduces to Eq. 10.4 For the modifications induced by nonzero thresholds, one can simply change the thresholds of \( E \) neurons from \( T^E \) to \( \frac{1}{2} W^{EE} + \theta_E \), to \( T^E = -\frac{1}{2} W^{EE} + \theta_E \), and of the \( I \) neurons from \( T^I = \frac{1}{2} W^{II} \) to \( T^I = \frac{1}{2} W^{II} + \theta_I \).

The rate dynamics that optimizes this objective is
\[
\begin{align*}
\tau_E \frac{du^E}{dt} &= -u^E + \sum_j W^{IE} j - \sum_j (W^{II} - \delta_j) r^I_j \\
\tau_I \frac{du^I}{dt} &= -u^I + \sum_j W^{IE} j + \delta_j - \sum_j W^{II} r^I_j + \sum_j F r^E j
\end{align*}
\]  
(11.16)

To see convergence, we can construct an energy function of the system assuming \( \tau_E = \tau_I = \tau \):
\[ L = \frac{1}{2} |r^E|^2 + \frac{1}{2} |r^I|^2 + \gamma S, \quad \gamma \in \mathbb{R}. \]  
(12.16)

Next we show that the energy function is decreasing. Except for a set of measure zero \((\forall i, u^E_{i} = 0)\), the ReLU function is differentiable and we have
\[ r^E_{i+1} = \frac{d u^E_{i+1}}{d u^E_{i+1}} r^E_{i}, \quad r^I_{i+1} = \frac{d u^E_{i+1}}{d u^E_{i+1}} r^I_{i} + \frac{d r^E_{i+1}}{d u^E_{i+1}} r^E_{i} = \frac{d r^E_{i+1}}{d u^E_{i+1}} u^E_{i}. \]  
(13.16)

We compute the time derivative of the energy function:
\[ \dot{L} = (\dot{r}^E, \dot{r}^I)^T (\ddot{r}^E, \ddot{r}^I) + \gamma \dot{S} \]
\[ = (\dot{u}^E, \dot{u}^I)^T \left[ \frac{d u^E_{i+1}}{d u^E_{i+1}} \begin{array}{c} 0 \\ 0 \end{array} \right] \left[ \begin{array}{c} -I + W^{EE} \frac{d r^E_{i+1}}{d u^E_{i+1}} \\ W^{IE} \frac{d r^E_{i+1}}{d u^E_{i+1}} \end{array} \right] (\dot{u}^E, \dot{u}^I) \]
\[ + \gamma (-\dot{u}^E, \dot{u}^I)^T \left[ \frac{d u^E_{i+1}}{d u^E_{i+1}} \begin{array}{c} 0 \\ 0 \end{array} \right] (\dot{u}^E, \dot{u}^I) \]
\[ = (\dot{u}^E, \dot{u}^I)^T \left[ \begin{array}{c} d u^E_{i+1} W^{EE} - d u^E_{i+1} \frac{d r^E_{i+1}}{d u^E_{i+1}} \end{array} \right] \left[ \begin{array}{c} 0 \\ 0 \end{array} \right] + \gamma \dot{S} \]
\[ \dot{L} = \dot{u}^E (W^{EE} - \gamma I) \dot{u}^E - \dot{u}^I (W^{II} - \gamma I) \dot{u}^I \]
If \( W^{II} - \gamma I \) is positive definite and \( W^{EE} - \gamma I \) is negative definite, then any bounded trajectory that does not cross \( u_i = 0 \) boundaries (and cause change in the set of active/inactive neurons) are convergent. If \( \min \lambda(W^{II}) > \max \lambda(W^{EE}) \), then there exist \( \gamma \) such that \( \min \lambda(W^{II}) > \gamma > \max \lambda(W^{EE}) \), and the condition is satisfied.

Now we proved that any bounded trajectories that do not cross the \( u_i = 0 \) boundaries are convergent, we need to further prove that the trajectories are bounded. We applying Thm 2 in [32].

**Theorem 2** Given a twice differentiable objective \( S \), suppose that \( \lambda_{inf}(S_{xx}) > \lambda_{sup}(S_{yy}) \). If either

1. \( \lambda_{inf}(S_{xx}) > 0 \) and \( -V(y) = -\min_x S(x, y) \) is radially unbounded, or
2. \( \lambda_{inf}(S_{yy}) < 0 \) and \( U(y) = \max_x S(x, y) \) is radially unbounded

is also satisfied, then any trajectory of gradient descent-ascent is bounded.

We see that the condition for boundedness is the same as the condition for second order optimality as discussed in Appendix B.

**C Attractor networks**

**C.1 Optimization problem for designing weights in fixed point attractor networks**

We design the network weights in order to have fixed point attractors by solving an optimization problem. For a given attractor state \( m \), we denote the set of active neurons as \( A_m \) and the set of silent neurons as \( \overline{A}_m \). We formulate the following optimization problem.

\[
\min \sum_{m} \sum_{i \in A_m} \| r^m_i \|^2, \quad \text{subject to}
\]

\[
\exists \kappa > 0, \ V^m_i \leq -\kappa, \ \forall i \in \overline{A}_m, \ \forall m \quad \text{(constraint 1)}
\]

\[
W^{EE}_m x^{II}_m - W^{IE}_m \gamma_i \geq 0, \ \forall m \quad \text{(constraint 2)}
\]

\[
\exists \gamma_m \in \mathbb{R}, \ \min \lambda(W^{II}_m) > \gamma_m > \max \lambda(W^{EE}_m), \ \forall m \quad \text{(constraint 3)}
\]

\[
W^{EE}_{ij} \geq 0, \ \forall i \neq j; \ W^{EE}_{ii} \leq 0, \ \forall i \quad \text{(constraint 4)}
\]

\[
W^{EE} = W^{EE^T}; \ W^{II} = W^{II^T}; \ W^{IE} = W^{IE^T} \quad \text{(constraint 5)}
\]

\[
W^{EE}/W^{EE^T}/W^{II}/W^{II^T} \quad \text{(constraint 6)}. \quad \text{(SI.15)}
\]

Here \( W^{EE}/W^{EE^T}/W^{II}/W^{II^T} \) denotes the submatrices with rows and columns corresponding to the active neurons in attractor state \( m \). The expression for \( V^m_i \) is given by plugging in \( r^E \) and \( r^I \) in the attractor state \( m \) into Eq.2 for E and I neurons respectively. Constraint 1 corresponds to the KKT condition on the inactive neurons. Constraints 2 and 3 guarantee convergence of rate dynamics when the trajectory does not cross the \( u_i = 0 \) boundaries. Constraints 4,5,6 are imposed for symmetry and nonnegativity of the matrices.

This is a nontrivial optimization problem with nonlinear constraints. We used the Sequential Quadratic Programming (SQP) algorithm for MATLAB function *fmincon* [50] to solve this optimization problem. We start with different initial values and only accept solutions that satisfy \( \min \sum_{m} \sum_{i \in A_m} \| r^m_i \|^2 = 0 \).

**C.2 Different parameter regimes in ring attractor network**

The energy function of a standard ring attractor model is given by

\[
L = -\frac{1}{2N_E} \sum_{i,j} r^E_i (w_0 + w_1 \cos(\theta_i - \theta_j)) r^E_j - \sum_i (h_0 + h_1 \cos(\theta_i - \theta_i)) r^E_i + \sum_i \int_0^{r^E_i} f^{-1}(r^E_i) \quad \text{(SI.16)}
\]

Here \( h_0 + h_1 \cos(\theta_0 - \theta_i) \) is the input term, and \( f(x) = [x]_+ \) is ReLU nonlinearity. This network has different parameter regimes for different behaviors [41]. When \( w_0 \geq 1 \), the network is unstable and the firing rate goes to infinity. When \( w_1 < 2 \), there is signal amplification, the network has larger response amplitude than its input. Finally, when \( w_1 \geq 2 \), there is symmetry breaking, the network can pick up inhomogeneous response even when the input is homogeneous (\( h_1 = 0 \)). The stable states of the network lie on a ring, and which steady state the network reaches depends on the initial conditions. The same parameter regimes and behaviors apply for our E-I network.

The parameters used in the simulations for Fig 3a&b are \( w_0 = 0.5, w_1 = 2.7, h_0 = 10 \). For the top panel, \( \gamma = 0 \). For 0-10s, \( h_1 = 5 \) and \( \theta_0 = 3\pi/4 \). For 20-30s \( h_1 = 5 \) and \( \theta_0 = 3\pi/4 \). \( h_1 = 0 \) otherwise, and the input is thus homogeneous.
Figure SI.1: The E and I inputs for $30\tau$, they are not balanced for network with travelling wave with the above parameters except $\hat{h}_1 = 0$ throughout the simulation and $\gamma = 0.08$.

For the travelling wave, we show that the E and I inputs are not balanced, because the system does not reach equilibrium. Fig. SI.1

C.3 Grid attractor

The energy function is given by

$$L = -\frac{1}{2} \sum_{i,j} r^E_i W_0(x_i - x_j) r^E_j - \sum_i A_i r^E_i + \sum_i \int_0^{r^E_i} f^{-1}(r^E_i).$$

(SI.17)

Here $A_i$ is the input to each neuron, in our simulation, we assume homogeneous input $A_i = A, \forall i$. $f(x) = [x]_+$ represents ReLU nonlinearity. Function $W_0(x)$ is given by $W_0(x) = ae^{-\gamma|x|^2} - e^{-\beta|x|^2}$. For simulation of the spiking network with grid attractor, the parameters we used are $N_E = N_I = 63^2$, neurons are arranged in a $63 \times 63$ square sheet, $\tau_E = 0.5$, $\tau_I = 0.2$, $\tau = 0.5$, $\lambda = 6$, $\beta = 3\lambda^2$, $\alpha = 1.1$, $\gamma = 1.2\beta$, $A = 2$.

For code used to reproduce results in this paper, see https://github.com/Pehlevan-Group/BalancedEIMinimax