Nonlinear stimulus representations in neural circuits with approximate excitatory-inhibitory balance

Cody Baker*, Vicky Zhu*, and Robert Rosenbaum*

Abstract

Balanced excitation and inhibition is widely observed in cortical recordings. How does this balance shape neural computations and stimulus representations? This problem is often studied using computational models of neuronal networks in a dynamically balanced state. However, these balanced network models predict a linear relationship between stimuli and population responses, in contrast to the nonlinearity of cortical computations. We show that every balanced network architecture admits some stimuli that break the balanced state and these breaks in balance push the network into a "semi-balanced state" characterized by excess inhibition to some neurons, but an absence of excess excitation. The semi-balanced state is unavoidable in networks driven by multiple stimuli, consistent with experimental data, has a direct mathematical relationship to artificial neural networks, and permits nonlinear stimulus representations and nonlinear computations.

Introduction

An approximate balance between excitatory and inhibitory synaptic currents is widely reported in cortical recordings [1, 2, 3, 4, 5, 6]. The implications of this balance are often studied using large networks of model neurons in a dynamically stable balanced state. Despite the complexity of spike timing dynamics in these models, their population-level firing rates [7, 8, 9, 10, 11] and correlations [12, 13, 14, 15, 16, 17] in response to a given stimulus can be derived using a simple mean-field theory.

This classical mean-field theory of balanced networks has two critical shortcomings. First, it predicts a linear relationship between stimuli and population responses, in contrast to the nonlinear computations that must be performed by cortical circuits. Secondly, parameters in balanced network models must be tuned so that the firing rates predicted by the mean-field theory are non-negative. In networks with many neural populations – such as multiple neuron subtypes, neural assemblies, or tuning preferences – the proportion of parameter space for which predicted rates are non-negative becomes exponentially small. Moreover, we show that for any network architecture, there are infinitely many excitatory stimuli for which the mean-field theory predicts negative rates.

We develop a theory of semi-balanced networks that quantifies network responses when the classical balanced network state is broken. In this semi-balanced state, balance is only enforced in one direction: neurons can receive excess inhibition, but not excess excitation. Neurons receiving excess inhibition are silenced and the remaining neurons form a balanced sub-network. Unlike balanced networks, semi-balanced networks implement nonlinear computations and stimulus representations. We establish a mathematical relationship between semi-balanced networks, artificial recurrent neural networks used for machine learning [18], and threshold-linear networks [19, 20, 21, 22]. We demonstrate that balance and semi-balance are achieved on a neuron-by-neuron basis in networks with large in-degrees and homeostatic inhibitory plasticity when exposed to a time-constant stimulus [23, 24, 25], but only semi-balance is achieved in the presence of time-varying stimuli. In this setting, semi-balanced networks implement richly nonlinear stimulus representations. We demonstrate the computational power of these representations using the hand-written digit classification benchmark, MNIST.

In summary, the large in-degrees typical of cortical neurons combined with the presence of time-varying stimuli imply that local cortical circuits are in a semi-balanced state. Our analysis of this state shows a direct correspondence to artificial neural networks used in machine learning and therefore has deep implications for the computational properties of cortical circuits.

*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 46556

Results

Balanced networks implement linear stimulus representations and computations

To review balanced network theory and its limitations, we consider a recurrent network of $N = 3 \times 10^4$ randomly connected adaptive exponential integrate-and-fire (adaptive EIF) neuron models. The network is composed of two excitatory populations and one inhibitory population (80% excitatory and 20% inhibitory neurons altogether) and receives feedforward synaptic input from two external populations of Poisson processes, modeling external synaptic input (Fig. 1A). The firing rates, $\boldsymbol{r}_x = [r_{x1} \;\; r_{x2}]$, of the external populations form a two-dimensional stimulus space (Fig. 1B).

Simulations of this model showed asynchronous-irregular spiking activity and excitatory-inhibitory balance (Fig. 1Ci-iii). How does connectivity between the populations determine the mapping from stimulus, $\boldsymbol{r}_x$, to firing rates, $\boldsymbol{r} = [r_{e1} \; r_{e2} \; r_i]$ in the recurrent network? Firing rate dynamics are often approximated using models of the form

$$\tau \dot{\boldsymbol{r}} = -\boldsymbol{r} + f(\overline{JK}[W\boldsymbol{r} + \boldsymbol{X}]) \tag{1}$$

where $\dot{\boldsymbol{r}}$ denotes the time derivative, f is a non-decreasing f-I curve, and W is the effective recurrent connectivity matrix. External input is quantified by $\boldsymbol{X} = W_x r_x$. Components of W and W_x are given by $w_{ab} = J_{ab}K_{ab}/\overline{JK}$ where K_{ab} is the mean number of connections from population b to a and J_{ab} is the average connection strength. The coefficient, $\overline{JK} = \text{average}(|J_{ab}|K_{ab})$, quantifies coupling strength in the network. Since $\overline{JK}$ is multiplied in the equation for $\dot{\boldsymbol{r}}$ and divided in the equation for w_{ab}, it does not affect dynamics, but serves as a notational tool in the calculations below, which require $\overline{JK} \sim J_{ab}K_{ab}$ so that $w_{ab} \sim \mathcal{O}(1)$ even when $J_{ab}K_{ab}$ is large.

The key idea underlying balanced network theory is that $\overline{JK}$ is typically large in cortical circuits because neurons receive thousands of synaptic inputs and each postsynaptic potential is moderate in magnitude. Total synaptic input,

$$\boldsymbol{I} = \overline{JK}[W\boldsymbol{r} + \boldsymbol{X}], \tag{2}$$

can only remain $\mathcal{O}(1)$ if there is a cancellation between excitation and inhibition. In particular, to have $\boldsymbol{I} \sim \mathcal{O}(1)$, we must have $W\boldsymbol{r} + \boldsymbol{X} \sim \mathcal{O}(1/\overline{JK})$ so, in the limit of large $\overline{JK}$, firing rates satisfy [8, 26, 11, 27]

$$\boldsymbol{r} = -W^{-1}\boldsymbol{X}. \tag{3}$$

In classical balanced network theory, one considers the $N \to \infty$ limit while taking $J_{ab} \sim 1/\sqrt{N}$ and $K_{ab} \sim N$ so that $\overline{JK} \to \infty$ and Eq. (3) is exact in the limit [8]. Experimental evidence for this scaling has been found in cortical cultures [6]. Note that, while Eq. (1) is a heuristic approximation to spiking networks, the conclusion that Eq. (3) must be satisfied to keep $\boldsymbol{I} \sim \mathcal{O}(1)$ as $\overline{JK} \to \infty$ does not depend on the approximation in Eq. (1), but is implied by Eq. (2) alone and is therefore mathematically valid for spiking networks [8] for which firing rates can depend on the variance, and higher order moments of neurons' synaptic input. Even though it is derived as a limit, Eq. (3) provides a simple approximation to firing rates in networks with finite $\overline{JK}$. Indeed, it accurately predicted firing rates in our spiking network simulations (Fig. 1Civ, compare dashed to solid) for which $\overline{JK} = 5.9$ mV/Hz.

While the simplicity of Eq. (3) is appealing, its linearity reveals a critical limitation of balanced networks as models of cortical circuits: Because $\boldsymbol{r}$ depends linearly on $\boldsymbol{X}$ and $\boldsymbol{r}_x$, balanced networks can only implement linear representations of stimuli and linear computations [8, 11, 27].

To demonstrate this linearity in our spiking network, we sampled a lattice of points in the two dimensional space of $\boldsymbol{r}_x = [r_{x1} \; r_{x2}]$ values and plotted the resulting neural manifold traced out in three dimensions by $\boldsymbol{r} = [r_{e1} \; r_{e2} \; r_i]$. The resulting manifold is approximately linear, *i.e.*, a plane (Fig. 1Di) because $\boldsymbol{r}$ depends linearly on $\boldsymbol{X}$, and therefore on $\boldsymbol{r}_x$, in Eq. (3). More generally, the neural manifold is an n_x-dimensional hyperplane in n-dimensional space where n and n_x are the number of populations in the recurrent and external populations respectively. In addition, any linear readout $R = \boldsymbol{w} \cdot \boldsymbol{r}$ is a linear function of $\boldsymbol{r}_x$ and therefore also planar (Fig. 1Dii).

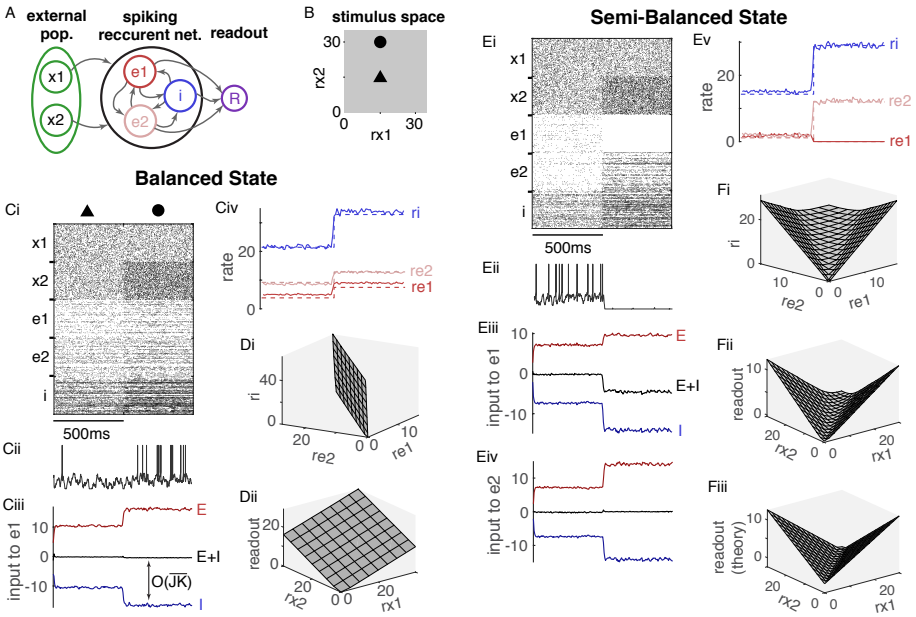

Figure 1: **Stimulus representations are linear in the balanced state and nonlinear in the semi-balanced state. A)** Network diagram. A recurrent spiking network of $N = 3\times10^4$ model neurons is composed of two excitatory populations ($e1$ and $e2$) and one inhibitory population (i) that receive input from two external spike train populations ($x1$ and $x2$). Recurrent network output is represented by a linear readout of firing rates (R). **B)** The two-dimensional space of external population firing rates represents a stimulus space. Filled triangle and circle show the two stimulus values used in Ci–iv. **Ci)** Raster plots of 200 randomly selected spike trains from each population across two stimuli. **Cii)** Membrane potential of one neuron from population $e1$. **Ciii)** Mean input current to population $e1$ from all excitatory sources ($e1$, $e2$, $x1$, and $x2$; red), from the inhibitory population (i; blue), and from all sources (black) showing approximate excitatory-inhibitory balance across stimuli. Mean input to i and $e2$ were similarly balanced. **Civ)** Firing rates of each population from simulations (solid) and predicted by Eq. (3) (dashed). **Di)** The neural manifold traced out by firing rates in each population in the recurrent network as external firing rates are varied across a square in stimulus space ($0 \le r_{x1}, r_{x2} \le 30$). **Dii)** The readout as a function of r_{x1} and r_{x2} from the same simulation as Di. **Ei–v)** Same as Ai–iv, but dashed lines in Dv are from Eq. (4) and input to $e2$ was additionally shown. D **Fi-iii)** Same as Di-ii except the theoretical readout predicted by Eq. (4) was additionally included. All firing rates are in Hz.

How do cortical circuits, which exhibit excitatory-inhibitory balance, implement nonlinear stimulus representations and computations? Below, we describe a parsimonious generalization of balanced network theory that allows for nonlinear stimulus representations by allowing excess inhibition without excess excitation.

Semi-balanced networks implement nonlinear representations in direct correspondence to artificial neural networks of rectified linear units

Note that Eq. (3) is only valid if all elements of $\boldsymbol{r}$ it predicts are non-negative. Early work considered a single excitatory and single inhibitory population, in which case positivity of $\boldsymbol{r}$ is assured by simple inequalities satisfied in a large proportion of parameter space [8, 10]. Similarly, in the simulations described above, we constructed W and W_x so all components of $\boldsymbol{r}$ were positive for all values of $r_{x1}, r_{x2} > 0$.

In networks with a large number of populations, conditions to assure $\boldsymbol{r} \ge 0$ become more complicated and the proportion of parameter space satisfying $\boldsymbol{r} \ge 0$ becomes exponentially small. In addition, we proved that connectivity structures, W, obeying Dale's law necessarily admit some positive external inputs, $\boldsymbol{X} > 0$, for which Eq. (3) predicts negative rates (Supplementary Materials S.1). Hence, the classical notion

of excitatory-inhibitory balance cannot be assured by conditions imposed on the recurrent connectivity structure, W, alone, but conditions on stimuli, $\boldsymbol{X}$, are also required.

While it is possible that cortical circuits somehow restrict themselves to the subsets of parameter space that maintain a positive solution to Eq. (3) across all salient stimuli, we consider the alternative hypothesis that Eq. (3) and the balanced network theory that underlies it do not capture the full spectrum of cortical circuit dynamics.

To explore spiking network dynamics when Eq. (3) predicts negative rates, we considered the same network as above, but changed the feedforward connection probabilities so that Eq. (3) predicts positive firing rates only when r_{x1} and r_{x2} are nearly equal. When r_{x2} is much larger than r_{x1}, Eq. (3) predicts negative firing rates for population $e1$, and vice versa, due to a competitive dynamic.

Simulating the network with $r_{x1} = r_{x2}$ produces positive rates, asynchronous-irregular spiking, and excitatory-inhibitory balance (Fig. 1Ei–v, first 500ms). Increasing r_{x2} to where Eq. (3) predicts negative rates for population $e1$ causes spiking to cease in $e1$ due to an excess of inhibition (Fig. 1Ei–v, last 500ms).

Notably, however, input currents to populations $e2$ and i remain balanced when $e1$ is silenced (Fig. 1Eiv) so the i and $e2$ populations form a balanced sub-network. These simulations demonstrate a network state that is not balanced in the classical sense because one population receives excess inhibition. However,

1. no population receives excess excitation,
2. any population with excess inhibition is silenced, and
3. the remaining populations form a balanced sub-network.

Here, an excess of excitation (inhibition) in population a should be interpreted as $\boldsymbol{I}_a \sim \mathcal{O}(\overline{JK})$ with $\boldsymbol{I}_a > 0$ ($\boldsymbol{I}_a < 0$). The three conditions above can be re-written mathematically in the large $\overline{JK}$ limit as two conditions,

1. $[W\boldsymbol{r} + \boldsymbol{X}]_a \leq 0$ for all populations, a, and
2. If $[W\boldsymbol{r} + \boldsymbol{X}]_a < 0$ then $\boldsymbol{r}_a = 0$.

These conditions, along with the implicit assumption that $\boldsymbol{r} \geq 0$, define a generalization of the balanced state. We refer to networks satisfying these conditions as "semi-balanced" since they require that strong excitation is canceled by inhibition, but they do not require that inhibition is similarly canceled. Note that the condition $[W\boldsymbol{r} + \boldsymbol{X}]_a \leq 0$ does not mean that $\boldsymbol{I}_a \leq 0$, but only that $\boldsymbol{I}_a \sim \mathcal{O}(1)$ whenever $\boldsymbol{I}_a \geq 0$ so that $[W\boldsymbol{r} + \boldsymbol{X}]_a = 0$ in the large $\overline{JK}$ limit, *i.e.*, no excess excitation.

How are firing rates related to connectivity in semi-balanced networks? In Supplementary Materials S.2, we prove that semi-balanced networks satisfy

$$\boldsymbol{r} = [W\boldsymbol{r} + \boldsymbol{X} + \boldsymbol{r}]^+ \tag{4}$$

in the limit of large $\overline{JK}$ where $[x]^+ = \max(0, x)$ is the positive part of x, sometimes called the rectified linear or threshold-linear function. Eq. (4) generalizes Eq. (3) to allow for excess inhibition. Even though it is derived in the limit of large $\overline{JK}$, Eq. (4) provides an accurate approximation to firing rates in our spiking network simulations (Fig. 1Ev, compare dashed to solid). Note that $\boldsymbol{r}$ satisfies Eq. (4) if and only if it satisfies $q\boldsymbol{r} = [W\boldsymbol{r} + \boldsymbol{X} + q\boldsymbol{r}]^+$ for any $q > 0$ (see Supplementary Materials S.2 for a proof), which explains why terms with different units can be summed together in Eq. (4).

Notably, Eq. (4) represents a piecewise linear, but globally nonlinear mapping from $\boldsymbol{X}$ to $\boldsymbol{r}$. Hence, unlike balanced networks, semi-balanced networks implement nonlinear stimulus representations (Fig. 1Fi). Eq. (4) also demonstrates a direct relationship between semi-balanced networks and recurrent artificial neural networks with rectified linear activations used in machine learning [18] and their continuous-time analogues studied by Curto and others under the label "threshold-linear networks" [19, 20, 21, 22]. These networks are defined by equations of the form

$$\tau\dot{\boldsymbol{r}} = -\boldsymbol{r} + [U\boldsymbol{r} + \boldsymbol{X}]^+.$$

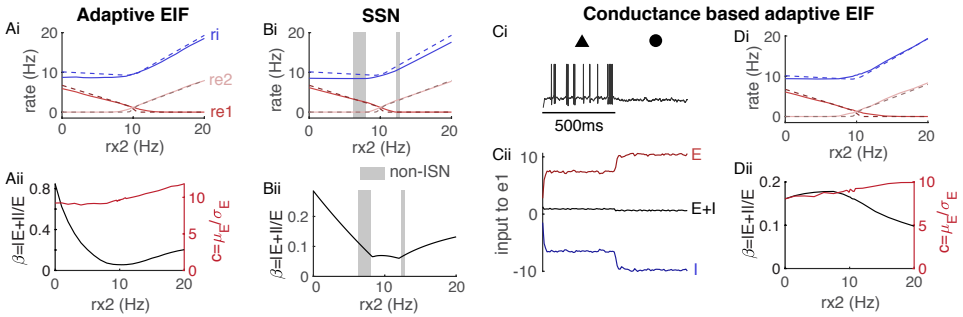

Figure 2: **The semi-balanced approximation is accurate across models and dynamical states. Ai)** Firing rates from simulations (solid) and Eq. (4) (dashed) as a function of r_{x2} when $r_{x1} = 10$Hz for the same model as in Fig. 1E-F. **Aii)** Balance ratio, β (black), and coupling strength coefficient, c (red), averaged across all neurons from the simulation in Ai. **Bi-ii)** Same as Ai and Aii, but using dynamical rate equations that implement a supralinear stabilized network (SSN). Gray shaded areas are states in which the network is not inhibitory stabilized. **Ci–ii)** Same as Fig. 1E except using a conductance-based model of synapses. **Di-ii)** Same as Ai-ii except using a conductance-based model of synapses.

Taking $U = W + Id$ where Id is the identity matrix establishes a one-to-one correspondence between solutions to Eq. (4) and fixed points of threshold-linear networks or recurrent artificial neural networks. Indeed, we used this correspondence to construct a semi-balanced spiking network that approximates a continuous exclusive-or (XOR) function (Fig. 1Fii–iii), which is widely known to be impossible with linear networks [18].

Previous work on threshold-linear networks shows that, despite the simplicity of Eq. (4), its solution space can be complicated [19, 20, 21, 22]: Any solution is partially specified by the subset of populations, a, at which $\boldsymbol{r}_a > 0$, called the "support" of the solution. There are 2^n potential supports in a network with n populations, there can be multiple supports that admit solutions, and these solutions can depend in complicated ways on the structure of W and $\boldsymbol{X}$. Hence, semi-balanced networks give rise to a rich mapping from stimuli, $\boldsymbol{X}$, to responses, $\boldsymbol{r}$.

In Supplementary Materials S.3, we proved that, under Eq. (2), the semi-balanced state is realized and Eq. (4) is satisfied only if firing rates remain moderate as $\overline{JK} \to \infty$. In other words, Eq. (4) and the semi-balanced state it describes are general properties of strongly and/or densely coupled networks with moderate firing rates. To the extent that cortical circuits have large $\overline{JK}$ values and moderate firing rates, therefore, Eq. (4) provides an accurate approximation to cortical circuit responses. In summary, our results establish a direct mapping from biologically realistic cortical circuit models to recurrent artificial neural networks used in machine learning and to the rich mathematical theory of threshold-linear networks.

Semi-balanced network theory is accurate across models and dynamical states

Recently, Ahmadian and Miller argued that cortical circuits may not be as tightly balanced or strongly coupled as assumed by classical balanced network theory [27]. They quantified the tightness of balance by the ratio of total synaptic input to excitatory synaptic input, $\beta = |E + I|/E$ (where E is the mean input current from e and x combined, and I is the mean input from i). Small values of β imply tight balance, for example $\beta \sim 1/\overline{JK}$ in classical balanced networks. They quantified coupling strength by the ratio of the mean to standard deviation of the excitatory synaptic current $c = \text{mean}(E)/\text{std}(E)$. Strongly coupled networks have large c, specifically $c \sim \overline{JK}$. Since Eq. (4) was derived in the limit of large $\overline{JK}$, it is only guaranteed to be accurate for sufficiently large c, but it is not immediately clear exactly how large c must be for Eq. (4) to be accurate.

In our spiking network simulations, Eq. (4) was accurate across a range of stimulus values even when β and c were in the range deemed to be biologically realistic by Ahmadian and Miller (Fig. 2Ai,ii). We

conclude that Eq. (4) can be a useful approximation for networks with biologically relevant levels of balance and coupling strength.

We next tested the accuracy of Eq. (4) against simulations of stabilized supralinear networks (SSNs) proposed and studied by Ahmadian, Miller, and colleagues [28, 29, 27]. In particular, we simulated the three-dimensional dynamical system

$$\tau\dot{\boldsymbol{r}} = -\boldsymbol{r} + k\overline{JK}[W\boldsymbol{r} + \boldsymbol{X}]_+^2$$

where $[x]_+^2 = ([x]^+)^2$ denotes the square of the positive part of x. Simulations of this network with parameters matched to our spiking network simulations show that the network transitioned between an inhibitory-stabilized network (ISN) state to a non-ISN state as r_{x2} varied (Fig. 2Bi), which is a defining property of SSNs. Simulations show agreement with Eq. (4), even when balance was relatively loose (Fig. 2Bi,ii).

A seemingly unrealistic property of semi-balanced networks is that the total mean synaptic current to some populations is $\mathcal{O}(\overline{JK})$ and negative (Fig. 1Eiii, black). In our simulations, this strong inhibitory input clamped the membrane potential to the lower bound we imposed at -85mV (Fig. 1Eii). The strong inhibitory current is an artifact of using a current-based model of synaptic transmission [30].

In a more realistic, conductance-based model, the magnitude of inhibitory current is limited by shunting at the inhibitory reversal potential. Repeating our simulations using a conductance-based synapse model shows similar overall trends to the current-based model (Fig. 2Ci,ii) except the mean synaptic input to population $e1$ is no longer so strongly inhibitory (Fig. 2Cii, compare to Fig. 1Eiii) and membrane potentials of $e1$ neurons still exhibit variability near the inhibitory reversal potential (Fig. 2Ci). Eq. (4) can be modified to account for conductance-based synapses (see Methods and [31, 32, 11]) and this corrected theory accurately predicted firing rates in our simulations across a range of c and β values (Fig. 2Di,ii).

Homeostatic plasticity achieves semi-balance at single neuron resolution, producing high-dimensional nonlinear representations

So far, we have only considered firing rates and excitatory-inhibitory balance averaged over discrete neural populations. Cortical circuits implement distributed neural representations that are not always captured by homogeneous population averages [33]. Balance is realized at the level of synaptic currents to individual neurons (as opposed to currents averaged over populations) is often referred to as "detailed balance" [34, 25]. Due to the use of this term for a different purpose in Markov process theory, we instead refer to it as balance "at single-neuron resolution."

To test for semi-balance above, we compared firing rates from simulations to those predicted by Eq. (4) (see Figs. 1Ev and 2Ai,Bi,Di). For semi-balance at single neuron resolution, or "detailed semi-balance," Eq. (4) becomes $[J\vec{r} + \vec{X} + \vec{r}]^+ = \vec{r}$. However, solving this equation is intractable for large networks because it would require searching for solutions across 2^N potential supports. Instead of comparing firing rates from simulations to those predicted by theory, we can test for semi-balance by verifying that synaptic currents to all neurons are only large in magnitude when they are negative (see Supplementary Materials S.2).

To explore balance and semi-balance at single-neuron resolution, we first considered the same randomly connected network of $N = 3 \times 10^4$ neurons considered above, but with only a single excitatory, inhibitory, and external population (Fig. 3A). We kept the firing rate of the external population fixed at $r_x = 10$Hz. To model a stimulus with a distributed representation, we added an extra external input perturbation that is constant in time, but randomly distributed across neurons. Specifically, the time-averaged synaptic input to each neuron was given by the $N \times 1$ vector

$$\vec{I} = \overline{JK}[J\vec{r} + \vec{X}] \tag{5}$$

where J is the $N \times N$ recurrent connectivity matrix and $\vec{r}$ is the $N \times 1$ vector of firing rates. External input is given by $\vec{X} = J_x\vec{r}_x + \vec{Z}$ where, J_x and $\vec{r}_x$ are the feedforward connectivity matrix and external rates. The distributed stimulus, $\vec{Z}$, is defined by

$$\vec{Z} = \sigma_1\vec{Z}_1 + \sigma_2\vec{Z}_2$$

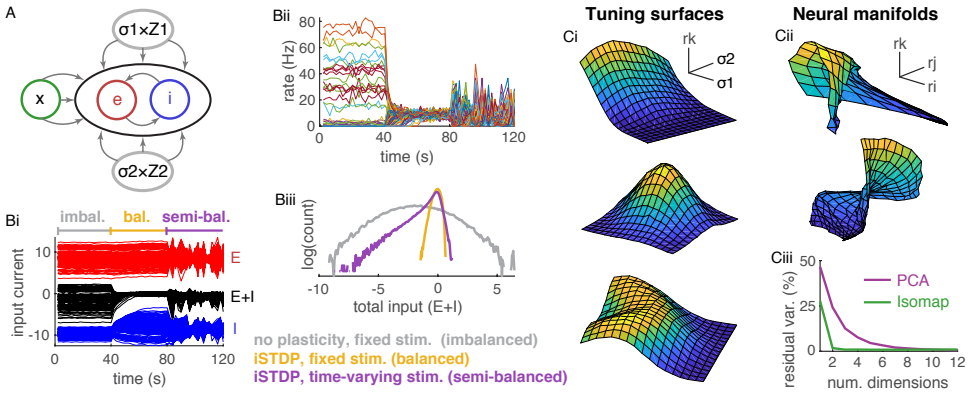

Figure 3: **Balance, semi-balance, and neural representations at single-neuron resolution with homeostatic plasticity. A)** Network diagram. Same as in Fig. 1A except there is just one excitatory and one external population and an additional input $\vec{Z} = \sigma_1\vec{Z}_1 + \sigma_2\vec{Z}_2$. **Bi)** Excitatory (red), inhibitory (blue), and total (black) input currents to 100 randomly selected excitatory neurons averaged over 2s time bins. During the first 40s, synaptic weights and $\sigma_1 = \sigma_2$ were fixed. During the next 40s, homeostatic iSTDP was turned on and $\sigma_1 = \sigma_2$ were fixed. During the last 40s, iSTDP was on and σ_1 and σ_2 were selected randomly every 2s. **Bii)** Firing rates of the same 100 neurons averaged over 2s bins. **Biii)** Histograms of input currents to all excitatory neurons averaged over the first 40s (gray, imbalanced), the next 40s (yellow, balanced), and the last 40s (purple, semi-balanced). **Ci)** Firing rates of three randomly selected excitatory neurons as a function of the two stimuli, σ_1 and σ_2 (the neuron's "tuning surface") in a network pre-trained by iSTDP. **Cii)** Three neural manifolds. Specifically, the surface traced out by the firing rates of the three randomly selected neurons as σ_1 and σ_2 are varied. **Ciii)** Percent variance unexplained by PCA (purple) and Isomap (green) applied to all excitatory neuron firing rates from the simulation in Ci-ii. Network size was $N = 3 \times 10^4$ in Bi-iii and reduced to $N = 5000$ in Ci-iii to save runtime (see Methods).

where $\vec{Z}_1$ and $\vec{Z}_2$ are standard normally distributed, $N \times 1$ vectors. The vector, $\vec{Z}$, lives on a two-dimensional hyperplane in N-dimensional space and the plane is parameterized by σ_1 and σ_2. Hence, $\vec{Z}$ models a two-dimensional stimulus whose representation is distributed randomly across the neural population.

Simulations show that this network does not achieve balance at single-neuron resolution: Some neurons receive excess inhibition and some receive excess excitation (Fig. 3Bi, first 40s), leading to large firing rates in some neurons (Fig. 3Biii) and a broad distribution of total input currents (Fig. 3Bii, blue). Indeed, it has been argued previously that randomly connected networks are imbalanced at single-neuron resolution when stimuli and connectivity are not co-tuned [9, 25]. This is consistent with previous results on "imbalanced amplification" in which connectivity matrices with small-magnitude eigenvalues values can break balance when external inputs are not orthogonal to the corresponding eigenvectors [11]. When J is large and random, it will have eigenvalues near the origin purely by chance, which can lead to imbalanced amplification if the corresponding eigenvectors are not orthogonal to $\vec{X}$.

Previous work shows that single-neuron resolution balance can be realized by a homeostatic inhibitory spike-timing dependent plasticity (iSTDP) rule [23, 25]. Indeed, when iSTDP was introduced in our simulations, balance was obtained and firing rates became more homogeneous (Fig. 3Bi-ii, second 40s) with a much narrower distribution of total input currents (Fig. 3Biii, red), at least while σ_1 and σ_2 were fixed.

Of course, real cortical circuits are exposed to multiple, time-varying stimuli. To simulate time-varying stimuli, we randomly selected new values of σ_1 and σ_2 every 2s (Fig. 3Bi-ii last 40s). This transition to a time-varying stimulus caused the total input to some neurons to become strongly inhibitory, but no neurons received excess excitation (Fig. 3Biii, yellow), indicating that the network was in a semi-balanced state at single-neuron resolution. These results show that the semi-balanced state is a natural state for cortical circuits exposed to time-varying stimuli. This is consistent with findings that inhibition dominates sensory responses in awake animals ([35], compare to dominance of inhibition in Fig. 3Bi, last 40s). Repeating these simulations in a model with conductance-based simulations shows that shunting inhibition prevents

an excess inhibitory currents in the semi-balanced state, but a measure of effective excitatory and inhibitory conductances recovers the imbalanced, balanced, and semi-balanced states observed in the current-based model (Supplemental Figure S.1).

We next investigated the properties of the mapping from the two-dimensional stimulus space to the N-dimensional firing rate space. We sampled a uniform lattice of $17 \times 17 = 289$ points in the two-dimensional space of σ_1 and σ_2 values, simulated a network of size $N = 5000$, then plotted the resulting firing rates of three randomly selected neurons as a function of σ_1 and σ_2. The resulting surfaces appear highly nonlinear and multi-modal (Fig. 3Ci). Next, we plotted two randomly selected neural manifolds, each defined by the firing rates of three random excitatory neurons. These manifolds also appear highly nonlinear with rich structure (Fig. 3Cii). Note that there are over 10^{10} such manifolds in the network, suggesting a rich representation of the two-dimensional stimulus.

To understand how these surfaces get their shape, note that semi-balance at single-neuron resolution is realized when $[J\vec{r} + \vec{X} + \vec{r}]^+ = \vec{r}$ in the $\overline{JK} \to \infty$ limit. This equation is piecewise linear in the sense that sufficiently small changes to $\vec{X}$ cause linear changes to $\vec{r}$. Nonlinearities occur whenever a change to $\vec{X}$ causes an individual r_j value to transition between zero and non-zero values, *i.e.*, at transitions between two of the 2^N "supports" (see above). The enormous number of supports in large networks (over 10^{1500} potential supports in the network from Fig. 3C) implies that nonlinearities are prevalent in the solution space and the underlying piecewise linearity is not visible in practice.

The nonlinearity of the stimulus representation is more precisely quantified by comparing the results of the dimension reduction techniques isometric feature mapping (Isomap) and principal component analysis (PCA) applied to the sampled firing rates. Both methods find a low-dimensional manifold in N-dimensional rate space near which the sampled rates lie. However, PCA is restricted to linear manifolds (hyperplanes) while Isomap finds nonlinear manifolds. We applied both methods to the set of all excitatory firing rates across all 289 stimuli from the simulations above.

Despite the fact that firing rates represent 289 points in a 4000-dimensional space, the points lie close to a two-dimensional manifold because they are approximately a function of the two-dimensional stimulus. Applying Isomap shows that the vast majority of the variance was explained by a two-dimensional manifold (Fig. 3Ciii, green; 1.76% residual variance at 2 dimensions). However, PCA required more than 8 dimensions to capture the same amount of variance and generally captured less variance per dimension (Fig. 3Ciii, purple). This implies that the two-dimensional neural manifold in 4000-dimensional space is nonlinear, *i.e.*, curved, so that it cannot be captured by a two-dimensional plane.

In summary, when networks are presented with time-varying stimuli, iSTDP produces a semi-balanced, but not balanced state at single neuron resolution. The mapping from stimuli to firing rates is richly nonlinear in this state. We next explore how this nonlinearity improves the computational capability of the network.

Nonlinear representations in semi-balanced networks improve computations.

To quantify the computational capabilities of our spiking networks, we used a network identical to the one from Fig. 3 except we replaced the random stimulus, $\vec{Z}$, with a linear projection of pixel values from images in the MNIST data set (Fig. 4A, layer 1; see below for description of layer 2). Unlike the 2-dimensional stimuli considered previously, the images live in a 400-dimensional space (20×20 pixels).

We first trained inhibitory synaptic weights with iSTDP using 100 MNIST images presented for 1s each. We then presented 2000 images to the trained network and recorded the firing rates over each stimulus presentation. Applying the same Isomap and PCA analysis used above to these 2000 firing rate vectors confirms that the network implements a nonlinear representation of the images (Supplementary Figure S.2).

We wondered if the nonlinearity of this representation imparted computational advantages over a linear representation. The 10 different digits (0-9) form ten clusters of points in the 4000-dimensional space of layer 1 excitatory neuron firing rates. Similarly, the raw images represent ten clusters of points in the 400-dimensional pixel space. Are these clusters of points linearly separable?

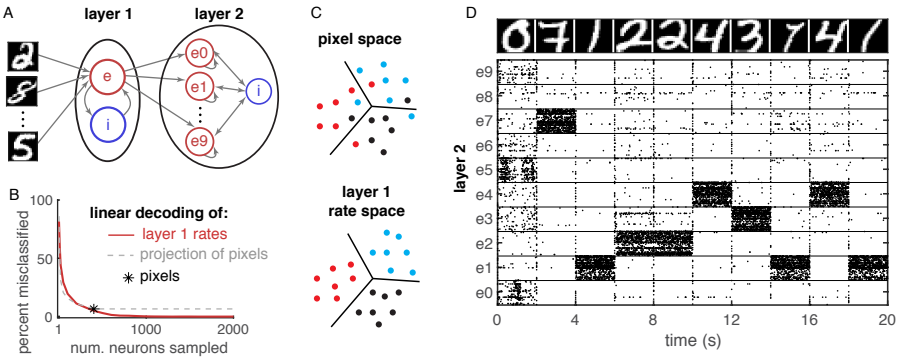

Figure 4: **Nonlinear representations in a semi-balanced network improve computations. A)** Network diagram. Pixel values provided external input to a semi-balanced network identical to the one in Fig. 3, representing layer 1. Layer 2 is a competitive, semi-balance network receiving external input from excitatory neurons in layer 1 with inter-laminar weights trained using a supervised Hebbian-like rule to classify the digits. **B)** Error rate (percent of 2000 images misclassified) of a linear readout of excitatory firing rates from layer 1 with readout weights optimized to classify the images, plotted as a function of the number, n, of neurons sampled (red). Black asterisk shows the error rate of an optimized readout of the $n = 400$ image pixels. Dashed gray shows the error rate of an optimized readout of a random projection of the pixels into n dimensions. The error rate of the rate readout (red curve) is zero for $n \geq 1600$. Hence, the digits are linearly separable in rate space, but not pixel space, which is only possible for nonlinear mappings. **C)** Diagram illustrating linear separability in rate space, but not pixel space. Different colors represent different digits and black lines are separating hyperplanes. **D)** Raster plot of 500 randomly selected neurons from layer 2 (50 from each population, ek) when images at top were provided as external input to layer 1.

To answer this, we trained an optimal linear readout of the 2000 firing rate vectors and found that the 10 different clusters of points in firing rate space were perfectly linearly separable. Specifically, we found a 10×4000 matrix, W_r, such that the 10-dimensional vector, $\vec{x} = W_r \vec{r}_e$, is maximized at the entry corresponding to the correct digit across all 2000 images. Here, $\vec{r}_e$ is the 4000×1 vector of excitatory neuron firing rates in layer 1.

For comparison, we used the same method to train an optimal linear readout of the 2000 raw MNIST images, treated as vectors in 400-dimensional pixel space. This analysis revealed that 6.6% of the images were misclassified (Fig. 4B, asterisk), implying that the digits are not linearly separable in pixel space. Hence, the digits are separable in rate space, but not in pixel space (Fig. 4C). Since the images are not linearly separable in pixel space, any linear representation of the images is not linearly separable. Hence, the separability in rate space is due to the nonlinearity of the neural representation.

We next investigated how many neurons or encoding dimensions were necessary to achieve linear separability. First, we trained an optimal linear readout on n randomly selected layer 1 excitatory neurons and computed the percentage of the 2000 images that were misclassified. The error decreased with n and perfect linear separation (zero error) was achieved for $n \geq 1600$ (Fig. 4B, red).

To compare this to pixel space, we projected each raw image randomly into n-dimensional space and trained a linear readout. The error of this readout for $n \leq 400$ was similar to the error in rate space (Fig. 4B, compare gray dashed to red). However, the error in pixel space saturated to 6.6% at $n = 400$ because a linear projection of pixels into a higher dimensional space cannot improve linear separability (Fig. 4B, gray dashed curve saturates at $n = 400$).

These results demonstrate that the nonlinearity of our network improves linear discriminability of stimuli, but they do not address how well the trained linear readout performs on images that were not used in training. Moreover, the readout weights have mixed sign and do not respect Dale's law. We next considered a downstream spiking network, layer 2, that receives synaptic input from excitatory neurons in layer 1 (Fig. 4A). Layer 2 has ten excitatory populations and one inhibitory population. Excitatory populations are coupled to themselves and bi-directionally with the inhibitory population, but do not connect to each other,

producing a competitive dynamic between the excitatory populations in layer 2.

Our goal was to train feedforward weights from excitatory neurons in layer 1 to those in layer 2 that are strictly positive and encourage the kth excitatory population in layer 2 to be most active when layer 1 receives the digit k as input. We used a simple, Hebbian like learning rule in which the weight from neuron i in layer 1 to neuron j in population ek of layer 2 is increased when neuron i is active during the presentation of digit k. This rule is not optimal, but maintains positive weights. We applied the rule to the same 2000 images mentioned above, then tested the performance of the learned weights on 200 images not previously presented to the network. In 72.5% of these 200 test images, the network guessed the correct digit in the sense that population ek in layer 2 had the highest firing rate when digit k was presented (Fig. 4D).

Discussion

We introduced the semi-balanced state, defined by an excess of inhibition without an excess of excitation. This state is realized naturally in networks for which the classical balanced state cannot be achieved and networks in this state implement nonlinear stimulus representations, which are not possible in classical balanced networks. We established a direct mathematical relationship between firing rates in semi-balanced networks, artificial neural networks, and the rich mathematical theory of threshold-linear networks. The semi-balanced state is realized at single-neuron resolution in networks with iSTDP, which implement high-dimensional nonlinear stimulus representations that improve the network's computational properties.

Previous work revealed multi-stability and nonlinear transformations at the level of population averages by balanced networks with short term synaptic plasticity [36]. Future work should consider how the non-linearities introduced by short term plasticity combine with the non-linearities introduced by semi-balance. Other work studied spike timing reliability and nonlinear representations at single-neuron resolution in non-plastic networks that satisfy balance at the level of population-averages [37, 38]. Since these studies did not implement iSTDP or similar mechanisms, our results suggest that their networks were not balanced at single-neuron resolution. Hence, these studies combined with our results support the general conclusion that while networks can only perform linear computations at the resolution over which they are balanced, they can perform non-linear computations at a finer resolution. A deeper mathematical understanding of this idea is a potential topic for future work.

An alternative theory of nonlinear computations in cortical circuits is given by the theory of SSNs with power-law f-I curves [28, 29, 27]. For large $\overline{JK}$, fixed point firing rates in these networks converge to the balanced fixed point, Eq. (3), when it is positive. At finite $\overline{JK}$, they implement nonlinearities that are not accounted for by Eq. (3). These nonlinearities are necessary to capture some experimentally observed response properties [29] and are distinct from the nonlinearities produced by semi-balance and discussed here. Indeed, fixed point firing rates in SSNs can be expanded in a series for which Eq. (3) is the first term [28]. This expansion is derived under the assumption that rates are positive, which implies that the nonlinearities produced by semi-balance are not present. Semi-balanced network theory (dashed) gives a piecewise-linear approximation of firing rates that captures the overall trends, but misses the curvature of firing rates as the stimulus changes (Fig. 2Bi, compare solid to dashed). Balanced network theory and the series expansion for SSNs are restricted to the regime in which all rates are positive. The fact that spiking networks and SSNs deviate from semi-balanced network theory in similar ways (solid and dashed differ similarly in Fig. 2Ai and Bi) suggests that SSNs can be used to refine the analysis of spiking network simulations beyond the more coarse-grained description provided by semi-balanced network theory. A key component of this analysis would be to generalize the series expansion for SSNs so that Eq. (4) is the first term instead of Eq. (3).

One limitation of our approach is that it focused on fixed point rates and did not consider their stability or the dynamics around fixed points. Previous work shows that balanced networks can exhibit spontaneous transitions between attractor states [39] which can be formed by iSTDP [23, 40]. Attractor states in those studies maintained strictly positive firing rates across populations, keeping the networks in the classical balanced state. This raises the question of whether similar attractors could arise in which some populations are silenced by excess inhibition, putting them in a semi-balanced state. Tools for studying these states

could potentially be developed from the mathematical theory of threshold-linear networks [19, 20, 21, 22].

Another limitation is that, in our network trained on MNIST digits, the recurrent connections in the network were only trained via an unsupervised iSTDP rule, which is agnostic to the image labels. Hence, the recurrent network did not learn a label-dependent representation of the stimuli. Moreover, recurrent excitatory weights were not trained. Gradient-descent based learning rules for excitatory weights are easy to derive using Eq. (4) since $\boldsymbol{r}$ depends linearly on $\boldsymbol{X}$ wherever $\boldsymbol{r}$ is positive and the gradient is zero elsewhere. Future work should consider excitatory synaptic plasticity in the recurrent network and supervised learning rules for recurrent weights to learn more informative representations.

We considered violations of the balanced state arising when Eq. (3) predicts negative rates. Previous work has shown that balance can also be broken when the connectivity matrix, W, in Eq. (3) is singular [9, 26, 11]. While singular W may at first seem contrived, it has been shown that singular or nearly singular W arise naturally when modeling structural heterogeneity of network architectures [9, 26], optogenetic stimulation [11], or connectivity structures that depend on continuous quantities like neuron distance or orientation tuning [10, 11]. In networks with singular W, Eq. (4) can admit a solution even when Eq. (3) does not. Hence, semi-balanced network theory can be applied to these models for which classical balanced network theory fails. Applying semi-balanced network theory to networks with spatially continuous connectivity structure would require extending the theory of spatially extended balanced networks [41, 10, 42, 11] to account for semi-balance, *i.e.*, for spatially localized regions of neurons with quenched firing rates, which could be a fruitful direction for future work.

The semi-balanced state is defined by an excess of inhibition without a corresponding excess of excitation. This is consistent with findings that inhibition dominates cortical responses in awake animals [35]. However, it should be noted that the dominance of inhibitory synaptic currents is reduced to some extent when shunting inhibition is accounted for (see Fig. 2Cii and Supplementary Figure S.1). A more precise prediction of our model is that time-varying stimuli will silence a subset of neurons through shunting inhibition and an effective imbalance between excitatory and inhibitory conductances (see Supplementary Figure S.1 and its caption). This is consistent with evidence that visual inputs evoke shunting inhibition in cat visual cortex [43]. These predictions should be tested more precisely using *in vivo* recordings.

Recurrent spiking neural networks are notoriously difficult to train in part because the mathematical analysis of firing rates in biologically realistic recurrent spiking neural networks is largely intractable, though some approximations have been developed for some models in some parameter regimes []. Hence, gradient-based methods for firing rates in recurrent spiking networks are difficult to derive because the firing rates themselves are unknown. The piecewise linearity of firing rates in the semi-balanced state (see Eq. (4)) could greatly simplify the training of recurrent spiking networks because the gradient of the firing rate with respect to the weights can be easily computed. Future work should consider the derivation of gradient-based learning rules from Eq. (4)

Artificial recurrent neural networks for machine learning often use sigmoidal activation functions instead of the rectified linear activations typically used in feedforward networks because the unboundedness of rectified linear units make recurrent networks susceptible to instabilities and large activations [18]. However, sigmoidal activations introduce the potential for vanishing gradients that can be problematic for training [18]. Our results suggest that a homeostatic learning rule akin to an iSTDP rule could help stabilize artificial recurrent neural networks with rectified linear activations while avoiding the problem of vanishing gradients.

In summary, semi-balanced networks are more biologically parsimonious and computationally powerful than widely studied balanced network models. The foundations of semi-balanced network theory presented here open the door to several directions for further research.

Methods

We modeled a network of N adaptive exponential integrate-and-fire (adaptive EIF) neurons with $0.8N$ excitatory neurons and $0.2N$ inhibitory neurons. For the current-based model used in all figures except Fig. 2B,C, the membrane

potential of neuron $j = 1, \ldots, N_a$ in population a obeyed

$$\tau_m \frac{dV_j^a}{dt} = -(V_j^a - E_L) + D_T e^{(V_j^a - V_T)/D_T} - w + I_j^a(t)$$
$$\tau_w \frac{dw_j^a}{dt} = -w_j^a$$

with the added condition that each time $V_j^a(t)$ crossed $V_{th} = -55$, a spike was recorded, it was reset to $V_{re} =$, and w_j^a was incremented by $B =$mV. A hard lower bound was imposed at $V_{lb} = -85$mV. Other neuron parameters were $\tau_m =$, $E_L =$, $D_T =$, $V_T =$, and $\tau_w =$. Input was given by

$$I_j^a(t) = \sum_b \sum_k J_{jk}^{ab} \sum_n \alpha_b(t - t_{k,n}^b)$$

where $t_{k,n}^b$ is the nth spike of neuron k in population b and $\alpha_b(t) = e^{-t/\tau_b}/\tau_b H(t)$ is an exponential postsynaptic current with $H(t)$ the Heaviside step function. Synaptic time constants, τ_b, were 8/4/10 ms for excitatory/inhibitory/external neurons. Synaptic weights were generated randomly and independently by

$$J_{jk}^{ab} = \begin{cases} j_{ab}/\sqrt{N} & \text{with probability } p_{ab} \\ 0 & \text{otherwise} \end{cases}.$$

In Fig. 1C,E and Fig. 2C, external input rates were $\boldsymbol{r}_x = [15\ 15]$Hz for the first 500ms and $\boldsymbol{r}_x = [15\ 30]$Hz for the next 500ms.

In Figs. 1 and 2, postsynaptic populations were $a = e1, e2, i$ and presynaptic populations were $b = e1, e2, i, x1, x2$ with $N_{e1} = N_{e2} = 1.2 \times 10^4$, $N_i = 6000$, and $N_{x1} = N_{x2} = 3000$ so that $N = N_{e1} + N_{e2} + N_i = 3 \times 10^4$. Neurons in external populations, $x1$ and $x2$, were not modeled directly, but spike times were generated as independent Poisson processes with firing rates r_{x1} and r_{x2}. Connection strength coefficients were $j_{ejek} = 37.5$, $j_{eji} = -225$, $j_{iek} = 168.75$, $j_{ii} = -375$, $j_{ejxk} = 2700$, and $j_{ixk} = 2025$mV/Hz for $j, k = 1, 2$. Note that these were scaled by $\sqrt{N}$ to get the actual synaptic weights as defined above. Connection probabilities in Fig. 1C,D were $p_{e1e1} = p_{e2e2} = 0.15$, $p_{e1e2} = p_{e2e1} = 0.05$, $p_{e1x1} = 0.08$, $p_{ix1} = p_{ix2} = 0.12$, and $p_{ab} = 0.1$ for all other connection probabilities. Connection probabilities in Fig. 1E,F and in Fig. 2 were the same except $p_{e1x1} = p_{e2x2} = 0.15$, $p_{e1x2} = p_{e2x1} = 0$, and $p_{ix1} = p_{ix2} = 0.15$.

For Fig. 2C,D, we used the model except

$$\tau_m \frac{dV_j^a}{dt} = -(V_j^a - E_L) + D_T e^{(V_j^a - V_T)/D_T} - w - g_{e,j}^a(t)(V - E_e) - g_{e,j}^a(t)(V - E_i)$$

where $E_e = 0$mV, $E_i = -75$mV,

$$g_{e,j}^a(t) = \sum_b \sum_k J_{jk}^{ab} \sum_n \alpha_b(t - t_{k,n}^b)$$

with the sum taken over excitatory presynaptic populations ($b = e1, e2, x1, x2$), and

$$g_{i,j}^a(t) = \sum_k J_{jk}^{ai} \sum_n \alpha_i(t - t_{k,n}^i).$$

The excitatory presynaptic weights (j_{ae1}, j_{ae2}, j_{ax1}, and j_{ax2}) were the same as above, but multiplied by $(E_e - V_0)$ to account for the change of units. Similarly, presynaptic weights (j_{ai}) were multiplied by $(E_i - V_0)$. We took $V_0 = V_T = -55$mV, but the accuracy of the theory did not depend sensitively on this choice. To obtain the dashed curves in Fig. 2Di, we used Eq. (4), but with the original values of W (those used for the current-based model). This is equivalent to dividing the conductance-based synaptic weights by $(E_e - V_0)$ and $(E_i - V_0)$, which is the approximation produced by a mean-field theory derived in previous work [31, 32, 11].

For Fig. 2B, we solved $\tau \dot{\boldsymbol{r}} = -\boldsymbol{r} + k\overline{JK}[W\boldsymbol{r} + \boldsymbol{X}]_+^2$ using the forward Euler method with $\boldsymbol{r} = [r_{e1}\ r_{e2}\ r_i]^T$, $\boldsymbol{X} = W_x \boldsymbol{r}_x$,

$$W = \begin{bmatrix} w_{e1e1} & w_{e1e2} & w_{e1i} \\ w_{e2e1} & w_{e2e2} & w_{e2i} \\ w_{ie1} & w_{ie2} & w_{ii} \end{bmatrix},$$

and

$$W_x = \begin{bmatrix} w_{e1x1} & w_{e1x2} \\ w_{e2x1} & w_{e2x2} \\ w_{ix1} & w_{ix2} \end{bmatrix}$$

where $w_{ab} = J_{ab}K_{ab}/\overline{JK} = j_{ab}p_{ab}N_b/\overline{JK}$. Note that $\overline{JK}$ is multiplied in the differential equation and divided in the definition of w_{ab}. We set $k = 44\text{Hz}^2/(\text{mV})^2$ which provided a rough match to the sample f-I curves in our spiking network while still exhibiting transitions between ISN and non-ISN regimes.

For Fig. 3, the model was the same as above except there was just one excitatory, one inhibitory, and one external population with $N_e = 0.8N$ and $N_i = N_x = 0.2N$ where $N = 3 \times 10^4$ in Fig. 3A,B. We reduced network size to $N = 5 \times 10^3$ for Fig. 3C because simulations for Fig. 3C required 289 simulations for 400s each. The long simulation time, 400s, was needed for accurate estimation of individual neuron's firing rates at each stimulus value, which requires a longer runtime than population averaged rates. The simulation for Fig. 3C took around 54 CPU hours and run time grows quadratically with N, so a simulation with $N = 3 \times 10^4$ would have taken prohibitively long. Stimulus coefficients in Fig. 3B were set to $\sigma_1 = \sigma_2 = 22.5$mV (about 1.4 times the rheobase) for the first 80s and randomly selected from a uniform distribution on $[-30, 30]$mV for the last 40s. In Fig. 3C, σ_1 and σ_2 values were sampled from a uniform 17×17 lattice on $[-18, 18] \times [-18, 18]$mV (-18mV to 18mV with a step size of 0.15 mV for each of σ_1 and σ_2). Connection probabilities between all populations in Fig. 3 were $p_{ab} = 0.1$. Initial synaptic weights were given by $j_{ee} = 37.5$, $j_{ei} = -225$, $j_{ie} = 168.75$, $j_{ii} = -375$, $j_{ex} = 2700$, and $j_{ix} = 2025$mV/Hz as above. Only inhibitory weights onto excitatory neurons (j_{ei}) changed, all others were plastic.

The inhibitory plasticity rule was taken directly from previous work [23]. The variables, $x_j^a(t)$, represent filtered spiking activity and are defined by $\tau_x dx_j^a/dt = -x_j^a$ with the added condition that $x_j^a(t)$ was incremented by one each time neuron j in population $a = e, i$ spiked. After each spike in excitatory neuron j, inhibitory synaptic connections onto that neuron were updated by $\Delta J_{jk}^{ei} = -\eta x_k^i(t)$ for all non-zero J_{jk}^{ei}. After each spike in inhibitory neuron, k, its outgoing synaptic connections were updated by $\Delta J_{jk}^{ei} = -\eta(x_j^e(t) - \alpha)$. We used $\tau_x = 200$ms and $\alpha = 2$ to get a "target rate" of $r_e^t = \alpha/(2\tau_x) = 5$Hz.

Layer 1 in Fig. 4 was identical to the model in Fig. 3C (with $N = 5000$) except the external input was replaced by $\vec{X}_i(t) = \overline{X}_i$ where $\overline{X}_i$ is the mean external input to inhibitory neurons in simulations with an external population (as in previous figures), so the time-varying input to inhibitory neurons was replaced by a time-constant input with the same mean. The external input to excitatory neurons was $\vec{X}_e(t) = \overline{X}_e + \vec{Z}$ where $\vec{Z} = Q\vec{x}$ where $\vec{x}$ is a 400×1 vector of pixel values in the presented MNIST digit and Q is a $N_e \times 400$ projection matrix where $N_e = 4000$. We constructed Q so that the kth pixel projected to 100 neurons, specifically to neuron indeices $j = 100(k-1)+1$ through $100k$ with strength σ. This corresponds to setting $Q_{jk} = \sigma$ for $100(k-1)+1 \le j \le 100k$ and $Q_{jk} = 0$ otherwise. We set $\sigma =?$.

We first trained the inhibitory synaptic weights by presenting 100 MNIST inputs for 1 s each with iSTDP turned on. We then froze the inhibitory weights and presented an additional 2000 MNIST digits for 10 s each and saved the resulting excitatory firing rates for each digit and each excitatory neuron. Weights were frozen for this simulation because the goal is to study the (fixed) representation of digits by the trained recurrent network.

To compute the optimal readout of firing rates from Layer 1, we defined a readout $Y = W_r R_1$ where $\vec{R}_1$ is the 4000×2000 matrix of the $N_e = 4000$ Layer 1 excitatory neuron firing rates for each of 2000 MNIST digit inputs, averaged over the 10 s that it was presented to the network. To train the 10×4000 readout matrix, W_r, we minimized the ℓ^2 (Euclidean) norm between the 10×2000 matrix, Y, and the binary matrix H for which $H(m, n) = 1$ only if digit $n = 1, \ldots, 2000$ was labeled with $m - 1 = 0, \ldots, 9$. In other words, H is a matrix of one-hot vectors encoding the labeled digit. Since the ℓ^2 loss is quadratic, the minimizing W_r can be found explicitly. Accuracy was then computed by checking if the maximum index of Y was at the correct digit, *i.e.*, by taking $\widetilde{Y}(m, n) = 1$ if $Y(m, n) \ge Y(m', n)$ for all $m = 1, \ldots, 10$. As reported in Results, we obtained perfect accuracy with this procedure, *i.e.*, we obtained $\widetilde{Y} = H$ exactly. To compute the optimal readout of pixel values, represented by an asterisk in Fig. 4B, we repeated these procedures except we used the 400×1 vector of pixel values in place of the 4000×1 vector of excitatory neuron firing rates. For the red curve in Fig. 4B, we performed the same procedure, but restricted to a randomly chosen subset of the 4000 excitatory neuron firing rates (subset size indicated on the horizontal axis). For the dashed gray curve in Fig. 4B, we used a random projection, $U\vec{x}$, of the pixel values where $\vec{x}$ is the 400×1 vector of pixel values and U is a $K \times 400$ matrix with K being the number on the horizontal axis of the plot.

Layer 2 in Fig. 4 had $N = 5000$ neurons. The inhibitory population contained $N_i = 1000$ neurons and there were ten excitatory populations each with 400 neurons. Neurons in the same excitatory population were connected with probability $p_{ejej} = 0.1$ and neurons in different excitatory populations were connected with probability $p_{ejek} = 0$ for $j \ne k$. Connection probabilities between the inhibitory population and each excitatory population were $p_{eji} = p_{iej} = 0.1$. Recurrent connection weights, j_{ab}, were the same as for all networks considered above. Layer 2 received feedforward input from Layer 1, *i.e.*, Layer 1 served as the external input population to Layer 2.

Connectivity from Layer 1 to Layer 2 was determined as follows. We first defined a 10×400 matrix, U, with entries $U_{mn} \ge 0$ representing connectivity from neurons in Layer 1 receiving input from pixel $k = n, \ldots, 400$ to

neurons in Layer 2 representing digit $m-1=0,\ldots,9$. We trained these weights on a simulation of Layer 1 with 2000 different MNIST digit inputs. For each digit, if the digit label was $m-1=0,\ldots,9$, we increased U_{mn} by the sum of all excitatory firing rates of neurons in Layer 1 receiving input from pixel m. In other words, $\Delta U_{mn}=\eta\vec{r}_1\cdot L$ where $\vec{r}_1$ is a vector of Layer 1 firing rates and $L=[0\ \cdots 1\ \cdots 0]$ is a 10×1 vector which is equal to 1 in the place of the labeled digit, *i.e.*, a one-hot vector [18]. We then normalized each column and row of U by its norm. This normalization makes the choice of η arbitrary, so we chose $\eta=1$. The 4000×4000 feedforward connection matrix, J^{21}, from excitatory neurons in Layer 1 to excitatory neurons in Layer 2 was then defined by $J^{21}_{jk}=U_{mn}$ where $m-1=0,\ldots,9$ is the population to which neuron $j=1,\ldots,4000$ belongs and $n=1,\ldots,400$ is the pixel from which neuron k receives input. Inhibitory neurons in Layer 2 did not receive feedforward synaptic input, only recurrent input. Since excitatory neurons in Layer 2 are only connected to other excitatory neurons within their population, but all excitatory populations connect reciprocally to the inhibitory population, this creates a winner-take-all dynamic in which the excitatory population with the strongest external input spikes at an elevated rate and suppresses other excitatory populations. Combined with the supervised Hebbian plasticity rule, this creates a dynamic where the network learns to activate population em when an image is presented that is similar to training images that were labeled with digit m. Fig. 4D and the accuracy reported in Results reflects spiking activity in Layer 2 after training of the feedforward weights is turned off.

Matlab code to produce all figures will be included with a revised submission. Until that time, code may be requested from the corresponding author via email.

Acknowledgments

This work was supported by NSF grants DMS-1654268 and Neuronex DBI-1707400.

References

[1] Haider, B., Duque, A., Hasenstaub, A.R., & McCormick, D.A. Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *J Neurosci* **26**, 4535–4545 (2006).

[2] Dorrn, A.L., Yuan, K., Barker, A.J., Schreiner, C.E., & Froemke, R.C. Developmental sensory experience balances cortical excitation and inhibition. *Nature* **465**, 932–936 (2010).

[3] Adesnik, H. & Scanziani, M. Lateral competition for cortical space by layer-specific horizontal circuits. *Nature* **464**, 1155 (2010).

[4] Okun, M. & Lampl, I. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat Neurosci* **11**, 535–537 (2008).

[5] Xue, M., Atallah, B.V., & Scanziani, M. Equalizing excitation-inhibition ratios across visual cortical neurons. *Nature* **511**, 596–600 (2014).

[6] Barral, J. & Reyes, A.D. Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics. *Nature neuroscience* **19**, 1690 (2016).

[7] van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724–1726 (1996).

[8] van Vreeswijk, C. & Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural Comput* **10**, 1321–1371 (1998).

[9] Landau, I.D., Egger, R., Dercksen, V.J., Oberlaender, M., & Sompolinsky, H. The impact of structural heterogeneity on excitation-inhibition balance in cortical networks. *Neuron* **92**, 1106–1121 (2016).

[10] Rosenbaum, R. & Doiron, B. Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys Rev X* **4**, 021039 (2014).

[11] Ebsch, C. & Rosenbaum, R. Imbalanced amplification: A mechanism of amplification and suppression from local imbalance of excitation and inhibition in cortical circuits. *PLoS computational biology* **14**, e1006048 (2018).

[12] Renart, A., et al. The Asynchronous State in Cortical Circuits. *Science* **327**, 587–590 (2010).

[13] Helias, M., Tetzlaff, T., & Diesmann, M. The correlation structure of local neuronal networks intrinsically results from recurrent dynamics. *PLoS Comput Biol* **10**, e1003428 (2014).

[14] Wimmer, K., et al. The dynamics of sensory integration in a hierarchical network explains choice probabilities in MT. *Nat Commun* **6**, 1–13 (2015).

[15] Rosenbaum, R., Smith, M.A., Kohn, A., Rubin, J.E., & Doiron, B. The spatial structure of correlated neuronal variability. *Nature Neurosci* **20**, 107 (2017).

[16] Darshan, R., van Vreeswijk, C., & Hansel, D. Strength of correlations in strongly recurrent neuronal networks. *Physical Review X* **8**, 031072 (2018).

[17] Dahmen, D., Grün, S., Diesmann, M., & Helias, M. Second type of criticality in the brain uncovers rich multiple-neuron dynamics. *Proceedings of the National Academy of Sciences* **116**, 13051–13060 (2019).

[18] Goodfellow, I., Bengio, Y., & Courville, A. *Deep learning* (MIT press, 2016).

[19] Hahnloser, R.H. & Seung, H.S. Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in neural information processing systems*, 217–223 (2001).

[20] Xie, X., Hahnloser, R.H., & Seung, H.S. Selectively grouping neurons in recurrent networks of lateral inhibition. *Neural computation* **14**, 2627–2646 (2002).

[21] Curto, C. & Morrison, K. Pattern completion in symmetric threshold-linear networks. *Neural computation* **28**, 2825–2852 (2016).

[22] Curto, C., Geneson, J., & Morrison, K. Fixed points of competitive threshold-linear networks. *Neural computation* **31**, 94–155 (2019).

[23] Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C., & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–73 (2011).

[24] Froemke, R.C. Plasticity of cortical excitatory-inhibitory balance. *Annual Review of Neuroscience* **38**, 195–219 (2015).

[25] Hennequin, G., Agnes, E.J., & Vogels, T.P. Inhibitory Plasticity: Balance, Control, and Codependence. *Annu. Rev. Neurosci.* **40**, 557–579 (2017).

[26] Pyle, R. & Rosenbaum, R. Highly connected neurons spike less frequently in balanced networks. *Phys Rev E* **93**, 040302 (2016).

[27] Ahmadian, Y. & Miller, K.D. What is the dynamical regime of cerebral cortex? *arXiv* (2020).

[28] Ahmadian, Y., Rubin, D.B., & Miller, K.D. Analysis of the stabilized supralinear network. *Neural computation* **25**, 1994–2037 (2013).

[29] Rubin, D.B., Hooser, S.D.V., & Miller, K.D. The stabilized supralinear network : A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* **85**, 1–51 (2015).

[30] Dayan, P. & Abbott, L.F. *Theoretical Neuroscience* (Cambridge, MA: MIT Press, 2001).

[31] Destexhe, A., Rudolph, M., & Paré, D. The high-conductance state of neocortical neurons in vivo. *Nature reviews neuroscience* **4**, 739 (2003).

[32] Kuhn, A., Aertsen, A., & Rotter, S. Neuronal integration of synaptic input in the fluctuation-driven regime. *Journal of Neuroscience* **24**, 2345–2356 (2004).

[33] Saxena, S. & Cunningham, J.P. Towards the neural population doctrine. *Current opinion in neurobiology* **55**, 103–111 (2019).

[34] Vogels, T.P. & Abbott, L.F. Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nat Neurosci* **12**, 483–491 (2009).

[35] Haider, B., Häusser, M., & Carandini, M. Inhibition dominates sensory responses in the awake cortex. *Nature* **493**, 97–100 (2013).

[36] Mongillo, G., Hansel, D., & Van Vreeswijk, C. Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission. *Physical review letters* **108**, 158101 (2012).

[37] Lajoie, G., Lin, K.K., Thivierge, J.P., & Shea-Brown, E. Encoding in balanced networks: Revisiting spike patterns and chaos in stimulus-driven systems. *PLoS computational biology* **12**, e1005258 (2016).

[38] Lajoie, G., Lin, K.K., & Shea-Brown, E. Chaos and reliability in balanced spiking networks with temporal drive. *Physical Review E* **87**, 052901 (2013).

[39] Litwin-Kumar, A. & Doiron, B. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature neuroscience* **15**, 1498 (2012).

[40] Litwin-Kumar, A. & Doiron, B. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature communications* **5**, 5319 (2014).

[41] Ben-Yishai, R., Bar-Or, R.L., & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci USA* **92**, 3844–3848 (1995).

[42] Lim, S. & Goldman, M.S. Balanced cortical microcircuitry for spatial working memory based on corrective feedback control. *J Neurosci.* **34**, 6790–6806 (2014).

[43] Borg-Graham, L.J., Monier, C., & Fregnac, Y. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* **393**, 369 (1998).