The correlated state in balanced neuronal networks

Cody Baker,[1] Christopher Ebsch,[1] Ilan Lampl,[2] and Robert Rosenbaum[1,3]

[1] *Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA*

[2] *Department of Neurobiology, Weizmann Institute of Science, Rehovot, 7610001, Israel*

[3] *Interdisciplinary Center for Network Science and Applications, University of Notre Dame, Notre Dame, IN, USA*

(Dated: August 17, 2018)

Understanding the magnitude and structure of inter-neuronal correlations and their relationship to synaptic connectivity structure is an important and difficult problem in computational neuroscience. Early studies show that neuronal network models with excitatory-inhibitory balance naturally create very weak spike train correlations. Later work showed that, under some connectivity structures, balanced networks can produce larger correlations between some neuron pairs, even when the average correlation is very small. All of these previous studies assume that the local neuronal network receives feedforward synaptic input from a population of uncorrelated spike trains. We show that when spike trains providing feedforward input are correlated, the downstream recurrent neuronal network produces much larger correlations. We provide an in-depth analysis of the resulting "correlated state" in balanced networks and show that, unlike the asynchronous state of previous work, it produces a tight excitatory-inhibitory balance consistent with in vivo cortical recordings.

I. INTRODUCTION

Correlations between the spiking activity of cortical neurons have important consequences for neural dynamics and coding [1–3]. A quantitative understanding of how spike train correlations are generated and shaped by the connectivity structure of neural circuits is made difficult by the noisy and nonlinear dynamics of recurrent neuronal network models [4–7]. Linear response and related techniques have been developed to overcome some of these difficulties [8–14], but their accuracy typically require an assumption of sparse and/or weak connectivity and, in some models, an additional assumption that neurons receive uncorrelated, feedforward Gaussian white noise input. However, cortical circuits are densely connected and receive spatially and temporally correlated synaptic input from outside the local circuit [15–18].

An alternative approach to analyzing correlated variability in recurrent neuronal network models is motivated in part by the widely observed balance between excitatory and inhibitory synaptic inputs in cortex [19–26]. When synaptic weights are scaled like $1/\sqrt{N}$ where N is the size of a model network, a cortex-like balance between excitation and inhibition arises naturally at large network size, which defines the "balanced state" [27, 28]. Early work on balanced networks assumed sparse connectivity to produce weak spike train correlations, but it was later shown that keeping connection probabilities $\mathcal{O}(1)$ naturally produces weak, $\mathcal{O}(1/N)$, spike train correlations, defining the "asynchronous state" [29]. While these extremely weak spike train correlations are consistent with some cortical recordings [30], the magnitude of correlations in cortex can depend on stimulus, cortical area, layer, and behavioral or cognitive state, and can be much larger than predicted by the asynchronous state [6, 31–35]. This raises the question of how larger correlation magnitudes can arise in balanced cortical circuits. Later theoretical work showed that larger correlations can be obtained between some cell pairs in densely connected networks with specially constructed connectivity structure [36–39], offering a potential explanation of the larger correlations often observed in recordings. These previous theoretical studies of correlated variability in balanced networks assume that the recurrent network receives feedforward synaptic input from an external population of uncorrelated spike trains, so feedforward input correlations arise solely from overlapping feedforward synaptic projections. In reality, feedforward synaptic input to a cortical population comes from thalamic projections, other cortical areas, or other cortical layers in which spike trains could be correlated.

We extend the theory of densely connected balanced networks to account for correlations between the spike trains of neurons in an external, feedforward input layer. We show that correlations between the feedforward synaptic input to neurons in the recurrent network are $\mathcal{O}(N)$ in this model, but cancel with $\mathcal{O}(N)$ correlations between recurrent synaptic input to produce $\mathcal{O}(1)$ total input correlation and $\mathcal{O}(1)$ spike train correlations on average, defining what we refer to as the "correlated state" in densely connected balanced networks. This correlated state offers an alternative explanation for the presence of moderately large spike train correlations in cortical recordings. We derive a simple, closed form approximation for the average cross-spectral density between neurons' spike trains in the correlated state in term of synaptic parameters alone, without requiring the use of linear response theory or any other knowledge of neurons' transfer functions. We show that the tracking of excitatory synaptic input currents by inhibitory currents is more precise and more similar to in vivo recordings [22] in the correlated state than in the asynchronous state. We also investigate the applicability of linear response approximations to correlated variability in densely connected balanced networks. Taken together, our results extend the theory of correlated variability in

balanced networks to the biologically realistic assumption that presynaptic neural populations are themselves correlated.

II. MODEL AND BACKGROUND

We consider recurrent networks of N integrate-and-fire model neurons, N_e of which are excitatory and N_i inhibitory. Neurons are randomly and recurrently interconnected and also receive random feedforward synaptic input from an external population of N_x neurons whose spike trains are homogeneous Poisson processes with rate r_x (Fig. 1A).

The membrane potential of neuron j in population $a = e, i$ obeys the exponential integrate-and-fire (EIF) dynamics

$$C_m \frac{dV_j^a}{dt} = -g_L(V_j^a - E_L) + D_T e^{(V_j^a - V_T)/D_T} + T_j^a(t)$$

with the added condition that each time $V_j^a(t)$ exceeds V_{th}, it is reset to V_{re} and a spike is recorded. We additionally set a lower bound on the membrane potential at $V_{lb} = -100$mV. Spike trains are represented as a sum of Dirac delta functions,

$$S_j^a(t) = \sum_n \delta(t - t_n^{a,j}),$$

where $t_n^{a,j}$ is the nth spike time of neuron j in population $a = e, i, x$. The total synaptic input current to neuron j in population $a = e, i$ is decomposed as

$$T_j^a(t) = E_j^a(t) + I_j^a(t) + X_j^a(t)$$

where

$$U_j^a(t) = \sum_{k=1}^{N_b} J_{jk}^{ab}(\alpha_b * S_j^b)(t) \tag{1}$$

for $U = E, I, X$ and $b = e, i, x$ respectively where $*$ denotes convolution, J_{jk}^{ab} is the synaptic weight from neuron k in population b to neuron j in population a, and $\alpha_b(t)$ is a postsynaptic current (PSC) waveform. Without loss of generality, we assume that $\int \alpha_b(t) = 1$. We use $\alpha_b(t) = \tau_b^{-1} e^{-t/\tau_b} H(t)$ where $H(t)$ is the Heaviside step function, though our results do not depend sensitively on the precise neuron model or PSC kernel used. For calculations, it is useful to decompose the total synaptic input into its recurrent and external sources,

$$T_j^a(t) = R_j^a(t) + X_j^a(t)$$

where

$$R_j^a(t) = E_j^a(t) + I_j^a(t)$$

is the recurrent synaptic input from the local circuit.

Local cortical circuits contain a large number of neurons and individual cortical neurons receive synaptic input from thousands of other neurons within their local circuit and from other layers or areas. Densely connected balanced networks have been proposed to model such large and densely interconnected neuronal networks [29, 38]. In such models, one considers the limit of large N (with N_x, N_e and N_i scaled proportionally) with fixed connection probabilities and where synaptic weights are scaled like $\mathcal{O}(1/\sqrt{N})$ [28, 29]. This scaling naturally captures the balance of mean excitatory and mean inhibitory synaptic input, as well as the tracking of excitation by inhibition, observed in cortical recordings [29]. In particular, we consider a random connectivity structure in which

$$J_{jk}^{ab} = \frac{1}{\sqrt{N}} \begin{cases} j_{ab} & \text{with probability } p_{ab} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where connections are statistically independent and $j_{ab}, p_{ab} \sim \mathcal{O}(1)$ for $b = e, i, x$ and $a = e, i$. We furthermore define the proportions

$$q_b = \frac{N_b}{N}$$

which are assumed $\mathcal{O}(1)$. For all examples we consider, $q_e = 0.8$ and $q_i = q_x = 0.2$.

We next introduce notational conventions for quantifying the statistics of spike trains and synaptic inputs in the network. The mean firing rates of neurons in population $a = e, i, x$ is defined by r_a for $a = e, i, x$ and it is useful to define the 2×1 vector, $\boldsymbol{r} = [r_e \;\; r_i]^T$. The mean is technically interpreted as the expectation over realizations of the network connectivity, but for large N it is approximately equal to the sample mean over all neurons the network. Similarly, mean-field synaptic inputs to neurons in populations $a = e, i$ are defined by

$$\overline{U}_a = \text{mean}_j U_j^a(t)$$

for $U = E, I, X, R, T$ and, in vector form, $\overline{\boldsymbol{U}} = [U_e \;\; U_i]^T$

For quantifying correlated variability, we use the cross-spectral density (CSD)

$$\langle U_j^a, Z_k^b \rangle(f) = \int_{-\infty}^{\infty} C_{U_j^a Z_k^b}(\tau) e^{-2\pi i f \tau} d\tau$$

between $U_j^a(t)$ and $Z_k^b(t)$ for $U, Z = E, I, X, S, R, T$ and $a, b = e, i, x$ where

$$C_{U_j^a, Z_k^b}(\tau) = \text{cov}(U_j^a(t), Z_k^b(t + \tau))$$

is the cross-covariance function. The argument, f, is the frequency at which the CSD is evaluated. The CSD is a convenient measure of correlated variability because it simplifies mathematical calculations due to the fact that it is a Hermitian operator and because most commonly used measures of correlated variability can be written as a function of the CSD. For example, the cross-covariance

is the inverse Fourier transform of the CSD. Spike count covariances over large time windows can be written in terms of the CSD by first noting that the spike count is an integral of the spike train [4],

$$\text{spike count over } [0,t_0] = \int_0^{t_0} S_j^a(t)dt.$$

For large t_0, the cross-spectrum between two integrals is related to the zero-frequency CSD,

$$\lim_{t_0\to\infty} \frac{1}{t_0}\text{cov}\left(\int_0^{t_0} U_j^a(t)dt, \int_0^{t_0} Z_k^b(t)dt\right) = \langle U_j^a, Z_k^b\rangle(f=0). \tag{3}$$

Hence, the spike count covariance over the interval $[0,t_0]$ is given asymptotically by $t_0\langle S_j^a, S_k^b\rangle(f=0)$ for large t_0. Following this result, we quantify covariability between spike trains and between synaptic currents using the zero-frequency CSD, which we estimate by taking the covariance between integrals as in Eq. (3) using $t_0 = 250$ms. This provides a simple, easily estimated quantity for quantifying covariance.

Most of our computations are performed at the level of population averages, so we define

$$\langle U_a, Z_b\rangle = \text{mean}_{j\neq k}\langle U_j^a, Z_k^b\rangle.$$

which is a scalar function of frequency, f, for each $a,b = e,i,x$ and $U,Z = E,I,X,S,R,T$. It is also convenient to define the 2×2 mean-field matrix form,

$$\langle \boldsymbol{U},\boldsymbol{Z}\rangle = \begin{bmatrix} \langle U_e, Z_e\rangle & \langle U_e, Z_i\rangle \\ \langle U_i, Z_e\rangle & \langle U_i, Z_i\rangle \end{bmatrix}.$$

for $\boldsymbol{U},\boldsymbol{Z} = \boldsymbol{E},\boldsymbol{I},\boldsymbol{X},\boldsymbol{S},\boldsymbol{R},\boldsymbol{T}$. We also define the recurrent and feedforward mean-field connectivity matrices,

$$W = \begin{bmatrix} w_{ee} & w_{ei} \\ w_{ie} & w_{ii} \end{bmatrix} \text{ and } W_x = \begin{bmatrix} w_{ex} \\ w_{ix} \end{bmatrix}$$

where $w_{ab}(f) = p_{ab}j_{ab}q_b\widetilde{\eta}_b(f) \sim \mathcal{O}(1)$ with $\widetilde{\eta}_b(f)$ the Fourier transform of $\eta_b(t)$. For the exponential kernels we use, $\widetilde{\eta}_b(f) = 1/(1+2\pi i f\tau_b)$. The zero-frequency values, $\overline{w}_{ab} = w_{ab}(0) = p_{ab}j_{ab}q_b$, define time-averaged interactions and mean-field connection matrices, $\overline{W} = W(0)$ and $\overline{W}_x = W_x(0)$.

This choice of notation allows us to perform computations on mean-field spike train and input statistics in a mathematically compact way. To demonstrate this, we first review the mean-field analysis of firing rates in the balanced state [27, 28, 40–42]. Mean external input is given by $\overline{\boldsymbol{X}} = \sqrt{N}\,\overline{W}_x r_x$ and mean recurrent input by $\overline{\boldsymbol{R}} = \sqrt{N}\,\overline{W}\boldsymbol{r}$ so that mean total synaptic input is given by

$$\overline{\boldsymbol{T}} = \sqrt{N}\left[\overline{W}\boldsymbol{r} + \overline{W}_x r_x\right].$$

In the balanced state, $\overline{\boldsymbol{T}}, \boldsymbol{r} \sim \mathcal{O}(1)$, which can only be obtained by a cancellation between external and recurrent synaptic inputs. This cancellation requires $\overline{W}\boldsymbol{r} + \overline{W}_x r_x \sim \mathcal{O}(1/\sqrt{N})$ so that [27, 28, 40–42]

$$\lim_{N\to\infty} \boldsymbol{r} = -\overline{W}^{-1}\overline{W}_x r_x \tag{4}$$

in the balanced state. Hence, the balanced state can only be realized when this solution has positive entries, $r_e, r_i > 0$, which requires that [27, 28, 40] $\overline{X}_e/\overline{X}_i > w_{ei}/w_{ii} > w_{ee}/w_{ie}$. Below, we perform analogous derivations of mean-field CSDs in balanced networks.

III. A REVIEW OF THE ASYNCHRONOUS BALANCED STATE

We first review prior theoretical work that derives mean-field CSDs when spike trains in the external population are uncorrelated Poisson processes (Fig. 1A,B), so

$$\langle S_x, S_x\rangle = 0.$$

Since the derivations of these results [38] and analogous derivations for networks of binary neuron models [29] are presented elsewhere and since the derivations are similar to those presented below, we only review the results here and give the details of the derivation in Appendix A.

Since spike trains in the external population are uncorrelated, correlations between the external input to neurons in the recurrent network arise solely from overlapping feedforward synaptic projections with [29, 38] (see Appendix A)

$$\langle \boldsymbol{X},\boldsymbol{X}\rangle = q_x^{-1}W_x r_x W_x^* \sim \mathcal{O}(1). \tag{5}$$

where W_x^* is the conjugate transpose of W_x. If the spike trains in the external layer were uncorrelated, but not Poisson, then r_x in this equation could be replaced by the power spectral density of the spike trains.

It would at first seem that this $\mathcal{O}(1)$ external input correlation would lead to $\mathcal{O}(1)$ correlations between neurons' spike trains, but this is prevented by a cancellation between positive and negative sources of input correlation. In particular, correlations between neurons' recurrent synaptic inputs, $\langle \boldsymbol{R},\boldsymbol{R}\rangle$, are also positive and $\mathcal{O}(1)$, but these positive sources of input correlations are canceled by negative correlations between neurons' recurrent and external inputs, $\langle \boldsymbol{X},\boldsymbol{R}\rangle$, in such a way that the total synaptic input correlation is weak,

$$\langle \boldsymbol{T},\boldsymbol{T}\rangle = \langle \boldsymbol{X},\boldsymbol{X}\rangle + \langle \boldsymbol{X},\boldsymbol{R}\rangle + \langle \boldsymbol{R},\boldsymbol{X}\rangle + \langle \boldsymbol{R},\boldsymbol{R}\rangle \sim \mathcal{O}(1/N)$$

where $\langle \boldsymbol{R},\boldsymbol{X}\rangle = \langle \boldsymbol{X},\boldsymbol{R}\rangle^*$. This cancellation is realized when the mean-field CSD between spike trains satisfies [38]

$$\langle \boldsymbol{S},\boldsymbol{S}\rangle = \frac{1}{N}W^{-1}\langle \boldsymbol{X},\boldsymbol{X}\rangle W^{-*} + \frac{1}{N}C_0 + o(1/N) \tag{6}$$

where $N\times o(1/N) \to 0$ as $N\to\infty$ and W^{-*} is the inverse of W^*. The first term in this equation represents spike

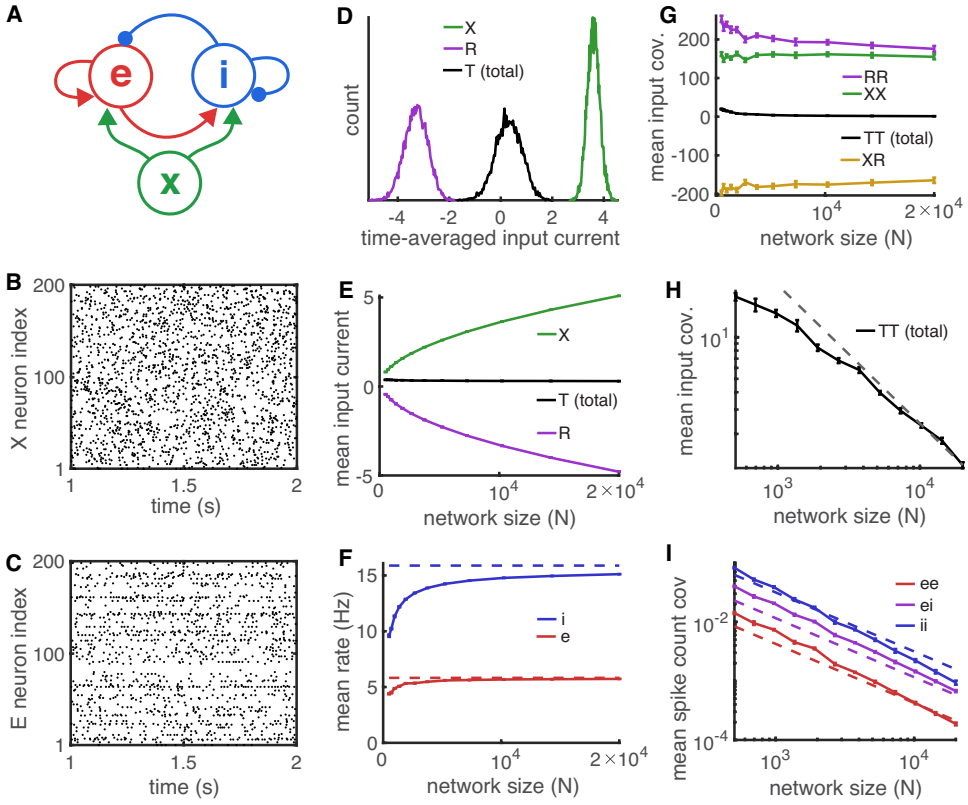

FIG. 1. **The asynchronous state in densely connected balanced networks. A)** Network diagram. An external population, X, of uncorrelated Poisson processes provides feedforward input to a randomly connected recurrent network of excitatory, E, and inhibitory, I, neurons. Feedforward input correlations are solely from overlapping projections from X. **B,C)** Raster plot of 200 randomly selected neurons from population X and E respectively. **D)** Histogram of external (X, green) recurrent ($R = E + I$, purple) and total ($T = X + E + I$, black) to all excitatory neurons. Currents here and elsewhere are reported in units $C_m V/s$ where C_m is the arbitrary membrane capacitance. **E)** Mean external (green), recurrent (purple), and total (black) input to excitatory neurons for networks of different sizes, N. **F)** Mean excitatory (red) and inhibitory (blue) neuron firing rates for difference network sizes. Solid curves are from simulations and dashed curves are from Eq. (4). **G)** Mean covariance between pairs of excitatory neurons' external inputs (green), recurrent inputs (purple), total inputs (black), and mean covariance between the recurrent input to one excitatory neuron and external input to the other (yellow) for different network sizes. Covariances were computed by integrating the inputs over 250ms windows then computing covariances between the integrals, which is proportional to zero-frequency CSD and has a closer relationship with spike count covariance (see Eq. (3) and surrounding discussion). Integrated currents have units $C_m mV$, so their covariances have units $C_m^2 mV^2$. **H)** Zoomed in view of black curve from E on a log-log axis (mean total input covariance, black) plotted alongside the function c/N (dashed gray) where c was chosen so that the two curves match at the largest N value. **I)** Mean spike count covariance between excitatory neuron spike trains (red), between inhibitory neuron spike trains (blue), and between excitatory-inhibitory pairs of spike trains (purple). Counts were computed over 250ms time windows. Solid curves are from simulations, dashed from Eq. (7) evaluated at zero frequency. Network size was $N = 10^4$ in B-D.

train correlations inherited from external inputs, namely $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$. The second term, C_0, represents correlations generated intrinsically by chaotic or chaos-like dynamics in the network [28, 38, 40, 43]. For example, a network with deterministic, constant external input, $X_j^a(t) = \overline{X}_a$, generates correlated variability [40] despite the fact that such networks are deterministic once an initial condition is specified and therefore $\langle \boldsymbol{X}, \boldsymbol{X} \rangle = 0$ for such networks. While an exact expression for C_0 is unknown, we show empirically below that its effects are small compared to correlations inherited from external input, at least at low frequencies for the network parameters that we consider. Therefore, the following approximation is accurate

$$\langle \boldsymbol{S}, \boldsymbol{S} \rangle \approx \frac{1}{N} W^{-1} \langle \boldsymbol{X}, \boldsymbol{X} \rangle W^{-*}. \tag{7}$$

To demonstrate these results, we first simulated a network of $N = 10^4$ randomly and recurrently connected neurons receiving feedforward input from a population of $N_x = 2000$ uncorrelated Poisson-spiking neurons (Fig. 1A,B). As predicted, spiking activity in the recurrent network was asynchronous and irregular (Fig. 1C; mean spike count correlation between neurons with rates at least 1 Hz was 5.2×10^{-4}) with approximate balance between external (X) and recurrent (R) synaptic input

sources (Fig. 1D). Varying the network size, N, demonstrates the $\mathcal{O}(\sqrt{N})$ growth of mean external ($\overline{\boldsymbol{X}}$) and recurrent ($\overline{\boldsymbol{R}}$) synaptic input currents that cancel to produce $\mathcal{O}(1)$ mean total input current ($\overline{\boldsymbol{T}}$) (Fig. 1E), as predicted by the mean-field theory of balance. As a result, firing rates converge to the limiting values predicted by Eq. (4) (Fig. 1F).

As predicted by the analysis of the asynchronous state, the mean covariances between individual sources of synaptic inputs appear $\mathcal{O}(1)$ (Fig. 1G), but cancel to produce much smaller, $\mathcal{O}(1/N)$, total input covariance (Fig. 1G,H). Mean spike count covariances also appear $\mathcal{O}(1/N)$ and show good agreement with the closed form approximation from Eq. (7) (Fig. 1I).

The fact that the approximation in Eq. (7) accurately captures the scaling of spike count covariances from simulations implies that correlated variability inherited by overlapping synaptic inputs dominate intrinsic correlations, C_0, which are ignored in Eq. (7). Similar results were found in previous work [38]. This contrasts with previous findings in networks of binary neurons, in which intrinsic variability appeared to dominate [44].

IV. THE CORRELATED STATE IN BALANCED NETWORKS

Above, we analyzed correlated variability when spike trains in the external population were uncorrelated, $\langle S_x, S_x\rangle = 0$, which produced asynchronous spiking in the recurrent network, $\langle \boldsymbol{S}, \boldsymbol{S}\rangle \sim \mathcal{O}(1/N)$. We now relax this assumption by considering moderate correlations between neurons in the external layer (Fig. 2A),

$$\langle S_x, S_x\rangle \sim \mathcal{O}(1).$$

Most previous work analyzing spike train correlations in recurrent networks relies on knowledge of the "correlation susceptibility" or "transfer" function, which quantifies the mapping from synaptic input covariance to spike train covariance, *e.g.*, the mapping from $\langle T_e, T_e\rangle$ to $\langle S_e, S_e\rangle$. However, susceptibility functions depend sensitively on the neuron model being used and their derivation typically relies on diffusion approximations that assume neurons' input currents can be approximated by Gaussian white noise and are weakly correlated [8, 11]. These assumptions are not valid for the densely connected, correlated networks with exponentially-decaying PSC kernels we consider here.

Instead of assuming knowledge of a covariance transfer or susceptibility function, our derivation relies only on an assumption that transfer of mean-field covariance is $\mathcal{O}(1)$, specifically that

$$\frac{\langle S_a, U_b\rangle}{\langle T_a, U_b\rangle} \sim \mathcal{O}(1). \tag{8}$$

In other words, mean-field spiking statistics are not drastically different in magnitude than mean-field input statistics because neurons transfer inputs to outputs in an $\mathcal{O}(1)$ fashion. Importantly, we do not need to know the value of the fraction in Eq. (8).

In addition to this assumption, all of our derivations follow from a few simple arithmetical rules that rely on the bilinearity of the operator $\langle\cdot,\cdot\rangle$. Specifically (see Appendix A for derivations),

$$\langle \boldsymbol{T}, \boldsymbol{U}\rangle = \langle \boldsymbol{R}, \boldsymbol{U}\rangle + \langle \boldsymbol{X}, \boldsymbol{U}\rangle \tag{9}$$

$$\langle \boldsymbol{R}, \boldsymbol{U}\rangle = \sqrt{N} W \langle \boldsymbol{S}, \boldsymbol{U}\rangle \tag{10}$$

$$\langle \boldsymbol{X}, \boldsymbol{U}\rangle = \sqrt{N} W_x \langle S_x, \boldsymbol{U}\rangle \tag{11}$$

$$\langle \boldsymbol{U}, \boldsymbol{Z}\rangle = \langle \boldsymbol{Z}, \boldsymbol{U}\rangle^* \tag{12}$$

for any $\boldsymbol{U}, \boldsymbol{Z} = \boldsymbol{E}, \boldsymbol{I}, \boldsymbol{X}, \boldsymbol{R}, \boldsymbol{S}, S_x, \boldsymbol{T}$ where $\boldsymbol{A}^*$ is the conjugate-transpose of $\boldsymbol{A}$ and where we omit smaller order terms here and below (see Appendix A for details). Eq. (9) follows from the fact that total input is composed of recurrent and external sources, $\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{X}$. Eqs. (10) and (11) follow from the fact that recurrent and external inputs are composed of linear combinations of $\mathcal{O}(N)$ spike trains, *c.f.* Eq. (1), and that synaptic weights are $\mathcal{O}(1/\sqrt{N})$. Eq. (12) is simply a property of the Hermitian cross-spectral operator.

We first derive the CSD between external inputs to neurons in the recurrent network. Applying of Eqs. (11) and (12) gives

$$\begin{aligned}\langle \boldsymbol{X}, \boldsymbol{X}\rangle &= \sqrt{N} W_x \langle S_x, \boldsymbol{X}\rangle \\ &= N W_x \langle S_x, S_x\rangle W_x^* \\ &\sim \mathcal{O}(N)\end{aligned}$$

Hence, $\mathcal{O}(1)$ covariance between the spike trains in the external population induces $\mathcal{O}(N)$ covariance between the external input currents to neurons in the recurrent network. This is a result of the effects of "pooling" on correlations and covariances, namely that the covariance between two sums of N correlated random variables is typically $\mathcal{O}(N)$ times larger than the covariances between the individual summed variables [29, 45, 46].

We next derive the CSD between spike trains and external inputs. First note that

$$\begin{aligned}\langle \boldsymbol{T}, \boldsymbol{X}\rangle &= \langle \boldsymbol{R}, \boldsymbol{X}\rangle + \langle \boldsymbol{X}, \boldsymbol{X}\rangle \\ &= \sqrt{N} W \langle \boldsymbol{S}, \boldsymbol{X}\rangle + \langle \boldsymbol{X}, \boldsymbol{X}\rangle.\end{aligned} \tag{13}$$

However, it follows from our assumption that neuronal transfer is $\mathcal{O}(1)$ (see Eq. (8)) that $\langle \boldsymbol{T}, \boldsymbol{X}\rangle \sim \langle \boldsymbol{S}, \boldsymbol{X}\rangle$. Therefore, we have

$$\langle \boldsymbol{S}, \boldsymbol{X}\rangle \sim \sqrt{N} W \langle \boldsymbol{S}, \boldsymbol{X}\rangle + \langle \boldsymbol{X}, \boldsymbol{X}\rangle,$$

which is only consistent if there is a cancellation between the two terms on the right hand side. Specifically, we must have that

$$\langle \boldsymbol{S}, \boldsymbol{X}\rangle = -\frac{1}{\sqrt{N}} W^{-1} \langle \boldsymbol{X}, \boldsymbol{X}\rangle \sim \mathcal{O}(\sqrt{N})$$

since $\langle \boldsymbol{X}, \boldsymbol{X}\rangle \sim \mathcal{O}(N)$ from above where we have omitted terms smaller than $\mathcal{O}(\sqrt{N})$. We can now calculate the

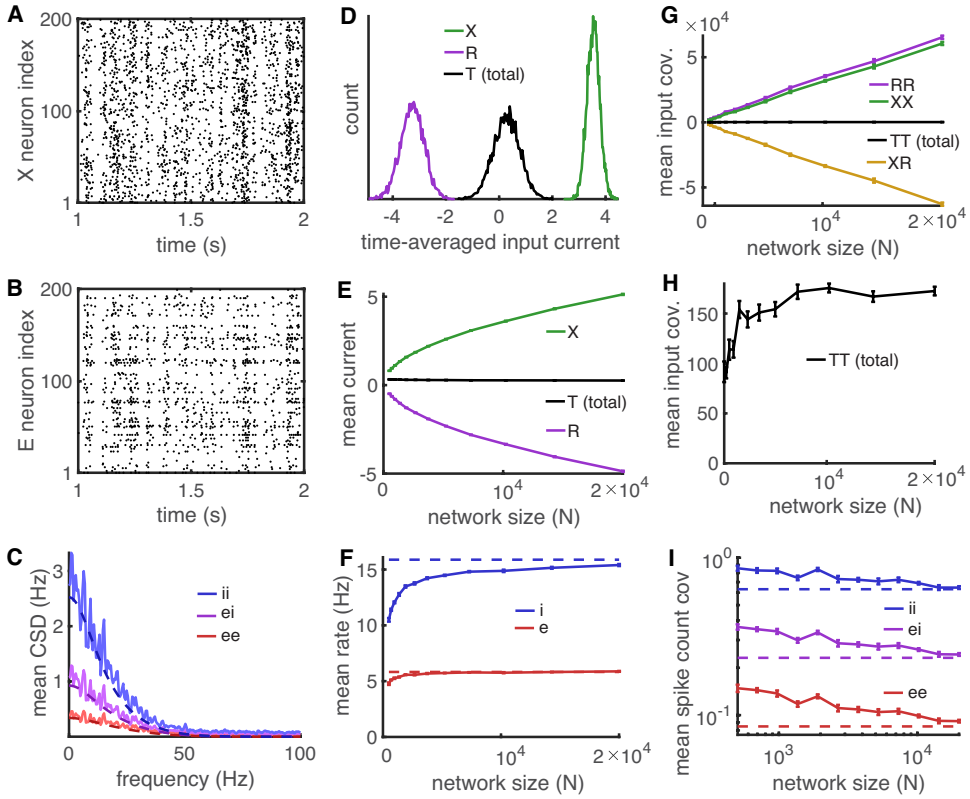

FIG. 2. **The correlated state in densely connected balanced networks. A,B)** Same as Fig. 1 B,C except spike trains in the external population, X, were correlated Poisson processes with spike count correlation $c = 0.1$. **C)** Mean CSD between excitatory neuron spike trains (red), between inhibitory neuron spike trains (blue), and between excitatory-inhibitory pairs (purple). Solid curves are from simulations (each CSD averaged over 10^5 pairs) and dashed are from Eq. (14). **D-I)** Same as Fig. 1D-I except spike trains in the external population, X, were correlated Poisson processes with spike count correlation $c = 0.1$. Dashed lines in I are from Eq. (14) evaluated at zero frequency. Network size was $N = 10^4$ in A-D.

CSD between spike trains in the recurrent network. First note that

$$\begin{aligned}\langle \boldsymbol{S}, \boldsymbol{T}\rangle &= \sqrt{N}\langle \boldsymbol{S}, \boldsymbol{S}\rangle W^* + \langle \boldsymbol{S}, \boldsymbol{X}\rangle \\ &= \sqrt{N}\langle \boldsymbol{S}, \boldsymbol{S}\rangle W^* - \frac{1}{\sqrt{N}} W^{-1}\langle \boldsymbol{X}, \boldsymbol{X}\rangle.\end{aligned}$$

However, our assumption of $\mathcal{O}(1)$ transfer implies that $\langle \boldsymbol{S}, \boldsymbol{T}\rangle \sim \langle \boldsymbol{S}, \boldsymbol{S}\rangle$ so

$$\langle \boldsymbol{S}, \boldsymbol{T}\rangle \sim \sqrt{N}\langle \boldsymbol{S}, \boldsymbol{S}\rangle W^* - \frac{1}{\sqrt{N}} W^{-1}\langle \boldsymbol{X}, \boldsymbol{X}\rangle,$$

which is only consistent if there is cancellation between the terms on the right hand side. This cancellation can only be realized if

$$\begin{aligned}\langle \boldsymbol{S}, \boldsymbol{S}\rangle &= \frac{1}{N} W^{-1}\langle \boldsymbol{X}, \boldsymbol{X}\rangle W^{-*} \\ &= W^{-1} W_x \langle S_x, S_x\rangle W_x^* W^{-*}\end{aligned} \tag{14}$$

which is $\mathcal{O}(1)$ and where we have omitted terms of smaller order. Notably, evaluating Eq. (14) does not depend on knowledge of neurons' correlation susceptibility functions or any other neuronal transfer properties, but only depends on synaptic parameters and input statistics.

In summary, $\mathcal{O}(1)$ covariance between spike trains in the external population produces $\mathcal{O}(N)$ covariance between neurons' external inputs, but $\mathcal{O}(1)$ covariance between spike trains in the recurrent network on average. We hereafter refer to this state as the "correlated state" since it produces moderately strong spike train correlations in contrast to the asynchronous state characterized by extremely weak spike train correlations. The reduction from $\mathcal{O}(N)$ external input covariance to $\mathcal{O}(1)$ spike train covariance arises from a cancellation mechanism analogous to the one that reduces $\mathcal{O}(1)$ external input correlation to $\mathcal{O}(1/N)$ spike train correlations in the asynchronous state (see above).

To demonstrate these results, we simulated a network of $N = 10^4$ neurons identical to the network from Fig. 1 except that spike trains in the external population were correlated Poisson processes (Fig. 2A) with

$$\langle S_x, S_x\rangle(f) = c r_x e^{-4f^2\pi^2\tau_c^2}. \tag{15}$$

Here, $r_x = 10$Hz is the same firing rate used in Fig. 1 and $c = 0.1$ quantifies the spike count correlation coefficient

between the spike trains in the external population over large counting windows. See Appendix B for a description of the algorithm used to generate the spike trains.

The recurrent network exhibited moderately correlated spike trains in contrast to spike trains in the asynchronous state (Fig. 2B, compare to Fig. 1C; mean spike count correlation between neurons with rates at least 1 Hz was 0.077) . The mean CSDs between spike trains in the recurrent network closely matched the theoretical predictions from Eq. (14) (Fig. 2C). As in the asynchronous state, external and recurrent synaptic input sources approximately canceled (Fig. 2D), as predicted by balanced network theory.

Varying N demonstrates that the network exhibits the same cancellation between $\mathcal{O}(\sqrt{N})$ mean external and recurrent synaptic input sources and that Eq. (4) for the mean firing rates is accurate (Fig. 2E,F). As predicted by the analysis of the correlated state, the covariance between individual sources of input currents appear $\mathcal{O}(N)$ (Fig. 2G), but cancel to produce much smaller, approximately $\mathcal{O}(1)$, total input covariance (Fig. 2G,H). Mean spike count covariances also appear $\mathcal{O}(1)$ and converge toward the limit predicted by Eq. (14) (Fig. 2I). Hence, despite the complexity of spike timing dynamics in densely connected balanced networks, mean-field spike train covariances is accurately predicted by a simple, linear equation in terms of synaptic parameters. The derivation of this equation does not require the use of linear response theory, which can be problematic for densely connected networks with synaptic kinetics and non-vanishing correlations.

V. THE CORRELATED STATE PRODUCES TIGHT BALANCE BETWEEN EXCITATORY AND INHIBITORY INPUT FLUCTUATIONS CONSISTENT WITH CORTICAL RECORDINGS

We have so far considered cancellation between positive and negative sources of input correlations at the mean-field level, *i.e.*, averaged over pairs of postsynaptic neurons (Figs. 1G,H and 2G,H). *In vivo* cortical recordings reveal that this cancellation occurs even at the level of single postsynaptic neuron pairs. When one neuron was clamped near its inhibitory reversal potential and another neuron is clamped near its excitatory reversal potential (with spiking suppressed), recorded membrane potential fluctuations are approximately mirror images of one another (Fig. 3A, top). Similarly, if both neurons are held near their excitatory reversal potential (Fig. 3A, middle) or both near their inhibitory reversal potential (Fig. 3A, bottom), recorded membrane potential fluctuations are highly correlated. This implies that fluctuations in the excitatory and inhibitory synaptic input to one neuron are strongly correlated with fluctuations in the excitatory and inhibitory input to other nearby neurons (see [22] for more details and interpretation).

To test whether this phenomenon occurred in our simulations, we randomly chose two neurons and decomposed their synaptic input into the total excitatory $(E+X)$ and the inhibitory (I) components. In the asynchronous state, input current fluctuations were fast and largely unshared between neurons or between current sources in the same neuron (Fig. 3B), in contrast to evidence from *in vivo* recordings. Input current fluctuations in the correlated state were slower, larger, and most importantly largely synchronized between neurons (Fig. 3C), consistent with evidence from *in vivo* recordings. This precise tracking of fluctuations in excitatory and inhibitory synaptic currents is referred to as "tight balance" [47] (as opposed to the "loose balance" of the asynchronous state). The results would be similar if we decomposed inputs into their external (X) and recurrent $(R = E + I)$ sources instead of excitatory $(E + X)$ and inhibitory (I).

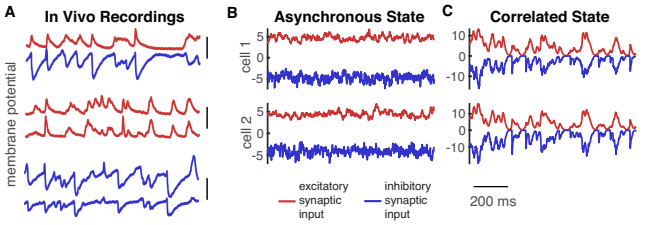

FIG. 3. **Excitatory-inhibitory tracking in vivo and in simulations. A)** In vivo membrane potential recordings from neurons in rat barrel cortex, reproduced from [22]. Each pair of traces are simultaneously recorded membrane potentials. Red traces were recorded in current clamp mode near the reversal potential of inhibition and blue traces near the reversal potential of excitation (with action potentials pharmacologically suppressed), so red traces are approximately proportional to excitatory input current fluctuations and blue traces approximate inhibitory input current fluctuations. Vertical scale bars are 20mV. **B,C)** Excitatory (red) and inhibitory (blue) synaptic input currents to two randomly selected excitatory neurons in the asynchronous (B) and correlated (C) states. Simulations were the same as those in Figs. 1B-D and 2A-D respectively. For ease of comparison with the current traces in panels B and D, the red voltage trace in the top of A was plotted above the corresponding blue trace.

To better understand this striking difference between input currents in the asynchronous and correlated states, we first computed the average covariance between the excitatory and inhibitory input to pairs of (excitatory) neurons in the network. These averages have the same dependence on network size, N, as they do when input currents are broken into external and recurrent sources (compare Fig. 4A,B to Figs. 1G and 2G). Specifically, in the asynchronous state, covariances between individual current sources are $\mathcal{O}(1)$ on average, but cancel to produce weak $\mathcal{O}(1/N)$ covariance between the total synaptic input to neurons on average (Fig. 4A). In the correlated state, the average covariance between individual input sources is $\mathcal{O}(N)$ and cancellation produces $\mathcal{O}(1)$ average total input covariance (Fig. 4B).

Hence, despite the precise cancellation of positive and

negative sources of input covariance at the mean-field level in the asynchronous state (Fig. 4A), this cancellation is apparently not observed at the level of individual neuron pairs (Fig. 3B). To see why this is the case, we computed the distribution of input current covariances across all pairs of excitatory neurons. We found that these distributions were broad and the distribution of total input covariance was highly overlapping with the distributions of individual input current sources (Fig. 4C, the black distribution overlaps with the others). This implies that cancellation does not reliably occur at the level of individual pairs since, for example, the total input covariance for a pair of neurons can be similar in magnitude or even larger than the covariance between those neurons' excitatory inputs.

The distributions of input covariances were strikingly difference in the correlated state. The distribution of total input covariances was far narrower than the distributions of individual input current sources and the distributions were virtually non-overlapping (Fig. 4D). Hence, a precise cancellation between positive and negative sources of input covariance must occur for every neuron pair, leading to the tight balance observed in Fig. 3C.

These results are more precisely understood by computing the variance across neuron pairs of input covariances as N is varied. In the asynchronous state, the variance of input covariances from all sources appear to scale like $\mathcal{O}(1)$ (Fig. 4E). Since the mean input covariance between individual sources are also $\mathcal{O}(1)$ (Fig. 4A), the overlap between distributions in Fig. 4C is expected. In the correlated state, the variances of input covariances appear to scale like $\mathcal{O}(N)$ except for the variance of the total input covariance, which appears to scale like $\mathcal{O}(1)$ (Fig. 4F). Since the variances scale like $\mathcal{O}(N)$, the standard deviations scale like $\mathcal{O}(\sqrt{N})$. This, combined with the fact that the mean input covariances between individual sources scale like $\mathcal{O}(N)$, implies that the distributions in Fig. 4E will be non-overlapping when N is large. The same conclusions would be reached if we decomposed inputs into their external (X) and recurrent ($R = E + I$) sources instead of total excitatory ($X + E$) and inhibitory (I).

VI. TESTING LINEAR RESPONSE APPROXIMATIONS TO PAIRWISE SPIKE TRAIN COVARIANCE

As discussed above, Eqs. (7) and (14) are appealing because, unlike many previous approximations of spike train CSD and spike count covariance in recurrent networks [8–11, 13], they do not require knowledge of neurons' correlation susceptibility functions or any other neuronal transfer properties. However, they are limited because they only give the population-averaged covariance or CSD.

Previous studies that provide approximations for the full $N \times N$ matrix of all spike train CSDs or spike count covariances use linear response approximations that rely on assumptions that the network is sparsely or weakly coupled and/or that external input is uncorrelated or weakly correlated Gaussian white noise [10, 11]. These assumptions permit the application of Fokker-Planck and linear response techniques. External input in our models is not weakly correlated and is not approximated well by Gaussian white noise, so these approaches are not fully justified in out model, but might nevertheless yield a useful approximation. We next derive and test a linear response approximation to spike train covariability in our model.

First define

$$\vec{S}(t) = [S_1^e(t) \ \dots \ S_{N_e}^e(t) \ S_1^i(t) \ \dots \ S_{N_i}^i(t)]^T$$

to be the full $N \times 1$ vector of spike trains in the recurrent network obtained by concatenating the excitatory and inhibitory spike train vectors. Define the $N \times 1$ synaptic input vectors $\vec{T}(t)$, $\vec{R}(t)$, and $\vec{X}(t)$ similarly. The total input vector can be written as

$$\vec{T} = \vec{R} + \vec{X}$$

with

$$\vec{R} = J * \vec{S}$$

and

$$\vec{X} = J_x * \vec{S}_x.$$

Here,

$$J = \begin{bmatrix} J_{ee} & J_{ei} \\ J_{ie} & J_{ii} \end{bmatrix},$$

is the $N \times N$ recurrent connectivity matrix and

$$J_x = \begin{bmatrix} J_{ex} \\ J_{ie} \end{bmatrix},$$

is the $N \times N_x$ feedforward connectivity matrix with $N_a \times N_b$ blocks defined by $[J_{ab}]_{jk}(t) = J_{jk}^{ab}\eta_b(t)$. Note that the connection weights are time dependent and $*$ denotes the matrix product with multiplication replaced by convolution [11].

The linearity of transfer from $\vec{S}$ to $\vec{R}$ and from $\vec{S}_x$ to $\vec{X}$ implies that

$$\langle \vec{R}, \vec{U} \rangle = \widetilde{J} \langle \vec{S}, \vec{U} \rangle \tag{16}$$

and

$$\langle \vec{X}, \vec{U} \rangle = \widetilde{J}_x \langle \vec{S}_x, \vec{U} \rangle$$

for $\vec{U} = \vec{S}, \vec{S}_x, \vec{R}, \vec{X}, \vec{T}$ where $\widetilde{J}(f)$ is the element-wise Fourier transform of the matrix $J(t)$ and similarly for $\widetilde{J}_x(f)$. This implies, for example, that the $N \times N$ matrix of external input CSDs is given by

$$\langle \vec{X}, \vec{X} \rangle = \widetilde{J}_x \langle \vec{S}_x, \vec{S}_x \rangle \widetilde{J}_x^*$$

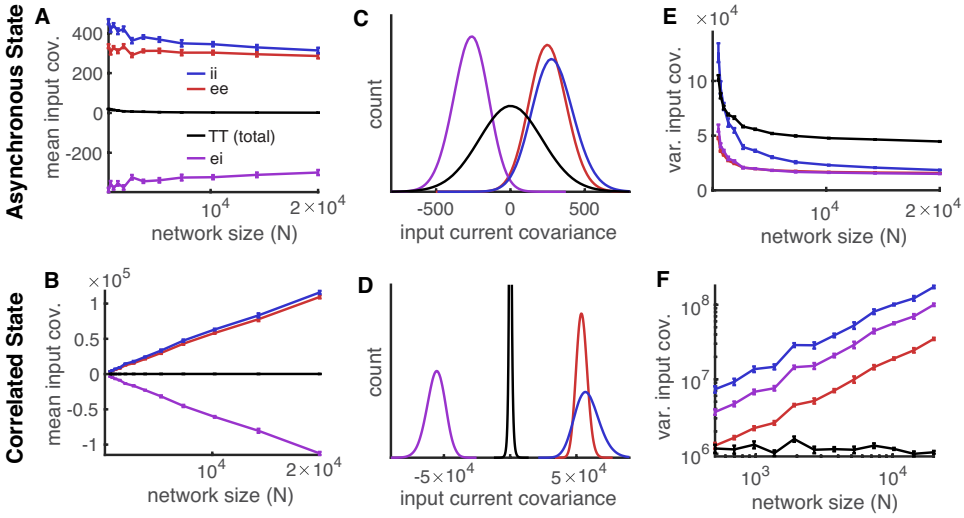

FIG. 4. **The scaling of mean and variance of excitatory and inhibitory input covariance in the asynchronous and correlated states. A,B)** Same as Figs. 1G and 2G, except inputs were decomposed into their excitatory ($E+X$), and inhibitory (I) components instead of external and recurrent. Red curves show mean excitatory-excitatory input covariance, blue show inhibitory-inhibitory, purple show excitatory-inhibitory, and black curves show total (same as black curves in Figs. 1G and 2G). **C,D)** Histogram of input current covariances across all excitatory cell pairs for a network of size $N = 10^4$. **E,F)** Same as A,B except we plotted the variance of covariances across cell pairs instead of the mean. As above, integrated currents have units $C_m mV$, so input covariances have units $C_m^2 mV^2$ and the variance of covariances have units $C_m^4 mV^4$ where C_m is the arbitrary membrane capacitance.

and similarly for the CSDs between recurrent inputs,

$$\langle \vec{R}, \vec{R} \rangle = \widetilde{J} \langle \vec{S}, \vec{S} \rangle \widetilde{J}^*$$

In summary, the transfer of spike train CSDs to input CSDs follows easily from the linearity of transfer from spike trains to inputs [4]. However, a closed expression for the matrix of spike train CSDs in the recurrent network, $\langle \vec{S}, \vec{S} \rangle$, requires knowledge of the transfer from total input to spike train CSDs.

If spike trains were related linearly to total input, $\vec{S} = A * \vec{T}$ for some diagonal matrix, $A(t)$, then we could derive an exact closed equation for $\langle \vec{S}, \vec{S} \rangle$. However, integrate-and-fire neuron models transfer their input currents to spike trains nonlinearly. Nevertheless, we can obtain an approximation to spike train CSDs by assuming an approximate linear transfer of CSDs, specifically that [11]

$$\begin{aligned} \langle \vec{S}, \vec{X} \rangle &\approx \widetilde{A} \langle \vec{T}, \vec{X} \rangle, \\ \langle \vec{S}, \vec{T} \rangle &\approx \widetilde{A} \langle \vec{T}, \vec{T} \rangle, \\ \langle \vec{S}, \vec{S} \rangle &\approx \widetilde{A} \langle \vec{T}, \vec{S} \rangle \end{aligned} \tag{17}$$

where $\widetilde{A}(f)$ is a diagonal matrix with $\widetilde{A}_{jj}(f)$ the "susceptibility function" of neuron j [8, 48–50]. Note that these equations would be exactly true if neural transfer were linear, $\vec{S} = A * \vec{T}$. It can be derived from these assumptions that (see Appendix C for derivation and [11] for similar derivations)

$$\langle \vec{S}, \vec{S} \rangle \approx [\widetilde{A}^{-1} - \widetilde{J}]^{-1} \langle \vec{X}, \vec{X} \rangle [\widetilde{A}^{-1} - \widetilde{J}]^{-*}. \tag{18}$$

One problem with testing the approximation in Eq. (18) is that we do not have an estimate of $\widetilde{A}_{jj}(f)$. Spike count covariances over large windows are proportional to zero-frequency CSD (see Eq. (3) and surrounding discussion). Evaluated at zero frequency, a neuron's susceptibility function is equal to its gain, *i.e.* the derivative of the neuron's firing rate with respect to its mean input [49, 50],

$$\widetilde{A}_j(0) = \frac{dr_j}{d\overline{T}_j}.$$

We therefore tested Eq. (18) for spike count covariances by estimating the gain empirically from simulations. In doing so, we found that Eq. (18) is only moderately accurate at approximating spike count covariances from simulations, both in the asynchronous state (Fig. 5A) and in the correlated state (Fig. 5B). We suspected that some of the error was due to numerical inaccuracies introduced by inverting the large, ill-conditioned matrices in Eq. (18). To test this hypothesis, we re-wrote Eq. (18) in a mathematically equivalent formulation that does not involve matrix inverses,

$$(Id - \widetilde{A}\widetilde{J}) \langle \vec{S}, \vec{S} \rangle (Id - \widetilde{A}\widetilde{J}) \approx \widetilde{A} \langle \vec{X}, \vec{X} \rangle \widetilde{A}^* \tag{19}$$

where Id is the $N \times N$ identity matrix. We tested the accuracy of Eq. (19) at zero frequency using the same empirically estimated values of the gains and using the matrix, $\langle \vec{S}, \vec{S} \rangle(0)$, of spike count covariances estimated from simulations and found that it was very accurate, especially in the correlated state (Fig. 5C,D). Since Eqs. (18) and (19) are mathematically equivalent, this suggests that

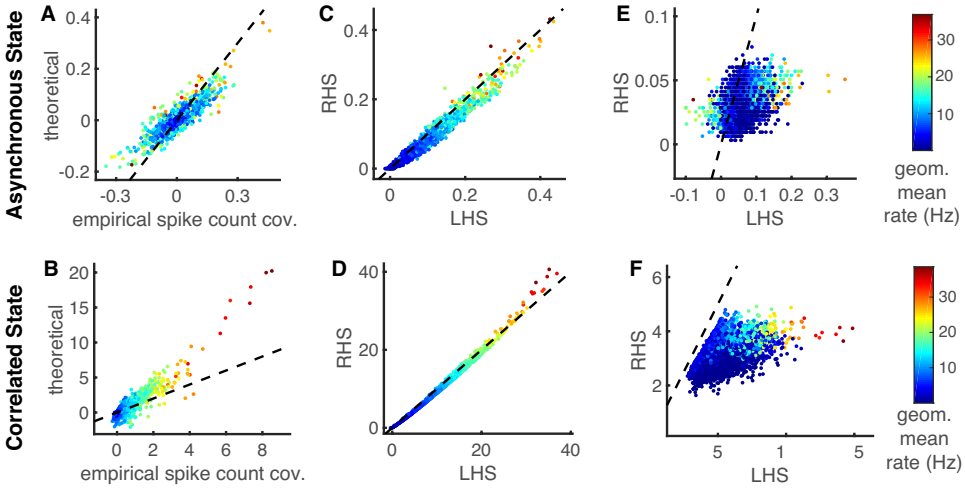

FIG. 5. **Testing the accuracy of linear response approximations in densely connected balanced networks in the asynchronous and correlated states. A,B)** Spike count covariance from simulations of a network with $N = 5000$ in the asynchronous (A) and correlated (B) states plotted against the theoretical value given by evaluating Eq. (18) evaluated at zero frequency, with empirically estimated gains used for $\widetilde{A}_j(0)$. **C,D)** The left-hand-side (LHS) of Eq. (19) plotted against the right-hand-side (RHS) using $\langle \vec{S}, \vec{S} \rangle$ estimated from the same simulations as in A,B and using the same empirically estimated gains. **E,F)** Same as C,D except we used the mean gain for all values of $\widetilde{A}_j(0)$.

much of the observed inaccuracy of Eq. (18) is due to the presence of inverses of large, ill-conditioned matrices.

One shortcoming of Eqs. (18) and (19) is that they require an estimate of $\widetilde{A}$. We obtained this estimate empirically by simulating the entire network and numerically estimating the gains, which greatly limits the utility of the equations. We next wondered whether the accuracy of the equations is sensitive to the accuracy of the gain estimate, or if a rough approximation to the gains would be sufficient. To answer this question, we tested Eq. (19) again, but replaced the individual estimated gains on the diagonal of $\widetilde{A}$ with their mean. In other words, we used $\widetilde{A}(0) = \overline{g} Id$ where $\overline{g}$ is the average gain estimated from simulations and Id is the $N \times N$ identity matrix. This replacement greatly decreased the accuracy of Eq. (19) (Fig. 5E,F), suggesting that an accurate estimate of the individual gains is necessary for applying and interpreting Eqs. (18) and (19).

In summary, linear response approximations are fairly accurate in the densely connected balanced networks with spatially and temporally correlated noisy feedforward input studied here (Fig. 5C,D). However, their utility is limited by two factors: First, that using linear response theory to approximate spike train CSDs or spike count covariances requires the inversion of large, ill-conditioned matrices, which introduces substantial error (Fig. 5A,B). Secondly, that the application of linear response approximation requires estimates of neurons' individual susceptibility functions or gains. The mean-field equations (7) and (14) do not have these shortcomings, but only give the population-averaged CSDs and covariances. It should also be noted that correlations were weak in our simulations, even in the correlated state (mean spike count correlations between neurons with firing rate above 1 Hz was 0.077). Stronger correlations can be obtained with alternative parameters, which could potentially make Eqs. (18) and (19) less accurate.

VII. CORRELATED VARIABILITY FROM SINGULAR MEAN-FIELD CONNECTIVITY STRUCTURE

We have shown that $\mathcal{O}(1)$ spike train correlations can be obtained in balanced networks by including correlations between neurons in an external layer ($\langle S_x, S_x \rangle \sim \mathcal{O}(1)$), defining what we refer to as the "correlated state." Previous work shows that $\mathcal{O}(1)$ spike train correlations can be obtained in the recurrent network with uncorrelated external spike trains ($\langle S_x, S_x \rangle = 0$) when the mean-field connectivity matrix is constructed in such a way that the recurrent network cannot achieve the cancellation required for these states to be realized [37–39]. This is most easily achieved using networks with several discrete sub-populations or networks with distance-dependent connectivity. For simplicity, we restrict our analysis to discrete sub-populations. We first extend the mean-field theory from above to such networks, then generalize and extend the analysis from previous work to understand the emergence of $\mathcal{O}(1)$ correlations when the mean-field connectivity matrix is singular.

The recurrent networks considered above have two statistically homogeneous sub-populations: one excitatory and one inhibitory and the external population is a single homogeneous population. Suppose instead that there are K sub-populations in the recurrent network, with

the kth popualtion containing $N_k = q_k N$ neurons where $\sum_k q_k = 1$. Connectivity is random with p_{jk} denoting the connection probability from population k to j, and $j_{jk}/\sqrt{N}$ denoting the strengths of the connections for $j, k = 1, \dots K$. All neurons in population k are assumed to have the PSC kernel $\eta_k(t)$ which is again assumed to have integral 1. Similarly, suppose that the external network contains K_x sub-populations each with $N_k^x = q_k^x N_x$ neurons where $q_k^x = \sum_k q_{x,k} = 1$. Feedforward connection probabilities and strengths are given by p_{jk}^x and $j_{jk}^x/\sqrt{N}$ for $j = 1, \dots K$ and $k = 1, \dots, K_x$. Assume that q_k, p_{jk}, j_{jk}, q_k^x, p_{jk}^x, and j_{jk}^x are all $\mathcal{O}(1)$. We then define the $K \times K$ mean-field recurrent connectivity matrix by $[W]_{jk} = p_{jk} j_{jk} q_k \widetilde{\eta}_k$ and the mean-field feedforward connectivity matrix by $[W_x]_{jk} = p_{jk}^x j_{jk}^x q_k^x \widetilde{\eta}_k^x$. For all of the networks considered above, we had $K = 2$ and $K_x = 1$.

When W is an invertible matrix, Eqs. (4), (7), and (14) are equally valid for networks with several subpopulations as they are for the simpler networks considered above. Hence, the mean-field theory of firing rates and correlations extends naturally to networks with several populations [38–42, 51]. However, when W is singular, Eqs. (4), (7), and (14) cannot be evaluated. Instead, Eq. (4) can be re-written as

$$W\boldsymbol{r} = -W_x \boldsymbol{r}_x. \tag{20}$$

When W is singular, this equation only has a solution, $\boldsymbol{r}$, when $\overline{\boldsymbol{X}} = -W_x \boldsymbol{r}_x$ is in the range or "column space" of W. Otherwise, balance is broken. An in-depth analysis of firing rates in such networks is provided in previous work [41, 42, 51] (and extended to spatially continuous networks in [40, 51]), so we hereafter assume that $\overline{\boldsymbol{X}}$ is in the range of W and balance is achieved.

A similar analysis may be applied to spike train CSDs. For simplicity, we assume here that spike trains in the external population are uncorrelated, $\langle S_x, S_x \rangle = 0$, since this is the case considered in previous work and since this is the case in which a singular W breaks the asynchronous state. Eq. (7) can be re-written as

$$W \langle \boldsymbol{S}, \boldsymbol{S} \rangle W^* = \frac{1}{N} \langle \boldsymbol{X}, \boldsymbol{X} \rangle. \tag{21}$$

where we have ignored smaller order terms. When W is singular, Eq. (21) is not guaranteed to have a solution, $\langle \boldsymbol{S}, \boldsymbol{S} \rangle$. More precisely, a solution can only exist when the $K \times K$ matrix, $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$, is in the range of the linear operator $\mathcal{L}$ defined by

$$\mathcal{L}U = WUW^*.$$

In that case, Eq. (21) has a solution so that $\langle \boldsymbol{S}, \boldsymbol{S} \rangle \sim \mathcal{O}(1/N)$ and the asynchronous state is still realized. However, if $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ is not in the range of $\mathcal{L}$, the asynchronous state cannot be realized because Eq. (21) does not have a solution.

Darshan et al. [39] looked at a similar scenario except the singularity of their networks made them incapable of cancelling *internally generated covariance*, in contrast to the *external input covariance* considered here. Other work [37, 38] analyzed the scenario with external input covariance and singular connectivity, as well as the extension to spatially extended networks. However, this previous work did not completely analyze the structure of correlations, but only showed that the asynchronous state was broken. We next show that correlations in the recurrent network are $\mathcal{O}(1)$ and derive their structure.

Using Eqs. (9), (10), and (12), we can write the mean-field total input CSD as

$$\begin{aligned}\langle \boldsymbol{T}, \boldsymbol{T} \rangle = {} & NW \langle \boldsymbol{S}, \boldsymbol{S} \rangle W^* + \sqrt{N}(W \langle \boldsymbol{S}, \boldsymbol{X} \rangle + \langle \boldsymbol{X}, \boldsymbol{S} \rangle W^*) \\ & + \langle \boldsymbol{X}, \boldsymbol{X} \rangle.\end{aligned} \tag{22}$$

If W is not invertible, then W^* has a non-trivial nullspace. Let $v_1, v_2, \dots, v_n$ be a basis for the nullspace of W^* and define

$$P = v_1 v_1^* + v_2 v_1^* + \cdots + v_n v_n^*$$

which is a self-adjoint matrix that defines the orthogonal projection onto the nullspace of W^*. Note that P is a Hermitian matrix ($P = P^*$) and $PW = W^*P = 0$ (the zero matrix). Define the projection $A_0 = PAP$ for any matrix A. Unless $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ is carefully constructed otherwise, we can expect that

$$\langle \boldsymbol{X}, \boldsymbol{X} \rangle_0 \sim \langle \boldsymbol{X}, \boldsymbol{X} \rangle \sim \mathcal{O}(1).$$

Then take the projection of both sides of Eq. (22) above to get

$$\langle \boldsymbol{T}, \boldsymbol{T} \rangle_0 = \langle \boldsymbol{X}, \boldsymbol{X} \rangle_0 \sim \mathcal{O}(1). \tag{23}$$

Hence, the total input CSD is $\mathcal{O}(1)$ when $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ is not in the range of $\mathcal{L}$, even though it is $\langle \boldsymbol{X}, \boldsymbol{X} \rangle / N$ when W is invertible (*i.e.*, in the asynchronous state). Moreover, the structure of $\langle \boldsymbol{T}, \boldsymbol{T} \rangle$ is given to highest order in N by $\langle \boldsymbol{X}, \boldsymbol{X} \rangle_0 = P \langle \boldsymbol{X}, \boldsymbol{X} \rangle P$, which can be computed exactly from knowledge of the structure of $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ and W.

When neural transfer from $\boldsymbol{T}$ to $\boldsymbol{S}$ is $\mathcal{O}(1)$ (see Eq. (8) and surrounding discussion), this implies that $\langle \boldsymbol{S}, \boldsymbol{S} \rangle \sim \mathcal{O}(1)$ so that the asynchronous state is broken when $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ is not in the range of $\mathcal{L}$. While we cannot be certain that $\langle \boldsymbol{S}, \boldsymbol{S} \rangle$ has the same structure as $\langle \boldsymbol{T}, \boldsymbol{T} \rangle$, it should have a similar structure as long as neural transfer of correlations is similar for each sub-population.

To demonstrate these results, we consider the same network from above with re-wired feedforward projections from the external population. Specifically, divide the excitatory, inhibitory, and external populations each into two equal-sized sub-populations, labeled e_1, i_1, x_1, e_2, i_2, and x_2 where population a_k contains $N_a/2$ neurons. Hence the network has the same total number of neurons as before, but we have simply sub-divided the populations. To distinguish this network from the one considered in Figs. 1 and 2, we refer to the previous network as the 3-population network and to this modified network as the 6-population network.

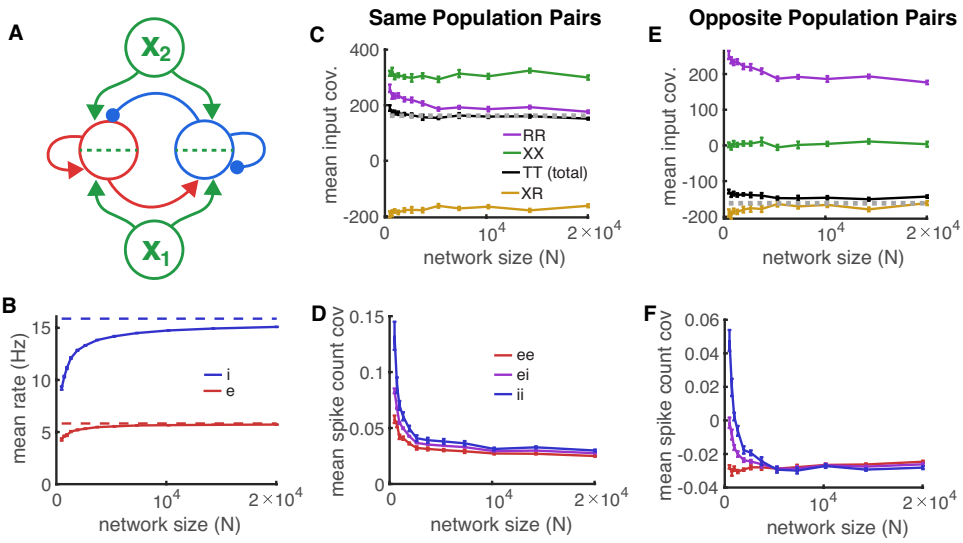

FIG. 6. **Correlated variability in a balanced network with singular mean-field connectivity matrix. A)** Network schematic. The recurrent network is statistically identical to the networks considered previously, but there are two external populations that each connect to a different half of the neurons in the recurrent network. **B)** Same as Fig. 1F, but for the multi-population network from A. **C)** Same as Fig. 1G, but for the network in A and where input covariances are only averaged over postsynaptic neurons in the same group (both postsynaptic cells in e_1 or both in e_2). The dashed gray curve shows the theoretical prediction for total input covariance (the black curve) from Eq. (24). **D)** Same as Fig. 1I, but for the network in A and where spike count covariances are only averaged over postsynaptic neurons in the same group (first cell in a_j and second cell in b_j for $a, b = e, i$ and $j = 1, 2$). **E,F)** Same as C and D, but covariances are computed between cells in opposite groups (one cell in a_1 and the other cell in b_2).

We re-wire the feedforward connections so that x_1 only connects to e_1 and i_1, and x_2 only projects to e_2 and i_2. Specifically, we set the connection probabilities to $p_{a_j x_k} = 2p_{ax}$ if $j = k$ and $p_{a_j x_k} = 0$ if $j \neq k$ for $a, b = e, i$ and $j, k = 1, 2$, where p_{ab} are the connection probabilities for the 3-population network and $p_{a_j b_k}$ for the 6-population network. This re-wiring assures that neurons in the recurrent network receive the same number of feedforward connections on average from the external population. The recurrent connectivity structure is not changed at all. Specifically, we set $p_{a_j b_k} = p_{ab}$ for $a, b = e, i$. All connection strengths are unchanged, $j_{a_j b_k} = j_{ab}$ for $a = e, i$ and $b = e, i, x$ and all neurons in the external population have the same firing rate, r_x, as before. See Fig. 6A for a schematic of this network.

The feedforward mean-field connectivity matrix can be written in block form as

$$W_x = \begin{bmatrix} W_x^{2\times 1} & \mathbf{0} \\ \mathbf{0} & W_x^{2\times 1} \end{bmatrix}$$

where $\mathbf{0}$ is the 2×1 zero-matrix and $W_x^{2\times 1} = [w_{ex} \;\; w_{ix}]^T$ is the 2×1 feedforward connectivity matrix for the 3-population network. Note that W_x is 4×2 since there are 4 populations in the recurrent network and 2 populations in the external population. The recurrent mean-field connectivity matrix is

$$W = \frac{1}{2} \begin{bmatrix} W^{2\times 2} & W^{2\times 2} \\ W^{2\times 2} & W^{2\times 2} \end{bmatrix}$$

where

$$W^{2\times 2} = \begin{bmatrix} w_{ee} & w_{ei} \\ w_{ie} & w_{ii} \end{bmatrix}$$

is the 2×2 recurrent connectivity matrix for the 3-population network. Note that W is 4×4. Here, $w_{ab} = p_{ab} j_{ab} q_b \widetilde{\eta}_b$ are the same values used above for analyzing the 3-population network.

Even though W is non-invertible, $\overline{\boldsymbol{X}} = W_x [r_x \;\; r_x]^T$ is in the range of W for this example, so firing rates in the balanced state can be computed using Eq. (20), and are identical to the firing rates for the 3-population networks considered above.

The nullspace of W^* is spanned by the orthonormal vectors

$$v_1 = \frac{1}{\sqrt{2}} [1 \;\; 0 \;\; -1 \;\; 0]^T$$

and

$$v_2 = \frac{1}{\sqrt{2}} [0 \;\; 1 \;\; 0 \;\; -1]^T$$

so the projection matrix is given in block form by

$$P = \frac{1}{2} \begin{bmatrix} I_2 & -I_2 \\ -I_2 & I_2 \end{bmatrix}$$

where I_2 is the 2×2 identity matrix.

The external input CSD is determined by the average number of overlapping feedforward projections to any pair of neurons in the recurrent network (multiplied by their connection strength and r_x), which gives (in block form)

$$\langle \boldsymbol{X}, \boldsymbol{X} \rangle = 2 \begin{bmatrix} \langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} & \mathbf{0} \\ \mathbf{0} & \langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} \end{bmatrix}$$

where $\mathbf{0}$ is the 2×2 zero matrix and $\langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2}$ is the external input CSD from the 3-population network, given by Eq. (5). Therefore, by Eq. (23),

$$\begin{aligned} \langle \boldsymbol{T}, \boldsymbol{T} \rangle_0 &= \langle \boldsymbol{X}, \boldsymbol{X} \rangle_0 = P \langle \boldsymbol{X}, \boldsymbol{X} \rangle P \\ &= \begin{bmatrix} \langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} & -\langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} \\ -\langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} & \langle \boldsymbol{X}, \boldsymbol{X} \rangle^{2\times 2} \end{bmatrix}. \end{aligned} \tag{24}$$

In other words, the mean total input CSD between excitatory neurons in the same subgroup (two neurons in e_1 or two neurons in e_2; diagonal blocks above) is positive and equal to half the mean external input between the same neurons. Hence, the cancellation by the recurrent network only reduces the external input CSD by a factor of 1/2, as opposed to the $\mathcal{O}(1/N)$ reduction realized in the asynchronous state (when W is invertible). In contrast, the mean total input CSD between excitatory neurons in opposite subgroups (one neuron in e_1 and the other in e_2; off-diagonal blocks above) has the same magnitude as for same-subgroup pairs, but is negative. This represents a competitive dynamic between the two groups since they inhibit one another (recurrent connections are net-inhibitory in balanced networks [27, 42]), but receive different feedforward input noise. Interestingly, the average CSD between all pairs of spike trains is still $\mathcal{O}(1/N)$ in this example, but it is easy to design examples with singular W in which this is not true. A similar example was considered in previous work [38], but external input was generated artificially instead of coming from an external population.

Simulating this network for varying values of N shows that firing rates approach those predicted by the balance equation (20) (Fig. 6B), confirming that balance is realized. Pairs of excitatory neurons in the same group (both neurons in e_1 or both neurons in e_2) receive positively correlated external input and recurrent input (Fig. 6C, purple and green curves) that are partially canceled by negative correlations between their recurrent and excitatory input (Fig. 6C, yellow curve). Because the cancellation is only partial, the correlation between the neurons' total inputs is $\mathcal{O}(1)$ (Fig. 6C, black curve) in contrast to the asynchronous state (compare to Fig. 1G,H where cancellation is perfect at large N). The total input covariance agrees well with the theoretical prediction from Eq. (24) (Fig. 6C, dashed gray line). As a result of this lack of cancellation between total input covariance, spike count covariances are also $\mathcal{O}(1)$ and positive between same-group pairs (Fig. 6D). For opposite group pairs (one neuron in e_1 and the other in e_2), cancellation is also imperfect, but this leads to negative total input covariance, in agreement with the theoretical prediction from Eq. (24) (Fig. 6E), and leads to negative spike count covariances between neurons in opposite populations (Fig. 6F).

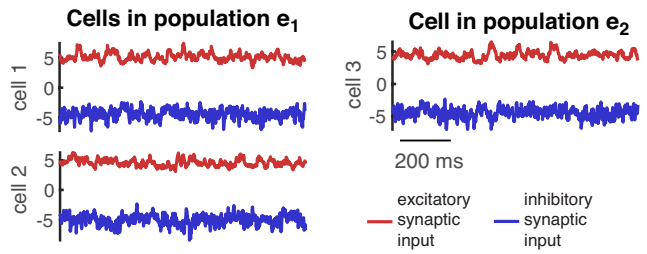

FIG. 7. **Synaptic input currents in a balanced network with correlations from singular mean-field connectivity.** Same as Fig. 3A except for the network from Fig. 6. The left two traces are input currents to two excitatory neurons in population e_1 (cells 1 and 2). The right traces are input currents to an excitatory neuron in population e_2 (cell 3).

In summary, we have analyzed two mechanisms to generate $\mathcal{O}(1)$ spike train correlations in balanced networks. For the first mechanism (Fig. 2), spike trains in the external population are correlated so that external input correlations are $\mathcal{O}(N)$. Cancellation is achieved so that spike train correlations are reduced to $\mathcal{O}(1)$. For the other mechanism (Fig. 6), external input correlation is $\mathcal{O}(1)$, but precise cancellation cannot be achieved so that spike trains inherit the $\mathcal{O}(1)$ correlations from the input. How could these two mechanisms be distinguished in cortical recordings? Under the first mechanism, we showed that fluctuations of inhibitory input to individual neurons closely tracks fluctuations of other neurons' excitatory inputs (Fig. 3C). This should not be the case under the second mechanism because precise cancellation is not realized. Indeed, plotting the excitatory and inhibitory input to three excitatory neurons (two in e_1 and one in e_2) shows that input fluctuations are not closely tracked (Fig. 7). This provides a way to distinguish the two mechanisms from paired intracellular recordings. Indeed, the first mechanism (which we refer to as the "correlated state") appears more consistent with the cortical recordings considered here (compare Fig. 3A to Figs. 3C and 7).

VIII. SUMMARY AND DISCUSSION

We analyzed correlated variability in recurrent, balanced networks of integrate-and-fire neurons receiving correlated feedforward input from an external population. We showed that correlations between spike trains in the recurrent network are small ($\mathcal{O}(1/N)$) when spike trains in the external population are uncorrelated, consistent with previous work on the asynchronous state [29, 38], but much larger ($\mathcal{O}(1)$) when spike trains in the external population are correlated, giving rise to a "correlated state." In both states, strong correlations in the feedforward input are canceled by recurrent synaptic input due

to the excitatory-inhibitory tracking that arises naturally in densely connected balanced networks. This cancellation allows for the derivation of a concise and accurate closed form expression for spike train CSDs in terms of synaptic parameters alone. Hence correlations in balanced networks are determined predominately by synaptic connectivity structure, not neuronal dynamics. The tracking of excitatory synaptic input by inhibition was observable on a pair-by-pair basis in the correlated state, but not the asynchronous state, suggesting that the correlated state is more consistent with in vivo recordings.

We only considered recurrent networks with two, statistically homogeneous neural populations: one excitatory and one inhibitory. Our analysis can be extended to multiple subpopulations as long as each sub-population contains $\mathcal{O}(N)$ neurons, and also extends to networks with connection probabilities that depend on distance, orientation tuning, or other continuous quantities. This analysis has been developed for the asynchronous state in previous work [38] and is easily extended to the correlated state as well. The primary difference is that $\langle \boldsymbol{X}, \boldsymbol{X} \rangle$ is $\mathcal{O}(N)$ instead of $\mathcal{O}(1)$.

Previous work has shown that networks with multiple sub-populations and networks with distance-dependent connectivity can break the asynchronous state in balanced networks when the network connectivity structure is constructed in such a way that the recurrent network cannot achieve the cancellation required for the asynchronous state [37–39], leading to $\mathcal{O}(1)$ correlations between some cell pairs. We showed that the precise tracking of excitation by inhibition provides an experimentally testable prediction for distinguishing this mechanism from the one underlying the correlated state (see Fig. 7 and surrounding discussion).

Another alternative mechanism for achieving larger correlations in balanced networks is through instabilities of the balanced state. Such instabilities can create pattern-forming dynamics that produce correlated spiking [40, 52–59]. Future studies should work towards experimentally testable predictions that distinguish correlations that arise from instabilities from those that arise through the mechanisms considered here. For example, since instabilities generate correlations internally, they should produce weak correlations between activity in the recurrent network and activity in the external population(s) providing input to that network [57], in contrast to the mechanisms we consider here. Indeed, some recordings show that local circuit connectivity can increase correlations [60], which is consistent with internally generated correlations, but inconsistent with the mechanisms that we consider here.

In the correlated state, spike train correlations in the recurrent network are essentially inherited from correlations between spike trains in the external population. Hence, the $\mathcal{O}(1)$ correlations realized by this mechanism require the presence of another local network with $\mathcal{O}(1)$ correlations. This raises the question of where the $\mathcal{O}(1)$ correlations are originally generated. One possibility is that they could be generated in a presynaptic cortical area or layer through the alternative mechanisms discussed in the previous paragraph. Another possibility is that they originate from a network that is not in the balanced state at all. Non-balanced networks can easily achieve $\mathcal{O}(1)$ spike train correlations simply from overlapping synaptic projections. While cortical circuits are commonly believed to operate in a balanced state, correlations could originate in thalamus, retina, or other sub-cortical neural populations then eventually propagate to cortex.

The cancellation between variances of covariances observed empirically in Fig. 4F is, to our knowledge, a novel observation, but we were unable to derive it analytically. Path integral approaches have recently been applied to compute variances of covariances in linear network models with uncorrelated external input [61] ($\langle \vec{X}, \vec{X} \rangle$ diagonal so $\langle \boldsymbol{X}, \boldsymbol{X} \rangle = 0$). This previous work derives a simple relationship between a network's criticality and a parameter, Δ, that represents the ratio between the standard deviation of spike count covariances and the mean spike count variance. Specifically, they showed that in their network model, the eigenvalue spectrum of the network dynamics is given to leading order in N by $\lambda_{max} = \sqrt{1 - \sqrt{1/(1 + N\Delta^2)}}$. In our model, where external input is correlated, the variance of covariances and spike count variances both appear to be $\mathcal{O}(1)$, so that $\Delta \sim \mathcal{O}(1)$ and their expression reduces to $\lambda_{max} = 1 - \mathcal{O}(1/\sqrt{N})$ regardless of the spectral radius of W. We interpret this to mean that their formula is not accurate for networks with correlated external input. Specifically, their formula would erroneously predict near-critical network dynamics if applied to a network that is far from criticality, but receives correlated external inputs. Future work should consider the possibility of extending their analysis of criticality to networks with external input correlation.

Our mean-field analysis applies to networks of integrate-and-fire neuron models, which are arguably more biologically realistic than networks of binary model neurons that are often used for mean-field analysis of neuronal networks. Binary neuron networks are appealing due to the mathematical tractability of their mean-field analysis, but our work demonstrates that integrate-and-fire networks are similarly tractable, calling into question the utility and appeal of binary network models.

Two unproven assumptions underly our mean-field analysis of the correlated state. The first assumption is that neural transfer is $\mathcal{O}(1)$ (Eq. (8) and surrounding discussion). The second assumption is that individual connection strengths are not strongly correlated with individual CSD values so that the step from Eq. (A2) to (A3) is valid when ignoring smaller order terms. These assumptions are made in other work, even if not stated explicitly. We have been unable to prove these assumptions rigorously for the model studied here, leaving an open problem for future work.

We showed that linear approximations to spike train covariance developed for small, sparsely coupled net-

works [10–12] can also be accurate for large, densely connected balanced networks (Fig. 5). However, their usefulness is limited by the need to invert large, ill-conditioned matrices and to approximate the susceptibility functions of individual neurons. The simpler equations we derived for mean-field spike train CSDs (Eqs. (7) and (14)) do not have these problems. Moreover, while linear response approximations require that neural transfer of input is approximately linear, our mean-field derivations did not depend on this assumption. Recent work has called for looking beyond linear analysis of neuronal networks [62]. Our analysis shows that, even in networks where neural transfer of inputs is nonlinear, linear mean-field analysis could still be accurate and useful.

In summary, we showed that correlations in balanced networks can be caused by feedforward input from a population of neurons with correlated spike trains, defining the "correlated state" that is quantitatively captured by a linear mean-field theory. In contrast to other mechanisms of correlation in balanced networks, the correlated state predicts a precise balance between the fluctuations in excitatory and inhibitory synaptic input to individual neuron pairs, consistent with some in vivo recordings [22].

ACKNOWLEDGMENTS

We thank Michael S. Okun for helpful comments on a draft of the manuscript and for his contribution to collecting the data used in Fig. 3A. RR was supported by National Science Foundation (https://www.nsf.gov/) grants DMS-1517828, DMS-1654268, and Neuronex DBI-1707400. IL was supported by Deutsche Forschungsgemeinschaft Sonderforschungsbereiche (www.dfg.de/sfb) grant number 1089, Israel Science Foundation (https://www.isf.org.il/) grant numbers 326/07 and 1539/17, and Minerva (http://www.minerva.mpg.de/weizmann/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A: Details for the derivation of mean-field CSDs.

We now provide details in the derivations of Eqs. (6) and (14), which can both be written as

$$\langle \boldsymbol{S}, \boldsymbol{S}\rangle = \frac{1}{N} W^{-1}\langle \boldsymbol{X}, \boldsymbol{X}\rangle W^{-*} + \frac{1}{N} C_0 + o(\langle \boldsymbol{X}, \boldsymbol{X}\rangle/N) \tag{A1}$$

where $o(\langle \boldsymbol{X}, \boldsymbol{X}\rangle/N)$ scales smaller than $\langle \boldsymbol{X}, \boldsymbol{X}\rangle/N$ as $N \to \infty$ and where $C_0 \sim \mathcal{O}(1)$. Note that in the correlated state, $\langle \boldsymbol{X}, \boldsymbol{X}\rangle \sim \mathcal{O}(N)$ so that the C_0/N term can be absorbed into the $o(\langle \boldsymbol{X}, \boldsymbol{X}\rangle/N)$ term. A sketch of the derivation for the correlated state is given in Results, and the derivation is similar in the asynchronous state. Here, we give the details of this derivation.

We first derive Eq. (5) for $\langle \boldsymbol{X}, \boldsymbol{X}\rangle$ in the asynchronous state, *i.e.* when spike trains in the external population are uncorrelated Poisson processes, so $\langle S_x, S_x\rangle = 0$. In that case, the CSD between two neurons' external input is given by

$$\begin{aligned}\langle X_j^a, X_k^b\rangle &= \left\langle \sum_{m=1}^{N_x} J_{jm}^{ax}\eta_x * S_m^x, \sum_{n=1}^{N_x} J_{kn}^{bx}\eta_x * S_n^x \right\rangle \\ &= \sum_{m,n=1}^{N_x} J_{jm}^{ax} J_{kn}^{bx} \widetilde{\eta}_x \widetilde{\eta}_x^* \langle S_m^x, S_n^x\rangle\end{aligned}$$

for $a, b = e, i$ where we have used the fact that the cross-spectral operator, $\langle\cdot,\cdot\rangle$ is a Hermitian operator. Since external spike trains are uncorrelated Poisson processes, $\langle S_m^x, S_n^x\rangle = 0$ when $m \neq n$ and $\langle S_m^x, S_m^x\rangle = r_x$. Therefore, we can rewrite the equation above as

$$\langle X_j^a, X_k^b\rangle = |\widetilde{\eta}_x|^2 r_x \sum_{m=1}^{N_x} J_{jm}^{ax} J_{km}^{bx}.$$

From Eq. (2), the expectation of the summand in this equation is

$$E[J_{jm}^{ax} J_{km}^{bx}] = \frac{p_{ax} p_{bx} j_{ax} j_{bx}}{N}.$$

Hence, taking expectations across the N_x elements of the sum and the coefficient in front of the sum gives

$$\langle X_a, X_b\rangle = q_x j_{ax} j_{bx} p_{ax} p_{bx} |\widetilde{\eta}_x|^2 r_x$$

which, in matrix form, is equivalent to Eq. (5).

We next derive Eq. (10). Noting that recurrent input to neuron j in population $a = e, i$ is composed of excitatory and inhibitory components, $R_j^a(t) = E_j^a(t) + I_j^a(t)$, we have

$$\langle R_j^a, U_k^b\rangle = \langle E_j^a, U_k^b\rangle + \langle I_j^a, U_k^b\rangle$$

where we can compute

$$\langle E_j^a, U_k^b\rangle = \left\langle \sum_{m=1}^{N_e} J_{jm}^{ae}\eta_e * S_m^e, U_k^b \right\rangle \tag{A2}$$

$$= \widetilde{\eta}_e \sum_{m=1}^{N_e} J_{jm}^{ae} \langle S_m^e, U_k^b\rangle \tag{A3}$$

$$= \widetilde{\eta}_e \sum_{m\neq k} J_{jm}^{ae} \langle S_m^e, U_k^b\rangle + \widetilde{\eta}_e J_{jk}^{ae} \langle S_k^e, U_k^b\rangle. \tag{A4}$$

Taking expectation over j and k as above gives

$$\langle E_a, U_b\rangle = \sqrt{N} p_{ae} j_{ae} q_e \widetilde{\eta}_e \langle S_e, U_b\rangle + \mathcal{O}(\mathrm{avg}_{k,b}\langle S_k^e, U_k^b\rangle/\sqrt{N}).$$

where $\mathcal{O}(\mathrm{avg}_{k,b}\langle S_k^e, U_k^b\rangle/\sqrt{N})$ accounts for the diagonal terms not counted in the definition of $\langle S_e, U_b\rangle$. Note that this step requires us to assume that individual values of the random variable, J_{jm}^{ae}, are not strongly correlated with

individual values of $\langle S_m^e, U_k^b\rangle$, so that the expectation of their product can be replaced by the product of their expectations. This assumption is implicit in derivations in other studies [29, 38, 61], even though it is never proven and often not made explicit.

Repeating this calculation for $\langle I_j^a, U_k^b\rangle$ and putting them together gives the average

$$\langle R_a, U_b\rangle = \sqrt{N}(w_{ae}\langle S_e, U_b\rangle + w_{ai}\langle S_i, U_b\rangle) + \mathcal{O}(\text{avg}_{k,b,c}\langle S_k^c, U_k^b\rangle/\sqrt{N}\rangle).$$

In matrix form, this becomes

$$\langle \boldsymbol{R}, \boldsymbol{U}\rangle = \sqrt{N}W\langle \boldsymbol{S}, \boldsymbol{U}\rangle + \mathcal{O}(\text{avg}_{k,b,c}\langle S_k^c, U_k^b\rangle/\sqrt{N}\rangle).$$

An identical calculation, replacing $\boldsymbol{R}$ with $\boldsymbol{X}$ and $\boldsymbol{S}$ with S_x, gives

$$\langle \boldsymbol{X}, \boldsymbol{U}\rangle = \sqrt{N}W_x\langle S_x, \boldsymbol{U}\rangle + \mathcal{O}(\text{avg}_{k,b}\langle S_k^x, U_k^b\rangle/\sqrt{N}\rangle)$$

For the correction terms, $\mathcal{O}(\text{avg}_{k,b,c}\langle S_k^c, U_k^b\rangle/\sqrt{N}\rangle)$ and $\mathcal{O}(\text{avg}_{k,b}\langle S_k^x, U_k^b\rangle/\sqrt{N}\rangle)$, to contribute at largest order in N, it needs to be true that

$$\text{avg}_{k,b}\langle S_k^c, U_k^b\rangle \geq N\mathcal{O}(\langle \boldsymbol{S}, \boldsymbol{U}\rangle)$$

for $c = e, i$, or x. In the correlated state, this is never the case, so the correction term can be ignored, giving Eqs. (10) and (11). In the asynchronous state, $\text{avg}_{k,b}\langle S_k^b, S_k^b\rangle \sim \mathcal{O}(1)$ since spike train power spectral densities are $\mathcal{O}(1)$ due to intrinsically generated variability, but $\langle \boldsymbol{S}, \boldsymbol{S}\rangle \sim \mathcal{O}(1/N)$. This causes the power spectral densities, *i.e.* auto-correlations, to contribute to correlated variability in the network at the largest order in N and ultimately leads to the presence of the C_0 term in Eq. (6) where $(1/N)C_0$ represents mean-field CSDs that would be obtained in the absence of external input correlations, $\langle \boldsymbol{X}, \boldsymbol{X}\rangle = 0$ (see other work [38, 39, 44] for an in-depth treatment of intrinsically generated correlations).

Appendix B: Generation of correlated spike trains for external inputs.

To generate correlated, Poisson spike trains for the external population in the correlated state we used the multiple interaction process (MIP) method [63] with jittering. Specifically, we generated one shared "mother" process with firing rate $r_m = r_x/c$. Then, for each of the N_x "daughter" processes, we randomly kept each spike in the mother process with probability c. As a result, each daughter process is a Poisson process with firing rate $cr_m = r_x$ and a proportion of c of the spikes are shared between any two daughter processes. To get rid of perfect synchrony between the daughter processes, we jittered each spike time in each daughter process by a normally distributed random variable with mean zero and standard deviation $\tau_c = 5$ms. Upon jittering, the daughter processes remain Poisson and the resulting CSD between daughter processes is given by Eq. (15). Spike count correlations between the daughter processes over large time windows are exactly c. The daughter processes were used as the spike trains, $S_j^x(t)$ in the external population in the correlated state. See [63] for a deeper analysis of this algorithm.

Appendix C: Derivation of linear response approximation to pairwise spike train CSDs.

We next give a derivation of Eq. (18) from Eqs. (16) and (17). Similar derivations have previously been given for integrate-and-fire networks [11, 12] and other models [10, 13, 44, 61, 64]. First compute

$$\begin{aligned}\langle \vec{X}, \vec{T}\rangle &= \langle \vec{X}, \vec{R}\rangle + \langle \vec{X}, \vec{X}\rangle \\ &= \langle \vec{X}, \vec{S}\rangle W^* + \langle \vec{X}, \vec{X}\rangle \\ &\approx \langle \vec{X}, \vec{T}\rangle \widetilde{A}^* W^* + \langle \vec{X}, \vec{X}\rangle\end{aligned}$$

from Eqs. (16) and (17). This can be solved for

$$\langle \vec{X}, \vec{T}\rangle = \langle \vec{X}, \vec{X}\rangle (Id - W\widetilde{A})^{-*}.$$

Similarly, compute

$$\begin{aligned}\langle \vec{T}, \vec{T}\rangle &= \langle \vec{R}, \vec{T}\rangle + \langle \vec{X}, \vec{T}\rangle \\ &= W\langle \vec{S}, \vec{T}\rangle + \langle \vec{X}, \vec{T}\rangle \\ &\approx W\widetilde{A}\langle \vec{T}, \vec{T}\rangle + \langle \vec{X}, \vec{X}\rangle (Id - W\widetilde{A})^{-*}\end{aligned}$$

which can be solved to obtain

$$\langle \vec{T}, \vec{T}\rangle = (Id - W\widetilde{A})^{-1}\langle \vec{X}, \vec{X}\rangle (Id - W\widetilde{A})^{-*}.$$

Finally, making the substitution $\langle \vec{S}, \vec{S}\rangle = A\langle \vec{T}, \vec{T}\rangle A^*$, which follows from Eqs. (17), gives Eq. (18).

Appendix D: Parameter values and details of computer simulations

All connection probabilities were $p_{ab} = 0.1$ for $a = e, i$ and $b = e, i, x$. Synaptic timescales were $\tau_e = 8$ms, $\tau_i = 4$ms, and $\tau_x = 10$ms. The firing rate of the external population was $r_x = 10$Hz and, in the correlated state, the correlation was $c = 0.1$ with a jitter of $\tau_c = 5$ms. All covariances and correlations were computed by counting spikes or integrating continuous processes over a window of length 250ms. Membrane capacitance, C_m, is arbitrary so we report all current-based parameters in relation to C_m. For convenience, one can therefore set $C_m = 1$. Unscaled connection strengths were $j_{ee}/C_m = 25$mV, $j_{ei}/C_m = -150$mV, $j_{ie}/C_m = 112.5$mV, $j_{ii}/C_m = -250$mV, $j_{ex}/C_m = 180$mV, and $j_{ix}C_m = 135$mV. Note that j_{ab} was scaled by $\sqrt{N}$ to produce the true connection strengths, as indicated in Results. Neuron parameters are $g_L = C_m/15$, $E_L = -72$mV, $V_{th} = -50$mV,

$V_{re} = -75$mV, $V_{lb} = -100$mV, $\Delta_T = 1$mV, and $V_T = -55$mV. Synaptic currents in figures are reported in units $C_m V/s$. Covariances between synaptic currents are computed between integrals of the currents (see Eq. (3) and surrounding discussion), so the covariances have units $C_m^2 mV^2$.

All simulations and numerical computations were performed on a MacBook Pro running OS X 10.9.5 with a 2.3 GHz Intel Core i7 processor. All simulations were written in Matlab (Matlab R 2018a, MathWorks). The differential equations defining the neuron model were solved using a forward Euler method with time step 0.1ms. Statistics in Figs. 1D, 2C,D, and 4C,D were computed from a simulation of duration 50s. Statistics in Figs. 1E-I, 2E–I, and 4A,B,E,F were computed by repeating a simulation of duration 50s over ten trials for each value of N, then averaging over trials. For each trial, network connectivity was generated with a different random seed, so the statistics are averaged over time and over realizations of the "quenched" variability of network connectivity. Statistics in Fig. 5 were computed by repeating a simulation of duration 100s for 50 trials, with network connectivity the same for each trial. Statistics were then averaged over trials. Gains were estimated by fitting a rectified quadratic function to the relationship between all neurons' firing rates and mean total inputs (r_j and $\overline{T}_j$), then computing the derivative of the fitted quadratic at the input value for each neuron. The same approach was used in previous work [38, 51] to estimate a mean-field gain.

[1] R. Moreno-Bote, A. Renart, and N. Parga, Neural computation **20**, 1651 (2008).
[2] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham, Nature neuroscience **16**, 1170 (2013).
[3] M. Shamir, Current opinion in neurobiology **25**, 140 (2014).
[4] T. Tetzlaff, S. Rotter, E. Stark, M. Abeles, A. Aertsen, and M. Diesmann, Neural Comput **20**, 2133 (2008).
[5] T. Tetzlaff, M. Helias, G. T. Einevoll, and M. Diesmann, PLoS Comput Biol **8**, e1002596 (2012).
[6] B. Doiron, A. Litwin-Kumar, R. Rosenbaum, G. K. Ocker, and K. Josić, Nature Neuroscience **19**, 383 (2016).
[7] G. K. Ocker, Y. Hu, M. A. Buice, B. Doiron, K. Josić, R. Rosenbaum, and E. Shea-Brown, Current opinion in neurobiology **46**, 109 (2017).
[8] B. Lindner, B. Doiron, and A. Longtin, Phys Rev E **72**, 061919 (2005).
[9] S. Ostojic, N. Brunel, and V. Hakim, Journal of Neuroscience **29**, 10234 (2009).
[10] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter, PLoS Comput. Biol. **7**, e1002059 (2011).
[11] J. Trousdale, Y. Hu, E. Shea-Brown, and K. Josić, PLoS Comput Biol **8**, e1002408 (2012).
[12] Y. Hu, J. Trousdale, K. Josić, and E. Shea-Brown, J Stat Mech **2013**, P03012 (2013).
[13] D. Grytskyy, T. Tetzlaff, M. Diesmann, and M. Helias, Front. Comput. Neurosci. **7**, 131 (2013).
[14] B. Dummer, S. Wieland, and B. Lindner, Frontiers in computational neuroscience **8**, 104 (2014).
[15] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, Nature **473**, 87 (2011).
[16] E. Fino and R. Yuste, Neuron **69**, 1188 (2011).
[17] R. B. Levy and A. D. Reyes, J Neurosci **32**, 5609 (2012).
[18] A.-M. Oswald, B. Doiron, J. Rinzel, and A. D. Reyes, J Neurosci **29**, 10321 (2009).
[19] Y. Shu, A. Hasenstaub, and D. A. McCormick, Nature **423**, 288 (2003).
[20] M. Wehr and A. M. Zador, Nature **426**, 442 (2003).
[21] B. Haider, A. Duque, A. R. Hasenstaub, and D. A. McCormick, J Neurosci **26**, 4535 (2006).
[22] M. Okun and I. Lampl, Nat Neurosci **11**, 535 (2008).
[23] A. L. Dorrn, K. Yuan, A. J. Barker, C. E. Schreiner, and R. C. Froemke, Nature **465**, 932 (2010).
[24] Y. J. Sun, G. K. Wu, B.-h. Liu, P. Li, M. Zhou, Z. Xiao, H. W. Tao, and L. I. Zhang, Nature **465**, 927 (2010).
[25] M. Zhou, F. Liang, X. R. Xiong, L. Li, H. Li, Z. Xiao, H. W. Tao, and L. I. Zhang, Nat Neurosci **17**, 841 (2014).
[26] P. C. Petersen, M. Vestergaard, K. H. R. Jensen, and R. W. Berg, J Neurosci **34**, 2774 (2014).
[27] C. van Vreeswijk and H. Sompolinsky, Neural Comput **10**, 1321 (1998).
[28] C. van Vreeswijk and H. Sompolinsky, Science **274**, 1724 (1996).
[29] A. Renart, J. de La Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. Harris, Science **327**, 587 (2010).
[30] A. Ecker, P. Berens, G. Keliris, M. Bethge, N. Logothetis, and A. Tolias, Science **327**, 584 (2010).
[31] M. R. Cohen and A. Kohn, Nat Neurosci **14**, 811 (2011).
[32] M. A. Smith, X. Jia, A. Zandvakili, and A. Kohn, J Neurophysiol **109**, 940 (2013).
[33] A. S. Ecker, P. Berens, R. J. Cotton, M. Subramaniyan, G. H. Denfield, C. R. Cadwell, S. M. Smirnakis, M. Bethge, and A. S. Tolias, Neuron **82**, 235 (2014).
[34] A. Y. Tan, Y. Chen, B. Scholl, E. Seidemann, and N. J. Priebe, Nature **509**, 226 (2014).
[35] M. J. McGinley, M. Vinck, J. Reimer, R. Batista-Brito, E. Zagha, C. R. Cadwell, A. S. Tolias, J. A. Cardin, and D. A. McCormick, Neuron **87**, 1143 (2015).
[36] G. Mochol, A. Hermoso-Mendizabal, S. Sakata, K. D. Harris, and J. de la Rocha, Proc Natl Acad Sci USA **112**, 201410509 (2015).
[37] K. Wimmer, A. Compte, A. Roxin, D. Peixoto, A. Renart, and J. de la Rocha, Nat Commun **6**, 1 (2015).
[38] R. Rosenbaum, M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron, Nature Neurosci **20**, 107 (2017).
[39] R. Darshan, C. van Vreeswijk, and D. Hansel, , 1 (2018), arXiv:1803.00643.
[40] R. Rosenbaum and B. Doiron, Phys Rev X **4**, 021039 (2014).
[41] I. D. Landau, R. Egger, V. J. Dercksen, M. Oberlaender, and H. Sompolinsky, Neuron **92**, 1106 (2016).
[42] R. Pyle and R. Rosenbaum, Phys Rev E **93**, 040302 (2016).
[43] M. Monteforte and F. Wolf, Phys Rev X **2**, 041007 (2012).
[44] M. Helias, T. Tetzlaff, and M. Diesmann, PLoS Comput

Biol **10**, e1003428 (2014).
[45] R. J. Rosenbaum, J. Trousdale, and K. Josić, Front Comp Neurosci **4**, 1 (2010).
[46] R. Rosenbaum, J. Trousdale, and K. Josić, Front Neurosci **5**, 1 (2011).
[47] G. Hennequin, E. J. Agnes, and T. P. Vogels, Annu. Rev. Neurosci. **40**, 557 (2017).
[48] H. Risken, in *The Fokker-Planck Equation* (Springer, 1996) pp. 63–95.
[49] B. Lindner, Phys. Rep. **392**, 321 (2004).
[50] R. Rosenbaum, Front. Comput. Neurosci. **10**, 1 (2016).
[51] C. Ebsch and R. Rosenbaum, PLoS computational biology **14**, e1006048 (2018).
[52] A. Roxin, N. Brunel, and D. Hansel, Physical review letters **94**, 238103 (2005).
[53] B. Kriener, M. Helias, S. Rotter, M. Diesmann, and G. T. Einevoll, Frontiers in Computational Neuroscience **7** (2014), 10.3389/fncom.2013.00187.
[54] S. Ostojic, Nature neuroscience **17**, 594 (2014).
[55] A. Keane and P. Gong, Journal of Neuroscience **35**, 1591 (2015).
[56] R. Pyle and R. Rosenbaum, Physical Rev Lett **118**, 018103 (2017).
[57] C. Huang, D. Ruff, R. Pyle, R. Rosenbaum, M. Cohen, and B. D. Doiron, bioRxiv , 217976 (2017).
[58] E. Ullner, A. Politi, and A. Torcini, bioRxiv , 389452 (2018).
[59] A. Politi, E. Ullner, and A. Torcini, arXiv preprint arXiv:1808.03359 (2018).
[60] K. C.-K. Malina, B. Mohar, A. N. Rappaport, and I. Lampl, Nat Comm **7**, 12740 (2016).
[61] D. Dahmen, S. Grün, M. Diesmann, and M. Helias, , 1 (2017), arXiv:1711.10930.
[62] T. Herfurth and T. Tchumatchenko, Current opinion in neurobiology **46**, 234 (2017).
[63] A. Kuhn, A. Aertsen, and S. Rotter, Neural Comput. **15**, 67 (2003).
[64] C. W. Gardiner, *Handbook of stochastic methods*, Vol. 3 (Springer, Berlin, 1985).