Exact analysis of the subthreshold variability for conductance-based neuronal models with synchronous synaptic inputs

Logan A. Becker,[1,2] Baowang Li,[1,2,3,4,5] Nicholas J. Priebe,[1,2,4]
Eyal Seidemann,[1,2,3,5] and Thibaud Taillefumier[1,2,6,*]

[1] *Center for Theoretical and Computational Neuroscience, The University of Texas at Austin*
[2] *Department of Neuroscience, The University of Texas at Austin*
[3] *Center for Perceptual Systems, The University of Texas at Austin*
[4] *Center for Learning and Memory, The University of Texas at Austin*
[5] *Department of Psychology, The University of Texas at Austin*
[6] *Department of Mathematics, The University of Texas at Austin*

The spiking activity of neocortical neurons exhibits a striking level of variability, even when these networks are driven by identical stimuli. The approximately Poisson firing of neurons has led to the hypothesis that these neural networks operate in the asynchronous state. In the asynchronous state neurons fire independently from one another, so that the probability that a neuron experience synchronous synaptic inputs is exceedingly low. While the models of asynchronous neurons lead to observed spiking variability, it is not clear whether the asynchronous state can also account for the level of subthreshold membrane potential variability. We propose a new analytical framework to rigorously quantify the subthreshold variability of a single conductance-based neuron in response to synaptic inputs with prescribed degrees of synchrony. Technically we leverage the theory of exchangeability to model input synchrony via jump-process-based synaptic drives; we then perform a moment analysis of the stationary response of a neuronal model with all-or-none conductances that neglects post-spiking reset. As a result, we produce exact, interpretable closed forms for the first two stationary moments of the membrane voltage, with explicit dependence on the input synaptic numbers, strengths, and synchrony. For biophysically relevant parameters, we find that the asynchronous regime only yields realistic subthreshold variability (voltage variance $\simeq 4-9\mathrm{mV}^2$) when driven by a restricted number of large synapses, compatible with strong thalamic drive. By contrast, we find that achieving realistic subthreshold variability with dense cortico-cortical inputs requires including weak but nonzero input synchrony, consistent with measured pairwise spiking correlations. We also show that without synchrony, the neural variability averages out to zero for all scaling limits with vanishing synaptic weights, independent of any balanced state hypothesis. This result challenges the theoretical basis for mean-field theories of the asynchronous state.

I. INTRODUCTION

A common and striking feature of cortical activity is the high degree of neuronal spiking variability [1]. This high variability is notably present in sensory cortex and motor cortex, as well as in regions with intermediate representations [2–5]. The prevalence of this variability has led to it being a major constraint for modeling cortical networks. Cortical networks may operate in distinct regimes depending on species, cortical area, and brain states. In asleep or anesthetized state, neurons tend to fire synchronously with strong correlations between the firing of distinct neurons [6–8]. In the awake state, although synchrony has been reported as well, stimulus drive, arousal, or attention tend to promote an irregular firing regime whereby neurons spike in a seemingly random manner, with decreased or little correlation [1, 8, 9]. This has lead to the hypothesis that cortex primarily operates asynchronously [10–12]. In the asynchronous state, neurons fire independently from one another, so that the probability that a neuron experiences synchronous synaptic inputs is exceedingly low. That said, the asynchronous state hypothesis appears at odds with the high-degree of observed spiking variability in cortex. Cortical neurons are thought to receive a large number of synaptic inputs ($\simeq 10^4$) [13]. Although the impact of these inputs may vary across synapses, the law of large numbers implies that variability should average out when integrated at the soma. In principle, this would lead to clock-like spiking responses, contrary to experimental observations [14].

A number of mechanisms have been proposed to explain how high spiking variability emerges in cortical networks [15]. The prevailing approach posits that excitatory and inhibitory inputs converge on cortical neurons in a balanced manner. In balanced models, the overall excitatory and inhibitory drives cancel each other so that transient imbalances in the drive can bring the neuron's membrane voltage across the spike-initiation threshold. Such balanced models result in spiking statistics that match those found in the neocortex [16, 17]. However, these statistics can emerge in distinct dynamical regimes depending on whether the balance between excitation and inhibition is tight or loose [18]. In tightly balanced networks, whereby the net neuronal drive is negligible compared to the antagonizing components, activity correlation is effectively zero, leading to a strictly asynchronous regime [19–21]. By contrast, in loosely-

* Corresponding author; ttaillef@austin.utexas.edu

balanced networks, the net neuronal drive remains of the same order as the antagonizing components, which allows for strong neuronal correlations during evoked activity, compatible with a synchronous regime [22–24].

While the high spiking variability is an important constraint for cortical network modeling, there are other biophysical signatures that may be employed. We now have access to the subthreshold membrane voltage fluctuations that underlie spikes in awake, behaving animals (see Fig. 1). Membrane voltage recordings reveal two main deviations from the asynchronous hypothesis: first, membrane voltage does not hover near spiking threshold and is modulated by the synaptic drive; second, it exhibits state- or stimulus dependent non-Gaussian fluctuation statistics with positive skewness [25–28]. In this work, we further argue that membrane voltage recordings reveal much larger voltage fluctuations than predicted by balanced cortical models [29, 30].

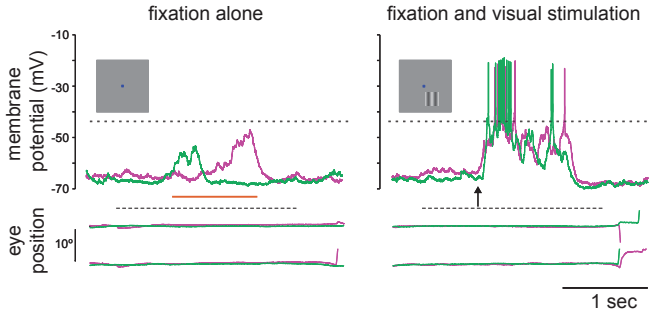

FIG. 1. **Large trial-by-trial membrane voltage fluctuations.** Membrane voltage responses are shown using whole cell recordings in awake behaving primates for both fixation alone trials (left) and visual stimulation trials (right). A drifting grating was presented for 1 second beginning at the arrow. Below the membrane voltage traces are records of horizontal and vertical eye movements, illustrating that the animal was fixating during the stimulus. Red and green traces indicate different trials under the same conditions. Adapted from [27].

How could such large subthreshold variations in membrane voltage emerge? One way that fluctuations could emerge, even for large numbers of input, is if there is synchrony in the driving inputs [31]. In practice, input synchrony is revealed by the presence of positive spiking correlations, which quantify the propensity of distinct synaptic inputs to coactivate. Measurements of spiking correlations between pairs of neurons vary across reports, but have generally been shown to be weak [10–12]. That said, even weak correlations can have a large impact when the population of correlated inputs is large [32, 33]. Further, the existence of input synchrony, supported by weak but persistent spiking correlations, is consistent with at least two other experimental observations. First, intracellular recordings from pairs of neurons in both anesthetized and awake animals reveal a high degree of membrane voltage correlations [7, 34, 35]. Second, excitatory and inhibitory conductance inputs are highly correlated with each other within the same neuron [35, 36]. These observations suggest that input synchrony could explain the observed level of subthreshold variability.

While our focus is on achieving realistic subthreshold variability, other challenges to asynchronous networks have been described. In particular, real neural networks exhibit distinct regimes of activity depending on the strength of their afferent drives [37]. In that respect, Zerlaut *et al.* showed that asynchronous networks can exhibit a spectrum of realistic regimes of activity if they have moderate recurrent connections and are driven by strong thalamic projections (see also [17]). Furthermore, it has been a challenge to identify the scaling rule that should apply to synaptic strengths for asynchrony to hold stably in idealized networks [38]. Recently, Sanzeni *et al.* proposed that a realistic asynchronous regime is achieved for a particular large-coupling rule, whereby synaptic strengths scale in keeping with the logarithmic size of the network. Both studies consider balanced networks with conductance-based neuronal models but neither focuses on the role of synchrony, consistent with the asynchronous state hypothesis. The asynchronous state hypothesis is theoretically attractive because it represents a naturally stable regime of activity in infinite-size, balanced networks of current-based neuronal models [16, 17, 20, 21]. Such neuronal models, however, neglect the voltage dependence of conductances and it remains unclear whether the asynchronous regime is asymptotically stable for infinite-size, conductance-based network models.

Here, independent of the constraint of network stability, we ask whether biophysically relevant neuronal models can achieve the observed subthreshold variability under realistic levels of input synchrony. To answer this question, we derive exact analytical expressions for the stationary voltage variance of a single conductance-based neuron in response to synchronous shot-noise drives [39, 40]. A benefit of shot-noise models compared to diffusion models is to allow for individual synaptic inputs to be temporally separated in distinct impulses, each corresponding to a transient positive conductance fluctuations [41–43]. We develop our shot-noise analysis for a variant of classically considered neuronal models. We call this variant the all-or-none-conductance-based (AONCB) model for which synaptic activation occurs as an all-or-none process rather than as an exponentially relaxing process. To perform an exact treatment of these models, we develop original probabilistic techniques inspired from Marcus' work about shot-noise driven dynamics [44, 45]. To model shot-noise drives with synchrony, we develop a statistical framework based on the property of input exchangeability, which assumes that no synaptic inputs play a particular role. In this framework, we show that input drives with varying degree of synchrony can be rigorously modeled via jump processes, while synchrony can be quantitatively related to measures of pairwise spiking correlations.

Our main results are biophysically interpretable for-

mulas for the voltage mean and variance in the limit of instantaneous synapses. Crucially, these formulas explicitly depend on the input numbers, weights, and synchrony, and hold without any forms of diffusion approximation. This is in contrast with analytical treatments which elaborate on the diffusion and effective-time-constant approximations [37, 38, 46, 47]. We leverage these exact, explicit formulas to determine under which synchrony conditions a neuron can achieve the experimentally observed subthreshold variability. For biophysically relevant synaptic numbers and weights, we find that achieving realistic variability is possible in response to a restricted number of large asynchronous connections, compatible with the dominance of thalamo-cortical projections in the input layers of the visual cortex. However, we find that achieving realistic variability in response to a large number of moderate cortical inputs, as in superficial cortical visual layers, necessitates nonzero input synchrony in amounts that are consistent with the weak levels of measured spiking correlations observed *in vivo*.

In practice, persistent synchrony may spontaneously emerge in large but finite neural networks, as nonzero correlations are the hallmark of finite-dimensional interacting dynamics. The network structural features responsible for the magnitude of such correlations remains unclear, and we do not address this question here (see [48, 49] for review). The persistence of synchrony is also problematic for theoretical approaches that consider networks in the infinite-size limits. Indeed, our analysis supports that in the absence of synchrony and for all scaling of the synaptic weights, subthreshold variability must vanish in the limit of arbitrary large numbers of synapses. This suggests that independent of any balanced condition, the mean-field dynamics that emerge in infinite-size networks of conductance-based neurons will not exhibit Poisson-like spiking variability, at least in the absence of additional constraints on the network structure or on the biophysical properties of the neurons. In current-based neuronal models, however, variability is not dampened by a conductance-dependent effective time constant. These findings therefore challenge the theoretical basis for the asynchronous state in conductance-based neuronal networks.

Our exact analysis, as well as its biophysical interpretations, is only possible at the cost of several caveats: First, we neglect the impact of the spike-generating mechanism (and of the post-spiking reset) in shaping the subthreshold variability. Second, we quantify synchrony under the assumption of input exchangeability, that is, for synapses having a typical strength as opposed to being heterogeneous. Third, we consider input drives that implement an instantaneous form of synchrony with temporally precise synaptic coactivations. Fourth, we do not consider slow temporal fluctations in the mean synaptic drive. Fifth, and perhaps most concerning, we do not account for the stable emergence of a synchronous regime in network models. We argue in the discussion that all the above caveats but the last one can be addressed without impacting our findings. Addressing the last caveat remains an open problem.

For reference, we list in Table I the main notations used in this work. These notations utilize the subscript $\{\}_{\mathrm{e}}$, $\{\}_{\mathrm{i}}$ to refer to excitation or inhibition, respectively. The notation $\{\}_{\mathrm{e/i}}$ means that the subscript can be either $\{\}_e$ or $\{\}_i$. The notation $\{\}_{\mathrm{ei}}$ is used to emphasize that a quantity depends jointly on excitation and inhibition.

II. STOCHASTIC MODELING AND ANALYSIS

A. All-or-none-conductance-based neurons

We consider the subthreshold dynamics of an original neuronal model, which we called the all-or-none-conductance-based (AONCB) model. In this model, the membrane voltage V obeys the first-order stochastic differential equation

$$C\dot{V} = G(V_{\mathrm{L}} - V) + g_{\mathrm{e}}(V_{\mathrm{e}} - V) + g_{\mathrm{i}}(V_{\mathrm{i}} - V) + I\,, \quad (1)$$

where randomness arises from the stochastically activating excitatory and inhibitory conductances, respectively denoted by g_{e} and g_{i} (see Fig. 2a). These conductances result from the action of K_{e} excitatory and K_{i} inhibitory synapses: $g_{\mathrm{e}}(t) = \sum_{k=1}^{K_{\mathrm{e}}} g_{\mathrm{e},k}(t)$ and $g_{\mathrm{i}}(t) = \sum_{k=1}^{K_{\mathrm{i}}} g_{\mathrm{i},k}(t)$. In the absence of synaptic inputs, i.e., when $g_{\mathrm{e}} = g_{\mathrm{i}} = 0$, and of external current I, the voltage exponentially relaxes toward its leak reversal potential V_{L} with passive time constant $\tau = C/G$, where C denotes the cell's membrane capacitance and G denotes the cellular passive conductance [50]. In the presence of synaptic inputs, the transient synaptic currents $I_{\mathrm{e}} = g_{\mathrm{e}}(V_{\mathrm{e}} - V)$ and $I_{\mathrm{i}} = g_{\mathrm{i}}(V_{\mathrm{i}} - V)$ cause the membrane voltage to fluctuate. Conductance-based models account for the voltage-dependence of synaptic currents via the driving forces $V_{\mathrm{e}} - V$ and $V_{\mathrm{i}} - V$, where V_{e} and V_{i} denotes the excitatory and inhibitory reversal potential, respectively. Without loss of generality, we assume in the following that $V_{\mathrm{L}} = 0$ and that $V_{\mathrm{i}} < V_{\mathrm{L}} = 0 < V_{\mathrm{e}}$.

We model the spiking activity of the $K_{\mathrm{e}} + K_{\mathrm{i}}$ upstream neurons as shot noise [39, 40], which can be generically modeled as a $(K_{\mathrm{e}} + K_{\mathrm{i}})$-dimensional stochastic point process [51, 52]. Let us denote by $\{N_{\mathrm{e},k}(t)\}_{1 \le k \le K_{\mathrm{e}}}$ its excitatory component and by $\{N_{\mathrm{i},k}(t)\}_{1 \le k \le K_{\mathrm{i}}}$ its inhibitory component, where t denotes time and k is the neuron index. For each neuron k, the process $N_{\mathrm{e/i},k}(t)$ is specified as the counting process registering the spiking occurrences of neuron k up to time t. In other words, $N_{\mathrm{e/i},k}(t) = \sum_k \mathbb{1}_{\{T_{\mathrm{e/i},k,n} \le t\}}$, where $\{T_{\mathrm{e/i},k,n}\}_{n \in \mathbb{Z}}$ denotes the full sequence of spiking times of neuron k and where $\mathbb{1}_A$ denotes the indicator function of set A. Note that by convention, we label spikes so that $T_{\mathrm{e/i},k,0} \le 0 < T_{\mathrm{e/i},k,1}$ for all neuron k. Given a point-process model for the upstream spiking activity, classical conductance-based models consider that a single input to a synapse causes an instantaneous increase of its conductance, followed by an exponential decay with typical time scale $\tau_{\mathrm{s}} > 0$. Here

TABLE I. Main notations.

$a_{\mathrm{e/i},1}$	first-order synaptic efficacies
$a_{\mathrm{e/i},2}$	second-order synaptic efficacies
$a_{\mathrm{e/i},12}$	auxiliary second-order synaptic efficacies
b,	rate of the driving Poisson process N
$b_{\mathrm{e/i}}$	rate of the excitatory/inhibitory Poisson process $N_{\mathrm{e/i}}$
C	membrane capacitance
c_{ei},	crosscorrelation synaptic efficacy
$\mathbb{C}[\cdot,\cdot]$	stationary covariance
$\mathbb{E}[\cdot]$	stationary expectation
$\mathbb{E}_{\mathrm{ei}}[\cdot]$	expectation w.r.t. the joint distribution p_{ei} or $p_{\mathrm{ei},kl}$
$\mathrm{E}_{\mathrm{e/i}}[\cdot]$	expectation w.r.t. the marginal distribution $p_{\mathrm{e/i}}$ or $p_{\mathrm{e/i},k}$
$\epsilon = \tau_{\mathrm{s}}/\tau$	fast-conductance small parameter
G	passive leak conductance
$g_{\mathrm{e/i}}$	overall excitatory/inhibitory conductance
$h_{\mathrm{e/i}} = g_{\mathrm{e/i}}/C$	reduced excitatory/inhibitory conductance
$k_{\mathrm{e/i}}$	number of coactivating excitatory/inhibitory synaptic inputs
$K_{\mathrm{e/i}}$	total number of excitatory/inhibitory synaptic inputs
N	driving Poisson process with rate b
$N_{\mathrm{e/i}}$	excitatory/inhibitory driving Poisson process with rate $b_{\mathrm{e/i}}$
p_{ei}	bivariate jump distribution of $(W_{\mathrm{e}}, W_{\mathrm{i}})$
$p_{\mathrm{e/i}}$	marginal jump distribution of $W_{\mathrm{e/i}}$
$p_{\mathrm{ei},kl}$	bivariate distribution for the numbers of coactivating synapses $(k_{\mathrm{e}}, k_{\mathrm{i}})$
$p_{\mathrm{e/i},k}$	marginal synaptic count distribution $k_{\mathrm{e/i}}$
$r_{\mathrm{e/i}}$	individual excitatory/inhibitory synaptic rate
ρ_{ei}	spiking correlation between excitatory and inhibitory inputs
$\rho_{\mathrm{e/i}}$	spiking correlation within excitatory/inhibitory inputs
τ	passive membrane time constant
τ_{s}	synaptic time constant
$\mathbb{V}[\cdot]$	stationary variance
$W_{\mathrm{e/i}}$	excitatory/inhibitory random jumps
$V_{\mathrm{e/i}}$	excitatory/inhibitory reversal potentials
$w_{\mathrm{e/i}}$	typical value for excitatory/inhibitory synaptic weights
X_k	binary variable indicating the activation of excitatory synapse k
Y_l	binary variable indicating the activation of inhibitory synapse l
Z	driving compound Poisson process with base rate b and jump distribution p_{ei}

we depart from this assumption and consider that the synaptic conductances $g_{\mathrm{e/i},k}$ operates all-or-none with a common activation time still referred to as τ_{s}. Specifically, we assume that the dynamics of the conductance $g_{\mathrm{e/i},k}$ follows

$$\dot{g}_{\mathrm{e/i},k}(t) = w_{\mathrm{e/i},k} \sum_n \left(\delta(t - T_{\mathrm{e/i},k,n}) - \delta(t - T_{\mathrm{e/i},k,n} - \tau_{\mathrm{s}})\right), \tag{2}$$

where $w_{\mathrm{e/i},k} \geq 0$ is the dimensionless synaptic weight.

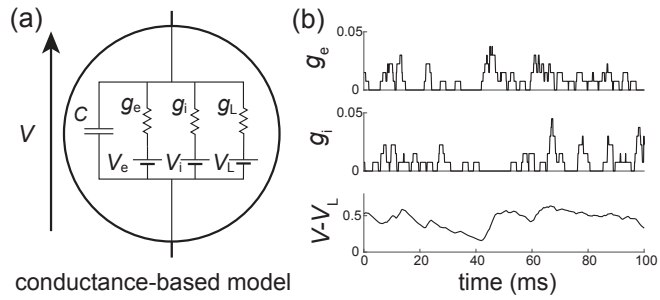

FIG. 2. **All-or-none-conductance-based models.** (a) Electrical diagram of conductance-based model for which the neuronal voltage V evolves in response to fluctuations of excitatory and inhibitory conductances g_{e} and g_{i}. (b) In all-or-none models, inputs delivered as Poisson processes transiently activate the excitatory and inhibitory conductances g_{e} and g_{i} during a finite, nonzero synaptic activation time $\tau_{\mathrm{s}} > 0$. Simulation parameters: $K_{\mathrm{e}} = K_{\mathrm{i}} = 50$, $r_{\mathrm{e}} = r_{\mathrm{i}} = 10\mathrm{Hz}$, $\tau = 15\mathrm{ms}$ $\tau_{\mathrm{s}} = 2\mathrm{ms} > 0$.

The above equation prescribes that the n-th spike delivery to synapse k at time $T_{\mathrm{e/i},k,n}$ is followed by an instantaneous increase of that synapse's conductance by an amount $w_{\mathrm{e/i},k}$ for a period τ_{s}. Thus, the synaptic response prescribed by Eq. (2) is all-or-none as opposed to being graded as in classical conductance-based models. Moreover, just as in classical models, Eq. (2) allows synapses to multi-activate, thereby neglecting nonlinear synaptic saturation (see Fig. 2b).

To be complete, AONCB neurons must in principle include a spike-generating mechanism. A customary choice is the integrate-and-fire mechanism [53, 54]: a neuron emits a spike whenever its voltage V exceeds a threshold value V_{T}, and reset instantaneously to some value V_R afterwards. Such a mechanism impacts the neuronal subthreshold voltage dynamics via post-spiking reset, which implements a nonlinear form of feedback. However, in this work we focus on the variability that is generated by fluctuating, possibly synchronous, synaptic inputs. For this reason, we will neglect the influence of the spiking reset in our analysis and actually, we will ignore the spike-generating mechanism altogether. Finally, although our analysis of AONCB neurons will apply to positive synaptic activation time $\tau_{\mathrm{s}} > 0$, we will only discuss our results in the limit of instantaneous synapses. This corresponds to taking $\tau_{\mathrm{s}} \to 0^+$ while adopting the scaling $g_{\mathrm{e/i}} \propto 1/\tau_s$ in order to maintain the charge transfer induced by a synaptic event. We will see that this limiting process preserves the response variability of AONCB neurons.

B. Quantifying the synchrony of exchangeable synaptic inputs

Our goal here is to introduce a discrete model for synaptic inputs, whereby synchrony can be rigorously

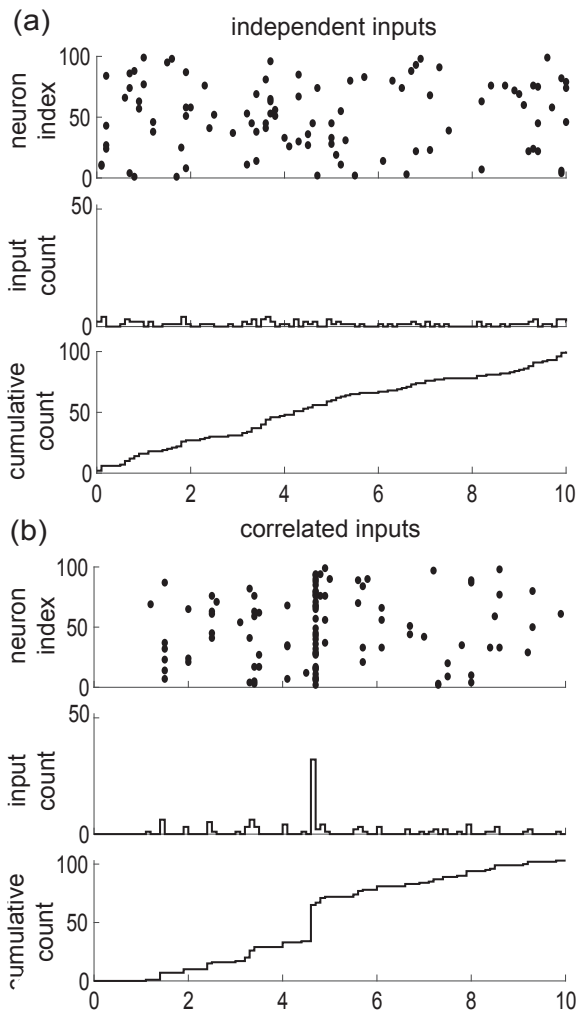

FIG. 3. **Parametrizing correlations via exchangeability.** The activity of $K_\mathrm{e} = 100$ exchangeable synaptic inputs collected over N consecutive time bins can be represented as $\{0,1\}$-valued array $\{X_{k,i}\}_{1\leq k\leq K_\mathrm{e}, 1\leq i\leq N}$, where $X_{k,i} = 1$ if input k activates in time bin i. Under assumptions of exchangeability, the input spiking correlation is entirely captured by the count statistics of how many inputs coactivate within a given time bin. In the limit $K_\mathrm{e} \to \infty$, the distribution of the fraction of coactivating inputs coincides with the directing de Finetti measure, which we consider as a parametric choice in our approach. In the absence of correlation, synapses tend to activate in isolation: $\rho_\mathrm{e} = 0$ in (a). In the presence of correlation, synapses tend to coactivate yielding disproportionately large synaptic activation event: $\rho_\mathrm{e} = 0.1$ in (b). Considering the associated cumulative counts specifies discrete-time jump processes that can be generalized to the continuous-time limit, i.e., for time bins of vanishing duration $\Delta t \to 0^+$.

quantified. To this end, let us suppose that the neuron under consideration receives inputs from K_e excitatory neurons and K_i inhibitory neurons, chosen from arbitrary large pools of $N_\mathrm{e} \gg K_\mathrm{e}$ excitatory neurons and $N_\mathrm{i} \gg K_\mathrm{i}$ inhibitory neurons. Adopting a discrete-time representation with elementary bin size Δt, we denote by $\{\{x_{1,n},\ldots,x_{K_\mathrm{e},n}\},\{y_{1,n},\ldots,y_{K_\mathrm{i},n}\}\}$ in $\{0,1\}^{K_\mathrm{e}} \times \{0,1\}^{K_\mathrm{i}}$ the input state within the n-th bin. Our main simplifying assumption consists in modeling the N_e excitatory inputs and the N_i inhibitory inputs as separately exchangeable random variables $\{X_{1,n},\ldots,X_{K_\mathrm{e},n}\}$ and $\{Y_{1,n},\ldots,Y_{K_\mathrm{i},n}\}$ that are distributed identically over $\{0,1\}^{N_\mathrm{e}}$ and $\{0,1\}^{N_\mathrm{i}}$, respectively, and independently across time. This warrants dropping the dependence on time index n. By separately exchangeable, we mean that no subset of excitatory inputs or inhibitory inputs plays a distinct role so that at all time, the respective distributions of $\{X_{1,n},\ldots,X_{K_\mathrm{e},n}\}$ and $\{Y_{1,n},\ldots,Y_{K_\mathrm{i},n}\}$ are independent of the inputs labelling. In other words, for all permutations σ_e of $\{1,\ldots,N_\mathrm{e}\}$ and σ_i of $\{1,\ldots,N_\mathrm{i}\}$, the joint distribution of $\{X_{\sigma_\mathrm{e}(1)},\ldots,X_{\sigma_\mathrm{e}(N_\mathrm{e})}\}$ and $\{Y_{\sigma_\mathrm{i}(1)},\ldots,Y_{\sigma_\mathrm{i}(N_\mathrm{i})}\}$ is identical to that of $\{X_1,\ldots,X_{N_\mathrm{e}}\}$ and $\{Y_1,\ldots,Y_{N_\mathrm{i}}\}$ [55, 56]. By contrast with independent random spiking variables, exchangeable ones can exhibit nonzero correlation structure. By symmetry, this structure is specified by three correlation coefficients

$$\rho_\mathrm{e} = \frac{\mathbb{C}\left[X_k, X_l\right]}{\mathbb{V}\left[X_k\right]}\,,\ \rho_\mathrm{i} = \frac{\mathbb{C}\left[Y_k, Y_l\right]}{\mathbb{V}\left[Y_k\right]}\,,\ \rho_\mathrm{ei} = \frac{\mathbb{C}\left[X_k, Y_l\right]}{\sqrt{\mathbb{V}\left[X_k\right]\mathbb{V}\left[Y_l\right]}}\,,$$

where $\mathbb{C}\left[X,Y\right]$ and $\mathbb{V}\left[X\right]$ denote the covariance and the variance of the binary variables X and Z, respectively.

Interestingly, a more explicit form for ρ_e, ρ_e, and ρ_ei can be obtained in the limit of an infinite-size pool $N_\mathrm{e}, N_\mathrm{i} \to \infty$. This follows from de Finetti's theorem [57], which states that the probability of observing a given input configuration for K_e excitatory neurons and K_i inhibitory neurons is given by

$$\mathbb{P}\left[X_1,\ldots,X_{K_\mathrm{e}},Y_1,\ldots,Y_{K_\mathrm{i}}\right] = \int \prod_{k=1}^{K_\mathrm{e}} \theta_\mathrm{e}^{X_k}(1-\theta_\mathrm{e})^{1-X_k} \prod_{l=1}^{K_\mathrm{i}} \theta_\mathrm{i}^{X_l}(1-\theta_\mathrm{i})^{1-X_l}\,\mathrm{d}F_\mathrm{ei}(\theta_\mathrm{e},\theta_\mathrm{i})\,.$$

where F_ei is the directing de Finetti measure, defined as a bivariate distribution over the unit square $[0,1]\times[0,1]$. In the equation above, the number θ_e and θ_i represent the (jointly fluctuating) probabilities that an excitatory neuron and an inhibitory neuron spike in a given time bin, respectively. The core message of de Finetti theorem is that the spiking activity of neurons from infinite exchangeable pools is obtained as a mixture of conditionally independent binomial laws. This mixture is specified by the directing measure F_ei, which fully parametrizes our synchronous input model. Independent spiking corresponds to choosing F_ei as a point-mass measure concentrated on some probabilities $\pi_\mathrm{e/i} = r_\mathrm{e/i}\Delta t$, where $r_\mathrm{e/i}$ denotes the individual spiking rate of a neuron: $\mathrm{d}F_\mathrm{ei}(\theta) = \delta(\theta_\mathrm{e}-\pi_\mathrm{e})\delta(\theta_\mathrm{i}-\pi_\mathrm{i})\mathrm{d}\theta_\mathrm{e}\mathrm{d}\theta_\mathrm{i}$ (see Fig 3a). By contrast, a dispersed directing measure F_ei corresponds to the existence of correlations among the inputs (see Fig 3b). Accordingly, we show in Appendix A that the spiking pairwise correlation $\rho_\mathrm{e/i}$ take the explicit form

$$\rho_\mathrm{e/i} = \frac{\mathbb{V}\left[\theta_\mathrm{e/i}\right]}{\mathbb{E}\left[\theta_\mathrm{e/i}\right]\left(1-\mathbb{E}\left[\theta_\mathrm{e/i}\right]\right)}\,, \tag{3}$$

whereas ρ_{ei}, the correlation between excitation and inhibition, is given by

$$\rho_{\mathrm{ei}} = \frac{\mathbb{C}\left[\theta_{\mathrm{e}}, \theta_{\mathrm{i}}\right]}{\sqrt{\mathbb{E}\left[\theta_{\mathrm{e}}\right]\mathbb{E}\left[\theta_{\mathrm{i}}\right](1-\mathbb{E}\left[\theta_{\mathrm{e}}\right])(1-\mathbb{E}\left[\theta_{\mathrm{i}}\right])}}\,. \tag{4}$$

In the above formulas, $\mathbb{E}\left[\theta_{\mathrm{e/i}}\right]$, $\mathbb{V}\left[\theta_{\mathrm{e/i}}\right]$, and $\mathbb{C}\left[\theta_{\mathrm{e}}, \theta_{\mathrm{i}}\right]$ denote expectation, variance, and covariance of $(\theta_{\mathrm{e}}, \theta_i) \sim F_{\mathrm{ei}}$, respectively. Note that these formulas show that nonzero correlations $\rho_{\mathrm{e/i}}$ correspond to nonzero variance, as is always the case for dispersed distribution. Independence between excitation and inhibition for which $\rho_{\mathrm{ei}} = 0$ corresponds to directing measure F_{ei} with product form, i.e., $F_{\mathrm{ei}}(\theta_{\mathrm{e}}, \theta_{\mathrm{i}}) = F_{\mathrm{e}}(\theta_{\mathrm{e}})F_{\mathrm{i}}(\theta_{\mathrm{i}})$, where F_{e} and F_{i} denote the marginal distributions. Alternative forms of the directed measure F_{ei} generally lead to nonzero cross correlation ρ_{ei}, which necessarily satisfies $0 < |\rho_{\mathrm{ei}}| \leq \sqrt{\rho_{\mathrm{e}}\rho_{\mathrm{i}}}$.

In this exchangeable setting, a reasonable parametric choice for the marginals F_{e} and F_{i} is given by beta distributions $\mathrm{Beta}(\alpha, \beta)$, where α and β denote shape parameters [58]. Practically, this choice is motivated by the ability of beta distributions to efficiently fit correlated spiking data generated by existing algorithms [59]. Formally, this choice is motivated by the fact that beta distributions are conjugate priors for the binomial likelihood functions, so that the resulting probabilistic models can be studied analytically [60–62]. For instance, for $F_{\mathrm{e}} \sim \mathrm{Beta}(\alpha_{\mathrm{e}}, \beta_{\mathrm{e}})$, the probability that k_{e} synapses among the K_{e} inputs are jointly active within the same time bin follows the beta-binomial distribution

$$P_{\mathrm{e},k} = \binom{K_{\mathrm{e}}}{k}\frac{B(\alpha_{\mathrm{e}}+k, \beta_{\mathrm{e}}+K_{\mathrm{e}}-k)}{B(\alpha_{\mathrm{e}}, \beta_{\mathrm{e}})}\,. \tag{5}$$

Accordingly, the mean number of active excitatory inputs is $\mathbb{E}\left[k_{\mathrm{e}}\right] = K_{\mathrm{e}}\alpha_{\mathrm{e}}/(\alpha_{\mathrm{e}}+\beta_{\mathrm{e}}) = K_{\mathrm{e}}r_{\mathrm{e}}\Delta t$. Utilizing Eq. (3), we also find that $\rho_{\mathrm{e}} = 1/(1+\alpha_{\mathrm{e}}+\beta_{\mathrm{e}})$. Note that the above results show that by changing de Finetti's measure, one can not only modify the spiking correlation but also the mean spiking rate.

In the following, we will exploit the above analytical results to illustrate that taking the continuous-time limit $\Delta t \to 0^+$ specifies synchronous input drives as compound Poisson processes [51, 52]. To do so, we will consider both excitation and inhibition, which in a discrete setting corresponds to considering bivariate probability distributions $P_{\mathrm{ei},kl}$ defined over $\{0, \ldots, K_{\mathrm{e}}\} \times \{0, \ldots, K_{\mathrm{i}}\}$. Ideally, these distributions $P_{\mathrm{ei},kl}$ should be such that its conditional marginals $P_{\mathrm{e},k}$ and $P_{\mathrm{i},l}$, with distributions given by Eq. (5). Unfortunately, there does not seem to be a simple low-dimensional parametrization for such distributions $P_{\mathrm{ei},kl}$, except in particular cases. To address this point, at least numerically, one can resort to a variety of methods including copulas [63, 64]. For analytical calculations, we will only consider consider two particular cases for which the marginals of F_{ei} are given by the beta distributions: (i) the case of maximum positive correlation for which $\theta_{\mathrm{e}} = \theta_{\mathrm{i}}$, i.e., $\mathrm{d}F_{\mathrm{ei}}(\theta_{\mathrm{e}}, \theta_{\mathrm{i}}) = \delta(\theta_{\mathrm{e}} - \theta_{\mathrm{i}})F(\theta_{\mathrm{e}})\,\mathrm{d}\theta_{\mathrm{e}}\mathrm{d}\theta_{\mathrm{i}}$ with $F_{\mathrm{e}} = F_{\mathrm{i}} = F$ and (ii) the case of zero correlation for which θ_{e} and θ_{i} are independent, i.e., $F_{\mathrm{ei}}(\theta_{\mathrm{e}}, \theta_{\mathrm{i}}) = F_{\mathrm{e}}(\theta_{\mathrm{e}})F_{\mathrm{i}}(\theta_{\mathrm{i}})$.

C. Synchronous synaptic drives as compound Poisson processes

Under assumption of input exchangeability and given typical excitatory and inhibitory synaptic weights $w_{\mathrm{e/i}}$, the overall synaptic drive to a neuron is determined by $(k_{\mathrm{e}}, k_{\mathrm{i}})$, the numbers of active excitatory and inhibitory inputs at each discrete time step. As AONCB dynamics unfolds in continuous time, we need to consider this discrete drive in the continuous-time limit as well, i.e., for vanishing time bins $\Delta t \to 0^+$. When $\Delta t \to 0^+$, we show in Appendix B that the overall synaptic drive specifies a compound Poisson process Z with bivariate jumps $(W_{\mathrm{e}}, W_{\mathrm{i}})$. Specifically, we have

$$Z(t) = \left(\sum_n^{N(t)} W_{\mathrm{e},n}, \sum_n^{N(t)} W_{\mathrm{i},n}\right)\,, \tag{6}$$

where $(W_{\mathrm{e},n}, W_{\mathrm{i},n})$ are i.i.d. samples with bivariate distribution denoted by p_{ei} and where the overall driving Poisson process N registers the number of synaptic events without mulitple counts (see Fig. 4). By synaptic events, we mean these times for which at least one excitatory synapse or one inhibitory synapse activates. We say that N registers these events without multiple count as it counts one event independent of the number of possibly coactivating synapses. Similarly, we denote by N_{e} and N_{i} the counting processes registering synaptic excitatory events and synaptic inhibitory events alone, respectively. These processes N_{e} and N_{i} are Poisson processes that are correlated in the presence of synchrony, as both N_{e} and N_{i} may register the same event. Note that this implies that $\max(N_{\mathrm{e}}(t), N_{\mathrm{i}}(t)) \leq N(t) \leq N_{\mathrm{e}}(t) + N_{\mathrm{i}}(t)$. More generally, denoting by b and $b_{\mathrm{e/i}}$ the rates of N and $N_{\mathrm{e/i}}$, respectively, the presence of synchrony implies that $\max(b_{\mathrm{e}}, b_{\mathrm{i}}) \leq b \leq b_{\mathrm{e}} + b_{\mathrm{i}}$ and $r_{\mathrm{e/i}} \leq b_{\mathrm{e/i}} \leq K_{\mathrm{e/i}}r_{\mathrm{e/i}}$, where $r_{\mathrm{e/i}}$ is the typical activation rates of a single synapse.

For simplicity, we explain how to obtain such limit compound Poisson processes by reasoning on the excitatory inputs alone. To this end, let us denote the marginal jump distribution of W_{e} as p_{e}. Given a fixed typical synaptic weight w_{e}, the jumps are quantized as $W_{\mathrm{e}} = kw_{\mathrm{e}}$, with k distributed on $\{1, \ldots, K_{\mathrm{e}}\}$, as by convention jumps cannot have zero size. These jumps are naturally defined in the discrete setting, i.e., with $\Delta t > 0$, and their discrete distribution is given via conditioning as $P_{\mathrm{e},k}/(1 - P_{\mathrm{e},0})$. For beta distributed marginals F_{e}, we show in Appendix B that considering $\Delta t \to 0^+$ yields the jump distribution

$$p_{\mathrm{e},k} = \lim_{\Delta t \to 0^+} \frac{P_{\mathrm{e},k}}{1 - P_{\mathrm{e},0}} = \binom{K_{\mathrm{e}}}{k}\frac{B(k, \beta_{\mathrm{e}}+K_{\mathrm{e}}-k)}{\psi(\beta_{\mathrm{e}}+K_{\mathrm{e}}) - \psi(\beta_{\mathrm{e}})}\,, \tag{7}$$

where ψ denotes the digamma function. In the following, we will explicitly index discrete count distributions, e.g. $p_{\mathrm{e},k}$, to distinguish them form the corresponding jump distributions, i.e., p_{e}. Eq. (7) follows from observing that the probability to find a spike within a bin is

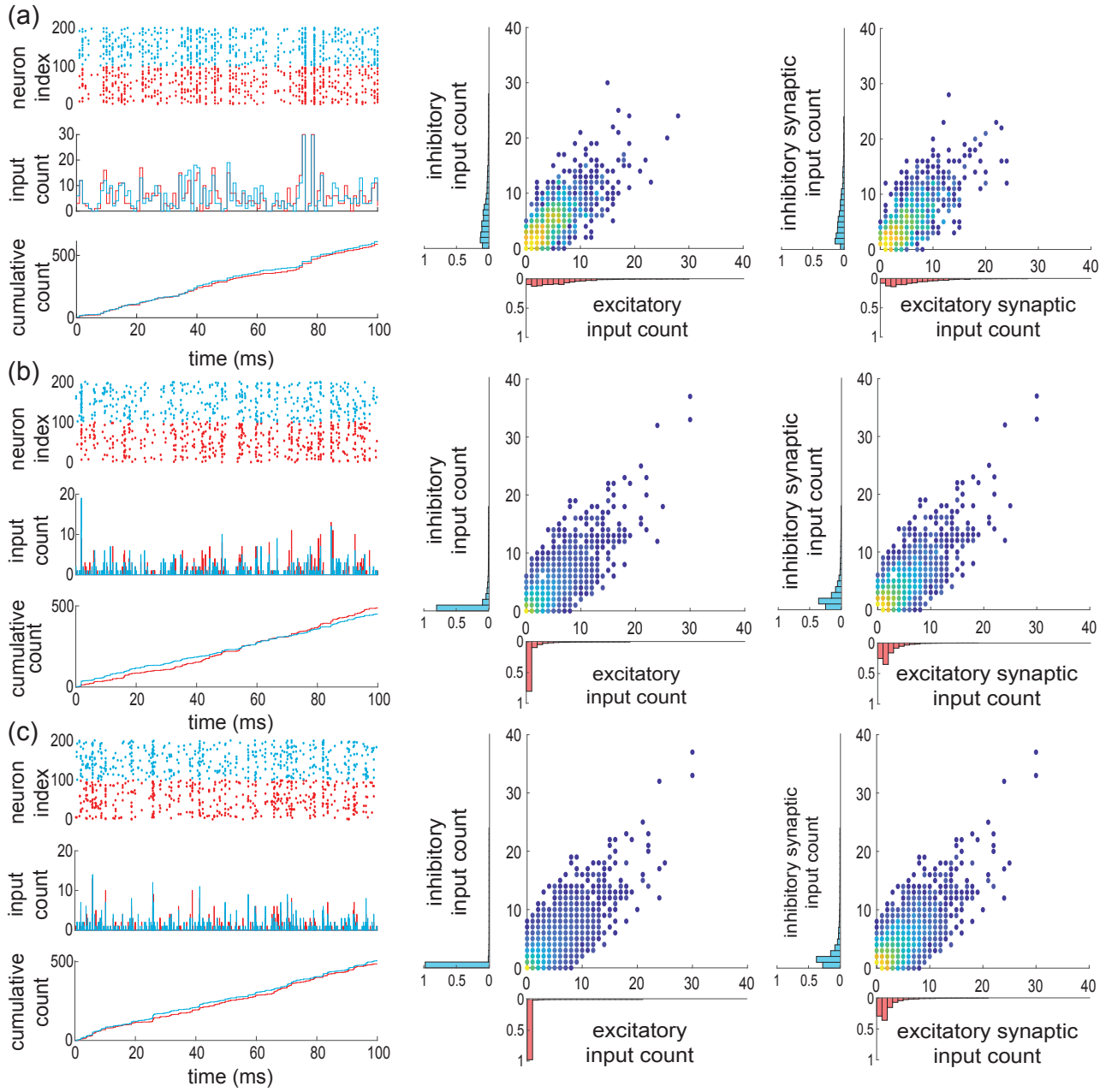

FIG. 4. **Limit compound Poisson process with excitation and inhibition.** (a) Under assumption of partial exchangeability, synaptic inputs can only be distinguished by the fact that they are either excitatory or inhibitory, which is marked by being colored in red or blue in the discrete representation of correlated synaptic inputs with bin size $\Delta t = 1$ms. Accordingly, considering excitation and inhibition separately specifies two associated input-count processes and two cumulative counting processes. For nonzero spiking correlation $\rho = 0.03$, these processes are themselves correlated as captured by the joint distribution of excitatory and inhibitory input counts $P_{\mathrm{ei},kl}$ (center) and by the joint distribution of excitatory and inhibitory jumps $P_{\mathrm{ei},kl}/(1-P_{00})$ (right). (b) The input count distribution $P_{\mathrm{ei},kl}$ is a finite-size approximation of the bivariate directing de Finetti measure F_{ei}, which we consider as a parameter as usual. For a smaller bin size $\Delta t = 0.1$ms, this distribution concentrates in $(0,0)$, as an increasing proportion of time bins does not register any synaptic events, be they excitatory or inhibitory. In the presence of correlation however, the conditioned jump distribution remains correlated but also dispersed. (c) In the limit $\Delta t \to 0$, the input-count distribution is concentrated in $(0,0)$, consistent with the fact that the average number of synaptic activations remains constant while the number of bins diverges. By contrast, the distribution of synaptic event size conditioned to distinct from $(0,0)$ converges toward a well-defined distribution: $p_{\mathrm{ei},kl} = \lim_{\Delta t \to 0^+} P_{\mathrm{ei},kl}/(1 - P_{\mathrm{ei},00})$. This distribution characterizes the jumps of a bivariate compound Poisson process, obtained as the limit of the cumulative count process when considering $\Delta t \to 0^+$.

$\mathbb{E}[X_i] = \alpha_e/(\alpha_e+\beta_e) = r_e\Delta t$, so that for fixed excitatory spiking rate r_e, $\alpha_e \to 0^+$ when $\Delta t \to 0^+$. As a result, the continuous-time spiking correlation is $\rho_e = 1/(1+\beta_e)$, so that we can interpret β_e as a parameter controlling correlations. More generally, we show in Appendix C that the limit correlation ρ_e only depends on the count distribution $p_{e,k}$ via

$$\rho_e = \frac{\mathbb{E}_e[k(k-1)]}{\mathbb{E}_e[k](K_e-1)}, \tag{8}$$

where $\mathbb{E}_e[\cdot]$ denotes expectations with respect to $p_{e,k}$. This shows that zero spiking correlation corresponds to single synaptic activations, i.e., to an input drive modeled as a Poisson process, as opposed to a compound Poisson process. For Poisson-process models, the overall rate of synaptic events is necessarily equal to the sum of the individual spiking rate: $b_e = K_e r_e$. This is no longer the case in the presence of synchronous spiking, when nonzero input correlation $\rho_e > 0$ arises from coincidental synaptic activations. Indeed, as the population spiking rate is conserved when $\Delta t \to 0^+$, the rate of excitatory synaptic events b_e governing N_e satisfies $K_e r_e = b_e \mathbb{E}_e[k]$ so that

$$b_e = \frac{K_e r_e}{\mathbb{E}_e[k]} = r_e\beta_e\left(\psi(\beta_e+K_e) - \psi(\beta_e)\right). \tag{9}$$

Let us reiterate for clarity, that if k_e synapses activate synchronously, this only counts as one synaptic event, which can come in variable size k. Consistently, we have in general $r_e \le b_e \le K_e r_e$. When $\beta_e \to 0$, we have perfect synchrony with $\rho_e = 1$ and $b_e \to r_e$, whereas the independent spiking regime with $\rho_e = 0$ is attained for $\beta_e \to \infty$, for which we have $b_e \to K_e r_e$.

It is possible to generalize the above construction to mixed excitation and inhibition but a closed-from treatment is only possible for special cases. For the independent case (i), in the limit $\Delta t \to 0^+$, jumps are either excitatory alone or inhibitory alone, i.e., the jump distribution p_{ei} has support on $\{1,\dots,K_e\}\times\{0\}\cup\{0\}\times\{1,\dots,K_i\}$. Accordingly, we show in Appendix D that

$$\begin{aligned} p_{ei,kl} &= \lim_{\Delta t\to 0^+} \frac{P_k P_l}{1-P_{e,0}P_{i,0}}, \\ &= \frac{b_e}{b_e+b_i} p_{e,k}\mathbb{1}_{\{l=0\}} + \frac{b_e}{b_e+b_i} p_{i,l}\mathbb{1}_{\{k=0\}}, \end{aligned} \tag{10}$$

where $p_{e/i,k}$ and $b_{e/i}$ are specified by Eq. (7) and Eq. (9) the parameters $\beta_{e/i}$ and $K_{e/i}$, respectively. This shows that as expected, in the absence of synchrony the driving compound poisson process Z with bidimensional jump is obtained as the direct sum of two independent compound Poisson processes. In particular, the driving processes are such that $N = N_e + N_i$, with rates satisfying $b = b_e + b_i$. By contrast, for the maximally correlated case with $r_e = r_i = r$ (ii), we show in Appendix D that the jumps are given as $(W_e, W_i) = (kw_e, lw_i)$, with (k,l) distributed on $\{0,\dots,K_e\}\times\{0,\dots,K_i\}\setminus\{0,0\}$ (see Fig. 4b and c) according to

$$\begin{aligned} p_{ei,kl} &= \lim_{\Delta t\to 0^+} \frac{P_{ei,kl}}{1-P_{ei,00}} \\ &= \binom{K_e}{k}\binom{K_i}{l}\frac{B(k+l,\beta+K_e+K_i-k-l)}{\psi(\beta+K_e+K_i)-\psi(\beta)}. \end{aligned} \tag{11}$$

Incidentally, the driving Poisson process N has a rate b determined by adapting Eq. (9)

$$b = r\beta\left(\psi(\beta+K_e+K_i) - \psi(\beta)\right),$$

for which one can check that $r \le b \le (K_e+K_i)r$.

All the closed-form results so far have been derived for synchrony parametrization in terms of beta distribution. There are other possible parametrizations and these would lead to different count distributions $p_{ei,kl}$, but without known closed-form. To address this limitation in the following, all our results will hold for arbitrary distributions p_{ei} of the jump sizes (W_e, W_i) on the positive orthant $(0,\infty)\times(0,\infty)$. In particular, our results will be given in terms of expectations with respect to p_{ei}, still denoted by $\mathbb{E}_{ei}[\cdot]$. Nonzero correlation between excitation and inhibition will correspond to those choices of p_{ei} for which $W_e W_i > 0$ with nonzero probability, which indicates the presence of synchronous excitatory and inhibitory inputs. Note that this modeling setting restricts nonzero correlations to be positive, which is an inherent limitation of our synchrony-based approach. When considering an arbitrary p_{ei}, the main caveat is understanding how such a distribution may correspond to a given input numbers $K_{e/i}$ and spiking correlations $\rho_{e/i}$ and ρ_{ei}. For this reason, we will always consider that $k_e = W_e/w_e$ and $k_i = W_i/w_i$ follows beta distributed marginal distributions when discussing the roles of $w_{e/i}$, $K_{e/i}$, $\rho_{e/i}$, and ρ_{ei} in shaping the voltage response of a neuron. In that respect, we show in Appendix C that the coefficient ρ_{ei} can always be deduced from the knowledge of a discrete count distribution $p_{ei,kl}$ on $\{0,\dots,K_e\}\times\{0,\dots,K_i\}\setminus\{0,0\}$ via

$$\rho_{ei} = \frac{\mathbb{E}_{ei}[k_e k_i]}{\sqrt{K_e\mathbb{E}_{ei}[k_e]\,K_i\mathbb{E}_{ei}[k_i]}} \ge 0,$$

where the expectations are with respect to $p_{ei,kl}$.

D. Instantaneous synapses and Marcus integrals

We are now in a position to formulate the mathematical problem at stake within the framework developed by Marcus to study shot-noise driven systems [44, 45]. Our goal is quantifying the subthreshold variability of an AONCB neuron subjected to synchronous inputs. Mathematically, this amounts to computing the first two moments of the stationary process solving the following stochastic dynamics

$$\dot{V} = -V/\tau + h_e(V_e - V) + h_i(V_i - V) + I/C, \tag{12}$$

where $V_\mathrm{i} < 0 < V_\mathrm{e}$ are constants and where the reduced conductances $h_\mathrm{e} = g_\mathrm{e}/C$ and $h_\mathrm{i} = g_\mathrm{i}/C$ follows stochastic processes defined in terms of a compound Poisson process Z with bivariate jumps. Formally, the compound Poisson process Z is specified by b, the rate of its governing Poisson process N, and by the joint distribution of its jumps p_ei. Each point of the Poisson process N represents a synaptic activation time T_n, where n is in $\mathbb{Z}$ with the convention that $T_0 \leq 0 \leq T_1$. At all these times, the synaptic input sizes are drawn as i.i.d. random variables $(W_{\mathrm{e},n}, W_{\mathrm{i},n})$ in $\mathbb{R}^+ \times \mathbb{R}^+$ with probability distribution p_ei.

At this point, it is important to observe that the driving process Z is distinct from the conductance process $h = (h_\mathrm{e}, h_\mathrm{i})$. The latter process is formally defined for AONCB neurons as

$$\begin{aligned} h(t) &= \frac{Z(t) - Z(t-\epsilon\tau)}{\epsilon\tau}\,, \\ &= \frac{1}{\epsilon\tau}\left(\sum_{n=N(t-\epsilon\tau)+1}^{N(t)} W_{\mathrm{e},n}\,,\ \sum_{n=N(t-\epsilon\tau)+1}^{N(t)} W_{\mathrm{i},n}\right)\,, \end{aligned}$$

where the dimensionless parameter $\epsilon = \tau_\mathrm{s}/\tau > 0$ is the ratio of the duration of synaptic activation relative to the passive membrane time constant. Note that the amplitude of h scales in inverse proportion to ϵ in order to maintain the overall charge transfer during synaptic events of varying durations. Such a scaling ensures that the voltage response of AONCB neurons has finite, nonzero variability for small or vanishing synaptic time constant, i.e., for $\epsilon \ll 1$ (see Fig 5). The simplifying limit of instantaneous synapses is obtained for $\epsilon = \tau_\mathrm{s}/\tau \to 0^+$, which corresponds to infinitely fast synaptic activation. By virtue of its construction, the conductance process h becomes a shot noise in the limit $\epsilon \to 0^+$, which can be formally identified to $\mathrm{d}Z/\mathrm{d}t$. This is consistent with the definition of shot-noise processes as temporal derivative of compound Poisson processes, i.e., as collections of randomly weighted Dirac-delta masses.

Due to their high degree of idealization, shot-noise models are often amenable to exact stochastic analysis, albeit with some caveats. For equations akin to Eq. (12) in the limit of instantaneous synapses, such a caveat follows from the multiplicative nature of the conductance shot noise h. In principle, one might expect to solve Eq. (12) with shot-noise drive via stochastic calculus, as for diffusion-based drive. This would involve interpreting the stochastic integral representations of solutions in terms of Stratonovich representations [65]. However, Stratonovich calculus is not well-defined for shot-noise drives [66]. To remedy this point, Marcus has proposed to study stochastic equations subjected to regularized versions of shot noises, whose regularity is controlled by a nonnegative parameter ϵ [44, 45]. For $\epsilon > 0$, the dynamical equations admit classical solutions, whereas the shot-noise-driven regime is recovered in the limit $\epsilon \to 0^+$. The hope is to be able to characterize analytically the shot-noise-driven solution, or at least some of its moments, by considering regular solutions in the limit $\epsilon \to 0^+$. We choose to refer to the control parameter as ϵ by design in the above. This is because AONCB models represent Marcus-type regularizations that are amenable to analysis in the limit of instantaneous synapses, i.e., when $\epsilon = \tau_\mathrm{s}/\tau \to 0^+$, for which the conductance process h converges toward a form of shot noise.

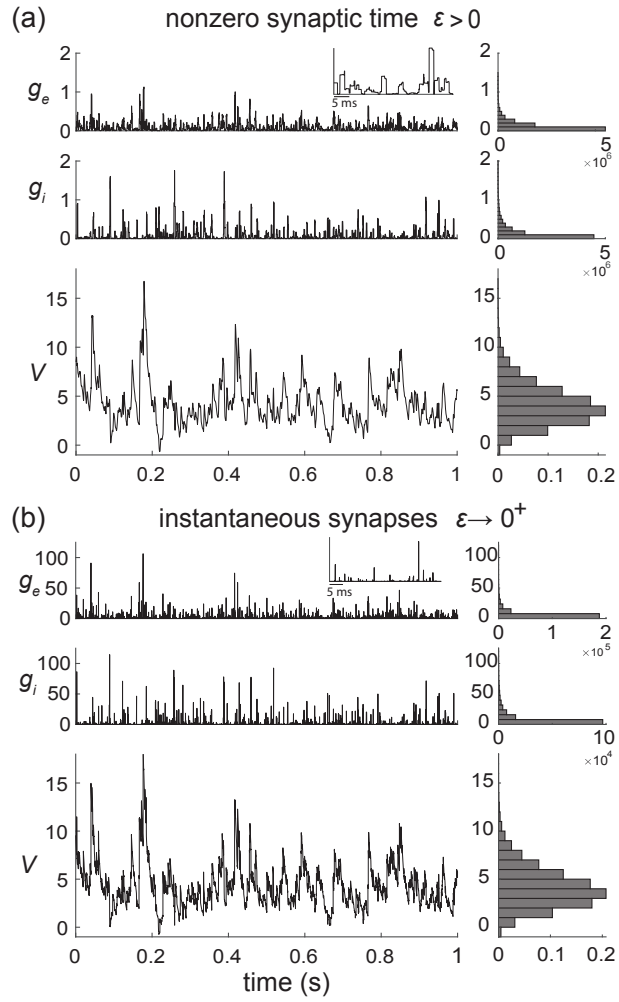

FIG. 5. **Limit of instantaneous synapses.** The voltage trace and the empirical voltage distribution are only marginally altered by taking the limit $\epsilon \to 0^+$ for short synaptic time constant: $\tau_\mathrm{s} = 2$ms in (a) and $\tau_\mathrm{s} = 0.02$ms in (b). In both (a) and (b), we consider the same compound Poisson-process drive with $\rho_\mathrm{e} = 0.03$, $\rho_\mathrm{i} = 0.06$, and $\rho_\mathrm{ei} = 0$, and the resulting fluctuating voltage V is simulated via a standard Euler discretization scheme. The corresponding empirical conductance and voltage distributions are shown on the right. The later voltage distribution asymptotically determines the stationary moments of V.

Marcus interpretation of stochastic integration has practical implications for numerical simulations with shot noise [41]. According to this interpretation, shot-noise-driven solutions shall be conceived as limits of regularized solutions for which standard numerical scheme applies.

Correspondingly, shot-noise-driven solutions to Eq. (12) can be simulated via a limit numerical scheme. We derive such a limit scheme in Appendix E. Specifically, we show that the voltage of shot-noise-driven AONCB neurons exponentially relaxes toward the leak reversal potential $V_{\mathrm{L}} = 0$, except when subjected to synaptic impulses at times $\{T_n\}_{n\in\mathbb{Z}}$. At these times, the voltage V updates discontinuously according to $V(T_n) = V(T_n^-) + J_n$, where the jumps are given in Appendix E via the Marcus rule

$$J_n = \left(\frac{W_{\mathrm{e},n}V_{\mathrm{e}} + W_{\mathrm{i},n}V_{\mathrm{i}}}{W_{\mathrm{e},n} + W_{\mathrm{i},n}} - V(T_n^-)\right)\left(1 - e^{-(W_{\mathrm{e},n}+W_{\mathrm{i},n})}\right) . \tag{13}$$

Observe that the above Marcus rule directly implies that no jump can cause the voltage to exit $(V_{\mathrm{i}}, V_{\mathrm{e}})$, the allowed range of variation for V. Moreover, note that this rule specifies an exact even-driven simulation scheme given knowledge of the synaptic activation times and sizes $\{T_n, W_{\mathrm{e},n}, W_{\mathrm{i},n}\}_{n\in\mathbb{Z}}$ [67]. We adopt the above Marcus-type numerical scheme in all the simulations that involve instantaneous synapses.

E. Moment calculations

When driven by stationary compound-Poisson processes, AONCB neurons exhibit ergodic voltage dynamics. As a result, the typical voltage state, obtained by sampling the voltage at random time, is captured by a unique stationary distribution. Our main analytical results, which we give here, consists in exact formulas for the first two voltage moment with respect to that stationary distribution. Specifically, we derive the stationary mean voltage Eq. (14) in Appendix F and the stationary voltage variance Eq. (16) in Appendix G. These results are obtained by a probabilistic treatment exploiting the properties of compound Poisson processes within Marcus' framework. This treatment yields compact, interpretable formulas in the limit of instantaneous synapses $\epsilon = \tau_{\mathrm{s}}/\tau \to 0^+$. Readers who are interested in the method of derivation for these results are encouraged to go over the calculations presented in Appendices F, G, H, I, J, K, and L.

In the limit of instantaneous synapses, $\epsilon \to 0^+$, we find that the stationary voltage mean is

$$\mathbb{E}[V] = \lim_{\epsilon\to 0^+} \mathbb{E}[V_\epsilon] = \frac{a_{\mathrm{e},1}V_{\mathrm{e}} + a_{\mathrm{i},1}V_{\mathrm{i}} + I/G}{1 + a_{\mathrm{e},1} + a_{\mathrm{i},1}} , \tag{14}$$

where we define the first-order synaptic efficacies as

$$\begin{aligned} a_{\mathrm{e},1} &= b\tau\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{W_{\mathrm{e}} + W_{\mathrm{i}}}\left(1 - e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right] , \\ a_{\mathrm{i},1} &= b\tau\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{i}}}{W_{\mathrm{e}} + W_{\mathrm{i}}}\left(1 - e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right] . \end{aligned} \tag{15}$$

Note the $\mathbb{E}_{\mathrm{ei}}[\cdot]$ refers to the expectation with respect to the jump distribution p_{ei} in Eq. (15), whereas $\mathbb{E}[\cdot]$ refers to the stationary expectation in Eq. (14). Eq. (14) has the same form as for deterministic dynamics with constant conductances, in the sense that the mean voltage is a weighted sum of the reversal potentials V_{e}, V_{i} and $V_{\mathrm{L}} = 0$. One can check that for such deterministic dynamics, the synaptic efficacies involved in the stationary mean simply read $a_{\mathrm{e/i},1} = K_{\mathrm{e/i}}r_{\mathrm{e/i}}w_{\mathrm{e/i}}$. Thus, the impact of synaptic variability, and in particular of synchrony, entirely lies in the definition of the efficacies in Eq. (15). In the absence of synchrony, one can check that accounting for the shot-noise nature of the synaptic conductances leads to synaptic efficacies under exponential form: $a_{\mathrm{e/i},1} = K_{\mathrm{e/i}}r_{\mathrm{e/i}}(1 - e^{-w_{\mathrm{e/i}}})$. In turn, accounting for input synchrony leads to synaptic efficacies expressed as expectation of these exponential forms in Eq. (15), consistent with the stochastic nature of the conductance jumps $(W_{\mathrm{e}}, W_{\mathrm{i}})$. Our other main result, the formula for the stationary voltage variance, involves synaptic efficacies of similar form. Specifically, we find that

$$\mathbb{V}[V] = \frac{1}{1 + a_{\mathrm{e},2} + a_{\mathrm{i},2}} \times \left(a_{\mathrm{e},12}(V_{\mathrm{e}} - \mathbb{E}[V])^2 + a_{\mathrm{i},12}(V_{\mathrm{i}} - \mathbb{E}[V])^2 - c_{\mathrm{ei}}(V_{\mathrm{e}} - V_{\mathrm{i}})^2\right) , \tag{16}$$

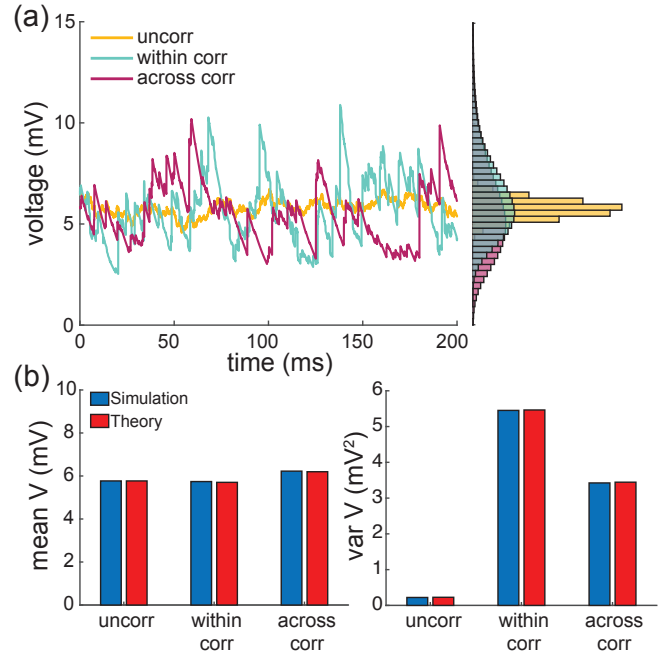

FIG. 6. **Comparison of simulation and theory.** (a) Examples of voltage traces obtained via Monte-Carlo simulations of an AONCB neuron for various type of synchrony-based input correlations: uncorrelated $\rho_{\mathrm{e}} = \rho_{\mathrm{i}} = \rho_{\mathrm{ei}} = 0$ (uncorr, yellow), within correlation $\rho_{\mathrm{e}}, \rho_{\mathrm{i}} > 0$ and $\rho_{\mathrm{ei}} = 0$ (within corr, cyan), within and across correlation $\rho_{\mathrm{e}}, \rho_{\mathrm{i}}, \rho_{\mathrm{ei}} > 0$ (across corr, magenta). (b) Comparison of the analytically derived expressions (14) and (16) with numerical estimates obtained via Monte-Carlo simulations for the synchrony conditions considered in (a).

where we define the second-order synaptic efficacies as

$$\begin{aligned} a_{\mathrm{e},2} &= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\left(1-e^{-2(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right], \\ a_{\mathrm{i},2} &= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{i}}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\left(1-e^{-2(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right]. \end{aligned} \tag{17}$$

Eq. (16) also prominently features auxiliary second-order efficacies defined by $a_{\mathrm{e/i},12} = a_{\mathrm{e/i},1} - a_{\mathrm{e/i},2}$. Owing to their prominent role, we also mention their explicit form:

$$\begin{aligned} a_{\mathrm{e},12} &= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right], \\ a_{\mathrm{i},12} &= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{i}}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]. \end{aligned} \tag{18}$$

The other quantity of interest featuring in Eq. (16) is the crosscorrelation coefficient

$$c_{\mathrm{ei}} = \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}W_{\mathrm{i}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right], \tag{19}$$

which entirely captures the (nonnegative) correlation between excitatory and inhibitory inputs and shall be seen as an efficacy as well.

To conclude, let us stress that for AONCB models, establishing the above exact expressions does not require any approximation other than taking the limit of instantaneous synapses. In particular, we neither resort to any diffusion approximations [37, 38] nor invoke the effective-time-constant approximation [41–43]. We give in Appendix L an alternative factorized form for $\mathbb{V}[V]$ to justify the nonnegativity of expression (16). In Fig. 6, we illustrate the excellent agreement of the analytically derived expressions (14) and (16) with numerical estimates obtained via Monte-Carlo simulations of the AONCB dynamics for various input synchrony conditions. Discussing and interpreting quantitatively Eq. (14) and Eq. (16) within a biophysically relevant context will be the main focus of the remaining of this work.

III. COMPARISON WITH EXPERIMENTAL DATA

A. Experimental measurements and parameter estimations

Cortical activity typically exhibits a high degree of variability in response to identical stimuli [68, 69], with individual neuronal spiking exhibiting Poissonian characteristics [3, 70]. Such variability is striking because neurons are thought to typically receive a large number ($\simeq 10^4$) of synaptic contacts [13]. As a result, in the absence of correlations, neuronal variability should average out, leading to quasi-deterministic neuronal voltage dynamics [71]. To explain how variability seemingly defeats averaging in large neural networks, it has been proposed that neurons operate in a special regime, whereby inhibitory and excitatory drive nearly cancel one another [16, 17, 19–21]. In such balanced networks, the voltage fluctuations become the main determinant of the dynamics, yielding a Poisson-like spiking activity [16, 17, 19–21]. However, depending upon the tightness of this balance, networks can exhibit distinct dynamical regimes with varying degree of synchrony [18].

In the following, we will exploit the analytical framework of AONCB neurons to argue that the asynchronous picture predicts voltage fluctuations are an order of magnitude smaller than experimental observations [1, 26–28]. Such observations indicate that the variability of the neuronal membrane voltage exhibits typical variance values of $\simeq 4-9\mathrm{mV}^2$. Then we will claim that achieving such variability requires input synchrony within the setting of AONCB neurons. Experimental estimates of the spiking correlations are typically thought as weak with coefficients ranging from 0.01 to 0.04 [10–12]. Such weak values do not warrant the neglect of correlations owing to the typically high number of synaptic connections. Actually, if K denotes the number of inputs, all assumed to play exchangeable roles, an empirical criterion to decide whether a correlation coefficient ρ is weak is that $\rho < 1/K$ [32, 33]. Assuming the lower estimate of $\rho \simeq 0.01$, this criterion is achieved for $K \simeq 100$ inputs, which is well below the typical number of excitatory synapses for cortical neurons. In the following, we will only consider the response of AONCB neurons to synchronous drive with biophysically realistic spiking correlations ($0 \leq \rho \leq 0.03$).

Two key parameters for our argument will be the excitatory and inhibitory synaptic weights denoted by w_{e} and w_{i}, respectively. Typical values for these weights can be estimated via biophysical considerations within the framework of AONCB neurons. In order to develop these considerations, we assume the values $V_{\mathrm{i}} = -10\mathrm{mV} < V_{\mathrm{L}} = 0 < V_{\mathrm{e}} = 60\mathrm{mV}$ for reversal potentials and $\tau = 15\mathrm{ms}$ for the passive membrane time constant. Given these assumptions, we set the upper range of excitatory synaptic weights so that when delivered to a neuron close to its resting state, unitary excitatory inputs cause peak membrane fluctuations of $\simeq 0.5\mathrm{mV}$ at the soma, attained after a peak time of $\simeq 5\mathrm{ms}$. Such fluctuations correspond to typically large *in-vivo* synaptic activations of thalamo-cortical projections in rats [72]. Although activations of similar amplitude have been reported for cortico-cortical connections [73, 74], recent large-scale *in vivo* studies have revealed that cortico-cortical excitatory connections are typically much weaker [75, 76]. At the same time, these studies have shown that inhibitory synaptic conductances are about fourfold larger than excitatory ones, but with similar timescales. Fitting these values within the framework of AONCB neurons for $\epsilon = \tau_{\mathrm{s}}/\tau \simeq 1/4$ reveals that the largest possible synaptic inputs correspond to dimensionless weights $w_{\mathrm{e}} \simeq 0.01$ and $w_{\mathrm{i}} \simeq 0.04$. Following on [75, 76], we consider that the comparatively moderate cortico-cortical recurrent connections are an order of magnitude weaker than typical thalamo-cortical

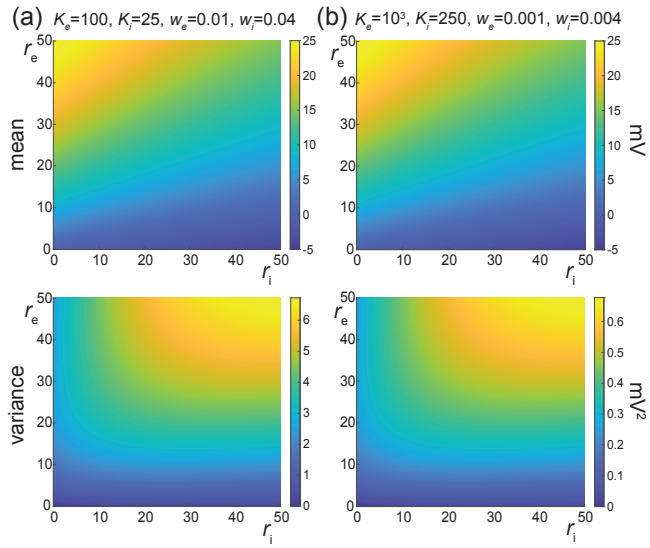

FIG. 7. **Voltage mean and variance in the absence of input correlations.** Column (a) depicts the stationary subthreshold response of an AONCB neurons driven by $K_\mathrm{e} = 100$ and $K_\mathrm{i} = 25$ synapses with large weights $w_\mathrm{e} = 0.01$ and $w_\mathrm{i} = 0.04$. Column (b) depicts the stationary subthreshold response of an AONCB neurons driven by $K_\mathrm{e} = 10^3$ and $K_\mathrm{i} = 250$ synapses with moderate weights $w_\mathrm{e} = 0.001$ and $w_\mathrm{i} = 0.004$. For synaptic weights $w_\mathrm{e}, w_\mathrm{i} \ll 1$, the mean response is identical as $K_\mathrm{e} w_\mathrm{e} = K_\mathrm{i} w_\mathrm{i} = 1$ for (a) and (b). By contrast, for $\rho_\mathrm{e} = \rho_\mathrm{i} = \rho_\mathrm{ei} = 0$, the variance is at least an order-of-magnitude smaller than that experimentally observed ($4 - 9\mathrm{mV}^2$) for moderate weights as shown in (a). Reaching the lower range of realistic neural variability requires driving the cell via large weights as shown in (b).

projections, i.e., $w_\mathrm{e} \simeq 0.001$ and $w_\mathrm{i} \simeq 0.004$. Such a range is in keeping with estimates used in [38].

B. The effective-time constant approximation holds in the asynchronous regime

Let us consider that neuronal inputs have zero (or negligible) correlation structure, which corresponds to assuming that all synapses are driven by independent Poisson processes. Incidentally, excitation and inhibition act independently. Within the framework of AONCB neurons, this latter assumption corresponds to choosing a joint jump distribution of the form

$$p_\mathrm{ei}(W_\mathrm{e}, W_\mathrm{i}) = \frac{b_\mathrm{e}}{b} p_\mathrm{e}(W_\mathrm{e})\delta(W_\mathrm{i}) + \frac{b_\mathrm{i}}{b} p_\mathrm{i}(W_\mathrm{i})\delta(W_\mathrm{e})\,.$$

where $\delta(\cdot)$ denotes the Dirac delta function so that $W_\mathrm{e} W_\mathrm{i} = 0$ with probability one. Moreover, b_e and b_i are independently specified via Eq. (9) and the overall rate of synaptic events is purely additive: $b = b_\mathrm{e} + b_\mathrm{i}$. Consequently, the crosscorrelation efficacy c_ei in Eq. (16) vanishes and the dimensionless efficacies simplify to

$$a_{\mathrm{e},1} = b_\mathrm{e}\tau\mathbb{E}_\mathrm{e}\left[1 - e^{-W_\mathrm{e}}\right] \quad \text{and} \quad a_{\mathrm{i},1} = b_\mathrm{i}\tau\mathbb{E}_\mathrm{i}\left[1 - e^{-W_\mathrm{i}}\right]\,.$$

Further assuming that individual excitatory and inhibitory synapses act independently leads to considering that p_e and p_i depict the size of individual synaptic inputs, as opposed to aggregate events. This corresponds to taking $\beta_\mathrm{e} \to \infty$ and $\beta_\mathrm{i} \to \infty$ in our parametric model based on beta distributions. Then, as intuition suggests, the overall rates of excitation and inhibition activation are recovered as $b_\mathrm{e} = K_\mathrm{e} r_\mathrm{e}$ and $b_\mathrm{i} = K_\mathrm{i} r_\mathrm{i}$, where r_e and r_i are the individual spiking rates.

Individual synaptic weights are small in the sense that $w_\mathrm{e}, w_\mathrm{i} \ll 1$, which warrants neglecting exponential corrections for the evaluation of the synaptic efficacies, at least in the absence of synchrony-based correlations. Accordingly, we have

$$a_{\mathrm{e},1} \simeq K_\mathrm{e} r_\mathrm{e} \tau w_\mathrm{e} \quad \text{and} \quad a_{\mathrm{e},12} \simeq K_\mathrm{e} r_\mathrm{e} \tau w_\mathrm{e}^2/2\,,$$

as well as symmetric expressions for inhibitory efficacies. Plugging these values in Eq. (16) yields the classical mean-field estimate for the stationary variance

$$\mathbb{V}[V] \simeq \frac{K_\mathrm{e} r_\mathrm{e} w_\mathrm{e}^2 (V_\mathrm{e} - \mathbb{E}[V])^2 + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}^2 (V_\mathrm{i} - \mathbb{E}[V])^2}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i})}\,,$$

which is exactly the same expression as that derived via the diffusion and effective-time-constant approximations in [46, 47]. However, observe that the only approximation we made in obtaining the above expression is to neglect exponential corrections due to the relative weakness of biophysically relevant synaptic weights, which we hereafter refer to as the small-weight approximation.

C. Asynchronous inputs yield exceedingly small neural variability

In Fig. 7, we represent the stationary mean $\mathbb{E}[V]$ and variance $\mathbb{V}[V]$ as a function of the neuronal spiking input rates r_e and r_i, but for distinct values of synaptic weights w_e and w_i. In Fig. 7a, we consider synaptic weights as large as biophysically admissible based on recent *in-vivo* studies [75, 76], i.e., $w_\mathrm{e} = 0.01$ and $w_\mathrm{i} = 0.04$. By contrast, in Fig. 7b, we consider moderate synaptic weights $w_\mathrm{e} = 0.001$ and $w_\mathrm{i} = 0.004$, which yield somatic postsynaptic deflections of typical amplitudes. In both cases, we consider input numbers K_e and K_i such that the mean voltage $\mathbb{E}[V]$ covers the same biophysical range of values as r_e and r_i varies between 0Hz and 50Hz. Given a zero resting potential, we set this biophysical range to be bounded by $\Delta\mathbb{E}[V] \leq 20\mathrm{mV}$ as typically observed experimentally in electrophysiological recordings. These conditions correspond to constant aggregate weights set to $K_\mathrm{e} w_\mathrm{e} = K_\mathrm{i} w_\mathrm{i} = 1$ so that

$$K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} = K_\mathrm{i} r_\mathrm{i} w_\mathrm{i} \leq 50\mathrm{Hz} \simeq 1/\tau\,.$$

This implies that the AONCB neurons under consideration do not reach the high-conductance regime for which the passive conductance can be neglected, i.e., $K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{e} r_\mathrm{e} w_\mathrm{i} \gg 1/\tau$ [77]. Away from the high-conductance

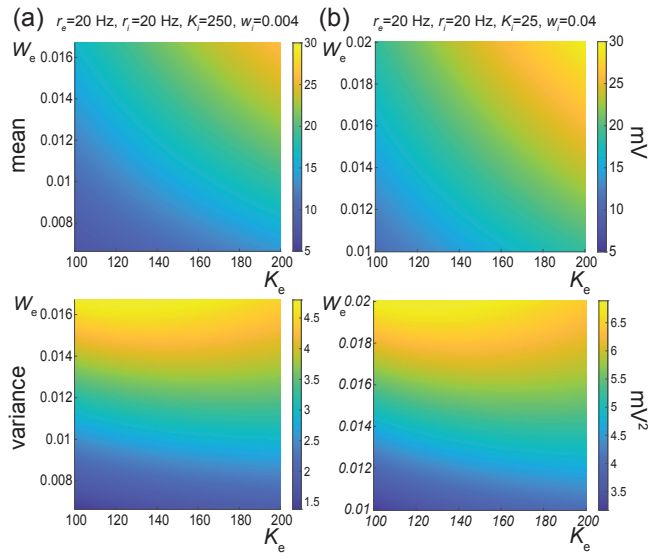

FIG. 8. **Dependence on the number of inputs and the synaptic weights in the absence of correlations.** Column (a) depicts the stationary subthreshold response of an AONCB neurons driven by a varying number of excitatory synapses K_e with varying weight w_e at rate $r_\mathrm{e} = 20$Hz, with background inhibitory drive given by $K_\mathrm{i} = 250$ with moderate weights $w_\mathrm{i} = 0.004$ and $r_\mathrm{i} = 20$Hz. Column (b) depicts the same as in column (a) but for a background inhibitory drive given by $K_\mathrm{i} = 25$ with large weights $w_\mathrm{i} = 0.04$ and $r_\mathrm{i} = 20$Hz. For both conditions, achieving realistic level of variance, i.e., $\mathbb{V}[V] \simeq 4 - 9\mathrm{mV}^2$, while ensuring a biophysically relevant mean range of variation, i.e., $\Delta\mathbb{E}[V] \simeq$ 10-20mV, is only possible for large weights: $w_\mathrm{e} \geq 0.015$ for moderate inhibitory weights in Fig. 8a and $w_\mathrm{e} \geq 0.01$ for large weights.

regime, the variance magnitude is controlled by the denominator of Eq. (20). Accordingly, the variance in both cases is primarily dependent on the excitatory rate r_e since for $K_\mathrm{e}w_\mathrm{e} = K_\mathrm{i}w_\mathrm{i} = 1$, the effective excitatory driving force $F_\mathrm{e} = K_\mathrm{e}w_\mathrm{e}^2(V_\mathrm{e} - \mathbb{E}[V])^2$ dominates the effective inhibitory driving force $F_\mathrm{i} = K_\mathrm{i}w_\mathrm{i}^2(V_\mathrm{i} - \mathbb{E}[V])^2$. This is because the neuronal voltage typically sits close to the inhibitory reversal potential but far from the excitatory reversal potential $V_\mathrm{e} - \mathbb{E}[V] > \mathbb{E}[V] - V_\mathrm{i}$. For instance, when close to rest $\mathbb{E}[V] \simeq 0$, the ratio of the effective driving forces is $(K_\mathrm{e}w_\mathrm{e}^2V_\mathrm{e}^2)/(K_\mathrm{i}w_\mathrm{i}^2V_\mathrm{i}^2) \simeq 9$ fold in favor of excitation. Importantly, the magnitude of the variance is distinct for moderate synapses and for large synapses. This is because for constant aggregate weights $K_\mathrm{e}w_\mathrm{e} = K_\mathrm{i}w_\mathrm{i} = 1$, the ratio of effective driving forces for large and moderate synapses scales in keeping with the ratio of the weights, and so does the ratio of variances away from the high conductance regime. Thus we have

$$F_\mathrm{e}|_{w_\mathrm{e}=10^{-2}}/F_\mathrm{e}|_{w_\mathrm{e}=10^{-3}} = F_\mathrm{i}|_{w_\mathrm{i}=10^{-2}}/F_\mathrm{i}|_{w_\mathrm{i}=10^{-3}} = 10\,,$$

and the variance decreases by one order of magnitude from large weights in Fig. 7a to moderate weights in Fig. 7b.

The above numerical analysis reveals that achieving realistic levels of subthreshold variability for a biophysical mean range of variation requires AONCB neurons to be exclusively driven by large synaptic weights. This is confirmed by considering the voltage mean $\mathbb{E}[V]$ and variance $\mathbb{V}[V]$ in Fig. 8 as a function of the number of inputs K_e and of the synaptic weights w_e for a given level of inhibition. We choose this level of inhibition to be set by $K_\mathrm{i} = 250$ moderate synapses $w_\mathrm{i} = 0.004$ with $r_\mathrm{i} = 20$Hz in Fig. 8a and by $K_\mathrm{i} = 25$ large synapses $w_\mathrm{i} = 0.04$ with $r_\mathrm{i} = 20$Hz in Fig. 8b. As expected, assuming that $r_\mathrm{e} = 20$Hz in the absence of input correlations, the voltage mean $\mathbb{E}[V]$ only depends on the product $K_\mathrm{e}w_\mathrm{e}$, which yields similar mean range of variations for K_e varying up to 2000 in Fig. 8a and up to 200 in Fig. 8b. Thus, it is possible to achieve the same range of variations as with moderate synaptic with a fewer number of larger synaptic weights. By contrast, the voltage variance $\mathbb{V}[V]$ only achieves realistic levels for large synaptic weights in both conditions, with $w_\mathrm{e} \geq 0.015$ for moderate inhibitory background synapses in Fig. 8a and $w_\mathrm{e} \geq 0.01$ for large inhibitory background synapses in Fig. 8b.

D. Including input correlations yields realistic subthreshold variability

Without synchrony, achieving the experimentally observed variability necessitates an excitatory drive mediated via synaptic weights $w_\mathrm{e} \simeq 0.01$, which corresponds to the upper bounds of the biophysically admissible range and is in agreement with numerical results presented in [38]. Albeit possible, this is unrealistic given the wide distribution of amplitudes observed experimentally, whereby the vast majority of synaptic events are small to moderate, at least for cortico-cortical connections [75, 76]. In principle, one can remedy this issue by allowing for synchronous activation of, say, $k_\mathrm{e} = 10$ synapses with moderate weight $w_\mathrm{e} = 0.001$, as it amounts to the activation of a single synapse with large weight $k_\mathrm{e}w_\mathrm{e} = 0.01$. A weaker assumption that yields a similar increase in neural variability is to only ask for synapses to tend to synchronize probabilistically, which amounts to require k_e to be a random variable with some distribution mass on $\{k_\mathrm{e} > 1\}$. This exactly amounts to model the input drive via a jump process as presented in Section II, with a jump distribution p_e that probabilistically captures this degree of input synchrony. In turn, this distribution p_e corresponds to a precise input correlation ρ_e via Eq. (8).

We quantify the impact of nonzero correlation in Fig. 9 where we consider the cases of moderate weights $w_\mathrm{e} = 0.001$ and $w_\mathrm{e} = 0.004$ and large weights $w_\mathrm{e} = 0.01$ and $w_\mathrm{i} = 0.04$ as in Fig. 7 but for $\rho_\mathrm{e} = \rho_\mathrm{i} = 0.03$. Specifically, we consider an AONCB neuron subjected to two independent beta-binomial-derived compound Poisson process drives with rate b_e and b_i, respectively. These rates b_e and b_i are obtained via Eq. (9) by setting $\beta_\mathrm{e} = \beta_\mathrm{i} = 1/\rho_\mathrm{e} - 1 = 1/\rho_\mathrm{i} - 1$ and for given input numbers K_e and

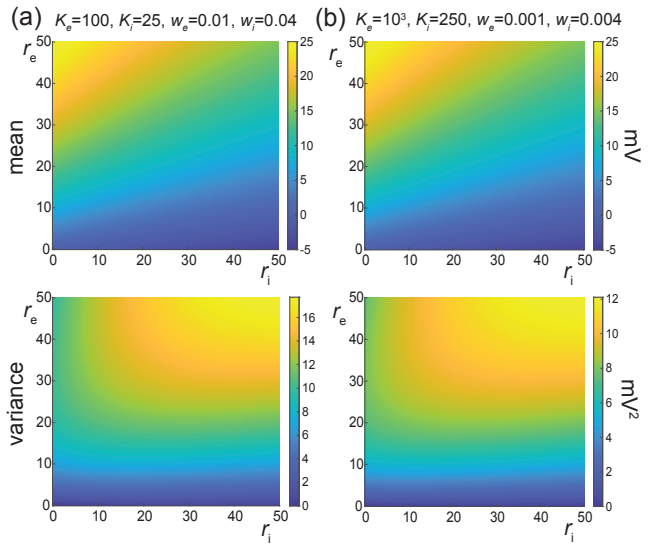

FIG. 9. **Voltage mean and variance in the presence of excitatory and inhibitory input correlations but without correlation across excitation and inhibition:** $\rho_\mathrm{e} = \rho_\mathrm{i} > \rho_\mathrm{ei} = 0$**.** Column (a) depicts the stationary subthreshold response of an AONCB neurons driven by $K_\mathrm{e} = 100$ and $K_\mathrm{i} = 25$ synapses with large weights $w_\mathrm{e} = 0.01$ and $w_\mathrm{i} = 0.04$. Column (b) depicts the stationary subthreshold response of an AONCB neurons driven by $K_\mathrm{e} = 10^3$ and $K_\mathrm{i} = 250$ synapses with moderate dimensionless weights $w_\mathrm{e} = 0.001$ and $w_\mathrm{i} = 0.004$. For synaptic weights $w_\mathrm{e}, w_\mathrm{i} \ll 1$, the mean response is identical as $K_\mathrm{e} w_\mathrm{e} = K_\mathrm{i} w_\mathrm{i} = 1$ for (a) and (b). By contrast with the case of no correlation in Fig. 7, for $\rho_\mathrm{e} = \rho_\mathrm{i} = 0.03$ and $\rho_\mathrm{ei} = 0$, the variance achieves similar levels as experimentally observed ($4 - 9\mathrm{mV}^2$) for moderate weights as shown in (b), but slightly larger levels for large weights as shown in (a).

K_i and spiking rates r_e and r_i. This ensures that the mean number of synaptic activations $b_\mathrm{e}\mathbb{E}_\mathrm{ei}[k_\mathrm{e}] = K_\mathrm{e} r_\mathrm{e}$ and $b_\mathrm{i}\mathbb{E}[k_\mathrm{i}] = K_\mathrm{i} r_\mathrm{i}$ remains constant when compared with Fig. 7. As a result, the mean response of the AONCB neuron is essentially left unchanged by the presence of correlations, with virtually identical biophysical range of variations $\Delta\mathbb{E}_\mathrm{ei}[V] \simeq$ 10-20mV. This is because for correlation $\rho_\mathrm{e} = \rho_\mathrm{i} \simeq 0.03$, the aggregate weights still satisfy $k_\mathrm{e} w_\mathrm{e}, k_\mathrm{i} w_\mathrm{i} < 1$ with probability close to one given that $K_\mathrm{e} w_\mathrm{e} = K_\mathrm{i} w_\mathrm{i} = 1$. Then, in the absence of crosscorrelation, i.e., $\rho_\mathrm{ei} = 0$, we still have

$$a_{\mathrm{e},1} = b_\mathrm{e}\tau\mathbb{E}_\mathrm{e}\left[1 - e^{-k_\mathrm{e} w_\mathrm{e}}\right] \simeq b_\mathrm{e}\tau w_\mathrm{e}\mathbb{E}_\mathrm{e}[k_\mathrm{e}] = K_\mathrm{e} r_\mathrm{e}\tau w_\mathrm{e}\,,$$

as well as $a_{\mathrm{i},1} \simeq K_\mathrm{i} r_\mathrm{i}\tau w_\mathrm{i}$ by symmetry. However, for both moderate and large synaptic weights, the voltage variance $\mathbb{V}[V]$ now exhibits slightly larger magnitudes than observed experimentally. This is because we show in Appendix M that in the small-weight approximation

$$\begin{aligned} a_{\mathrm{e},12} &= \frac{b_\mathrm{e}\tau}{2}\mathbb{E}_\mathrm{e}\left[\left(1 - e^{-k_\mathrm{e} w_\mathrm{e}}\right)^2\right]\,, \\ &\simeq \left(1 + \rho_\mathrm{e}(K_\mathrm{e} - 1)\right)\frac{K_\mathrm{e} r_\mathrm{e}\tau w_\mathrm{e}^2}{2}\,, \end{aligned}$$

where we recognize $K_\mathrm{e} r_\mathrm{e}\tau w_\mathrm{e}^2/2 = a_{\mathrm{e},12}|_{\rho_\mathrm{e}=0}$ as the second-order efficacy in the absence of correlations from Fig. 7. A similar statement holds for $a_{\mathrm{i},12}$. This shows that correlations increase neural variability whenever $\rho_\mathrm{e} > 1/K_\mathrm{e}$ or $\rho_\mathrm{i} > 1/K_\mathrm{i}$, which coincides with our previously given criterion to assess the relative weakness of correlations. Accordingly, when excitation and inhibition act independently, i.e., $\rho_\mathrm{ei} = 0$, we find that the increase in variability due to input synchrony $\Delta_{\rho_{\mathrm{e/i}}} = \mathbb{V}[V]|_{\rho_\mathrm{ei}=0} - \mathbb{V}[V]|_{\rho_{\mathrm{e/i}}=\rho_\mathrm{ei}=0}$ satisfies

$$\begin{aligned} \Delta_{\rho_{\mathrm{e/i}}} \simeq\ & \frac{\rho_\mathrm{e}(K_\mathrm{e} - 1)K_\mathrm{e} r_\mathrm{e} w_\mathrm{e}^2 (V_\mathrm{e} - \mathbb{E}[V])^2}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}) +} \\ & \frac{\rho_\mathrm{i}(K_\mathrm{i} - 1)K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}^2 (V_\mathrm{i} - \mathbb{E}[V])^2}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i})}\,, \end{aligned} \tag{20}$$

The above relation follows from the fact that the small-weight approximation for $\mathbb{E}[V]$ is independent of correlations and from neglecting the exponential corrections due to the nonzero size of the synaptic weights. The above formula remains valid as long as the correlations ρ_e and ρ_i are weak enough so that the aggregate weights satisfy $k_\mathrm{e} w_\mathrm{e}, k_\mathrm{i} w_\mathrm{i} < 1$ with probability close to one. To inspect the relevance of exponential corrections, we estimate in Appendix N the error incurred by neglecting exponential corrections. Focusing on the case of excitatory inputs, we find that for correlation coefficients $\rho_\mathrm{e} \leq 0.05$, neglecting exponential corrections incurs less than a 3% error if the number of inputs is smaller than $K_\mathrm{e} \leq 1000$ for moderate synaptic weight $w_\mathrm{e} = 0.001$ or than $K_\mathrm{e} \leq 100$ for large synaptic weight $w_\mathrm{e} = 0.01$.

E. Including correlations between excitation and inhibition reduces subthreshold variability

The voltage variance estimated for realistic excitatory and inhibitory correlations, e.g., $\rho_\mathrm{e} = \rho_\mathrm{i} = 0.03$ and $\rho_\mathrm{ei} = 0$, exceeds the typical levels measured *in vivo*, i.e., $4 - 9\mathrm{mV}^2$, for large synaptic weights. The inclusion of correlations between excitation an inhibition, i.e., $\rho_\mathrm{ei} > 0$ can reduce the voltage variance to more realistic levels. We confirm this point in Fig. 10 where we consider the cases of moderate weights $w_\mathrm{e} = 0.001$ and $w_\mathrm{e} = 0.004$ and large weights $w_\mathrm{e} = 0.01$ and $w_\mathrm{i} = 0.04$ as in Fig. 9 but for $\rho_\mathrm{e} = \rho_\mathrm{i} = \rho_\mathrm{ei} = 0.03$. Positive crosscorrelation between excitation and inhibition only marginally impacts the mean voltage response. This is due to the fact that exponential corrections become slightly more relevant as the presence of crosscorrelation leads to larger aggregate weights: $W_\mathrm{e} + W_\mathrm{i}$ with W_e and W_i possibly being jointly positive. By contrast with this marginal impact on the mean response, the voltage variance is significantly reduced when excitation and inhibition are correlated. This is in keeping with the intuition that the net effect of such crosscorrelation is to cancel excitatory and inhibitory synaptic inputs with one another, before they can cause voltage fluctuations. The amount by which

the voltage variance is reduced can be quantified in the small-weight approximation. In this approximation, we show in Appendix M that the efficacy $c_{\rm ei}$ capturing the impact of crosscorrelations simplifies to

$$c_{\rm ei} \simeq \frac{b\tau}{2}\mathbb{E}_{\rm ei}\left[W_{\rm e}W_{\rm i}\right] = (\rho_{\rm ei}\sqrt{r_{\rm e}r_{\rm i}}\tau/2)(K_{\rm e}w_{\rm e})(K_{\rm i}w_{\rm i})\,.$$

Using the above simplified expression and invoking the fact that the small-weight approximation for $\mathbb{E}\left[V\right]$ is independent of correlations show a decrease in the amount $\Delta_{\rho_{\rm ei}} = \mathbb{V}\left[V\right] - \mathbb{V}\left[V\right]|_{\rho_{\rm ei}=0}$ with

$$\Delta_{\rho_{\rm ei}} \simeq \tag{21}$$
$$-\frac{\rho_{\rm ei}\sqrt{r_{\rm e}r_{\rm i}}(K_{\rm e}w_{\rm e})(K_{\rm i}w_{\rm i})(V_{\rm e}-\mathbb{E}\left[V\right])(\mathbb{E}\left[V\right]-V_{\rm i})}{1/\tau + K_{\rm e}r_{\rm e}w_{\rm e} + K_{\rm i}r_{\rm i}w_{\rm i}} \leq 0\,.$$

Despite the above reduction in variance, we also show in Appendix M that positive input correlations always cause an overall increase of neural variability:

$$0 \leq \mathbb{V}\left[V\right]|_{\rho_{\rm e/i}=\rho_{\rm ei}=0} \leq \mathbb{V}\left[V\right] \leq \mathbb{V}\left[V\right]|_{\rho_{\rm ei}=0}\,.$$

Note that the reduction of variability due to $\rho_{\rm ei} > 0$ crucially depends on the instantaneous nature of correlations between excitation and inhibition. To see this, observe that Marcus rule Eq. (13) specifies instantaneous jumps via a weighted average of the reversal potentials $V_{\rm e}$ and $V_{\rm i}$, which represent extreme values for voltage updates. Thus, perfectly synchronous excitation and inhibition updates the voltage toward an intermediary value rather than extreme ones, leading to smaller jumps on average, Such an effect can vanish or even reverse when synchrony breaks down, e.g., when inhibition substantially lags behind excitation.

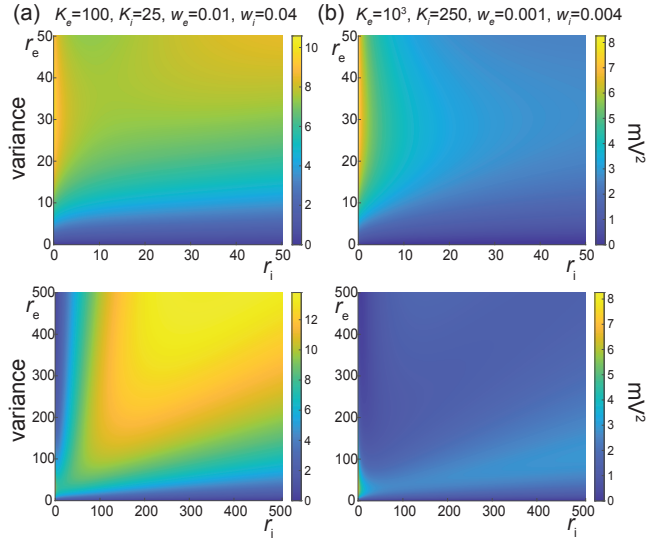

FIG. 10. **Voltage mean and variance in the presence of excitatory and inhibitory input correlations and with correlation across excitation and inhibition:** $\rho_{\rm e} = \rho_{\rm i} = \rho_{\rm ei} > 0$**.** Column (a) depicts the stationary subthreshold response of an AONCB neurons driven by $K_{\rm e} = 100$ and $K_{\rm i} = 25$ synapses with large weights $w_{\rm e} = 0.01$ and $w_{\rm i} = 0.04$. Column (b) depicts the stationary subthreshold response of an AONCB neurons driven by $K_{\rm e} = 10^3$ and $K_{\rm i} = 250$ synapses with moderate dimensionless weights $w_{\rm e} = 0.001$ and $w_{\rm i} = 0.004$. For synaptic weights $w_{\rm e}, w_{\rm i} \ll 1$, the mean response is identical as $K_{\rm e}w_{\rm e} = K_{\rm i}w_{\rm i} = 1$ for (a) and (b). Compared with the case of no crosscorrelation in Fig. 9, for $\rho_{\rm e} = \rho_{\rm i} = \rho_{\rm ei} = 0.03$, the variance is reduced to a biophysical range similar to that experimentally observed ($4-9{\rm mV}^2$) for moderate weights as shown in (a), as well as for large weights as shown in (b).

F. Asynchronous scaling limits require fixed-size synaptic weights

Our analysis reveals that the correlations must significantly impact the voltage variability whenever the number of inputs are such that $K_{\rm e} > 1/\rho_{\rm e}$ or $K_{\rm i} > 1/\rho_{\rm i}$. Spiking correlations are typically measured *in vivo* to be larger than 0.01. Therefore, synchrony must shape the response of neurons that are driven by more than 100 active inputs, which is presumably allowed by the typically high number of synaptic contacts ($\simeq 10^4$) in cortex [13]. In practice, we find that synchrony can explain the relatively high level of neural variability observed in the subthreshold neuronal responses. Beyond these practical findings, we predict that input synchrony also have significant theoretical implications with respect to modeling spiking networks. Analytically tractable models for cortical activity are generally obtained by considering spiking networks in the infinite-size limit. Such infinite-size networks are tractable because the neurons they comprise only interact via population averages, erasing any role for nonzero correlation structure. Distinct mean-field models assume that synaptic weights vanish according to distinct scalings with respect to the number of synapses, i.e., $w_{\rm e/i} \to 0$ as $K_{\rm e/i} \to \infty$. In particular, classical mean-field limits consider the scaling $w_{\rm e/i} \sim 1/K_{\rm e/i}$, balanced mean-field limits consider the scaling $w_{\rm e/i} \sim 1/\sqrt{K_{\rm e/i}}$, with $K_{\rm e}w_{\rm e} - K_{\rm i}w_{\rm i} = O(1)$, and strong coupling limits consider the scaling $w_{\rm e/i} \sim 1/\ln K_{\rm e/i}$, with $K_{\rm e}w_{\rm e} - K_{\rm i}w_{\rm i} = O(1)$ as well. Importantly, all these mean-field limits assume no correlation, and in particular, no synchrony.

Our analysis of AONCB neurons shows that the neglect of synchrony-based correlations is incompatible with the maintenance of neural variability in the infinite-size limit. Indeed, Eq. (20) shows that for any scaling with $1/w_{\rm e} = o(K_{\rm e})$ and $1/w_{\rm i} = o(K_{\rm i})$, as for all the mean-field limits mentioned above, we have

$$\mathbb{V}\left[V\right] = O(w_{\rm e}) + O(w_{\rm i}) \xrightarrow{K_e,K_i\to\infty} 0\,.$$

Thus, in the absence of correlation and independent of the synaptic weight scaling, the subthreshold voltage variance of AONCB neurons must vanish in the limit of arbitrary large numbers of synapses. We expect such decay of the voltage variability to be characteristic of conductance-based models in the absence of input correlation. Indeed, dimensional analysis suggests that voltage variance for both current-based and conductance-

based models are generically obtained via normalization by the reciprocal of the membrane time constant. However, by contrast with current-based models, the reciprocal of the membrane time constant for conductance-based models, i.e., $1/\tau + K_\mathrm{e} w_\mathrm{e} r_\mathrm{e} + K_\mathrm{i} w_\mathrm{i} r_\mathrm{i}$, involves contributions from synaptic conductances. Thus, to ensure nonzero asymptotic variability, the denominator scaling $O(K_\mathrm{e} w_\mathrm{e}) + O(K_\mathrm{i} w_\mathrm{i})$ must be balanced by the natural scaling of the Poissonian input drives, i.e., $O(K_\mathrm{e} w_\mathrm{e}^2) + O(K_\mathrm{i} w_\mathrm{i}^2)$. In the absence of input correlations, this is only possible for fixed-size weights, which is incompatible with any scaling assumptions.

G. Synchrony allows for variability-preserving scaling limits with vanishing weights

Infinite-size networks with fixed-size synaptic weights are problematic for restricting modeled neurons to operate in the high-conductance regime, whereby the intrinsic conductance properties of the cell play no role. Such a regime is biophysically unrealistic as it implies that the cell would respond to perturbations infinitely fast. We propose to address this issue by considering a new type of variability-preserving limit models obtained for the classical scaling but in the presence of synchrony-based correlations. For simplicity, let us consider our correlated input model with excitation alone in the limit of an arbitrary large number of inputs $K_\mathrm{e} \to \infty$. When $\rho_\mathrm{e} > 0$, the small-weight approximation Eq. (20) suggests that adopting the scaling $w_\mathrm{e} \sim \Omega_\mathrm{e}/K_\mathrm{e}$, where Ω_e denotes the aggregate synaptic weight, yields a nonzero contribution when $K_\mathrm{e} \to \infty$ as the numerator scales as $O(K_\mathrm{e}^2 w_\mathrm{e}^2)$. It turns out that this choice can be shown to be valid without resorting to any approximations. Indeed, under the classical scaling assumption, we show in Appendix O that the discrete jump distribution $p_{\mathrm{e},k}$ weakly converges to the continuous density $\mathrm{d}\nu_\mathrm{e}/\mathrm{d}w$ in the sense that

$$b_\mathrm{e} \sum_{k=1}^{K_\mathrm{e}} p_{\mathrm{e},k} \delta\left(\frac{w}{\Omega_\mathrm{e}} - \frac{k}{K_\mathrm{e}}\right) \mathrm{d}w \xrightarrow{K_\mathrm{e}\to\infty} \tag{22}$$
$$\nu_\mathrm{e}(\mathrm{d}w) = \frac{r_\mathrm{e}\beta_\mathrm{e}}{w}\left(1 - \frac{W_\mathrm{e}}{w}\right)^{\beta_\mathrm{e}-1} \mathrm{d}w\,.$$

The above density has infinite mass over $[0, \Omega_\mathrm{e}]$ owing to its diverging behavior in zero and is referred to as a degenerate beta distribution. In spite of its degenerate nature, it is known that densities of the above form define well-posed processes, the so-called beta processes, which have been studied extensively in the field of nonparametric Bayesian inference [61, 62]. These beta processes represent generalizations of our compound Poisson process drives in so far as they allow for a countable infinity of jumps to occur within a finite time window. This is a natural requirement to impose when considering an infinite pool of synchronous synaptic inputs, the overwhelming majority of which having nearly zero amplitude.

The above arguments show that one can defined a generalized class of synchronous input models that can serve as the drive of AONCB neurons as well. Such generalizations are obtained as limits of compound Poisson processes and are specified via their Lévy-Khintchine measures, which formalize the role of ν_e [78, 79]. Our results naturally extend to this generalized class. Concretely, for excitation alone, our results extend by replacing all expectations of the form $b_\mathrm{e}\mathbb{E}_\mathrm{e}[\cdot]$ by integral with respect to the measure ν_e. One can easily check that these expectations, which feature prominently in the definition of the various synaptic efficacies, all remain finite for Lévy-Khintchine measures. In particular, the voltage mean and variance of AONCB neurons remain finite with

$$\mathbb{E}[V] = \frac{V_\mathrm{e}\int_0^{\Omega_\mathrm{e}}(1-e^{-w})\nu_\mathrm{e}(\mathrm{d}w)}{1/\tau + \int_0^{\Omega_\mathrm{e}}(1-e^{-w})\nu_\mathrm{e}(\mathrm{d}w)}\,,$$
$$\mathbb{V}[V] = \frac{(V_\mathrm{e} - \mathbb{E}[V])^2\int_0^{\Omega_\mathrm{e}}(1-e^{-w})^2\nu_\mathrm{e}(\mathrm{d}w)}{2/\tau + \int_0^{\Omega_\mathrm{e}}(1-e^{-2w})\nu_\mathrm{e}(\mathrm{d}w)}\,.$$

Thus, considering the classical scaling limit $w_\mathrm{e} \propto 1/K_\mathrm{e}$ preserves nonzero subthreshold variability in the infinite size limit $K_\mathrm{e} \to \infty$ as long as ν_e puts mass away from zero, i.e., for $\beta_\mathrm{e} < \infty \Leftrightarrow \rho_\mathrm{e} > 0$. Furthermore, we show in Appendix O that $\mathbb{V}[V] = O(\rho_\mathrm{e})$ so that voltage variability consistently vanishes in the absence of spiking correlation, for which ν_ϵ concentrates in zero, i.e., when $\beta_\mathrm{e} \to \infty \Leftrightarrow \rho_\mathrm{e} = 0$.

IV. DISCUSSION

A. Synchrony modeling

We have presented a parametric representation of the neuronal drives resulting from a finite number of asynchronous or (weakly) synchronous synaptic inputs. Several parametric statistical models have been proposed for generating correlated spiking activities in a discrete setting [59, 80–82]. Such models have been used to analyze the activity of neural populations via Bayesian inference methods [83–85], as well as maximum entropy methods [86, 87]. Our approach is not to simulate or analyze complex neural dependencies but rather to derive from first principles the synchronous input models that could drive conductance-based neuronal models. This approach primarily relies on extending the definition of discrete-time correlated spiking models akin to [59] to the continuous-time setting. To do so, the main tenet of our approach is to realize that input synchrony and spiking correlation represent equivalent measures under the assumption of input exchangeabilty.

Input exchangeabilty posits that the driving inputs form a subset of an arbitrarily large pool of exchangeable random variables [55, 56]. In particular, this implies that the main determinant of the neuronal drive is the number of active inputs, as opposed to the magnitude

of these synaptic inputs. Then, de Finetti theorem [57] states that the probability of observing a given input configuration can be represented in the discrete setting under an integral form (see Eq. (3)) involving a directing probability measure F. Intuitively, F represents the probability distribution of the fraction of coactivating inputs at any discrete time. Our approach identifies the directing measure F as a free parameter that captures input synchrony. The more dispersed the distribution F, the more synchronous the inputs, as previously noted in [88, 89]. Our work elaborates on this observation to develop computationally tractable statistical models for synchronous spiking in the continuous-time limit, i.e., for vanishing discrete time step $\Delta t \to 0^+$.

We derive our results using a discrete-time directing measure chosen as beta distribution $F \sim B(\alpha, \beta)$, where the parameters α and β can be related to the individual spiking rate r and the spiking correlation ρ via $r\Delta t = \alpha/(\alpha+\beta)$ and $\rho = 1/(1+\alpha+\beta)$. For this specific choice of distribution, we are able to construct statistical models of the correlated spiking activity as generalized beta-binomial processes [60], which play an important role in statistical Bayesian inference [61, 62]. This construction allows us to fully parametrize the synchronous activity of a finite number of inputs via the jump distribution of a compound Poisson process, which depends explicitly on the spiking correlation. For being continuously indexed in time, stationary compound Poisson processes can naturally serve as the drive to biophysically relevant neuronal models. The idea to utilize compound Poisson processes to model input synchrony was originally proposed in [90–92], but without constructing these processes as limits of discrete spiking models and without providing explicit functional form for their jump distributions. More generally, our synchrony modeling can be interpreted as a limit case of the formalism proposed in [93, 94] to model correlated spiking activity via multidimensional Poisson processes.

B. Moment analysis

We analytically characterize the subthreshold variability of a tractable conductance-based neuronal model, the AONCB neurons, when driven by synchronous synaptic inputs. The analytical characterization of a neuron's voltage fluctuations has been the focus of intense research [46, 47, 95–97]. These attempts have considered neuronal models that already incorporate some diffusion scaling hypotheses [98, 99], formally obtained by assuming an infinite number of synaptic inputs. The primary benefit of these diffusion approximations is that one can treat the corresponding Fokker-Planck equations to quantify neuronal variability in conductance-based integrate-and-fire models, while also including the effect of post-spiking reset [37, 38]. In practice, subthreshold variability is often estimated in the effective-time-constant approximation, while neglecting the multiplicative noise contributions due to voltage-dependent membrane fluctuations [46, 95, 96], although an exact treatment is also possible without this simplifying assumption [38]. By contrast, the analysis of conductance-based models has resisted exact treatments when driven by shot noise, as for compound Poisson input processes, rather than by Gaussian white noise, as in the diffusion approximation [41–43].

The exact treatment of shot-noise-driven neuronal dynamics is primarily hindered by the limitations of the Ito/Stratonovich integrals [65, 100] to capture the effects of point-process-based noise sources, even without including a reset mechanism. These limitations were originally identified by Marcus, who proposed to approach the problem via a new type of stochastic equation [44, 45]. The key to Marcus equation is to define shot noise as limits of regularized, well-behaved approximations of that shot noise, for which classical calculus applies [66]. In practice, these approximations are canonically obtained as the solutions of shot-noise-driven Langevin equations with relaxation time scale $\tau_{\rm s}$, and shot noise is formally recovered in the limit $\tau_{\rm s} \to 0^+$. Our assertion here is that all-or-none conductances implement such a form of shot-noise regularization for which a natural limiting process can be defined when synapses operate instantaneously, i.e., $\tau_{\rm s} \to 0^+$. The main difference with the canonical Marcus approach is that our regularization is all-or-none, substituting each Dirac delta impulse with a finite step-like impulse of duration $\tau_{\rm s}$ and magnitude $1/\tau_{\rm s}$, thereby introducing a synaptic timescale but without any relaxation mechanism.

The above assertion is the basis for introducing AONCB neurons, which is supported by our ability to obtain exact formulas for the first two moments of their stationary voltage dynamics (see Eq. (14) and Eq. (16)). For $\tau_{\rm s} > 0$, these moments can be expressed in terms of synaptic efficacies that takes exact but rather intricate integral forms. Fortunately, these efficacies drastically simplify in the instantaneous synapse limit $\tau_{\rm s} \to 0^+$, for which the canonical shot-noise drive is recovered. These resulting formulas mirror those obtained in the diffusion and effective-time-constant approximations [46, 47], except that they involve synaptic efficacies whose expressions are oirginal in three ways (see Eq. (15) Eq. (G4), Eq. (G7), and Eq. (G8)): First, independent of input synchrony, these efficacies all have exponential forms and saturate in the limit of large synaptic weights. Such saturation is a general characteristic of shot-noise-driven, continuously-relaxing systems [101–103]. Second, these efficacies are defined as expectations with respect to the jump distribution $p_{\rm ei}$ of the driving compound Poisson process (see Eq. (11) and Appendix B). A nonzero dispersion of $p_{\rm ei}$, indicating that synaptic activation is truly modeled via random variables $W_{\rm e}$ and $W_{\rm i}$, is the hallmark of input synchrony [91, 92]. Third, these efficacies involve the overall rate of synaptic events b (see Eq. (12)), which also depends on input synchrony. Such dependence can be naturally understood within the framework of Palm

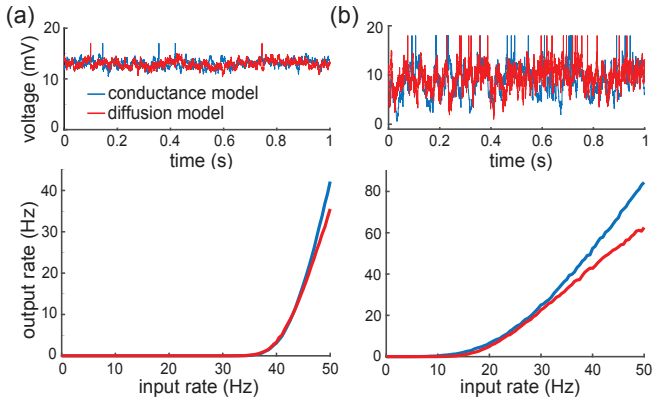

FIG. 11. **Diffusion approximations in the presence of synchrony.** (a) Comparison of an asynchronously driven integrate-and-fire AONCB neuron (blue trace) with its diffusion approximation obtained via the effective-time-constant approximation (red trace). (b) Comparison of a synchronously driven integrate-and-fire AONCB neuron (blue trace) with its diffusion approximation obtained by our exact analysis (red trace). Parameters: $K_e = 1000$, $K_i = 350$, $\tau = 15$ ms, $w_e = 0.001$, $w_i = 0.004$, $r_e = r_i = 25$ Hz, $\rho_e = \rho_i = 0.03$, $\rho_{ei} = 0$, $V_\mathrm{T} = 15$ mV, and $V_\mathrm{R} = 12$ mV.

calculus [104], a form of calculus specially developed for stationary point processes.

C. Biophysical relevance

Our analysis allows us to investigate quantitatively how subthreshold variability depends on the numbers and strength of the synaptic contacts. This approach requires that we infer synaptic weights from the typical peak time and peak amplitude of the somatic membrane fluctuations caused by post-synaptic potentials [72, 75, 76]. Within our modeling framework, these weights are dimensionless quantities that we estimate by fitting the AONCB neuronal response to a single all-or-none synaptic activation at rest. For biophysically relevant parameters, this yields typically small synaptic weights in the sense that $w_\mathrm{e}, w_\mathrm{i} \ll 1$. These small values warrant adopting the small-weight approximation, for which expressions Eq. (14) and Eq. (16) simplify.

In the small-weight approximation, the mean voltage becomes independent of input synchrony, whereas the simplified voltage variance Eq. (20) only depends on input synchrony via the spiking correlation coefficients ρ_e, ρ_i, and ρ_ei, as opposed to depending on a full jump distribution. Spike-count correlations have been experimentally shown to be weak in cortical circuits [10–12] and for this reason, most theoretical approaches argued for asynchronous activity [17, 105–109]. A putative role for synchrony in neural computations remains a matter of debate [110–112]. In modeled networks, although the tight balance regime implies asynchronous activity, [19–21], the loosely balance regime is compatible with the establishment of strong neuronal correlations [22–24]. When distributed over large networks, weak correlations can still give rise to precise synchrony, once information is pooled from a large enough number of synaptic inputs [32, 33]. In this view, and assuming that distinct inputs play comparable roles, correlations measure the propensity of distinct synaptic inputs impinging on a neuron to coactivate, which represents a clear form of synchrony. Our analysis shows that considering synchrony in amounts consistent with the levels of observed spiking correlation is enough to account for the surprisingly large magnitude of subthreshold neuronal variability [1, 26–28]. In contrast, the asynchronous regime yields unrealistically low variability, an observation the challenges the basis for the asynchronous state hypothesis.

Recent theoretical works [37, 38] have also noted that the asynchronous state hypothesis seems at odds with certain features of the cortical activity such as the emergence of spontaneous activity or the maintenance of significant average polarization during evoked activity. Zerlaut *et al.* have analyzed under which conditions conductance-based networks can achieve a spectrum of asynchronous states with realistic neural features. In their work, a key variable to achieve this spectrum is a strong afferent drive that modulates a balanced network with moderate recurrent connections. Moderate recurrent conductances are inferred from allowing for up to 2mV somatic deflections at rest, whereas the afferent drive is provided via even stronger synaptic conductances that can activate synchronously. These inferred conductances appear large in light of recent *in-vivo* measurements [72, 75, 76], and the corresponding synaptic weights all satisfy $w_\mathrm{e}, w_\mathrm{i} \geq 0.01$ within our framework. Correspondingly, the typical connectivity numbers considered are small with $K_\mathrm{e} = 200$, $K_\mathrm{i} = 50$ for recurrent connections and $K_\mathrm{e} = 10$ for the coactivating afferent projections. Thus, results from [37] appear consistent with our observation that realistic subthreshold variability can only be achieved asynchronously for a restricted number of large synaptic weights. Our findings, however, predict that these results follow from connectivity sparseness and will not hold in denser networks, for which the pairwise spiking correlation will exceed the empirical criteria for asynchrony, e.g., $\rho_\mathrm{e} > 1/K_\mathrm{e}$ ($\rho_\mathrm{e} < 0.005 \leq 1/K_\mathrm{e}$ in [37]). Sanzeni *et al.* have pointed out that implementing the effective-time-constant approximation in conductance-based models suppresses subthreshold variability, especially in the high-conductance state [77]. As mentioned here, this suppression causes the voltage variability to decay as $O(w_\mathrm{e}) + O(w_\mathrm{i})$ in any scaling limit with vanishing synaptic weights. Sanzeni *et al.* observe that such decay is too fast to yield realistic variability for the balanced scaling, which assumes $w_\mathrm{e} \sim 1/\sqrt{K_\mathrm{e}}$ and $w_\mathrm{i} \sim 1/\sqrt{K_\mathrm{i}}$. To remedy this point, these authors propose to adopt a slower scaling of the weights, i.e., $w_\mathrm{e} \sim 1/\ln K_\mathrm{e}$ and $w_\mathrm{i} \sim 1/\ln K_\mathrm{i}$, which can be derived from the principle of rate conservation in neural net-

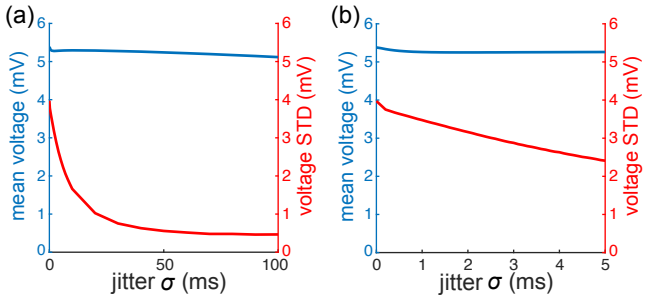

FIG. 12. **Impact of jittering synchronous inputs.** Voltage mean and variance of AONCB neurons in response to synchronous inputs that have been jittered for parameters: $K_\mathrm{e} = 1000$, $w_\mathrm{e} = 0.001$, $\rho_\mathrm{e} = 0.03$. Specifically, each input timing has been independently shifted in time by a centered Gaussian random variable with standard deviation σ. (a) If the mean response is largely independent of jittering, the variance steadily decreases with the jittering, which erases synchrony-based correlation over the timescale σ. Accordingly, for large timescale, $\sigma \geq \tau$, we recover variance values obtained for asynchronous drive with $\rho_\mathrm{e} = \rho_\mathrm{i} = \rho_\mathrm{ei} = 0$. (b) Variability estimates are reliable for jittering at timescales $\sigma \leq 2$ms.

works. Such a scaling is sufficiently slow for variability to persist in networks with large connectivity number ($\simeq 10^5$). However, as any scaling with vanishing weights, our exact analysis shows that such scaling must eventually lead to decaying variability, thereby challenging the basis for the synchronous state hypothesis.

Both of these studies focus on the network dynamics of conductance-based networks under the diffusion and effective-time-constant approximations. The effective-time-constant approximation follows the prior assumption that the diffusion approximation is valid [46, 47, 95–97]. In turn, diffusive behaviors only rigorously emerge under some scaling limit with vanishing weights [98, 99]. By focusing on the single-cell level rather than the network level, we are able to demonstrate that the effective-time-constant approximation holds exactly for shot-noise driven, conductance-based neurons, without any diffusive approximations. Consequently, suppression of variability must occur independent of any scaling choice, except in the presence of input synchrony. Although this observation poses a serious theoretical challenge to the asynchronous state hypothesis, observe that it does not invalidate the practical usefulness of the diffusion approximation. For instance, we show in Fig. 11 that the mean spiking response of an a shot-noise driven AONCB neuron with an integrate-and-fire mechanism can be satisfactorily captured via the diffusion approximation. In addition, our analysis allows one to extend the diffusion approximation to include input synchrony.

D. Limitations of the approach

A first limitation of our analysis is that we neglect the spike-generating mechanism as a source of neural variability. Most diffusion-based approaches model spike generation via the integrate-and-fire mechanism, whereby the membrane voltages reset to fixed value upon reaching a spike-initiation threshold [37, 38, 46, 47, 95–97]. Accounting for such a mechanism can impact our findings in two ways: (*i*) By confining voltage below the spiking threshold, the spiking mechanism may suppress the mean response enough for the neuron to operate well in the high-conductance regime for large input drives. Such a scenario will still produce exceedingly low variability due to variability quenching in the high-conductance regime, consistent with [1]. (*ii*) The additional variability due to post-spiking resets may dominate the synaptic variability, so that a large overall subthreshold variability can be achieved in spite of low synaptic variability. This possibility also seems unlikely as dominant yet stereotypical resets would imply a quasi-deterministic neural response [71]. Addressing the above limitations quantitatively requires extending our exact analysis to include the integrate-and-fire mechanism using technique from queueing theory [104]. This is beyond the scope of this work. We note, however, that implementing a post-spiking reset to a fixed voltage level yields simulated trajectories that markedly differ from physiological ones (see Fig. 1), for which the post-spiking voltage varies across conditions [26–28].

A second limitation of our analysis is our assumption of exchangeability, which is the lens through which we operate a link between spiking correlations and input drives. Taken literally, the exchangeability assumption states that synapses all have a typical strength and that conductance variability primarily stems from the variable numbers of coactivating synapses. This is certainly an oversimplification as synapses exhibit heterogeneity [113], which likely plays a role in shaping neural variability [114]. Distinguishing between heterogeneity and correlation contributions, however, is a fundamentally ambiguous task [115]. For instance, considering K_e synchronous inputs with weight w_e at rate b_e and with jump probability p_e (see Eq. (5) and Eq. (9)) is indistinguishable from considering K_e independent inputs with heterogeneous weights $\{w_\mathrm{e}, 2w_\mathrm{e}, \ldots, K_\mathrm{e}w_\mathrm{e}\}$ and rates $K_\mathrm{e} r_\mathrm{e} p_{\mathrm{e},k}$. Within our modeling approach, accounting for synaptic heterogeneity, with dispersed distribution for synaptic weights $q_\mathrm{e}(w)$, can be done by taking the jump distribution p_e as

$$p_\mathrm{e}(w) = \sum_{k=1}^{K} q_\mathrm{e}^{(\star k)}(w) p_{\mathrm{e},k} \,,$$

where $q_\mathrm{e}^{(\star k)}$ refers to the k-fold convolution of $q_\mathrm{e}(w)$. This leads to an overdispersion of the jump distribution p_e, and thus increased subthreshold neural variability. Therefore, while we have assumed exchangeability, our

approach can accommodate weight heterogeneity. The interpretation of our results in term of synchrony rather than heterogeneity is supported by recent experimental evidence that cortical response selectivity derives from strength in numbers of synapses, rather than difference in synaptic weights [116].

A third limitation of our analysis is to consider a perfect form of synchrony, with exactly simultaneous synaptic activations. Although seemingly unrealistic, we argue that perfect input synchrony can still yield biologically relevant estimates of the voltage variability. This is because voltage fluctuations result from the integration of inputs over a time scale set by the passive membrane time constant $\tau \sim 20$ms. As a result, synaptic activation times that differ by significantly less than τ can be considered as synchronous inputs. To illustrate this point, we show in Fig. 12 the dependence of the voltage variance on the degree of synchrony by gradually jittering initially synchronous synaptic inputs. Assuming $K_{\mathrm{e}} = 1000$ excitatory inputs alone with spiking correlation $\rho_{\mathrm{e}} = 0.03$, one can check that the neural variability is left unchanged by jittering synaptic activations over time scales $\sigma \leq 2$ms. One can also check that jittering over larger timescales than the synaptic time constant yields neural variability similar to that obtained in the absence of correlation in the inputs. This supports that our finding are robust to including temporal variability on timescales $\sigma \leq 2$ms, which is consistent with typical heterogeneities with axonal or dendritic conduction delays. Extending our analytic results to include jittering will require modeling spiking correlations via multidimensional Poisson processes rather than via compound Poisson processes [93, 94]. However, this is is beyond the scope of this work. Here, we point out that weakened forms of synchrony will yield lower variability, so that our challenge to the asynchronous state will remain. One remaining limitation of our synchrony modeling is that our analysis can only account for instantaneous correlations between excitation and inhibition, while in reality such correlations are expected to peak at a non-zero time lag.

A fourth limitation of our analysis is that it is restricted to a form of synchrony that ignores temporal heterogeneity. This is a limitation because a leading hypothesis for the emergence of variability is that neurons generate spikes as if through a doubly stochastic process, i.e., as a Poisson process with temporally fluctuating rate [117]. To better understand this limitation, let us interpret our exchangeability-based modeling approach within the framework of doubly stochastic processes [51, 52]. This can be done most conveniently by reasoning on the discrete correlated spiking model specified by Eq. (3). Specifically, given fixed bin size $\Delta t > 0$, one can interpret the collection of *i.i.d.* variables $\theta \sim F$ as an instantaneously fluctuating rate. In this interpretation, nonzero correlations can be seen as emerging from a doubly stochastic process for which the rate fluctuates as uncorrelated noise, i.e., with zero correlation time. This zero correlation time is potentially a serious limitation as it has been argued that shared variability is best modeled by a low-dimensional latent process evolving with slow, or even smooth, dynamics [82]. Addressing this limitation will require developing limit spiking model with nonzero correlation time using probabilistic techniques that are beyond the scope of this work [56].

A final limitation of our analysis is that it does not explain the consistent emergence of synchrony in network dynamics. It remains conceptually unclear how synchrony can emerge and persist in neural networks that are fundamentally plagued by noise and exhibit large degrees of temporal and cellular heterogeneity. It may well be that carefully taking into account the finite-size of networks will be enough to produce the desired level of synchrony-based correlation, which is rather weak after all. Still, one would have to check wether achieving a given degree of synchrony requires the tuning of certain network features, such as the degree of shared input or the propensity of certain recurrent motifs [118] or the relative width of recurrent connections with respect to feedforward projections [119]. From a theoretical standpoint, the asynchronous state hypothesis answers the consistency problem by assuming no spiking correlations, and thus no synchrony. One can justify this assumption in idealized mathematical models by demonstrating the so-called "propagation-of-chaos" property [120], which rigorously holds for certain scaling limits with vanishing weights and under the assumption of exchangeability [107–109]. In this light, the main theoretical challenge posed by our analysis is extending the latter exchangeability-based property to include nonzero correlations [121], and hopefully characterize irregular synchronous state in some scaling limits.

ACKNOWLEDGMENTS

LB, BL, NP, ES, and TT were supported by the Vision Research program of the National Institutes of Health under award number R01EY024071. LB and TT were also supported by the CRCNS program of the National Science Foundation under award number DMS-2113213. We would like to thank François Baccelli, David Hansel, and Nicolas Brunel for insightful discussions.

Appendix A: Discrete-time spiking correlation

In this appendix, we consider first the discrete-time version of our model for possibly correlated excitatory synaptic inputs. In this model, we consider that observing K_{e} synaptic inputs during N time steps specifies a $\{0, 1\}$-valued

matrix $\{X_{k,i}\}_{1\leq k\leq K_\mathrm{e},1\leq i\leq N}$, where 1 indicates that an input is received and 0 indicates an absence of inputs. For simplicity, we further assume that the inputs are independent across time

$$\mathbb{P}\left[\{X_{k,i}\}_{1\leq k\leq K_\mathrm{e},1\leq i\leq N}\right] = \prod_{i=1}^{N}\mathbb{P}\left[\{X_{k,i}\}_{1\leq k\leq K_\mathrm{e}}\right] ,$$

so that we can drop the time index and consider the population vector $\{X_k\}_{1\leq k\leq K_\mathrm{e}}$. Consequently, given the individual spiking rate r_e, we have $\mathbb{E}[X_k] = \mathbb{P}[X_k = 1] = r_i\Delta t$, where Δt is the duration of the time step where a spike may or may not occur. Under the assumptions that $\{X_k\}_{1\leq k\leq K_\mathrm{e}}$ belongs to an infinitely exchangeable set of random variables, de Finetti theorem states that there exists a probability measure F_e on $[0,1]$ such that

$$\mathbb{P}\left[\{X_k\}_{1\leq k\leq K_\mathrm{e}}\right] = \int\prod_{k=1}^{K_\mathrm{e}}\theta_\mathrm{e}^{X_k}(1-\theta_\mathrm{e})^{1-X_k}\,\mathrm{d}F_\mathrm{e}(\theta_\mathrm{e}) .$$

Assuming the directing measure F_e known, we can compute the spiking correlation attached to our model. To see this, first observe that specifying the above probabilistic model for $K_\mathrm{e} = 1$, we have

$$\mathbb{E}[X_k] = \mathbb{E}[\mathbb{E}[X_k\,|\,\theta_\mathrm{e}]] = \mathbb{E}[\theta_\mathrm{e}] = \int\theta_\mathrm{e}\mathrm{d}F_\mathrm{e}(\theta_\mathrm{e}) .$$

Then, using the total law of covariance and specifying the above probabilistic model for $K = 2$, we have

$$\begin{aligned}\mathbb{C}[X_k, X_l] &= \mathbb{E}[\mathbb{C}[X_k, X_l\,|\,\theta_\mathrm{e}]] + \mathbb{C}[\mathbb{E}[X_k\,|\theta_\mathrm{e}], \mathbb{E}[X_l\,|\,\theta_\mathrm{e}]] ,\\ &= \mathbb{1}_{\{k=l\}}\mathbb{E}[\mathbb{V}[X_k\,|\,\theta_\mathrm{e}]] + \mathbb{C}[\theta_\mathrm{e},\theta_\mathrm{e}] ,\\ &= \mathbb{1}_{\{k=l\}}\mathbb{E}[\theta_\mathrm{e}(1-\theta_\mathrm{e})] + \mathbb{V}[\theta_\mathrm{e}] ,\\ &= \mathbb{1}_{\{k=l\}}\mathbb{E}[\theta_\mathrm{e}](1-\mathbb{E}[\theta_\mathrm{e}]) + \mathbb{1}_{\{k\neq l\}}\mathbb{V}[\theta_\mathrm{e}] .\end{aligned}$$

This directly yields that the spiking correlation reads

$$\rho_\mathrm{e} = \frac{\mathbb{C}[X_k, X_l]}{\mathbb{V}[X_k]} = \frac{\mathbb{V}[\theta_\mathrm{e}]}{\mathbb{E}[\theta_\mathrm{e}](1-\mathbb{E}[\theta_\mathrm{e}])} \tag{A1}$$

The exact same calculations can be performed for the partially exchangeable case of mixed excitation and inhibition. The assumption of partial exchangeability requires that when considered separately, the $\{0,1\}$-valued vectors $\{X_1,\ldots,X_{K_\mathrm{e}}\}$ and $\{Y_1,\ldots,Y_{K_\mathrm{e}}\}$ each belong to an infinitely exchangeable sequence of random variables. Then, de Finetti's theorem states that the probability to find the full vector of inputs $\{X_1,\ldots,X_{K_\mathrm{e}},Y_1,\ldots,Y_{K_\mathrm{e}}\}$ in any particular configuration is given by

$$\mathbb{P}[X_1,\ldots,X_{K_\mathrm{e}},Y_1,\ldots,Y_{K_\mathrm{i}}] = \int\prod_{k=1}^{K_\mathrm{e}}\theta_\mathrm{e}^{X_k}(1-\theta_\mathrm{e})^{1-X_k}\prod_{l=1}^{K_\mathrm{i}}\theta_\mathrm{i}^{Y_l}(1-\theta_\mathrm{i})^{1-Y_l}\,\mathrm{d}F_\mathrm{ei}(\theta_\mathrm{e},\theta_\mathrm{i}) , \tag{A2}$$

where the directing measure F_ei fully parametrizes our probabilistic model. Performing similar calculations as for the case of excitation alone within this partially exchangeable setting yields

$$\rho_\mathrm{ei} = \frac{\mathbb{C}[X_k, Y_l]}{\sqrt{\mathbb{V}[X_k]\,\mathbb{V}[Y_l]}} = \frac{\mathbb{C}[\theta_\mathrm{e},\theta_\mathrm{i}]}{\sqrt{\mathbb{E}[\theta_\mathrm{e}](1-\mathbb{E}[\theta_\mathrm{e}])\mathbb{E}[\theta_\mathrm{i}](1-\mathbb{E}[\theta_\mathrm{i}])}} . \tag{A3}$$

Appendix B: Compound Poisson processes as continuous-time limits

Let us consider the discrete-time model specified by Eq. (A2), which is obtained under the assumption of partial infinite exchangeability. Under this assumption, the probability laws of the inputs is entirely determined by the distribution of $(k_\mathrm{e}, k_\mathrm{i})$, where k_e denotes the number of active excitatory inputs and k_i denotes the number of inhibitory inputs. This distribution can be computed as

$$P_{\mathrm{ei},kl} = \mathbb{P}[k_\mathrm{e} = k, k_\mathrm{i} = l] = \binom{K_\mathrm{e}}{k}\binom{K_\mathrm{i}}{l}\int\theta_\mathrm{e}^k(1-\theta_\mathrm{e})^{K_\mathrm{e}-k}\theta_\mathrm{i}^l(1-\theta_\mathrm{i})^{K_\mathrm{i}-l}\,\mathrm{d}F_\mathrm{ei}(\theta_\mathrm{e},\theta_\mathrm{i}) .$$

It is convenient to choose the directing measure as beta distributions since these are conjugate to the binomial distributions. Such a choice yields a class of probabilistic models referred to as beta-binomial models, which have

been studied extensively [61, 62]. In this appendix, we always assume that the marginals $F_{\rm e}$ and $F_{\rm i}$ have the form $F_{\rm e} \sim \text{Beta}(\alpha_{\rm e}, \beta_{\rm e})$ and $F_{\rm i} \sim \text{Beta}(\alpha_{\rm i}, \beta_{\rm i})$. Then, direct integrations shows that the marginal distributions for the number of excitatory inputs and inhibitory inputs are

$$P_{{\rm e},k} = \sum_{l=0}^{K_{\rm i}} P_{{\rm ei},kl} = \binom{K_{\rm e}}{k} \frac{B(\alpha_{\rm e}+k, \beta_{\rm e}+K_{\rm e}-k)}{B(\alpha_{\rm e}, \beta_{\rm e})} \quad \text{and} \quad P_{{\rm i},l} = \sum_{k=0}^{K_{\rm e}} P_{{\rm ei},kl} = \binom{K_{\rm i}}{l} \frac{B(\alpha_{\rm i}+l, \beta_{\rm i}+K_{\rm i}-l)}{B(\alpha_{\rm i}, \beta_{\rm i})} \,.$$

Moreover, given individual spiking rates $r_{\rm e}$ and $r_{\rm i}$ within a time step Δt, we have

$$r_{\rm e}\Delta t = \mathbb{E}\left[X_k\right] = \mathbb{P}\left[X_k = 1\right] = \mathbb{E}\left[\theta_{\rm e}\right] = \frac{\alpha_{\rm e}}{\alpha_{\rm e}+\beta_{\rm e}} \quad \text{and} \quad r_{\rm i}\Delta t = \mathbb{E}\left[Y_l\right] = \mathbb{P}\left[Y_l = 1\right] = \mathbb{E}\left[\theta_{\rm i}\right] = \frac{\alpha_{\rm i}}{\alpha_{\rm i}+\beta_{\rm i}} \,.$$

The continuous-time limit is obtained by taking $\Delta t \to 0^+$, which implies that the parameters $\alpha_{\rm e}$ and $\alpha_{\rm i}$ jointly vanish. When $\alpha_{\rm e}, \alpha_{\rm i} \to 0^+$, the beta distributions $F_{\rm e}$ and $F_{\rm i}$ becomes deficient and we have $P_{{\rm e},0}, P_{{\rm i},0} \to 1$. In other words, time bins of size Δt almost surely have no active inputs in the limit $\Delta t \to 0^+$. Actually, one can show that

$$1 - P_{{\rm e},0} \sim \left(\psi(K_{\rm e}+\beta_{\rm e}) - \psi(\beta_{\rm e})\right)\alpha_{\rm e} \quad \text{and} \quad 1 - P_{{\rm i},0} \sim \left(\psi(K_{\rm i}+\beta_{\rm i}) - \psi(\beta_{\rm i})\right)\alpha_{\rm i} \,,$$

where ψ denotes the digamma function. This indicates in the limit $\Delta t \to 0^+$, the times at which some excitatory inputs or some inhibitory inputs are active define a point process. Moreover, owing to the assumption of independence across time, this point process will actually be a Poisson point process. Specifically, consider a time $T > 0$ and set $\Delta t = T/N$ for some large integer N. Define the sequence of times

$$T_{{\rm e},n} = \frac{T}{N} \cdot \inf\left\{i > NT_{{\rm e},n-1}/T \,|k_{{\rm e},i} \geq 1\right\} \quad \text{with} \quad T_{{\rm e},1} = \frac{T}{N} \cdot \inf\left\{i \geq 0 \,|k_{{\rm e},i} \geq 1\right\} \,,$$
$$T_{{\rm i},n} = \frac{T}{N} \cdot \inf\left\{i > NT_{{\rm i},n-1}/T \,|k_{{\rm i},i} \geq 1\right\} \quad \text{with} \quad T_{{\rm i},1} = \frac{T}{N} \cdot \inf\left\{i \geq 0 \,|k_{{\rm i},i} \geq 1\right\} \,.$$

Considered separately, the sequences of times $\{T_{{\rm e},n}\}_{n\geq 1}$ and $\{T_{{\rm i},n}\}_{n\geq 1}$ constitute binomial approximations of Poisson processes which we denote by $N_{\rm e}$ and $N_{\rm i}$, respectively. It is a classical result that these limit Poisson processes are recovered exactly when $N \to \infty$ and that their rates are respectively given by

$$b_{\rm e} = \lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm e},0}}{\Delta t} = \left(\psi(K_{\rm e}+\beta_{\rm e}) - \psi(\beta_{\rm e})\right)\left(\lim_{\Delta t \to 0^+} \frac{\alpha_{\rm e}}{\Delta t}\right) = \left(\psi(K_{\rm e}+\beta_{\rm e}) - \psi(\beta_{\rm e})\right)\beta_{\rm e} r_{\rm e} \,,$$
$$b_{\rm i} = \lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm i},0}}{\Delta t} = \left(\psi(K_{\rm i}+\beta_{\rm i}) - \psi(\beta_{\rm i})\right)\left(\lim_{\Delta t \to 0^+} \frac{\alpha_{\rm i}}{\Delta t}\right) = \left(\psi(K_{\rm i}+\beta_{\rm i}) - \psi(\beta_{\rm i})\right)\beta_{\rm i} r_{\rm i} \,.$$

For all integer $K > 1$, the function $\beta \mapsto \beta\left(\psi(K+\beta) - \psi(\beta)\right)$ is an increasing analytic functions on the domain $\mathbb{R}^+$ with range $[1, K]$. Thus, we always have $r_{\rm e} \leq b_{\rm e} \leq K_{\rm e} r_{\rm e}$ and $r_{\rm i} \leq b_{\rm i} \leq K_{\rm i} r_{\rm i}$ and the extreme cases are achieved for perfect or zero correlations. Perfect correlations are achieved when $\rho_{\rm e} = 1$ or $\rho_{\rm i} = 1$, which corresponds to $\beta_{\rm e} \to 0$ or $\beta_{\rm i} \to 0$. This implies that $b_{\rm e} = r_{\rm e}$ and $b_{\rm i} = r_{\rm i}$, consistent with all synapses activating simultaneously. Zero correlations are achieved when $\rho_{\rm e} = 0$ or $\rho_{\rm i} = 0$, which corresponds to $\beta_{\rm e} \to \infty$ or $\beta_{\rm i} \to \infty$. This implies that $b_{\rm e} = K_{\rm e} r_{\rm e}$ and $b_{\rm i} = K_{\rm i} r_{\rm i}$, consistent with all synapses activating asynchronously, so that no inputs simultaneously activate. Observe that in all generality, the rates $b_{\rm e}$ and $b_{\rm i}$ are such that the mean number of spikes over the duration T is conserved in the limit $\Delta t \to 0^+$. For instance, one can check that

$$K_{\rm e} r_{\rm e} T = \mathbb{E}\left[\sum_{T_{{\rm e},n} \leq T} k_{{\rm e},NT_{{\rm e},n}/T}\right] = \mathbb{E}\left[\sum_{n=1}^{N_{\rm e}(T)} k_{{\rm e},n}\right] = \mathbb{E}\left[N_{\rm e}(T)\right]\mathbb{E}\left[k_{\rm e}\right] = b_{\rm e} T \mathbb{E}\left[k_{\rm e}\right]$$

When excitation and inhibition are considered separately, the limit process $\Delta t \to 0^+$ specifies two compound Poisson processes

$$t \mapsto \sum_{n=1}^{N_{\rm e}(t)} k_{{\rm e},n} \quad \text{and} \quad t \mapsto \sum_{n=1}^{N_{\rm i}(t)} k_{{\rm i},n} \,,$$

where $N_{\rm e}$ and $N_{\rm i}$ are Poisson processes with rate $b_{\rm e}$ and $b_{\rm i}$ and where $\{k_{{\rm e},n}\}_{n\geq 1}$ are i.i.d according to $p_{\rm e}$ and $\{k_{{\rm e},n}\}_{n\geq 1}$ are i.i.d according to $p_{\rm i}$. Nonzero correlations between excitation and inhibition emerge when the Poisson processes $N_{\rm e}$ and $N_{\rm i}$ are not independent. This corresponds to the processes $N_{\rm e}$ and $N_{\rm i}$ sharing times, so excitation and inhibition occur simultaneously at these times. To understand this point intuitively, let us consider the limit Poisson process N obtained by considering synaptic events without distinguishing excitation and inhibition. For perfect correlation, i.e.,

$\rho_{\rm ei} = 1$, all synapses activate synchronously and we have $N = N_{\rm e} = N_{\rm i}$: all times are shared. By contrast, for zero correlation, i.e., $\rho_{\rm ei} = 0$, no synapses activate simultaneously and we have $N = N_{\rm e} + N_{\rm i}$: no times are shared. For intermediary regime of correlations, a nonzero fraction of times will be shared resulting in a driving Poisson process N with overall rate b satisfying $\min(b_{\rm e}, b_{\rm i}) \leq b < b_{\rm e} + b_{\rm i}$. We investigate the above intuitive statements quantitatively in Appendix D by inspecting two key examples.

Let us conclude this appendix by recapitulating the general form of the limit compound process Z obtained in the continuous-time limit $\Delta t \to 0^+$ when jointly considering excitation and inhibition. This compound Poisson process can be represented as

$$t \mapsto Y(t) = \left(\sum_n^{N(t)} W_{{\rm e},n}, \sum_n^{N(t)} W_{{\rm i},n} \right) ,$$

where N is that Poisson process registering all synaptic events without distinguishing excitation and inhibition and where the pairs $(W_{{\rm e},n,W_{{\rm i},n}})$ are i.i.d. random jumps in $\mathbb{R} \times \mathbb{R} \setminus \{0,0\}$. Formally, such a process is specified by the rate of N, denoted by b, and the bivariate distribution of the jumps $(W_{{\rm e},n}, W_{{\rm i},n})$, denoted by $p_{\rm ei}$. These are defined as

$$b = \lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm ei},00}}{\Delta t} \quad \text{and} \quad p_{{\rm ei},kl} = \lim_{\Delta t \to 0^+} \frac{P_{{\rm ei},kl}}{1 - P_{{\rm ei},00}} \quad \text{for} \quad (k,l) \neq (0,0) , \tag{B1}$$

where $P_{{\rm ei},00}$ is the probability to register no synaptic activation during a time step Δt. According to these definitions, b is the infinitesimal likelihood that an input is active within a time bin, whereas $p_{\rm ei}$ is the probability that k excitatory inputs and l inhibitory inputs are active given that at least one input is active. One can similarly define the excitatory and inhibitory rates of events $b_{\rm e}$ and $b_{\rm i}$, as well as the excitatory jump distribution $p_{\rm e}$ and the inhibitory jump distribution $p_{{\rm i},l}$. Specifically, we have

$$\begin{aligned} b_{\rm e} &= \lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm e},0}}{\Delta t} \quad \text{and} \quad p_{{\rm e},k} = \lim_{\Delta t \to 0^+} \frac{P_{{\rm e},k}}{1 - P_{{\rm e},0}} \quad \text{for} \quad k \neq 0 , \\ b_{\rm i} &= \lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm i},0}}{\Delta t} \quad \text{and} \quad p_{{\rm i},l} = \lim_{\Delta t \to 0^+} \frac{P_{{\rm i},l}}{1 - P_{{\rm i},0}} \quad \text{for} \quad l \neq 0 , \end{aligned} \tag{B2}$$

with $P_{{\rm e},k} = \sum_{l=0}^{K_{\rm i}} P_{{\rm ei},k,l}$ and $P_{{\rm i},k} = \sum_{k=0}^{K_{\rm e}} P_{{\rm ei},k,l}$. Observe that thus defined, the jump distribution $p_{\rm e}$ and $p_{\rm i}$ are specified as conditional marginal distributions of the joint jump distribution $p_{\rm ei}$ on the events $\{k_{\rm e} > 0\}$ and $\{k_{\rm i} > 0\}$, respectively. These are such that $p_{{\rm e},k} = (b/b_{\rm e}) \sum_{l=0}^{K_{\rm i}} p_{{\rm ei},kl}$ and $p_{{\rm i},l} = (b/b_{\rm i}) \sum_{k=0}^{K_{\rm e}} p_{{\rm ei},kl}$. To see why, observe for instance that

$$p_{{\rm e},k} = \lim_{\Delta t \to 0^+} \frac{P_{{\rm e},k}}{1 - P_{{\rm e},0}} = \lim_{\Delta t \to 0^+} \sum_{l=0}^{K_{\rm i}} \frac{P_{{\rm ei},kl}}{1 - P_{{\rm ei},00}} \frac{1 - P_{{\rm ei},00}}{1 - P_{{\rm e},0}} = \left(\sum_{l=0}^{K_{\rm i}} p_{{\rm ei},kl} \right) \left(\lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm ei},00}}{1 - P_{{\rm e},0}} \right) = \frac{b}{b_{\rm e}} \sum_{l=0}^{K_{\rm i}} p_{{\rm ei},kl} \tag{B3}$$

where we have used the definitions of the rates b and $b_{\rm e}$ given in Eq. (B1) and Eq. (B2) to establish that

$$\lim_{\Delta t \to 0^+} \frac{1 - P_{{\rm ei},00}}{1 - P_{{\rm e},0}} = \frac{\lim_{\Delta t \to 0^+} (1 - P_{{\rm ei},00})/\Delta t}{\lim_{\Delta t \to 0^+} (1 - P_{{\rm e},0})/\Delta t} = \frac{b}{b_{\rm e}} .$$

Appendix C: Continuous-time spiking correlation

Eq. (A1) and Eq. (A3) carry over to the continuous time limit $\Delta t \to 0^+$ by observing that for limit compound Poisson processes to emerge, one must have that $\mathbb{E}[X_k] = \mathbb{E}[\theta_{\rm e}] = O(\Delta t)$ and $\mathbb{E}[Y_l] = \mathbb{E}[\theta_{\rm i}] = O(\Delta t)$. This directly implies that when $\Delta t \to 0^+$, we have

$$\rho_{\rm e} = \frac{\mathbb{C}[X_k, X_l]}{\mathbb{V}[X_k]} \sim \frac{\mathbb{E}[X_k X_l]}{\mathbb{E}[X_k^2]} = \frac{\mathbb{E}[X_k X_l]}{\mathbb{E}[X_k]} \quad \text{and} \quad \rho_{\rm ei} = \frac{\mathbb{C}[X_k, Y_l]}{\sqrt{\mathbb{V}[X_k]\,\mathbb{V}[Y_l]}} \sim \frac{\mathbb{E}[X_k Y_l]}{\sqrt{\mathbb{E}[X_k^2]\,\mathbb{E}[Y_l^2]}} = \frac{\mathbb{E}[X_k Y_l]}{\sqrt{\mathbb{E}[X_k]\,\mathbb{E}[Y_l]}} . \tag{C1}$$

All the stationary expectations appearing above can be computed via the jump distribution of the limit point process emerging in the limit $\Delta t \to 0^+$ [104]. Because this limit process is a compound Poisson process with discrete bivariate jumps, the resulting jump distribution $p_{\rm ei}$ is specified over $\{1, \ldots, K_{\rm e}\} \times \{1, \ldots, K_{\rm i}\} \setminus \{0,0\}$. Denoting by b the

overall rate of synaptic events, one has $\lim_{\Delta t\to 0^+}\mathbb{E}\left[X_kY_l\right]/\Delta t = b\mathbb{E}_{\rm ei}\left[X_kY_l\right]$. Then by partial exchangeability of the $\{0,1\}$-valued population vectors $\{X_k\}_{1\le k\le K_{\rm e}}$ and $\{Y_l\}_{1\le l\le K_{\rm i}}$, we have

$$\mathbb{E}_{\rm ei}\left[X_kY_l\right] = \mathbb{E}_{\rm ei}\left[\mathbb{E}\left[X_kY_l|(k_{\rm e},k_{\rm i})\right]\right] = \mathbb{E}_{\rm ei}\left[\frac{k_{\rm e}}{K_{\rm e}}\frac{k_{\rm i}}{K_{\rm i}}\right] = \sum_{k=0}^{K_{\rm e}}\sum_{l=0}^{K_{\rm i}}\frac{k}{K_{\rm e}}\frac{l}{K_{\rm i}}p_{{\rm ei},kl} = \frac{\mathbb{E}_{\rm ei}\left[k_{\rm e}k_{\rm i}\right]}{K_{\rm e}K_{\rm i}}\,. \tag{C2}$$

where the bivariate jump $(k_{\rm e},k_{\rm i})$ is distributed as $p_{\rm ei}$.

To further proceed, it is important to note the relation between the expectation $\mathbb{E}_{\rm ei}\left[\cdot\right]$, which is tied to the overall input process with rate b, and the expectation $\mathbb{E}_{\rm e}\left[\cdot\right]$ which is tied to the excitatory input process with rate $b_{\rm e}$. This relation is best captured by remarking that $p_{\rm e}$ are not defined as the marginals of $p_{\rm ei}$, but as only as conditional marginals on $\{k_{\rm e}>0\}$. In other words, we have $p_{{\rm e},k} = (b/b_{\rm e})\sum_{l=0}^{K_{\rm i}}p_{{\rm ei},kl}$, which implies that $b\mathbb{E}_{\rm ei}\left[X_kX_l\right] = b_{\rm e}\mathbb{E}_{\rm e}\left[X_kX_l\right]$ and $\mathbb{E}\left[X_k\right] = b\mathbb{E}_{\rm ei}\left[X_k\right] = b_{\rm e}\mathbb{E}_{\rm e}\left[X_k\right]$ with

$$\mathbb{E}_{\rm e}\left[X_kX_l\right] = \mathbb{E}_{\rm e}\left[\mathbb{E}\left[X_kX_l|k_{\rm e}\right]\right] = \mathbb{E}_{\rm e}\left[\frac{k_{\rm e}(k_{\rm e}-1)}{K_{\rm e}(K_{\rm e}-1)}\right] = \sum_{k=0}^{K_{\rm e}}\frac{k(k-1)}{K_{\rm e}(K_{\rm e}-1)}p_{{\rm e},k} = \frac{\mathbb{E}_{\rm e}\left[k_{\rm e}(k_{\rm e}-1)\right]}{K_{\rm e}(K_{\rm e}-1)}\,, \tag{C3}$$

$$\mathbb{E}_{\rm e}\left[X_k\right] = \mathbb{E}_{\rm e}\left[\mathbb{E}\left[X_k|k_{\rm e}\right]\right] = \mathbb{E}_{\rm e}\left[\frac{k_{\rm e}}{K_{\rm e}}\right] = \sum_{k=0}^{K_{\rm e}}\sum_{l=0}^{K_{\rm i}}\frac{k}{K_{\rm e}}p_{{\rm e},k} = \frac{\mathbb{E}_{\rm e}\left[k_{\rm e}\right]}{K_{\rm e}}\,, \tag{C4}$$

with similar expressions for the inhibition-related quantities. Injecting Eq. (C2), Eq. (C3), and Eq. (C4) in Eq. (C1) yields

$$\rho_{\rm e} = \frac{\mathbb{E}_{\rm e}\left[k_{\rm e}(k_{\rm e}-1)\right]}{\mathbb{E}_{\rm e}\left[k_{\rm e}\right](K_{\rm e}-1)} \quad\text{and}\quad \rho_{\rm ei} = \frac{b\mathbb{E}_{\rm ei}\left[k_{\rm e}k_{\rm i}\right]}{\sqrt{K_{\rm e}b_{\rm e}\mathbb{E}_{\rm e}\left[k_{\rm e}\right]K_{\rm i}b_{\rm i}\mathbb{E}_{\rm i}\left[k_{\rm i}\right]}} = \frac{\mathbb{E}_{\rm ei}\left[k_{\rm e}k_{\rm i}\right]}{\sqrt{K_{\rm e}\mathbb{E}_{\rm ei}\left[k_{\rm e}\right]K_{\rm i}\mathbb{E}_{\rm ei}\left[k_{\rm i}\right]}}\,.$$

Appendix D: Two examples of limit compound Poisson processes

The probability $P_{{\rm ei},00}$ that plays a central role in Appendix B can be easily computed for zero correlation, i.e., $\rho_{\rm ei}=0$, by considering a directing measure under product form $F_{\rm ei}(\theta_{\rm e},\theta_{\rm i}) = F_{\rm e}(\theta_{\rm e})F_{\rm i}(\theta_{\rm i})$. Then integration with respect to the separable variables $\theta_{\rm e}$ and $\theta_{\rm i}$ yields

$$P_{{\rm ei},kl} = P_{{\rm e},k}P_{{\rm i},l} = \binom{K_{\rm e}}{k}\frac{B(\alpha_{\rm e}+k,\beta_{\rm e}+K_{\rm e}-k)}{B(\alpha_{\rm e},\beta_{\rm e})}\binom{K_{\rm i}}{l}\frac{B(\alpha_{\rm i}+l,\beta_{\rm i}+K_{\rm i}-l)}{B(\alpha_{\rm i},\beta_{\rm i})}\,.$$

In turn, the limit compound Poisson process can be obtain in the limit $\Delta t\to 0^+$ by observing that

$$1-P_{{\rm e},0} = b_{\rm e}\Delta t + o(\Delta t)\,,\quad 1-P_{{\rm e},0}P_{{\rm i},0} = b_{\rm i}\Delta t + o(\Delta t)\,,\quad\text{and}\quad 1-P_{{\rm e},0}P_{{\rm i},0} = (b_{\rm e}+b_{\rm i})\Delta t + o(\Delta t)\,,$$

which implies that the overall rate is determined as $b = \lim_{\Delta t\to 0^+}(1-P_{{\rm e},0}P_{{\rm i},0})/\Delta t = b_{\rm e}+b_{\rm i}$, as expected. To characterize the limit compound Poisson process, it remains to exhibit $p_{{\rm ei},kl}$, the distribution of the jumps $k_{\rm e}$ and $k_{\rm i}$, Suppose that $k\ge 1$, then we have

$$\begin{aligned}
p_{{\rm ei},kl} &= \lim_{\Delta t\to 0^+}\frac{P_{{\rm e},k}P_{{\rm i},l}}{1-P_{{\rm e},0}P_{{\rm i},0}}\,,\\
&= \lim_{\Delta t\to 0^+}\left[\left(\frac{1-P_{{\rm e},0}}{1-P_{{\rm e},0}P_{{\rm i},0}}\right)P_{{\rm i},l}\left(\frac{P_{{\rm e},k}}{1-P_{{\rm e},0}}\right)\right]\,,\\
&= \left(\lim_{\Delta t\to 0^+}\frac{1-P_{{\rm e},0}}{1-P_{{\rm e},0}P_{{\rm i},0}}\right)\left(\lim_{\Delta t\to 0^+}P_{{\rm i},l}\right)\left(\lim_{\Delta t\to 0^+}\frac{P_{{\rm e},k}}{1-P_{{\rm e},0}}\right)\,.
\end{aligned}$$

Then one can use the limit behaviors

$$\lim_{\Delta t\to 0^+}\frac{1-P_{{\rm e},0}}{1-P_{{\rm e},0}P_{{\rm i},0}} = \frac{b_{\rm e}}{b_{\rm e}+b_{\rm i}}\quad\text{and}\quad \lim_{\Delta t\to 0^+}P_{{\rm i},l} = \mathbb{1}_{\{l=0\}}\,.$$

so that for $k\ge 1$, we have

$$p_{{\rm ei},kl} = \frac{b_{\rm e}}{b_{\rm e}+b_{\rm i}}\mathbb{1}_{\{l=0\}}p_{{\rm e},k}\quad\text{with}\quad p_{{\rm e},k} = \lim_{\Delta t\to 0^+}\frac{P_{{\rm e},k}}{1-P_{{\rm e},0}} = \binom{K_{\rm e}}{k}\frac{B(k,\beta_{\rm e}+K_{\rm e}-k)}{\psi(K_{\rm e}+\beta_{\rm e})-\psi(\beta)}\,.$$

A similar calculation shows that for all $l \geq 1$, we have $p_{\mathrm{ei},kl} = b_\mathrm{i}/(b_\mathrm{e}+b_\mathrm{i})\mathbb{1}_{\{k=0\}}p_{\mathrm{i},l}$. Thus $p_{\mathrm{ei},kl} = 0$ whenever $k, l \geq 1$, so that the support of $p_{\mathrm{ei},kl}$ is $\{1,\ldots,K_\mathrm{e}\}\times\{0\}\cup\{0\}\times\{1,\ldots,K_\mathrm{i}\}$. This is consistent with the intuition that excitation and inhibition happen at distinct times in the absence of correlations.

Let us now consider the case of maximum correlation for $F_\mathrm{e} = F_\mathrm{i} = F$, where F is a beta distribution with parameters α and β. Moreover, let us assume the deterministic coupling $\theta_\mathrm{e} = \theta_\mathrm{i}$ such that $F_\mathrm{ei}(\theta_\mathrm{e},\theta_\mathrm{i}) = F(\theta_\mathrm{e})\delta(\theta_\mathrm{i}-\theta_\mathrm{e})$. Then, the joint distribution of the jumps $(k_\mathrm{e}, k_\mathrm{i})$ can be evaluated via direct integration as

$$\begin{aligned}
P_{\mathrm{ei},kl} &= \binom{K_\mathrm{e}}{k}\binom{K_\mathrm{i}}{l}\int \theta_\mathrm{e}^k(1-\theta_\mathrm{e})^{K_\mathrm{e}-k}\theta_\mathrm{e}^l(1-\theta_\mathrm{i})^{K_\mathrm{i}-l}\,dF(\theta_\mathrm{e})\delta(\theta_\mathrm{i}-\theta_\mathrm{e})\,,\\
&= \binom{K_\mathrm{e}}{k}\binom{K_\mathrm{i}}{l}\int \theta^{k+l}(1-\theta)^{K_\mathrm{e}+K_\mathrm{i}-k-l}\,dF(\theta)\,,\\
&= \binom{K_\mathrm{e}}{k}\binom{K_\mathrm{i}}{l}\frac{B(\alpha+k+l,\beta+K_\mathrm{e}+K_\mathrm{i}-k-l)}{B(\alpha,\beta)}\,.
\end{aligned}$$

As excitation and inhibition are captured separately by the same marginal functions $F_\mathrm{e} = F_\mathrm{i} = F$, we necessarily have $\alpha/(\alpha+\beta) = \mathbb{E}[X_k] = \mathbb{E}[Y_l] = r_\mathrm{e}\Delta t = r_\mathrm{i}\Delta t$ and we refer to the common spiking rate as r. Then the overall rate of synaptic events is obtained as

$$b = \lim_{\Delta t\to 0^+}\frac{1-P_{\mathrm{ei},00}}{\Delta t} = \lim_{\alpha\to 0^+}\frac{1-P_{\mathrm{ei},00}}{\alpha}\lim_{\Delta t\to 0^+}\frac{\alpha}{\Delta t} = (\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta))\,\beta r\,, \tag{D1}$$

and one can check that b differs from the excitatory- and inhibitory-specific rates b_e and b_i, which satisfy

$$b_\mathrm{e} = \lim_{\Delta t\to 0^+}\frac{1-P_{\mathrm{e},0}}{\Delta t} = (\psi(K_\mathrm{e}+\beta)-\psi(\beta))\,\beta r \quad\text{and}\quad b_\mathrm{i} = \lim_{\Delta t\to 0^+}\frac{1-P_{\mathrm{i},0}}{\Delta t} = (\psi(K_\mathrm{i}+\beta)-\psi(\beta))\,\beta r\,. \tag{D2}$$

To characterize the limit compound Poisson process, it remains to exhibit $p_{\mathrm{ei},kl}$, the joint distribution of the jumps $(k_\mathrm{e}, k_\mathrm{i})$. A similar calculation as for the case of excitation alone yields

$$p_{\mathrm{ei},kl} = \lim_{\Delta t\to 0^+}\frac{P_{\mathrm{ei},kl}}{1-P_{\mathrm{ei},00}} = \binom{K_\mathrm{e}}{k}\binom{K_\mathrm{i}}{l}\frac{B(k+l,\beta+K_\mathrm{e}+K_\mathrm{i}-k-l)}{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta)}\,.$$

Remember that within our model, spiking correlations do not depends on the number of neurons and that by construction we have $\rho_\mathrm{ei} \leq \sqrt{\rho_\mathrm{e}\rho_\mathrm{i}}$. Thus, for the symmetric case under consideration, maximum correlation corresponds to $\rho_\mathrm{ei} = \rho_\mathrm{e} = \rho_\mathrm{i} = 1/(1+\beta)$. In particular perfect correlation between excitation and inhibition can only be attained for $\beta \to 0$. When $\beta > 0$, i.e., for partial correlations, the Poisson processes N_e and N_i only share a fraction of their times, yielding an aggregate Poisson process N such that $\min(b_\mathrm{e}, b_\mathrm{i}) < b < b_\mathrm{e}+b_\mathrm{i}$. The relations between b, b_e, and b_i can be directly recovered from the knowledge of p_ei by observing that

$$\begin{aligned}
\mathbb{P}[k_\mathrm{e}=0,k_\mathrm{i}>0] &= \sum_{l=1}^{K_\mathrm{i}} p_{\mathrm{ei},0l} = \frac{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(K_\mathrm{e}+\beta)}{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta)}\\
\mathbb{P}[k_\mathrm{i}=0,k_\mathrm{e}>0] &= \sum_{k=1}^{K_\mathrm{e}} p_{\mathrm{ei},k0} = \frac{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(K_\mathrm{i}+\beta)}{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta)}\,.\\
\mathbb{P}[k_\mathrm{i}>0,k_\mathrm{e}>0] &= \sum_{k=1}^{K_\mathrm{e}}\sum_{l=1}^{K_\mathrm{i}} p_{\mathrm{ei},kl} = 1-\frac{2\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(K_\mathrm{e}+\beta)-\psi(K_\mathrm{i}+\beta)}{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta)}\,.
\end{aligned}$$

This implies that the fraction of times with nonzero excitation is given by

$$\mathbb{P}[k_\mathrm{e}>0] = \mathbb{P}[k_\mathrm{e}>0,k_\mathrm{i}=0] + \mathbb{P}_0[k_\mathrm{e}>0,k_\mathrm{i}>0] = \frac{\psi(K_\mathrm{e}+\beta)-\psi(\beta)}{\psi(K_\mathrm{e}+K_\mathrm{i}+\beta)-\psi(\beta)}\,,$$

so that we consistently recover the value of b_e already obtained in Eq. (9) and Eq. (D2) via

$$b_\mathrm{e}T = \mathbb{E}[N_\mathrm{e}(T)] = \mathbb{E}\left[\mathbb{1}_{\{k_\mathrm{e}>0\}}N(T)\right] = bT\mathbb{E}_\mathrm{ei}\left[\mathbb{1}_{\{k_\mathrm{e}>0\}}\right] = bT\mathbb{P}_0[k_\mathrm{e}>0]\,.$$

Appendix E: Marcus jump rule

The goal of this appendix is to justify the Marcus-type update rule given in Eq. (13). To do so let us first remark that given a finite time interval $[0,T]$, the number of synaptic activation times $\{T_n\}_{n\in\mathbb{Z}}$ falling in this interval is almost surely finite. In particular, we have $\Delta = \inf_{0\le T_n\neq T_m\le T}|T_n - T_m| > 0$ almost surely. Consequently, taking $\epsilon < \Delta/\tau_{\rm s}$ ensures that synaptic activation events do not overlap in time, so that it is enough to consider a single synaptic activation triggered with no lack of generality in $T_0 = 0$. Let us denote the voltage just before the impulse onset as $V(T_0^-) = V_0$, which will serve as initial condition for the ensuing voltage dynamics. As the synaptic conductances remains equals to $W_e/(\epsilon\tau)$ and $W_i/(\epsilon\tau)$ for a duration $[0,\epsilon\tau]$, the voltage V_ϵ satisfies

$$\tau\dot{V}_\epsilon = -V_\epsilon + (W_{\rm e}/\epsilon)(V_{\rm e} - V_\epsilon) + (W_{\rm i}/\epsilon)(V_{\rm i} - V_\epsilon)\,, \quad 0\le t\le \epsilon\tau\,,$$

where we assume $I = 0$ for simplicity. The unique solution satisfying $V(0^-) = V_0$ is

$$V_\epsilon(t) = V_0 e^{-t/\tau(1+W_{\rm e}/\epsilon+W_{\rm i}/\epsilon)} + \frac{W_{\rm e}V_{\rm e} + W_{\rm i}V_{\rm i}}{\epsilon + W_{\rm e} + W_{\rm i}}\left(1 - e^{-t/\tau(1+W_{\rm e}/\epsilon+W_{\rm i}/\epsilon)}\right)\,, \quad 0\le t\le\epsilon\tau\,.$$

The Marcus-type rule follows from evaluating the jump update as the limit

$$\begin{aligned}\lim_{\epsilon\to 0^+} V_\epsilon(\epsilon\tau) - V_0 &= \lim_{\epsilon\to 0^+}\left\{V_0\left(e^{-(\epsilon+W_{\rm e}+W_{\rm i})} - 1\right) + \frac{W_{\rm e}V_{\rm e} + W_{\rm i}V_{\rm i}}{\epsilon + W_{\rm e} + W_{\rm i}}\left(1 - e^{-(\epsilon+W_{\rm e}+W_{\rm i})}\right)\right\}\,,\\ &= \left(\frac{W_{\rm e}V_{\rm e} + W_{\rm i}V_{\rm i}}{W_{\rm e} + W_{\rm i}} - V_0\right)\left(1 - e^{-(W_{\rm e}+W_{\rm i})}\right)\,,\end{aligned}$$

which has the same form as the rule announced in Eq. (13). Otherwise, at fixed $\epsilon > 0$, the fraction of time for which the voltage V_ϵ is exponentially relaxing toward the leak reversal potential $V_{\rm L} = 0$ is larger than $1 - N\epsilon/T$, where N denotes the almost surely finite number of synaptic activations, which does not depends on ϵ. Thus, the voltage $V = \lim_{\epsilon\to 0^+} V_\epsilon$ exponentially relaxes toward $V_{\rm L} = 0$, except when it has jump discontinuities in $\{T_n\}_{n\in\mathbb{Z}}$.

Appendix F: Stationary voltage mean

For a positive synaptic activation time $t > 0$, the classical method of the variation of the constant applies to solve Eq. (1). This yields an expression for $V_\epsilon(t)$ in terms of regular Riemann-Stieltjes integrals where the conductance traces $h_{\rm e}(t)$ and $h_{\rm i}(t)$ are treated as a form of deterministic quenched disorder. Specifically, given an initial condition $V_\epsilon(0)$, we have

$$V_\epsilon(t) = V_\epsilon(0)e^{-\int_0^t \frac{1}{\tau}+h_{\rm e}(u)+h_{\rm i}(u)\,{\rm d}u} + \int_0^t \left(V_{\rm e}h_{\rm e}(u) + V_{\rm i}h_{\rm i}(u) + I/C\right) e^{-\int_u^t \frac{1}{\tau}+h_{\rm e}(v)+h_{\rm i}(v)\,{\rm d}v}\,{\rm d}u\,.$$

where $V_\epsilon(t)$ depends on ϵ via the all-or-none-conductance processes $h_{\rm e}$ and $h_{\rm i}$. As usual, the stationary dynamics of the voltage V_ϵ is recovered by considering the limit of arbitrary large times $t\to\infty$, for which one can neglect the influence of the initial condition $V_\epsilon(0)$. Introducing the cumulative input processes $H = (H_{\rm e}, H_{\rm i})$ defined by

$$\big(H_{\rm e}(t), H_{\rm i}(t)\big) = \left(\int_0^t h_{\rm e}(u)\,du, \int_0^t h_{\rm i}(u)\,du\right)\,.$$

and satisfying $\tau{\rm d}H_{\rm e}(t) = h_{\rm e}(t)\,{\rm d}t$ and $\tau{\rm d}H_{\rm i}(t) = h_{\rm i}(t)\,{\rm d}t$, we have

$$V_\epsilon = \int_{-\infty}^0 e^{\frac{t}{\tau}+H_{\rm e}(t)+H_{\rm i}(t)}\left({\rm d}[V_{\rm e}H_{\rm e}(t) + V_{\rm i}H_{\rm i}(t)] + \frac{I}{G}\frac{{\rm d}t}{\tau}\right)\,. \tag{F1}$$

In turn, expanding the integrand above yields the following expression for the stationary expectation of the voltage

$$\mathbb{E}\left[V_\epsilon\right] = V_{\rm e}\int_{-\infty}^0 e^{\frac{t}{\tau}}\mathbb{E}\left[e^{H_{\rm e}(t)+H_{\rm i}(t)}\,{\rm d}H_{\rm e}(t)\right] + V_{\rm i}\int_{-\infty}^0 e^{\frac{t}{\tau}}\mathbb{E}\left[e^{H_{\rm e}(t)+H_{\rm i}(t)}\,{\rm d}H_{\rm i}(t)\right] + \frac{I}{G}\int_{-\infty}^0 e^{\frac{t}{\tau}}\mathbb{E}\left[e^{H_{\rm e}(t)+H_{\rm i}(t)}\right]\frac{{\rm d}t}{\tau}\,. \tag{F2}$$

Our primary task is to evaluate the various stationary expectations appearing in the above formula. Such a goal can be achieved analytically for AONCB models. As the involved calculations tend to be cumbersome, we only give a detailed account in Appendix. Here we account for the key steps of the calculation, which ultimately produces an interpretable compact formula for $\mathbb{E}\left[V_\epsilon\right]$ in the limit of instantaneous synapses, i.e., when $\epsilon\to 0$.

In order to establish this compact formula, it is worth introducing the stationary bivariate function

$$Q_\epsilon(t,s) = \mathbb{E}\left[e^{H_\mathrm{e}(t)+H_\mathrm{i}(s)}\right] , \tag{F3}$$

which naturally depends on ϵ via $H_\mathrm{e}(t)$ and $H_\mathrm{i}(s)$. The function Q_ϵ is of great interest because all the stationary expectations at stake in Eq. (F2) can be derived from it. Before justifying this point, an important observation is that the expectation defining $Q_\epsilon(t,s)$ only bears on the cumulative input processes H_e and H_i, which specify bounded, piecewise continuous functions with probability one, independent of ϵ. As a result of this regular behavior, the expectation commute with the limit of instantaneous synapses allowing one to write

$$Q(t,s) = \lim_{\epsilon\to 0^+} Q_\epsilon(t,s) = \mathbb{E}\left[e^{\lim_{\epsilon\to 0} H_\mathrm{e}(t)+H_\mathrm{i}(s)}\right] , = \mathbb{E}\left[e^{-Z_\mathrm{e}(t)-Z_\mathrm{i}(s)}\right] ,$$

where we exploit the fact that the cumulative input processes H_e and H_i converge toward the coupled compound Poisson processes Z_e and Z_i when $\epsilon \to 0^+$:

$$Z_\mathrm{e}(t) = \sum_n^{N_\mathrm{e}(t)} W_{\mathrm{e},n} \quad \text{and} \quad Z_\mathrm{i}(t) = \sum_n^{N_\mathrm{i}(t)} W_{\mathrm{i},n} . \tag{F4}$$

The above remark allows one to compute the term due to current injection I in Eq. (F2), where the expectation can be identified to $Q_\epsilon(t,t)$. Indeed, utilizing the standard form for the moment-generating function for compound Poisson processes [51], we find that

$$Q(t,t) = e^{a_{\mathrm{ei},1}t/\tau} ,$$

where we introduce the first-order aggregate efficacy

$$a_{\mathrm{ei},1} = b\tau\left(1 - \mathbb{E}_\mathrm{ei}\left[e^{-(W_\mathrm{e}+W_\mathrm{i})}\right]\right) .$$

Remember that in the above definition, $\mathbb{E}_\mathrm{ei}[\cdot]$ denotes the expectation with respect to the joint probability of the conductance jumps, i.e., p_ei.

It remains to evaluate the expectations associated to excitation and inhibition reversal potentials in Eq. (F2). These terms differ from the current-associated term in that they involve expectations of stochastic integrals with respect to the cumulative input processes $H_\mathrm{e/i}$. This is by contrast with evaluating Eq. (F3), which only involves expectations of functions that depends on $H_\mathrm{e/i}$. In principle, one could still hope to adopt a similar route as for the current associated term, exploiting the compound Poisson process Z obtained in the limit of instantaneous synapses. However, such an approach would require that the operations of taking the limit of instantaneous synapses and evaluating the stationary expectation still commute. This is a major caveat as such a commuting relation generally fails for point-process-based stochastic integrals. Therefore, one has to analytically evaluate the expectations at stake for positive synaptic activation time $\epsilon > 0$, without resorting to the simplifying limit of instantaneous synapses. This analytical requirement is the primary motivation to consider AONCB models.

The first step in the calculation is to realize that for $\epsilon > 0$, the conductance traces $h_\mathrm{e}(t) = \tau \mathrm{d}H_\mathrm{e}(t)/\mathrm{d}t$ and $h_\mathrm{i}(t) = \tau \mathrm{d}H_\mathrm{i}(t)/\mathrm{d}t$ are bounded, piecewise continuous functions with probability one. Under these conditions, it then holds that

$$\lim_{s\to t} \partial_t Q_\epsilon(t,s) = \mathbb{E}\left[\frac{\mathrm{d}H_\mathrm{e}(t)}{\mathrm{d}t}\, e^{H_\mathrm{e}(t)+H_\mathrm{i}(t)}\right] \quad \text{and} \quad \lim_{s\to t} \partial_s Q_\epsilon(t,s) = \mathbb{E}\left[\frac{\mathrm{d}H_\mathrm{i}(t)}{\mathrm{d}t}\, e^{H_\mathrm{e}(t)+H_\mathrm{i}(t)}\right] ,$$

so that the sought-after expectations can be deduced from the closed-form knowledge of $Q_\epsilon(t,s)$ for positive $\epsilon > 0$. The analytical expression of $Q_\epsilon(t,s)$ can be obtained via careful manipulation of the processes H_e and H_i featured in the exponent of Eq. (F3) (see Appendix H). In a nutshell, these manipulations hinge on splitting the integrals defining $H_\mathrm{e}(t)$ and $H_\mathrm{i}(s)$ into independent contributions arising from spiking events occurring in the five nonoverlapping, contiguous intervals bounded by the times $0 \geq -\epsilon\tau \geq t \geq s \geq t - \epsilon\tau \geq s - \epsilon\tau$. There is no loss of generality in assuming the latter ordering and from the corresponding analytical expression, we can compute

$$\lim_{\epsilon\to 0^+}\lim_{s\to t} \partial_t Q_\epsilon(t,s) = b a_{\mathrm{e},1} e^{a_{\mathrm{ei},1}t/\tau} \quad \text{and} \quad \lim_{\epsilon\to 0^+}\lim_{s\to t} \partial_s Q_\epsilon(t,s) = b a_{\mathrm{i},1} e^{a_{\mathrm{ei},1}t/\tau} ,$$

where the effective first-order synaptic efficacies via Eq. (15) as

$$a_{\mathrm{e},1} = b\tau\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{e}}{W_\mathrm{e}+W_\mathrm{i}}\left(1 - e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)\right] \quad \text{and} \quad a_{\mathrm{i},1} = b\tau\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{i}}{W_\mathrm{e}+W_\mathrm{i}}\left(1 - e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)\right] .$$

Observe that by definition, $a_{\mathrm{e},1}$ and $a_{\mathrm{i},1}$ satisfy $a_{\mathrm{e},1} + a_{\mathrm{i},1} = a_{\mathrm{ei},1}$.

Altogether, upon evaluation of the integrals featured in Eq. (F2), these results allow one to produce the compact expression Eq. (14) for the stationary voltage mean in the limit of instantaneous synapses:

$$\mathbb{E}\left[V\right] = \lim_{\epsilon\to 0^+} \mathbb{E}\left[V_\epsilon\right] = \frac{a_{\mathrm{e},1}V_\mathrm{e} + a_{\mathrm{i},1}V_\mathrm{i} + I/G}{1 + a_{\mathrm{e},1} + a_{\mathrm{i},1}} \,.$$

Appendix G: Stationary voltage variance

The calculation of the stationary voltage variance is more challenging than that of the stationary voltage mean. However, in the limit of instantaneous synapses, this calculation produces a compact, interpretable formula as well. Adopting a similar approach as for the stationary mean calculation, we start by expressing V_ϵ^2 in the stationary limit in terms of a stochastic integrals involving the cumulative input processes H_e and H_i. Specifically, using Eq. (F1), we have

$$\begin{aligned} V_\epsilon^2 &= \left(\int_{-\infty}^{0} e^{\frac{t}{\tau}+H_\mathrm{e}(t)+H_\mathrm{i}(t)} \left(\mathrm{d}(V_\mathrm{e}H_\mathrm{e}(t) + V_\mathrm{i}H_\mathrm{i}(t)) + \frac{I}{G}\frac{\mathrm{d}t}{\tau}\right)\right)^2 , \\ &= \iint_{\mathbb{R}_-^2} e^{\frac{t+s}{\tau}+H_\mathrm{e}(t)+H_\mathrm{i}(t)+H_\mathrm{e}(s)+H_\mathrm{i}(s)} \left(\mathrm{d}(V_\mathrm{e}H_\mathrm{e}(t) + V_\mathrm{i}H_\mathrm{i}(t)) + \frac{I}{G}\frac{\mathrm{d}t}{\tau}\right)\left(\mathrm{d}(V_\mathrm{e}H_\mathrm{e}(s) + V_\mathrm{i}H_\mathrm{i}(s)) + \frac{I}{G}\frac{ds}{\tau}\right) . \end{aligned} \quad \text{(G1)}$$

Our main goal is to compute the stationary expectation of the above quantity. As for the stationary voltage mean, our strategy is (i) to derive the exact stationary expectation of the integrands for finite synaptic activation time, (ii) to evaluate these integrands in the simplifying limit of instantaneous synapses, and (iii) to rearrange the terms obtained after integration into an interpretable final form. Enacting the above strategy is a rather tedious task, and as for the calculation of the mean voltage, we only present the key steps of the calculation in the following.

The integrand terms at stake are obtained by expanding Eq. (G1), which yields the following quadratic expression for the stationary second moment of the voltage

$$\mathbb{E}\left[V_\epsilon^2\right] = A_{\mathrm{e},\epsilon}V_\mathrm{e}^2 + B_{\mathrm{ei},\epsilon}V_\mathrm{e}V_\mathrm{i} + A_{\mathrm{i},\epsilon}V_\mathrm{i}^2 + \left(V_\mathrm{e}B_{\mathrm{e}I,\epsilon} + V_\mathrm{i}B_{\mathrm{i}I,\epsilon}\right)(I/G) + A_{I,\epsilon}(I/G)^2 \,,$$

whose various coefficients need to be evaluated. These coefficients are conveniently specified in terms of the following symmetric random function

$$\mathcal{E}_\mathrm{ei}(t,s) = e^{H_\mathrm{e}(t)+H_\mathrm{i}(t)+H_\mathrm{e}(s)+H_\mathrm{i}(s)} \,.$$

which features prominently in Eq. (G1). Moreover, drawing on the calculation of the stationary mean voltage, we anticipate that the quadrivariate version of $\mathcal{E}_\mathrm{ei}(t,s)$ will play a central role in the calculation via its stationary expectation. Owing to this central role, we denote this expectation as

$$R_\epsilon(t,u,s,v) = \mathbb{E}\left[e^{H_\mathrm{e}(t)+H_\mathrm{i}(u)+H_\mathrm{e}(s)+H_\mathrm{i}(v)}\right] .$$

where we make the ϵ-dependence explicit. As a mere expectation with respect to the cumulative input processes $(H_\mathrm{e}, H_\mathrm{i})$, the expectation can be evaluated in closed form for AONCB models. This again requires careful manipulations of the processes H_e and H_i, which need to split into independent contributions arising from spiking events occurring in nonoverlapping intervals. By contrast with the bivariate case, the quadrivariate case requires to consider nine contiguous intervals. There is no loss of generality to consider these interval bounds to be determined by the two following time orderings:

O-order: $0 \geq -\epsilon\tau \geq t \geq u \geq t-\epsilon\tau \geq u-\epsilon\tau \geq s \geq v \geq s-\epsilon\tau \geq v-\epsilon\tau$,

D-order: $0 \geq -\epsilon\tau \geq t \geq u \geq s \geq v \geq t-\epsilon\tau \geq u-\epsilon\tau \geq s-\epsilon\tau \geq v-\epsilon\tau$,

where O stands for off-diagonal ordering and D for diagonal ordering.

The reason to only consider the O/D-orders is that all the relevant calculations will be made in the limit $(u,v) \to (t,s)$. By symmetry of $R_\epsilon(t,u,s,v)$, it is then enough to restrict our consideration to the limit $(u,v) \to (t^-,s^-)$, which leaves the choice of $t,s \leq 0$ to be determined. By symmetry, one can always choose $t > s$, so that the only remaining alternative is to decide wether (t,s) belong to the diagonal region $\mathcal{D}_\epsilon = \{t,s \leq 0 \,|\, \epsilon\tau \geq |t-s|\}$ or the off-diagonal region $\mathcal{O}_\epsilon = \{t,s \leq 0 \,|\, \epsilon\tau < |t-s|\}$. For the sake of completeness, we give the two expressions of $R_\epsilon(t,u,s,v)$ on the regions $\mathcal{O}_\epsilon$ and $\mathcal{D}_\epsilon$ in Appendix I. Owing to their tediousness, we do not give the detailed

calculations leading to these expressions, which are lengthy but straightforward elaborations on those used in Appendix H. Here we stress that for $\epsilon > 0$, these expressions reveal that $R_\epsilon(t,u,s,v)$ is defined as a twice-differentiable quadrivariate function.

With these remarks in mind, the coefficients featured in Eq. (G2) can be categorized in three classes:

1. There is a single current-dependent inhomogeneous coefficient

$$A_{I,\epsilon} = \iint_{\mathbb{R}_-^2} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\right] \frac{\mathrm{d}t\, ds}{\tau^2}\,,$$

where we recognize that $\mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\right] = R_\epsilon(t,t,s,s) \stackrel{\rm def}{=} R_\epsilon(t,s)$. As $R_\epsilon(t,s)$ is merely a stationary expectation with respect to the cumulative input processes $(H_{\rm e}, H_{\rm i})$, it can be directly evaluated in the limit of instantaneous synapses. In other word, step (ii) can be performed before step (i), similarly as for the stationary voltage mean calculation. However, having a general analytical expression for $R_\epsilon(t,u,s,v)$ on $\mathcal{O}_\epsilon$ (see Appendix I), we can directly evaluate for all $t \neq s$ that

$$R(t,s) = \lim_{\epsilon\to 0^+} R_\epsilon(t,s) = e^{(2a_{\rm ei,2}\max(t,s) - a_{\rm ei,1}|t-s|)/\tau}\,, \tag{G2}$$

where we define the second-order aggregate efficacy

$$a_{\rm ei,2} = \frac{b\tau}{2}\left(1 - \mathbb{E}_{\rm ei}\left[e^{-2(W_{\rm e}+W_{\rm i})}\right]\right)\,.$$

It is clear that the continuous function $R(t,s)$ is smooth everywhere except on the diagonal where it admits a slope discontinuity. As we shall see, this slope discontinuity is the reason why one needs to consider the $\mathcal{D}_\epsilon$ region carefully, even when only concerned with the limit $\epsilon \to 0^+$. That being said, the diagonal behavior plays no role here and straightforward integration of $R(t,s)$ on the negative orthant gives

$$A_I = \lim_{\epsilon\to 0^+} A_{I,\epsilon} = \frac{1}{(1+a_{\rm ei,1})(1+a_{\rm ei,2})}\,.$$

2. There are two current-dependent linear coefficients

$$B_{{\rm e}I,\epsilon} = 2\iint_{\mathbb{R}_-^2} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\mathrm{d}H_{\rm e}(t)\right]\frac{\mathrm{d}s}{\tau} \quad\text{and}\quad B_{{\rm i}I,\epsilon} = 2\iint_{\mathbb{R}_-^2} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\mathrm{d}H_{\rm i}(t)\right]\frac{\mathrm{d}s}{\tau}\,,$$

where the coefficient 2 above comes from the fact that $B_{{\rm e}I,\epsilon}$ and $B_{{\rm i}I,\epsilon}$ are actually resulting from the contributions of two symmetric terms in the expansion of Eq. (G1). Both $B_{{\rm e}I,\epsilon}$ and $B_{{\rm i}I,\epsilon}$ involve expectations of stochastic integrals akin to those evaluated for the stationary mean calculation. Therefore, these terms can be treated similarly by implementing step (i) and (ii) sequentially. The trick is to realize that for positive ϵ and $t \neq s \leq 0$, it holds that

$$\mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\frac{\mathrm{d}H_{\rm e}(t)}{\mathrm{d}t}\right] = \lim_{u\to t}\partial_t R_\epsilon(t,u,s,s) \quad\text{and}\quad \mathbb{E}\left[\mathcal{E}_{\rm ei}(t,s)\frac{\mathrm{d}H_{\rm i}(t)}{\mathrm{d}t}\right] = \lim_{v\to s}\partial_s R_\epsilon(t,t,s,v)\,.$$

Thus for any (t,s) in the off-diagonal region O_ϵ, the analytical knowledge of $R_\epsilon(t,u,s,v)$ (see Appendix I) allows one to evaluate

$$\lim_{u\to t^-}\tau\frac{\partial_t R_\epsilon(t,u,s,s)}{R_\epsilon(t,s)} = \begin{cases} a_{\rm e,1} & \text{if } t > s\,,\\ a_{\rm e,2} - a_{\rm e,1} & \text{if } t < s\,,\end{cases} \quad\text{and}\quad \lim_{v\to s^-}\tau\frac{\partial_s R_\epsilon(t,u,s,s)}{R_\epsilon(t,s)} = \begin{cases} a_{\rm i,1} & \text{if } t > s\,,\\ a_{\rm i,2} - a_{\rm i,1} & \text{if } t < s\,,\end{cases} \tag{G3}$$

where the second-order synaptic efficacies are defined as

$$a_{\rm e,2} = \frac{b\tau}{2}\mathbb{E}_{\rm ei}\left[\frac{W_{\rm e}}{W_{\rm e}+W_{\rm i}}\left(1 - e^{-2(W_{\rm e}+W_{\rm i})}\right)\right] \quad\text{and}\quad a_{\rm i,2} = \frac{b\tau}{2}\mathbb{E}_{\rm ei}\left[\frac{W_{\rm i}}{W_{\rm e}+W_{\rm i}}\left(1 - e^{-2(W_{\rm e}+W_{\rm i})}\right)\right]\,. \tag{G4}$$

Observe that these efficacies satisfy the familiar relation $a_{\rm e,2} + a_{\rm i,2} = a_{\rm ei,2}$. Taking the limits of Eq. (G3) when $\epsilon \to 0^+$ specify two bivariate functions that are continuous everywhere, except on the diagonal $t = s$, where these functions present a jump discontinuity. This behavior is still regular enough to discard any potential contributions from diagonal terms, so that we can restrict ourselves to the region O_ϵ. Then, taking the limit $\epsilon \to 0^+$ after integration of over O_ϵ, we find that

$$B_{{\rm e}I} = \lim_{\epsilon\to 0^+} B_{{\rm e}I,\epsilon} = \frac{a_{\rm e,2}}{(1+a_{\rm ei,1})(1+a_{\rm ei,2})} \quad\text{and}\quad B_{{\rm i}I} = \lim_{\epsilon\to 0^+} B_{{\rm i}I,\epsilon} = \frac{b\tau a_{\rm i,2}}{(1+a_{\rm ei,1})(1+a_{\rm ei,2})}\,.$$

3. There are four quadratic coefficients associated to the reversal-potential V_e and V_i, including two diagonal terms

$$A_{\mathrm{e},\epsilon} = \iint_{\mathbb{R}^2_-} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{e}(s)\right] \quad \text{and} \quad A_{\mathrm{i},\epsilon} = \iint_{\mathbb{R}^2_-} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{i}(t)\mathrm{d}H_\mathrm{i}(s)\right] ,$$

and two symmetric cross terms contributing

$$B_{\mathrm{ei},\epsilon} = 2\iint_{\mathbb{R}^2_-} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{i}(s)\right] .$$

Notice that it is enough to compute only one diagonal term as the other term can be deduced by symmetry. Following the same method as for the linear terms, we start by remarking that for all (t,s) in the off-diagonal region $\mathcal{O}_\epsilon$, it holds that

$$\mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\,\frac{\mathrm{d}H_\mathrm{e}(t)}{\mathrm{d}t}\frac{\mathrm{d}H_\mathrm{e}(s)}{\mathrm{d}s}\right] = \lim_{(u,v)\to(t,s)} \partial_t\partial_s R_\epsilon(t,u,s,v) ,$$
$$\mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\,\frac{\mathrm{d}H_\mathrm{e}(t)}{\mathrm{d}t}\frac{\mathrm{d}H_\mathrm{i}(s)}{\mathrm{d}s}\right] = \lim_{(u,v)\to(t,s)} \partial_t\partial_v R_\epsilon(t,u,s,v) ,$$

As before, the analytical knowledge of $R_\epsilon(t,u,s,v)$ on the O_ϵ region (see Appendix I) allows one to evaluate

$$\lim_{(u,v)\to(t,s)^-} \tau^2 \frac{\partial_t\partial_u R_\epsilon(t,u,s,s)}{R_\epsilon(t,s)} = a_{\mathrm{e},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1}) ,$$
$$\lim_{(u,v)\to(t,s)^-} \tau^2 \frac{\partial_t\partial_s R_\epsilon(t,u,s,v)}{R_\epsilon(t,s)} = \frac{1}{2}(a_{\mathrm{e},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + a_{\mathrm{i},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})) .$$

The above closed-form expressions allow one to compute $A'_{\mathrm{e},\epsilon}$ and $B'_{\mathrm{ei},\epsilon}$, the part of the coefficients $A_{\mathrm{e},\epsilon}$ and $B_{\mathrm{ei},\epsilon}$ resulting from integration over the off-diagonal region O_ϵ, which admit well-defined limit values $A'_\mathrm{e} = \lim_{\epsilon\to0^+} A'_{\mathrm{e},\epsilon}$ and $B'_\mathrm{ei} = \lim_{\epsilon\to0^+} B'_{\mathrm{ei},\epsilon}$ with:

$$A'_\mathrm{e} = \lim_{\epsilon\to0^+} \iint_{\mathcal{O}_\epsilon} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{e}(s)\right] = \frac{a_{\mathrm{e},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+b\tau a_{\mathrm{ei},2})} ,$$
$$B'_\mathrm{ei} = 2\lim_{\epsilon\to0^+} \iint_{\mathcal{O}_\epsilon} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{i}(s)\right] = \frac{a_{\mathrm{e},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + a_{\mathrm{i},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})} .$$

However, for quadratic terms, one also needs to include the contributions arising from the diagonal region $\mathcal{D}_\epsilon$, as suggested be the first-order jump discontinuity of $R(t,s) = \lim_{\epsilon\to0^+} R_\epsilon(t,s)$ on the diagonal $t=s$. To confirm this point, one can show from the analytical expression of $R_\epsilon(t,u,s,v)$ on $\mathcal{D}_\epsilon$ (see Appendix I), that all relevant second-order derivative terms scale as $1/\epsilon$ over $\mathcal{D}_\epsilon$. This scaling leads to the nonzero contributions $A''_{\mathrm{e},\epsilon}$ and $B''_{\mathrm{ei},\epsilon}$ resulting form the integration of these second-order derivative terms over the diagonal region $\mathcal{D}_\epsilon$, even in the limit $\epsilon\to0^+$. Actually, we find that these contributions also admit well-defined limit values $A''_\mathrm{e} = \lim_{\epsilon\to0^+} A''_{\mathrm{e},\epsilon}$ and $B''_\mathrm{ei} = \lim_{\epsilon\to0^+} B''_{\mathrm{ei},\epsilon}$ with: (see Appendix J)

$$A''_\mathrm{e} = \lim_{\epsilon\to0^+} \iint_{\mathcal{D}_\epsilon} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{e}(s)\right] = \frac{a_{\mathrm{e},12} - c_\mathrm{ei}}{1+a_{\mathrm{ei},2}} , \tag{G5}$$
$$B''_\mathrm{ei} = 2\lim_{\epsilon\to0^+} \iint_{\mathcal{D}_\epsilon} e^{\frac{t+s}{\tau}} \mathbb{E}\left[\mathcal{E}_\mathrm{ei}(t,s)\mathrm{d}H_\mathrm{e}(t)\mathrm{d}H_\mathrm{i}(s)\right] = \frac{2c_\mathrm{ei}}{1+a_{\mathrm{ei},2}} . \tag{G6}$$

Remembering that the expression of A''_i can be deduced from that of A''_e by symmetry, Eq. (G5) defines A''_e, and thus A''_i, in terms of the useful auxiliary second-order efficacies $a_{\mathrm{e},12} = a_{\mathrm{e},1} - a_{\mathrm{e},2}$ and $a_{\mathrm{i},12} = a_{\mathrm{i},1} - a_{\mathrm{i},2}$. These efficacies will feature prominently in the final variance expression and it is worth mentioning their explicit definitions as

$$a_{\mathrm{e},12} = \frac{b\tau}{2}\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{e}}{W_\mathrm{e}+W_\mathrm{i}}\left(1-e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)^2\right] \quad \text{and} \quad a_{\mathrm{i},12} = \frac{b\tau}{2}\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{i}}{W_\mathrm{e}+W_\mathrm{i}}\left(1-e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)^2\right] . \tag{G7}$$

The other quantity of interest is the coefficient c_ei, which appears in both Eq. (G5) and Eq. (G6). This nonnegative coefficient, defined as

$$c_\mathrm{ei} = \frac{b\tau}{2}\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{e}W_\mathrm{i}}{(W_\mathrm{e}+W_\mathrm{i})^2}\left(1-e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)^2\right] , \tag{G8}$$

entirely captures the (nonnegative) correlation between excitatory and inhibitory inputs and shall be seen as an efficacy as well. Keeping these definitions in mind, the full quadratic coefficients are finally obtained as $A_{\rm e} = A'_{\rm e} + A''_{\rm e}$, $A_{\rm i} = A'_{\rm i} + A''_{\rm i}$, and $B_{\rm ei} = B'_{\rm ei} + B''_{\rm ei}$.

From there, injecting the analytical expressions of the various coefficients in the quadratic form Eq. (G2) leads to an explicit formula for the stationary voltage variance in the limit of instantaneous synapses. Then, one is only left with step (iii), which aims at exhibiting a compact, interpretable form for this formula. We show in Appendix K that lengthy but straightforward algebraic manipulations lead to the simplified form given in Eq. (16):

$$\mathbb{V}\left[V\right] = \lim_{\epsilon\to 0^+} \mathbb{V}\left[V_\epsilon\right] == \frac{1}{1+a_{\rm ei,2}} \left(a_{\rm e,12}(V_{\rm e}-\mathbb{E}\left[V\right])^2 + a_{\rm i,12}(V_{\rm i}-\mathbb{E}\left[V\right])^2 - c_{\rm ei}(V_{\rm e}-V_{\rm i})^2\right) .$$

Appendix H: Evaluation of $Q_\epsilon(t,s)$ for $\epsilon > 0$

The goal here is to justify the closed-form expression of $Q_\epsilon(t,s) = \mathbb{E}\left[e^{H_{\rm e}(t)+H_{\rm i}(s)}\right]$ via standard manipulation of exponential functionals of Poisson processes. By definition, assuming with no loss of generality the order $0 \geq t \geq s$, we have

$$\begin{aligned} H_{\rm e}(t) + H_{\rm i}(s) &= -\frac{1}{\tau}\left(\int_t^0 h_{\rm e}(u)\,{\rm d}u + \int_s^0 h_{\rm i}(u)\,{\rm d}u\right) , \\ &= -\frac{1}{\epsilon\tau}\left(\int_t^0 {\rm d}u \sum_{N(u-\epsilon\tau)+1}^{N(u)} W_{{\rm e},k} + \int_s^0 {\rm d}u \sum_{N(u-\epsilon\tau)+1}^{N(u)} W_{{\rm i},k}\right) , \\ &= -\frac{1}{\epsilon\tau}\left(\int_t^0 {\rm d}u \sum_{N(u-\epsilon\tau)+1}^{N(u)} (W_{{\rm e},k} + W_{{\rm i},k}) + \int_s^t {\rm d}u \sum_{N(u-\epsilon\tau)+1}^{N(u)} W_{{\rm i},k}\right) . \end{aligned} \tag{H1}$$

We will evaluate $Q_\epsilon(t,s) = \mathbb{E}\left[e^{H_{\rm e}(t)+H_{\rm i}(s)}\right]$ as a product of independent integral contributions.

Isolating these independent contributions from Eq. (H1) requires to establish two preliminary results about the quantity

$$I(t,s) = \int_s^t \sum_{k=N(u-\Delta)+1}^{N(u)} X_k\,{\rm d}u\,, \tag{H2}$$

where N denotes a Poisson process, X_k denotes i.i.d. nonnegative random variables, and where Δ is positive activation time. Assume $t-s \geq \Delta$, then given some real $w < u-\Delta$, we have

$$\begin{aligned} I(t,s) &= \int_s^t {\rm d}u \sum_{k=N(v)+1}^{N(u)} X_k - \int_s^t {\rm d}u \sum_{k=N(v)+1}^{N(u-\Delta)} X_k\,, \\ &= \int_s^t {\rm d}u \sum_{k=N(v)+1}^{N(u)} X_k - \int_{s-\Delta}^{t-\Delta} {\rm d}u \sum_{k=N(v)+1}^{N(u)} X_k\,, \\ &= \int_{t-\Delta}^t {\rm d}u \sum_{k=N(v)+1}^{N(u)} X_k - \int_{s-\Delta}^{s} {\rm d}u \sum_{k=N(v)+1}^{N(u)} X_k\,, \\ &= \left(\int_{t-\Delta}^t {\rm d}u \sum_{k=N(v)+1}^{N(t-\Delta)} X_k + \int_{t-\Delta}^t {\rm d}u \sum_{k=N(t-\Delta)+1}^{N(u)} X_k\right) - \left(\int_{s-\Delta}^s {\rm d}u \sum_{k=N(v)+1}^{N(s)} X_k - \int_{s-\Delta}^s {\rm d}u \sum_{k=N(u)+1}^{N(s)} X_k\right) , \\ &= \int_{t-\Delta}^t {\rm d}u \sum_{k=N(t-\Delta)+1}^{N(u)} X_k + \Delta \sum_{k=N(s)+1}^{N(t-\Delta)} X_k + \int_{s-\Delta}^s {\rm d}u \sum_{k=N(u)+1}^{N(s)} X_k\,. \end{aligned} \tag{H3}$$

One can check that the three terms in Eq. (H3) above are independent for involving independent numbers of i.i.d. draws over the intervals $(t-\Delta, t]$, $(s, t-\Delta]$, and $(s-\Delta, s]$, respectively. Similar manipulations for the order for

$t-s \leq \Delta$ yields

$$I(t,s) = \int_s^t \mathrm{d}u \sum_{k=N(s)+1}^{N(u)} X_k + (t-s) \sum_{k=N(t-\Delta)+1}^{N(s)} X_k + \int_{s-\Delta}^{t-\Delta} \mathrm{d}u \sum_{k=N(u)+1}^{N(t-\Delta)} X_k \,, \tag{H4}$$

where that three independent contributions corresponds to independent numbers of i.i.d. draws over the intervals $(s,t]$, $(t-\Delta,s]$, and $(s-\Delta,t-\Delta]$, respectively.

As evaluating Q_ϵ only involves taking the limit $s \to t^-$ at fixed $\epsilon > 0$, it is enough to consider the order $0 \geq -\epsilon\tau \geq t \geq s \geq t-\epsilon\tau$ With that in mind, we can apply Eq. (H3) Eq. (H4) with $\Delta = \epsilon\tau$ and $X_k = W_{\mathrm{e},k} + W_{\mathrm{i},k}$ or $X_k = W_{\mathrm{i},k}$, to decompose the two terms of Eq. (H1) in six contributions

$$\begin{aligned} I(t,s) = &\int_{-\epsilon\tau}^{0} \mathrm{d}u \sum_{k=N(t-\epsilon\tau)+1}^{N(u)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) + \epsilon\tau \sum_{k=N(t)+1}^{N(-\epsilon\tau)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) + \int_{t-\epsilon\tau}^{t} \mathrm{d}u \sum_{k=N(u)+1}^{N(t)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) \\ &+ \int_s^t \mathrm{d}u \sum_{k=N(s)+1}^{N(u)} W_{\mathrm{i},k} + (t-s) \sum_{k=N(t-\epsilon\tau)+1}^{N(s)} W_{\mathrm{i},k} + \int_{s-\epsilon\tau}^{t-\epsilon\tau} \mathrm{d}u \sum_{k=N(u)+1}^{N(t-\epsilon\tau)} W_{\mathrm{i},k} \,. \end{aligned}$$

It turns out that the contribution of the third term overlaps with that of the fourth and fifth terms. Further splitting of that third term produces the following expression

$$\begin{aligned} I(t,s) = &\underbrace{\int_{-\epsilon\tau}^{0} \mathrm{d}u \sum_{k=N(t-\epsilon\tau)+1}^{N(u)} (W_{\mathrm{e},k} + W_{\mathrm{i},k})}_{I_1} + \underbrace{\epsilon\tau \sum_{k=N(t)+1}^{N(-\epsilon\tau)} (W_{\mathrm{e},k} + W_{\mathrm{i},k})}_{I_2(t)} \\ &+ \underbrace{\int_s^t \mathrm{d}u \left(\sum_{k=N(u)+1}^{N(t)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) + \sum_{k=N(s)+1}^{N(u)} W_{\mathrm{i},k} \right) + (s-t+\epsilon\tau) \sum_{k=N(s)+1}^{N(t)} (W_{\mathrm{e},k} + W_{\mathrm{i},k})}_{I_3(t,s)} \\ &+ \underbrace{\left(\int_{t-\epsilon\tau}^{s} \mathrm{d}u \sum_{k=N(u)+1}^{N(s)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) + (t-s) \sum_{k=N(t-\epsilon\tau)+1}^{N(s)} W_{\mathrm{i},k} \right)}_{I_4(s,t)} + \underbrace{\int_{s-\epsilon\tau}^{t-\epsilon\tau} \mathrm{d}u \sum_{k=N(u)+1}^{N(t-\epsilon\tau)} W_{\mathrm{i},k}}_{I_5(t,s)} \,, \end{aligned}$$

where all five terms correspond to independent numbers of i.i.d. draws over the intervals $(-\epsilon\tau,0]$, $(t,-\epsilon\tau]$, $(s,t]$, $(t-\epsilon\tau,s]$, and $(s-\epsilon\tau,t-\epsilon\tau]$. Then, we have

$$Q_\epsilon(t,s) = \mathbb{E}\left[e^{H_\mathrm{e}(t)+H_\mathrm{i}(s)}\right] = \mathbb{E}\left[e^{-I_1/(\epsilon\tau)}\right] \mathbb{E}\left[e^{-I_2(t)/(\epsilon\tau)}\right] \mathbb{E}\left[e^{-I_3(t,s)/(\epsilon\tau)}\right] \mathbb{E}\left[e^{-I_4(s,t)/(\epsilon\tau)}\right] \mathbb{E}\left[e^{-I_5(t,s)/(\epsilon\tau)}\right] \,,$$

where all expectation terms can be computed via standard manipulation of the moment-generating function of Poisson processes [51]. The trick is to remember that for all $t \geq s$, given that a Poisson process admits $K = N(t) - N(s)$ points in $(s,t]$, all these K points are uniformly i.i.d. over $(s,t]$. This trick allows one to simply represent all integral terms in terms of uniform random variables, whose expectations are easily computable. To see this, let us consider $I_3(t,s)$ for instance. We have

$$\begin{aligned} I_3(t,s) &= (t-s) \sum_{k=N(s)+1}^{N(t)} \left[(1-U_k)(W_{\mathrm{e},k} + W_{\mathrm{i},k}) + U_k W_{\mathrm{i},k}\right] + (s-t+\epsilon\tau) \sum_{k=N(s)+1}^{N(t)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) \,, \\ &= (t-s) \sum_{k=N(s)+1}^{N(t)} U_k W_{\mathrm{e},k} + \epsilon\tau \sum_{k=N(s)+1}^{N(t)} (W_{\mathrm{e},k} + W_{\mathrm{i},k}) \,, \end{aligned}$$

where $\{U_k\}_{N(s)+1 \leq k \leq N(t)}$ are uniformly i.i.d. on $[0,1]$. From the knowledge of the moment-generating function of

Poisson random variables [51], one can evaluate

$$\begin{aligned}\mathbb{E}\left[e^{-I_3(t,s)/(\epsilon\tau)}\right] &= \mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}\sum_{k=N(s)+1}^{N(t)} U_k W_{\mathrm{e},k} - \sum_{k=N(s)+1}^{N(t)}(W_{\mathrm{e},k}+W_{\mathrm{i},k})}\right] , \\ &= \mathbb{E}\left[\mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}UW_{\mathrm{e}}-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right]^{N(t)-N(s)} \,\middle|\, N(t)-N(s)\right] , \\ &= \exp\left(b(t-s)\left(\mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}UW_{\mathrm{e}}-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right]-1\right)\right) ,\end{aligned}$$

where $(W_{\mathrm{e}}, W_{\mathrm{i}})$ denotes exemplary conductance jumps and U denotes an independent uniform random variable. Furthermore we have

$$\begin{aligned}\mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}UW_{\mathrm{e}}-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}UW_{\mathrm{e}}-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right] \,\middle|\, W_{\mathrm{e}}, W_{\mathrm{i}}\right] , \\ &= \mathbb{E}_{\mathrm{ei}}\left[e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\mathbb{E}\left[e^{-\frac{t-s}{\epsilon\tau}UW_{\mathrm{e}}}\right]\right] , \\ &= \mathbb{E}_{\mathrm{ei}}\left[e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\frac{\left(1-e^{-\frac{t-s}{\epsilon\tau}W_{\mathrm{e}}}\right)}{\frac{t-s}{\epsilon\tau}W_{\mathrm{e}}}\right] ,\end{aligned}$$

so that we finally obtain

$$\ln\mathbb{E}\left[e^{-I_3(t,s)/(\epsilon\tau)}\right] = \epsilon b\tau\left(\mathbb{E}_{\mathrm{ei}}\left[e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\frac{\left(1-e^{-\frac{t-s}{\epsilon\tau}W_{\mathrm{e}}}\right)}{W_{\mathrm{e}}}\right]-\frac{t-s}{\epsilon\tau}\right) .$$

Similar calculations show that we have

$$\ln\mathbb{E}\left[e^{-I_1/(\epsilon\tau)}\right] = \epsilon b\tau\left(\mathbb{E}_{\mathrm{ei}}\left[\frac{1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\right]-1\right) ,$$

$$\ln\mathbb{E}\left[e^{-I_2(t)/(\epsilon\tau)}\right] = b(\epsilon\tau+t)\left(1-\mathbb{E}_{\mathrm{ei}}\left[e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right]\right) ,$$

$$\ln\mathbb{E}\left[e^{-I_4(s,t)/(\epsilon\tau)}\right] = \epsilon b\tau\left(\mathbb{E}_{\mathrm{ei}}\left[e^{-\frac{t-s}{\epsilon\tau}W_{\mathrm{i}}}\frac{\left(1-e^{-\left(1+\frac{s-t}{\epsilon\tau}\right)(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)}{W_{\mathrm{e}}+W_{\mathrm{i}}}\right]-\left(1+\frac{s-t}{\epsilon\tau}\right)\right) ,$$

$$\ln\mathbb{E}\left[e^{-I_5(t,s)/(\epsilon\tau)}\right] = \epsilon b\tau\left(\mathbb{E}_{\mathrm{ei}}\left[\frac{1-e^{-\frac{t-s}{\epsilon\tau}W_{\mathrm{i}}}}{W_{\mathrm{i}}}\right]-\frac{t-s}{\epsilon\tau}\right) .$$

Appendix I: Expression of $R_\epsilon(t,u,s,v)$ on $\mathcal{O}_\epsilon$ and $\mathcal{D}_\epsilon$

Using the similar calculations as in Appendix H, we can evaluate the quadrivariate expectation $R_\epsilon(t,u,s,v)$ on the region $\mathcal{O}_\epsilon$, for which the O-order holds: $0 \geq -\epsilon\tau \geq t \geq u \geq t-\epsilon\tau \geq u-\epsilon\tau \geq s \geq v \geq s-\epsilon\tau \geq v-\epsilon\tau$. This requires to isolate consider 9 independent contributions, corresponding to the 9 contiguous intervals specified by the O-order. We find

$$\ln R_\epsilon(t,u,s,v) = A_1 + A_2(t) + A_3(t,u) + A_4(u,t) + A_5(t,u) + A_6(u,s) + A_7(s,v) + A_8(v,s) + A_9(s,v) ,$$

where the nonnegative terms making up the above sum are defined as

$$A_1 = \epsilon b\tau\left(\mathbb{E}_{\mathrm{ei}}\left[\frac{1-e^{-2(W_{\mathrm{e}}+W_{\mathrm{i}})}}{2(W_{\mathrm{e}}+W_{\mathrm{i}})}\right]-1\right) ,$$

$$A_2(t) = b(\epsilon\tau+t)\left(1-\mathbb{E}_{\mathrm{ei}}\left[e^{-2(W_{\mathrm{e}}+W_{\mathrm{i}})}\right]\right) ,$$

$$A_3(t,u) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-2(W_{\rm e}+W_{\rm i})} \frac{\left(1 - e^{-\frac{t-u}{\epsilon\tau}W_{\rm e}}\right)}{W_{\rm e}} \right] - \frac{t-u}{\epsilon\tau} \right) ,$$

$$A_4(u,t) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-W_{\rm e}-\left(1+\frac{t-u}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1 - e^{-\left(1+\frac{u-t}{\epsilon\tau}\right)(W_{\rm e}+W_{\rm i})}\right)}{W_{\rm e}+W_{\rm i}} \right] - \left(1 + \frac{u-t}{\epsilon\tau}\right) \right) ,$$

$$A_5(t,u) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-(W_{\rm e}+W_{\rm i})} \frac{\left(1 - e^{-\frac{t-u}{\epsilon\tau}W_{\rm i}}\right)}{W_{\rm i}} \right] - \frac{t-u}{\epsilon\tau} \right) ,$$

$$A_6(u,s) = b(s+\epsilon\tau-u) \left(1 - \mathbb{E}_{\rm ei}\left[e^{-(W_{\rm e}+W_{\rm i})}\right]\right)$$

$$A_7(s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-(W_{\rm e}+W_{\rm i})} \frac{\left(1 - e^{-\frac{s-v}{\epsilon\tau}W_{\rm e}}\right)}{W_{\rm e}} \right] - \frac{s-v}{\epsilon\tau} \right) ,$$

$$A_8(v,s) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-\frac{s-v}{\epsilon\tau}W_{\rm i}} \frac{\left(1 - e^{-\left(1-\frac{s-v}{\epsilon\tau}\right)(W_{\rm e}+W_{\rm i})}\right)}{W_{\rm e}+W_{\rm i}} \right] - \left(1 - \frac{s-v}{\epsilon\tau}\right) \right) ,$$

$$A_9(s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[\frac{\left(1 - e^{-\frac{s-v}{\epsilon\tau}W_{\rm i}}\right)}{W_{\rm i}} \right] - \frac{s-v}{\epsilon\tau} \right) .$$

One can check that $A_3(t,t) = A_5(t,t) = 0$ and $A_7(s,s) = A_9(s,s) = 0$ and that A_1, $A_4(u,t)$, and $A_8(v,s)$ are all uniformly $O(\epsilon)$ on the region $\mathcal{O}_\epsilon$. This implies that for all (t,s) in $\mathcal{O}_\epsilon$, we have

$$R(t,s) = \lim_{\epsilon\to 0^+} R_\epsilon(t,t,s,s) = \lim_{\epsilon\to 0^+} e^{A_2(t)+A_6(t,s)} = e^{2bta_{\rm ei,2}-b|t-s|a_{\rm ei,1}} .$$

Using the similar calculations as in Appendix H, we can evaluate the quadrivariate expectation $R_\epsilon(t,u,s,v)$ on the region $\mathcal{D}_\epsilon$, for which the D-order holds: $0 \geq -\epsilon\tau \geq t \geq u \geq s \geq v \geq t-\epsilon\tau \geq u-\epsilon\tau \geq s-\epsilon\tau \geq v-\epsilon\tau$. This requires to isolate consider 9 independent contributions, corresponding to the 9 contiguous intervals specified by the O-order. We find

$$\begin{aligned} &\ln R_\epsilon(t,u,s,v) = \\ &B_1 + B_2(t) + B_3(t,u) + B_4(t,u,s) + B_5(t,u,s,v) + B_6(t,u,s,v) + B_7(t,u,s,v) + B_8(u,s,v) + B_9(s,v) \end{aligned} \tag{I1}$$

where the nonnegative terms making up the above sum are defined as

$$B_1 = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[\frac{1 - e^{-2(W_{\rm e}+W_{\rm i})}}{2(W_{\rm e}+W_{\rm i})} \right] - 1 \right) ,$$

$$B_2(t) = b(\epsilon\tau + t) \left(1 - \mathbb{E}_{\rm ei}\left[e^{-2(W_{\rm e}+W_{\rm i})}\right]\right) ,$$

$$B_3(t,u) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-2(W_{\rm e}+W_{\rm i})} \frac{\left(1 - e^{-\frac{t-u}{\epsilon\tau}W_{\rm e}}\right)}{W_{\rm e}} \right] - \frac{t-u}{\epsilon\tau} \right) ,$$

$$B_4(t,u,s) = \epsilon b\tau \left(\mathbb{E}_{\rm ei} \left[e^{-\left(2-\frac{t-s}{\epsilon\tau}\right)W_{\rm e}-\left(2-\frac{u-s}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1 - e^{-\frac{u-s}{\epsilon\tau}(W_{\rm e}+W_{\rm i})}\right)}{W_{\rm e}+W_{\rm i}} \right] - \frac{u-s}{\epsilon\tau} \right) ,$$

$$B_5(t,u,s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei}\left[e^{-\left(2-\frac{t-v}{\epsilon\tau}\right)W_{\rm e}-\left(2-\frac{u-v}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1-e^{-\frac{s-v}{\epsilon\tau}(2W_{\rm e}+W_{\rm i})}\right)}{2W_{\rm e}+W_{\rm i}} \right] - \frac{s-v}{\epsilon\tau} \right),$$

$$B_6(t,u,s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei}\left[e^{-\left(\frac{t-s}{\epsilon\tau}\right)W_{\rm e}-\left(\frac{2t-(u+v)}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1-e^{-\left(1-\frac{t-v}{\epsilon\tau}\right)2(W_{\rm e}+W_{\rm i})}\right)}{2(W_{\rm e}+W_{\rm i})} \right] - \left(1-\frac{t-v}{\epsilon\tau}\right) \right),$$

$$B_7(t,u,s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei}\left[e^{-\left(\frac{u-s}{\epsilon\tau}\right)W_{\rm e}-\left(\frac{u-v}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1-e^{-\frac{t-u}{\epsilon\tau}(W_{\rm e}+2W_{\rm i})}\right)}{W_{\rm e}+2W_{\rm i}} \right] - \frac{t-u}{\epsilon\tau} \right),$$

$$B_8(u,s,v) = \epsilon b\tau \left(\mathbb{E}_{\rm ei}\left[e^{-\left(\frac{s-v}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1-e^{-\frac{u-s}{\epsilon\tau}(W_{\rm e}+W_{\rm i})}\right)}{W_{\rm e}+W_{\rm i}} \right] - \frac{u-s}{\epsilon\tau} \right),$$

$$B_9(s,v) = b\epsilon\tau \left(\mathbb{E}_{\rm ei}\left[e^{-\left(\frac{s-v}{\epsilon\tau}\right)W_{\rm i}} \frac{\left(1-e^{-\frac{s-v}{\epsilon\tau}W_{\rm i}}\right)}{W_{\rm i}} \right] - \frac{s-v}{\epsilon\tau} \right).$$

Observe that $B_1 = A_1$ and $B_2(t) = A_2(t)$ and that $B_3(t,t) = B_7(t,t,s,v) = 0$ and $B_5(t,u,s,s) = B_9(s,s) = 0$. Moreover, one can see that $R(t,s)$ is continuous over the whole negative orthant by checking that:

$$\lim_{s\to(t-\epsilon\tau)^-} B_4(t,t,s) = \lim_{s\to(t-\epsilon\tau)^+} A_4(t,s)\,,$$
$$\lim_{s\to(t-\epsilon\tau)^-} B_6(t,t,s,s) = \lim_{s\to(t-\epsilon\tau)^+} A_6(t,s)\,,$$
$$\lim_{s\to(t-\epsilon\tau)^-} B_8(t,s,s) = \lim_{s\to(t-\epsilon\tau)^+} A_8(t,s)\,.$$

Actually, by computing the appropriate limit values of the relevant first- and second-order derivatives of $R_\epsilon(t,u,s,v)$, one can check that for $\epsilon > 0$, all the integrands involved in specifying the coefficients of the quadratic form Eq. (G2) define continuous functions.

Appendix J: Integrals of the quadratic terms on $\mathcal{D}_\epsilon$

Here, we only treat the quadratic term $A_{\rm e}$ as the other quadratic terms $A_{\rm i}$ and $B_{\rm ei}$ involve a similar treatment. The goal is to compute $A''_{\rm e}$, which is defined as the contribution to $A_{\rm e}$ resulting from integrating $\lim_{(u,v)\to(t,s)^-} \partial_t\partial_s R_\epsilon(t,u,s,v)$ over the diagonal region $\mathcal{D}_\epsilon = \{t,s \le 0 \,|\, \tau\epsilon \ge |t-s|\}$, in the limit $\epsilon \to 0^+$. To this end we first remark that

$$\frac{\partial_t\partial_s R_\epsilon(t,u,s,v)}{R_\epsilon(t,u,s,v)} = \partial_t\partial_s \ln R_\epsilon(t,u,s,v) + \big(\partial_t \ln R_\epsilon(t,u,s,v)\big)\big(\partial_s \ln R_\epsilon(t,u,s,v)\big)\,.$$

Injecting the analytical expression Eq. (I1) into the above relation and evaluating $I_\epsilon(t,s) = \lim_{(u,v\to(t,s)^-)} \partial_t\partial_s R_\epsilon(t,u,s,v)$ reveals that $I_\epsilon(t,s)$ scales as $1/\epsilon$, so that one expects that

$$A''_{\rm e} = \lim_{\epsilon\to 0^+} \iint_{\mathcal{D}_\epsilon} e^{\frac{t+s}{\tau}} I_\epsilon(t,s)\,{\rm d}t{\rm d}s > 0\,.$$

To compute the exact value of $A''_{\rm e}$, we perform the change of variable $x = (t-s)/(\epsilon\tau) \Leftrightarrow s = t - \epsilon\tau x$ to write

$$\iint_{\mathcal{D}_\epsilon} e^{\frac{t+s}{\tau}} I_\epsilon(t,s)\,{\rm d}t{\rm d}s = 2\int_{-\infty}^0 \left(\int_0^1 \epsilon\tau e^{-\epsilon x} I_\epsilon(t,t+\epsilon\tau x)\,{\rm d}x\right) e^{\frac{2t}{\tau}}\,{\rm d}t\,,$$

where the function $\epsilon e^{-\frac{\epsilon x}{\tau}} I_\epsilon(t,t+\epsilon x)$ remains of order one on $\mathcal{D}_\epsilon$ in the limit of instantaneous synapses. Actually, one can compute that

$$\lim_{\epsilon\to 0^+} \epsilon e^{-\epsilon x} I_\epsilon(t,t+\epsilon\tau x) = \frac{b}{2\tau}\mathbb{E}_{\rm ei}\left[\frac{W_{\rm e}^2}{W_{\rm e}+W_{\rm i}} e^{-x(W_{\rm e}+W_{\rm i})}\left(1-e^{-2(1-x)(W_{\rm e}+W_{\rm i})}\right)\right] e^{2bta_{\rm ei,2}}\,.$$

Then, for dealing with positive, continuous, uniformly bounded functions, one can safely exchange the integral and limit operations to get

$$\begin{aligned}
A_{\mathrm{e}}'' &= 2\int_{-\infty}^{0}\left(\int_0^1 \lim_{\epsilon\to 0^+} \epsilon\tau e^{-\epsilon x} I_\epsilon(t, t+\epsilon\tau x)\,\mathrm{d}x\right) e^{\frac{2t}{\tau}}\,\mathrm{d}t\,,\\
&= \left(\int_{-\infty}^{0} e^{2\frac{t}{\tau}(1+a_{\mathrm{ei},2})}\,\mathrm{d}t\right)\left(\int_0^1 b\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}^2}{W_{\mathrm{e}}+W_{\mathrm{i}}} e^{-\frac{x}{\tau}(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-2\left(1-\frac{x}{\tau}\right)(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right]\mathrm{d}x\right)\,,\\
&= \frac{b\tau}{2(1+a_{\mathrm{ei},2})}\,\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}^2}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\right)\right]\,.
\end{aligned}$$

A similar calculation for the quadratic cross term B_{ei}'' yields

$$B_{\mathrm{ei}}'' = \frac{2c_{\mathrm{ei}}}{1+a_{\mathrm{ei},2}} \quad \text{with} \quad c_{\mathrm{ei}} = \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}W_{\mathrm{i}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\right)\right]\,.$$

In order to express A_{e}'' in term of c_{ei}, we need to introduce the quantity $a_{\mathrm{e},12} = a_{\mathrm{e},1} - a_{\mathrm{e},2}$ which satisfies

$$\begin{aligned}
a_{\mathrm{e},12} &= b\tau\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right] - \frac{1}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,,\\
&= b\tau\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\left(1-\frac{1}{2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right)\right]\,,\\
&= b\tau\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\left(\frac{\left(1+e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)}{2}\right)\right]\,,\\
&= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1+e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,,
\end{aligned}$$

With the above observation, we remark that

$$\begin{aligned}
(1+a_{\mathrm{ei},2})A_{\mathrm{e}}'' - a_{\mathrm{e},12} &= \frac{b\tau}{2}\left(\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}^2}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right] - \mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\right)\,,\\
&= \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}^2 - W_{\mathrm{e}}(W_{\mathrm{e}}+W_{\mathrm{i}})}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,,\\
&= -\frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}W_{\mathrm{i}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,,\\
&= -c_{\mathrm{ei}}
\end{aligned}$$

so that we have the following compact expression for the quadratic diagonal term

$$A_{\mathrm{e}}'' = \frac{a_{\mathrm{e},12} - c_{\mathrm{ei}}}{1+a_{\mathrm{ei},2}}\,.$$

Appendix K: Compact variance expression

Our goal is to find a compact, interpretable formula for the stationary variance $\mathbb{V}[V]$ from the knowledge of the quadratic form

$$\mathbb{E}\left[V^2\right] = A_{\mathrm{e}}V_{\mathrm{e}}^2 + B_{\mathrm{ei}}V_{\mathrm{e}}V_{\mathrm{i}} + A_{\mathrm{i}}V_{\mathrm{i}}^2 + \left(V_{\mathrm{e}}B_{\mathrm{e}I} + V_{\mathrm{i}}B_{\mathrm{i}I}\right)(I/G) + A_I(I/G)^2\,.$$

Let us first assume no current injection, $I = 0$, so that one only has to keep track of the quadratic terms. Specifying the quadratic coefficient $A_\mathrm{e} = A'_\mathrm{e} + A''_\mathrm{e}$, $A_\mathrm{i} = A'_\mathrm{i} + A''_\mathrm{i}$ and $B_\mathrm{ei} = B'_\mathrm{ei} + B''_\mathrm{ei}$ in Eq. (K1), we get

$$\begin{aligned}
\mathbb{E}\left[V^2\right] &= \left(\frac{a_{\mathrm{e},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})} + \frac{a_{\mathrm{e},12} - c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}\right) V_\mathrm{e}^2 \\
&\quad + \left(\frac{a_{\mathrm{e},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + a_{\mathrm{i},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})} + \frac{2c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}\right) V_\mathrm{e} V_\mathrm{i} \\
&\quad + \left(\frac{a_{\mathrm{i},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})} + \frac{a_{\mathrm{i},12} - c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}\right) V_\mathrm{i}^2 \,, \\
&= \left(\frac{a_{\mathrm{e},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1}) + (1 + a_{\mathrm{e},1} + a_{\mathrm{i},1})(a_{\mathrm{e},1} - a_{\mathrm{e}2})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e}^2 \\
&\quad + \left(\frac{a_{\mathrm{e},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + a_{\mathrm{i},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e} V_\mathrm{i} \\
&\quad + \left(\frac{a_{\mathrm{i},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + (1 + a_{\mathrm{e},1} + a_{\mathrm{i},1})(a_{\mathrm{i},1} - a_{\mathrm{i}2})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{i}^2 - \frac{c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}(V_\mathrm{e} - V_\mathrm{i})^2 \,,
\end{aligned}$$

where we collect separately all the terms containing the coefficient c_ei and where we use the facts that by definition $a_{\mathrm{e},12} = a_{\mathrm{e},1} - a_{\mathrm{e},2}$, $a_{\mathrm{i},12} = a_{\mathrm{i},1} - a_{\mathrm{i},2}$, and $a_{\mathrm{ei},1} = a_{\mathrm{e},1} + a_{\mathrm{i},1}$. Expanding and simplifying the coefficients of V_e^2 and V_i^2 above yield

$$\begin{aligned}
\mathbb{E}\left[V^2\right] &= \left(\frac{a_{\mathrm{e},1}a_{\mathrm{e},2} + (1 + a_{\mathrm{i},1})(a_{\mathrm{e},1} - a_{\mathrm{e}2})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e}^2 \\
&\quad + \left(\frac{a_{\mathrm{e},1}(2a_{\mathrm{i},2} - a_{\mathrm{i},1}) + a_{\mathrm{i},1}(2a_{\mathrm{e},2} - a_{\mathrm{e},1})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e} V_\mathrm{i} \\
&\quad + \left(\frac{a_{\mathrm{i},1}a_{\mathrm{i},2} + (1 + a_{\mathrm{e},1})(a_{\mathrm{i},1} - a_{\mathrm{i}2})}{(1+a_{\mathrm{ei},1})(1+a_{\mathrm{ei},2})}\right) V_\mathrm{i}^2 - \frac{c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}(V_\mathrm{e} - V_\mathrm{i})^2 \,.
\end{aligned}$$

Then, we can utilize the expression above for $\mathbb{E}\left[V^2\right]$ together with the stationary mean formula

$$\mathbb{E}\left[V\right] = \frac{a_{\mathrm{e},1}V_\mathrm{e} + a_{\mathrm{i},1}V_\mathrm{i}}{1 + a_{\mathrm{ei},1}} \,, \tag{K1}$$

to write the variance $\mathbb{V}\left[V\right] = \mathbb{E}\left[V^2\right] - \mathbb{E}\left[V\right]^2$ as

$$\begin{aligned}
\mathbb{V}\left[V\right] &= \left(\frac{(a_{\mathrm{e},1} - a_{\mathrm{e},2})(1 + a_{\mathrm{i},1})^2 + (a_{\mathrm{i},1} - a_{\mathrm{i},2})a_{\mathrm{e},1}^2}{(1+a_{\mathrm{ei},1})^2(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e}^2 \\
&\quad - \left(\frac{(a_{\mathrm{e},1} - a_{\mathrm{e},2})a_{\mathrm{e},1}(1 + a_{\mathrm{e},1}) + a_{\mathrm{i},1}(a_{\mathrm{i},1} - a_{\mathrm{i},2})(1 + a_{\mathrm{i},1})}{(1+a_{\mathrm{ei},1})^2(1+a_{\mathrm{ei},2})}\right) V_\mathrm{e} V_\mathrm{i} \\
&\quad + \left(\frac{(a_{\mathrm{i},1} - a_{\mathrm{i},2})(1 + a_{\mathrm{e},1})^2 + (a_{\mathrm{e},1} - a_{\mathrm{e},2})a_{\mathrm{i},1}^2}{(1+a_{\mathrm{ei},1})^2(1+a_{\mathrm{ei},2})}\right) V_\mathrm{i}^2 - \frac{c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}(V_\mathrm{e} - V_\mathrm{i})^2 \,.
\end{aligned}$$

To factorize the above expression, let us reintroduce $a_{\mathrm{e},12} = a_{\mathrm{e},1} - a_{\mathrm{e},2}$ and $a_{\mathrm{i},12} = a_{\mathrm{i},1} - a_{\mathrm{i},2}$ and collect the terms where these two coefficients occur. This yields

$$\begin{aligned}
\mathbb{V}\left[V\right] &= \frac{a_{\mathrm{e},12}}{(1+a_{\mathrm{ei},1})^2(1+a_{\mathrm{ei},2})}\left((1 + a_{\mathrm{i},1})^2 V_\mathrm{e}^2 - a_{\mathrm{i},1}(1 + a_{\mathrm{e},1})^2 V_\mathrm{e} V_\mathrm{i} + (a_{\mathrm{e},1})^2 V_\mathrm{i}^2\right) \\
&\quad + \frac{a_{\mathrm{i},12}}{(1+a_{\mathrm{ei},1})^2(1+a_{\mathrm{ei},2})}\left((1 + a_{\mathrm{e},1})^2 V_\mathrm{i}^2 - a_{\mathrm{e},1}(1 + a_{\mathrm{i},1})^2 V_\mathrm{e} V_\mathrm{i} + (a_{\mathrm{i},1})^2 V_\mathrm{e}^2\right) \\
&\quad - \frac{c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}(V_\mathrm{e} - V_\mathrm{i})^2 \,, \\
&= \frac{a_{\mathrm{e},12}}{1+a_{\mathrm{ei},2}}\left(\frac{(1 + a_{\mathrm{i},1})V_\mathrm{e} - a_{\mathrm{e},1}V_\mathrm{i}}{1+a_{\mathrm{ei},1}}\right)^2 + \frac{a_{\mathrm{i},12}}{1+a_{\mathrm{ei},2}}\left(\frac{(1 + a_{\mathrm{i},e})V_\mathrm{i} - a_{\mathrm{i},1}V_\mathrm{e}}{1+a_{\mathrm{ei},1}}\right)^2 \\
&\quad - \frac{c_\mathrm{ei}}{1+a_{\mathrm{ei},2}}(V_\mathrm{e} - V_\mathrm{i})^2 \,.
\end{aligned}$$

Finally, injecting the expression of stationary mean Eq. (K1) in both parentheses above produces the compact formula

$$\mathbb{V}[V] = \frac{a_{\mathrm{e},12}}{1+a_{\mathrm{ei},2}}\left(V_{\mathrm{e}} - \mathbb{E}[V]\right)^2 + \frac{a_{\mathrm{i},12}}{1+a_{\mathrm{ei},2}}\left(V_{\mathrm{i}} - \mathbb{E}[V]\right)^2 - \frac{c_{\mathrm{ei}}}{1+a_{\mathrm{ei},2}}(V_{\mathrm{e}} - V_{\mathrm{i}})^2\,, \tag{K2}$$

which is the same as the one given in Eq. (16).

Appendix L: Factorized variance expression

In this appendix, we reshape the variance expression given in Eq. (K2) under a form that is clearly nonnegative. To this end, let us first remark that the calculation in Appendix J shows that

$$a_{\mathrm{e},12} - c_{\mathrm{ei}} = \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}^2}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1+e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right].$$

Then, setting $(V_{\mathrm{e}} - V_{\mathrm{i}})^2 = ((V_{\mathrm{e}} - \mathbb{E}[V]) - (V_{\mathrm{i}} - \mathbb{E}[V]))^2 = (V_{\mathrm{e}} - \mathbb{E}[V])^2 - 2(V_{\mathrm{e}} - \mathbb{E}[V])(V_{\mathrm{i}} - \mathbb{E}[V]) + (V_{\mathrm{i}} - \mathbb{E}[V])^2$ in Eq. (K2), we obtain

$$\begin{aligned}
\mathbb{V}[V] &= \frac{1}{1+a_{\mathrm{ei},2}}\left(a_{\mathrm{e},12}(V_{\mathrm{e}} - \mathbb{E}[V])^2 + a_{\mathrm{i},12}(V_{\mathrm{i}} - \mathbb{E}[V])^2 - c_{\mathrm{ei}}(V_{\mathrm{e}} - V_{\mathrm{i}})^2\right)\,,\\
&= \frac{1}{1+a_{\mathrm{ei},2}}\left((a_{\mathrm{e},12} - c_{\mathrm{ei}})(V_{\mathrm{e}} - \mathbb{E}[V])^2 + 2c_{\mathrm{ei}}(V_{\mathrm{e}} - \mathbb{E}[V])(V_{\mathrm{i}} - \mathbb{E}[V]) + (a_{\mathrm{i},12} - c_{\mathrm{ei}})(V_{\mathrm{i}} - \mathbb{E}[V])^2\right)\,,\\
&= \frac{b\tau}{1+a_{\mathrm{ei},2}}\\
&\qquad \mathbb{E}_{\mathrm{ei}}\left[\left(\frac{W_{\mathrm{e}}^2(V_{\mathrm{e}} - \mathbb{E}[V])^2}{2(W_{\mathrm{e}}+W_{\mathrm{i}})^2} + \frac{2W_{\mathrm{e}}(V_{\mathrm{e}} - \mathbb{E}[V])W_{\mathrm{i}}(V_{\mathrm{i}} - \mathbb{E}[V])}{2(W_{\mathrm{e}}+W_{\mathrm{i}})^2} + \frac{W_{\mathrm{i}}^2(V_{\mathrm{i}} - \mathbb{E}[V])^2}{2(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\right)\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,,\\
&= \frac{b\tau}{2(1+a_{\mathrm{ei},2})}\mathbb{E}_{\mathrm{ei}}\left[\left(\frac{\left[W_{\mathrm{e}}(V_{\mathrm{e}} - \mathbb{E}[V]) + W_{\mathrm{i}}(V_{\mathrm{i}} - \mathbb{E}[V])\right]^2}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\right)\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]\,.
\end{aligned} \tag{L1}$$

Note that the above quantity is clearly non negative as any variance shall be. From there, one can include the impact of the injected current I by further considering all the terms in Eq. (K1), including the linear and inhomogeneous current-dependent terms. Similar algebraic manipulations confirm that Eq. (L1) remains valid so that the only impact of I is via altering the expression $\mathbb{E}[V]$, so that we ultimately obtain the following explicit compact form:

$$\mathbb{V}[V] = \frac{\mathbb{E}_{\mathrm{ei}}\left[\left(\frac{W_{\mathrm{e}}V_{\mathrm{e}}+W_{\mathrm{i}}V_{\mathrm{i}}}{W_{\mathrm{e}}+W_{\mathrm{i}}} - \mathbb{E}[V]\right)^2\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right]}{2/(b\tau) + \mathbb{E}_{\mathrm{ei}}\left[\left(1-e^{-2(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right]} \quad \text{with} \quad \mathbb{E}[V] = \frac{b\tau\mathbb{E}_{\mathrm{ei}}\left[\left(\frac{W_{\mathrm{e}}V_{\mathrm{e}}+W_{\mathrm{i}}V_{\mathrm{i}}}{W_{\mathrm{e}}+W_{\mathrm{i}}}\right)\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right] + I/G}{1 + b\tau\mathbb{E}_{\mathrm{ei}}\left[\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)\right]}\,.$$

The above expression shows that as expected $\mathbb{V}[V] \geq 0$ and that the variability vanishes if and only if $W_{\mathrm{e}}/W_{\mathrm{i}} = (\mathbb{E}[V] - V_{\mathrm{i}})/(V_{\mathrm{e}} - \mathbb{E}[V])$ with probability one. In turn plugging this relation into the mean voltage expression and solving for $\mathbb{E}[V]$ reveals that we necessarily have $\mathbb{E}[V] = I/G$. This is consistent with the intuition that variability can only vanish if excitation and inhibition perfectly cancel one another.

Appendix M: Variance in the small-weight approximation

In this appendix, we compute the simplified expression for the variance $\mathbb{V}[V]$ obtained via the small-weight approximation. Second, let us compute the small-weight approximation of the second-order efficacy

$$c_{\mathrm{ei}} = \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[\frac{W_{\mathrm{e}}W_{\mathrm{i}}}{(W_{\mathrm{e}}+W_{\mathrm{i}})^2}\left(1-e^{-(W_{\mathrm{e}}+W_{\mathrm{i}})}\right)^2\right] \simeq \frac{b\tau}{2}\mathbb{E}_{\mathrm{ei}}\left[W_{\mathrm{e}}W_{\mathrm{i}}\right] = \frac{b\tau}{2}w_{\mathrm{e}}w_{\mathrm{i}}\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{e}}k_{\mathrm{i}}\right]\,,$$

which amounts to compute the expectation of the crossproduct of the jumps k_{e} and k_{i}. To estimate the above approximation, it is important to remember that first that p_{e} and p_{i} are not defined as the marginals of p_{ei}, but as conditional marginals, for which we have $p_{\mathrm{e},k} = (b/b_{\mathrm{e}})\sum_{l=0}^{K_{\mathrm{i}}} p_{\mathrm{ei},kl}$ and $p_{\mathrm{i},l} = (b/b_{\mathrm{i}})\sum_{k=0}^{K_{\mathrm{e}}} p_{\mathrm{ei},kl}$. Then by the definition of the correlation coefficient ρ_{ei} in Eq. (4), we have

$$\rho_{\mathrm{ei}} = \frac{b\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{e}}k_{\mathrm{i}}\right]}{\sqrt{K_{\mathrm{e}}b\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{e}}\right]K_{\mathrm{i}}b\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{i}}\right]}} = \frac{b\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{e}}k_{\mathrm{i}}\right]}{\sqrt{K_{\mathrm{e}}b_{\mathrm{e}}\mathbb{E}_{\mathrm{e}}\left[k_{\mathrm{e}}\right]K_{\mathrm{i}}b_{\mathrm{i}}\mathbb{E}_{\mathrm{i}}\left[k_{\mathrm{i}}\right]}} = \frac{b\mathbb{E}_{\mathrm{ei}}\left[k_{\mathrm{e}}k_{\mathrm{i}}\right]}{K_{\mathrm{e}}K_{\mathrm{i}}\sqrt{r_{\mathrm{e}}r_{\mathrm{i}}}}\,,$$

as the rates b_e and b_i are such that $b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}\right] = K_\mathrm{e} r_\mathrm{e}$ and $b_\mathrm{i}\mathbb{E}_\mathrm{e}\left[k_\mathrm{i}\right] = K_\mathrm{i} r_\mathrm{i}$. As a result, we obtain a simplified expression for the cross-correlation coefficient:

$$c_\mathrm{ei} = (\rho_\mathrm{ei}\sqrt{r_\mathrm{e} r_\mathrm{i}}\tau/2)(K_\mathrm{e} w_\mathrm{e})(K_\mathrm{i} w_\mathrm{i})\,.$$

Observe that as expected, c_ei vanishes when $\rho_\mathrm{ei} = 0$. Second, let us compute the small-weight approximation of the second-order efficacy

$$a_{\mathrm{e},12} = \frac{b\tau}{2}\mathbb{E}_\mathrm{ei}\left[\frac{W_\mathrm{e}}{W_\mathrm{e}+W_\mathrm{i}}\left(1-e^{-(W_\mathrm{e}+W_\mathrm{i})}\right)^2\right] \simeq \frac{b\tau}{2}\mathbb{E}_\mathrm{ei}\left[W_\mathrm{e}(W_\mathrm{e}+W_\mathrm{i})\right] = \frac{b\tau}{2}\left(w_\mathrm{e}^2\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}^2\right] + w_\mathrm{e} w_\mathrm{i}\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e} k_\mathrm{i}\right]\right)\,.$$

To estimate the above approximation, we use the definition of the correlation coefficient ρ_e in Eq. (8),

$$\rho_\mathrm{e} = \frac{b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}(k_\mathrm{e}-1)\right]}{b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}\right](K_\mathrm{e}-1)} = \frac{b\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}(k_\mathrm{e}-1)\right]}{K_\mathrm{e}(K_\mathrm{e}-1)r_\mathrm{e}}\,,$$

as the rate b_e is such that $b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}\right] = K_\mathrm{e} r_\mathrm{e}$. This directly implies that

$$b\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}^2\right] = b\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}(k_\mathrm{e}-1)\right] + b\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}\right] = \rho_\mathrm{e} K_\mathrm{e}(K_\mathrm{e}-1)r_\mathrm{e} + K_\mathrm{e} r_\mathrm{e} = K_\mathrm{e} r_\mathrm{e}(1+\rho_\mathrm{e}(K_\mathrm{e}-1))\,.$$

so that we evaluate

$$a_{\mathrm{e},12} = \frac{b\tau}{2}\left(w_\mathrm{e}^2\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e}^2\right] + w_\mathrm{e} w_\mathrm{i}\mathbb{E}_\mathrm{ei}\left[k_\mathrm{e} k_\mathrm{i}\right]\right) = \frac{r_\mathrm{e}\tau}{2}K_\mathrm{e}(1+\rho_\mathrm{e}(K_\mathrm{e}-1))w_\mathrm{e}^2 + \rho_\mathrm{ei}\frac{\sqrt{r_\mathrm{i} r_\mathrm{e}}\tau}{2}(K_\mathrm{e} w_\mathrm{e})(K_\mathrm{i} w_\mathrm{i})\,,$$

which simplifies to $a_{\mathrm{e},12} = (r_\mathrm{e}\tau/2)K_\mathrm{e}(1+\rho_\mathrm{e}(K_\mathrm{e}-1))w_\mathrm{e}^2$ when excitation and inhibition act independently. A symmetric expression holds for the inhibitory efficacy $a_{\mathrm{i},12}$. Plugging the above expressions for synaptic efficacies in the variance expression Eq. (16) yields the small-weight approximation

$$\begin{aligned}\mathbb{V}\left[V\right] \simeq\ & \frac{(1+\rho_\mathrm{e}(K_\mathrm{e}-1))K_\mathrm{e} r_\mathrm{e} w_\mathrm{e}^2(V_\mathrm{e}-\mathbb{E}\left[V\right])^2 + (1+\rho_\mathrm{i}(K_\mathrm{i}-1))K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}^2(V_\mathrm{i}-\mathbb{E}\left[V\right])^2}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i})} \\ & + \frac{\rho_\mathrm{ei}\sqrt{r_\mathrm{e} r_\mathrm{i}}(K_\mathrm{e} w_\mathrm{e})(K_\mathrm{i} w_\mathrm{i})\left[(V_\mathrm{e}-\mathbb{E}\left[V\right])^2 + (V_\mathrm{i}-\mathbb{E}\left[V\right])^2 - (V_\mathrm{e}-V_\mathrm{i})^2\right]}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i})}\,.\end{aligned}$$

Let us note that the first-term in the right-hand side above represents the small-weight approximation of the voltage variance in the absence of correlation between excitation and inhibition, i.e., for $\rho_\mathrm{ei} = 0$. Denoting the latter approximation by $\mathbb{V}\left[V\right]|_{\rho_\mathrm{ei}=0}$ and using the fact that the small-weight expression for the mean voltage

$$\mathbb{E}\left[V\right] = \frac{K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} V_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i} V_\mathrm{i}}{1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}}\,,$$

is independent of correlations, we observe that as intuition suggests, synchrony-based correlation between excitation and inhibition results in a decrease of the neural variability:

$$\Delta\mathbb{V}\left[V\right]_{\rho_\mathrm{ei}} = \mathbb{V}\left[V\right] - \mathbb{V}\left[V\right]|_{\rho_\mathrm{ei}=0} \simeq -\frac{\rho_\mathrm{ei}\sqrt{r_\mathrm{e} r_\mathrm{i}}(K_\mathrm{e} w_\mathrm{e})(K_\mathrm{i} w_\mathrm{i})(V_\mathrm{e}-\mathbb{E}\left[V\right])(\mathbb{E}\left[V\right]-V_\mathrm{i})}{1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}} \leq 0\,.$$

However, the overall contribution of correlation is to increase variability in the small-weight approximation. This can be shown under the assumptions that $K_\mathrm{e} \gg 1$ and $K_\mathrm{i} \gg 1$, by observing that

$$\begin{aligned}\Delta\mathbb{V}\left[V\right]_{\rho_\mathrm{ei},\rho_{\mathrm{e/i}}} = \mathbb{V}\left[V\right] - \mathbb{V}\left[V\right]|_{\rho_{\mathrm{e/i}}=\rho_\mathrm{ei}=0} \simeq\ & \frac{\left(\sqrt{\rho_\mathrm{e} r_\mathrm{e}}K_\mathrm{e} w_\mathrm{e}(V_\mathrm{e}-\mathbb{E}\left[V\right]) - \sqrt{\rho_\mathrm{i} r_\mathrm{i}}K_\mathrm{i} w_\mathrm{i}(V_\mathrm{i}-\mathbb{E}\left[V\right])\right)^2}{2(1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i})} \\ & + (\sqrt{\rho_\mathrm{e}\rho_\mathrm{i}} - \rho_\mathrm{ei})\frac{\sqrt{r_\mathrm{e} r_\mathrm{i}}(K_\mathrm{e} w_\mathrm{e})(K_\mathrm{i} w_\mathrm{i})(V_\mathrm{e}-\mathbb{E}\left[V\right])(\mathbb{E}\left[V\right]-V_\mathrm{i})}{1/\tau + K_\mathrm{e} r_\mathrm{e} w_\mathrm{e} + K_\mathrm{i} r_\mathrm{i} w_\mathrm{i}} \geq 0\,,\end{aligned}$$

where both terms are positive since we always have $0 \leq \rho_\mathrm{ei} \leq \sqrt{\rho_\mathrm{e}\rho_\mathrm{i}}$.

Appendix N: Validity of the small-weight approximation

Biophysical estimates of the synaptic weights $w_\mathrm{e} < 0.01$, $w_\mathrm{i} < 0.04$ and the synaptic input numbers $K_\mathrm{e} < 10000$, $K_\mathrm{i} < 2500$, suggest that neurons operates in the small-weight regime. In this regime, we claim that exponential corrections due to finite-size effect can be neglected in the evaluation of synaptic efficacies, as long as the spiking correlations remains weak. Here, we make this latter statement quantitative by focusing on the first-order efficacies in the case of excitation alone. The relative error due to neglecting exponential corrections can be quantified as

$$\mathcal{E} = \frac{\mathbb{E}_\mathrm{e}\left[W_\mathrm{e}\right] - \mathbb{E}_\mathrm{e}\left[1 - e^{-W_\mathrm{e}}\right]}{\mathbb{E}_\mathrm{e}\left[1 - e^{-W_\mathrm{e}}\right]} \geq 0\,.$$

Let us evaluate this relative error, assumed to be small, when correlations are parametrized via beta distributions with parameter $\beta_\mathrm{e} = 1/\rho_\mathrm{e} - 1$. Assuming correlations to be weak, $\rho_\mathrm{e} \ll 1$, amounts to assuming large, $\beta_\mathrm{e} \gg 1$ Under the assumptions of small error, we can compute

$$\mathbb{E}_\mathrm{e}\left[1 - e^{-W_\mathrm{e}}\right] \simeq \mathbb{E}_\mathrm{e}\left[W_\mathrm{e}\right] = w_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}\right] \quad \text{and} \quad \mathbb{E}_\mathrm{e}\left[W_\mathrm{e} - 1 + e^{-W_\mathrm{e}}\right] \simeq \mathbb{E}_\mathrm{e}\left[W_\mathrm{e}^2\right]/2 = w_\mathrm{e}^2\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}^2\right]/2\,,$$

By the calculations carried out in Appendix N, we have

$$b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}\right] = K_\mathrm{e} r_\mathrm{e} \quad \text{and} \quad b_\mathrm{e}\mathbb{E}_\mathrm{e}\left[k_\mathrm{e}^2\right] = K_\mathrm{e} r_\mathrm{e}(1 + \rho_\mathrm{e}(K_\mathrm{e} - 1))\,.$$

Remembering that $\beta_\mathrm{e} = 1/\rho_\mathrm{e} - 1$, this implies that we have

$$\mathcal{E} \simeq \frac{\mathbb{E}_\mathrm{e}\left[W_\mathrm{e}^2\right]/2}{\mathbb{E}_\mathrm{e}\left[W_\mathrm{e}\right] - \mathbb{E}_\mathrm{e}\left[W_\mathrm{e}^2\right]/2} \simeq \frac{w_\mathrm{e}(1 + \rho_\mathrm{e}(K_\mathrm{e} - 1))/2}{1 - w_\mathrm{e}(1 + \rho_\mathrm{e}(K_\mathrm{e} - 1))/2}\,,$$

For a correlation coefficient $\rho_\mathrm{e} \leq 0.05$, this means that neglecting exponential corrections incurs less than a $e = 3\%$ error if the number of inputs is smaller than $K_\mathrm{e} \leq 1000$ for moderate synaptic weight $w_\mathrm{e} = 0.001$ or than $K_\mathrm{e} \leq 100$ for large synaptic weight $w_\mathrm{e} = 0.01$.

Appendix O: Infinite-size limit with spiking correlations

The computation of the first two moments $\mathbb{E}\left[V\right]$ and $\mathbb{E}\left[V^2\right]$ requires to evaluate various efficacies as expectations. Upon inspection, these expectations are all of the form $b\mathbb{E}_\mathrm{ei}\left[f(W_\mathrm{e}, W_\mathrm{i})\right]$, where f is a smooth positive function that is bounded on $\mathbb{R}^+ \times \mathbb{R}^+$ with $f(0,0) = 0$. Just as for the Lévy-Khintchine decomposition of stable jump processes [78, 79], this observation allows one to generalize our results to processes that exhibit and countable infinity of jumps over finite, nonzero time intervals. For our parametric forms based on beta distributions, such processes emerge in the limit of an arbitrary large number of inputs, i.e., for $K_\mathrm{e}, K_\mathrm{i} \to \infty$. Let us consider the case of excitation alone for simplicity. Then, we need to make sure that all expectations of the form $b_\mathrm{e}\mathbb{E}_\mathrm{ei}\left[f(W_\mathrm{e})\right]$ remain well-posed in the limit $K_\mathrm{e} \to \infty$ for smooth, bounded test function f with $f(0) = 0$. To check this, observe that for all $0 < k \leq K_\mathrm{e}$, we have by Eq. (7) and Eq. (9) that

$$b_\mathrm{e} p_{\mathrm{e},k} = \beta r_\mathrm{e}\binom{K_e}{k} B(k, \beta + K_\mathrm{e} - k) = \beta r_\mathrm{e}\frac{\Gamma(K_\mathrm{e}+1)}{\Gamma(k+1)\Gamma(K_\mathrm{e}-k+1)}\frac{\Gamma(k)\Gamma(\beta + K_\mathrm{e} - k + 1)}{\Gamma(\beta + K_\mathrm{e})}\,,$$

where we have introduce the Gamma function Γ. Rearranging terms and using the fact that $\Gamma(z+1) = z\Gamma(z)$ for all $z > 0$, we obtain

$$b_\mathrm{e} p_{\mathrm{e},k} = \frac{\beta r_\mathrm{e}}{k}\frac{K_\mathrm{e}\Gamma(K_\mathrm{e})}{\Gamma(\beta + K_\mathrm{e})}\frac{\Gamma(\beta + K_\mathrm{e} - k)}{(K_\mathrm{e} - k)\Gamma(K_\mathrm{e} - k)} = \frac{\beta r_\mathrm{e}}{k}\left(1 - \frac{k}{K_\mathrm{e}}\right)^{\beta - 1} + o\left(\frac{1}{K_\mathrm{e}}\right)\,,$$

where the last equality is uniform in k and follows from the fact that for all $x > 0$, we have

$$\lim_{z\to\infty}\frac{\Gamma(z+x)}{\Gamma(z)} = z^x\left(1 + \binom{x}{2}\frac{1}{z} + o\left(\frac{1}{z}\right)\right)$$

From there, given a test function f, let us consider

$$\begin{aligned} b_{\mathrm{e}}\mathbb{E}_{\mathrm{e}}\left[f(W_{\mathrm{e}})\right] &= \int \sum_{k=1}^{K_{\mathrm{e}}} b_{\mathrm{e}} p_{\mathrm{e},k} \delta\left(W_{\mathrm{e}} - \frac{k\Omega_{\mathrm{e}}}{K_{\mathrm{e}}}\right) f(W_{\mathrm{e}})\,\mathrm{d}W_{\mathrm{e}}\,, \\ &= \sum_{k=1}^{K_{\mathrm{e}}} b_{\mathrm{e}} p_{\mathrm{e},k} f\left(\frac{k\Omega_{\mathrm{e}}}{K_{\mathrm{e}}}\right)\,, \\ &= r_{\mathrm{e}} \sum_{k=1}^{K_{\mathrm{e}}} \frac{\beta}{k}\left(1 - \frac{k}{K_{\mathrm{e}}}\right)^{\beta-1} f\left(\frac{k\Omega_{\mathrm{e}}}{K_{\mathrm{e}}}\right) + o(1)\,. \end{aligned}$$

The order zero term above can be interpreted as a Riemann sum so that one has

$$\begin{aligned} \lim_{K_{\mathrm{e}}\to\infty} b_{\mathrm{e}}\mathbb{E}_{\mathrm{e}}\left[f(W_{\mathrm{e}})\right] &= r_{\mathrm{e}} \lim_{K_{\mathrm{e}}\to\infty} \frac{1}{K_{\mathrm{e}}} \sum_{k=1}^{K_{\mathrm{e}}} \frac{\beta K_{\mathrm{e}}}{k}\left(1 - \frac{k}{K_{\mathrm{e}}}\right) f\left(\frac{k\Omega_{\mathrm{e}}}{K_{\mathrm{e}}}\right)\,, \\ &= r_{\mathrm{e}} \int_0^1 \beta\theta^{-1}(1-\theta)^{\beta-1} f(\theta\Omega_{\mathrm{e}})\,\mathrm{d}\theta\,, \\ &= r_{\mathrm{e}} \int_0^{\Omega_{\mathrm{e}}} \frac{\beta}{w}\left(1 - \frac{w}{\Omega_{\mathrm{e}}}\right)^{\beta-1} f(w)\,\mathrm{d}w\,. \end{aligned}$$

Thus, the jump densities is specified via the Lévy-Khintchine measure

$$\nu_{\mathrm{e}}(w) = \frac{\beta}{w}\left(1 - \frac{w}{\Omega_{\mathrm{e}}}\right)^{\beta-1}\,,$$

which is a deficient measure for admitting a pole in zero. This singular behavior indicates that the limit jump process obtained when $K_{\mathrm{e}} \to \infty$ has a countable infinity of jumps within any finite, nonempty time interval. Generic stationary jump processes with independent increments, as is the case here, are entirely specified by their Lévy-Khintchine measure ν_{e} [78, 79]. Moreover, one can check that given knowledge of ν_{e}, one can consistently estimate the corresponding pairwise spiking correlation as

$$\rho_{\mathrm{e}} = \lim_{K_{\mathrm{e}}\to\infty} \frac{\mathbb{E}_{\mathrm{e}}\left[k_{\mathrm{e}}(k_{\mathrm{e}}-1)\right]}{\mathbb{E}_{\mathrm{e}}\left[k_{\mathrm{e}}\right](K_{\mathrm{e}}-1)} = \lim_{K_{\mathrm{e}}\to\infty} \frac{b_{\mathrm{e}}\mathbb{E}_{\mathrm{e}}\left[(k_{\mathrm{e}}/K_{\mathrm{e}})^2\right]}{b_{\mathrm{e}}\mathbb{E}_{\mathrm{e}}\left[k_{\mathrm{e}}/K_{\mathrm{e}}\right]} = \frac{\int_0^{\Omega_{\mathrm{e}}} w^2 \nu_{\mathrm{e}}(w)\,\mathrm{d}w}{\Omega_{\mathrm{e}} \int_0^{\Omega_{\mathrm{e}}} w\nu_{\mathrm{e}}(w)\,\mathrm{d}w}\,.$$

Performing integral with respect to the Lévy-Khintchine measure ν_{e} instead of the evaluating the expectation $\mathbb{E}_{\mathrm{e}}\left[\cdot\right]$ in Eq. (14) and Eq. (16) yields

$$\mathbb{E}\left[V\right] = \frac{V_{\mathrm{e}} \int_0^{\Omega_{\mathrm{e}}} (1-e^{-w})\nu_{\mathrm{e}}(\mathrm{d}w)}{1/\tau + \int_0^{\Omega_{\mathrm{e}}} (1-e^{-w})\nu_{\mathrm{e}}(\mathrm{d}w)} \quad \text{and} \quad \mathbb{V}\left[V\right] = \frac{(V_{\mathrm{e}} - \mathbb{E}\left[V\right])^2 \int_0^{\Omega_{\mathrm{e}}} (1-e^{-w})^2 \nu_{\mathrm{e}}(\mathrm{d}w)}{2/\tau + \int_0^{\Omega_{\mathrm{e}}} (1-e^{-2w})\nu_{\mathrm{e}}(\mathrm{d}w)}\,.$$

Observe that as $(1-e^{-w})^2 \leq w^2$ for all $w \geq 0$, the definition of the spiking correlation and voltage variance implies that we have $\mathbb{V}\left[V\right] = O(\rho_{\mathrm{e}})$ so that neural variability consistently vanishes in the absence of correlations.

[1] M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy, Stimulus onset quenches neural variability: a widespread cortical phenomenon, Nature Neuroscience **13**, 369 EP (2010).

[2] D. Tolhurst, J. A. Movshon, and I. Thompson, The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast, Experimental brain research **41**, 414 (1981).

[3] D. J. Tolhurst, J. A. Movshon, and A. F. Dean, The statistical reliability of signals in single neurons in cat and monkey visual cortex, Vision research **23**, 775 (1983).

[4] M. M. Churchland, M. Y. Byron, S. I. Ryu, G. Santhanam, and K. V. Shenoy, Neural variability in premotor cortex provides a signature of motor preparation, Journal of Neuroscience **26**, 3697 (2006).

[5] J. Rickert, A. Riehle, A. Aertsen, S. Rotter, and M. P. Nawrot, Dynamic encoding of movement direction in motor cortical neurons, Journal of Neuroscience **29**, 13870 (2009).

[6] C. F. Stevens and A. M. Zador, Input synchrony and the irregular firing of cortical neurons, Nature neuroscience **1**, 210 (1998).

[7] I. Lampl, I. Reichova, and D. Ferster, Synchronous membrane potential fluctuations in neurons of the cat visual cortex, Neuron **22**, 361 (1999).

[8] 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, State dependence of noise correlations in macaque primary visual cortex, Neuron **82**, 235 (2014).

[9] J. F. Poulet and C. C. Petersen, Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice, Nature **454**, 881 (2008).

[10] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, The asynchronous state in cortical circuits, Science **327**, 587 (2010).

[11] A. S. Ecker, P. Berens, G. A. Keliris, M. Bethge, N. K. Logothetis, and A. S. Tolias, Decorrelated neuronal firing in cortical microcircuits, science **327**, 584 (2010).

[12] M. R. Cohen and A. Kohn, Measuring and interpreting neuronal correlations, Nature neuroscience **14**, 811 (2011).

[13] V. Braitenberg and A. Schüz, *Cortex: statistics and geometry of neuronal connectivity* (Springer Science & Business Media, 2013).

[14] W. R. Softky and C. Koch, Cortical cells should fire regularly, but do not (1992).

[15] A. Bell, Z. F. Mainen, M. Tsodyks, and T. J. Sejnowski, "balancing" of conductances may explain irregular cortical spiking, La Jolla, CA: Institute for Neural Computation Technical Report INC-9502 (1995).

[16] D. J. Amit and N. Brunel, Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex., Cerebral cortex (New York, NY: 1991) **7**, 237 (1997).

[17] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, Journal of computational neuroscience **8**, 183 (2000).

[18] Y. Ahmadian and K. D. Miller, What is the dynamical regime of cerebral cortex?, Neuron **109**, 3373 (2021).

[19] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Chaos in random neural networks, Phys. Rev. Lett. **61**, 259 (1988).

[20] van Vreeswijk C. and S. H., Chaos in neuronal networks with balanced excitatory and inhibitory activity, Science **274**, 1724 (1996).

[21] C. v. Vreeswijk and H. Sompolinsky, Chaotic balanced state in a model of cortical circuits, *Neural Computation*, Neural Computation **10**, 1321 (1998).

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

[23] D. B. Rubin, S. D. Van Hooser, and K. D. Miller, The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex, Neuron **85**, 402 (2015).

[24] G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, and K. D. Miller, The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability, Neuron **98**, 846 (2018).

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

[26] A. Y. Y. Tan, S. Andoni, and N. J. Priebe, A spontaneous state of weakly correlated synaptic excitation and inhibition in visual cortex, Neuroscience **247**, 364 (2013).

[27] A. Y. Y. Tan, Y. Chen, B. Scholl, E. Seidemann, and N. J. Priebe, Sensory stimulation shifts visual cortex from synchronous to asynchronous states, Nature **509**, 226 EP (2014).

[28] M. Okun, N. A. Steinmetz, L. Cossell, M. F. Iacaruso, H. Ko, P. Barthó, T. Moore, S. B. Hofer, T. D. Mrsic-Flogel, M. Carandini, *et al.*, Diverse coupling of neurons to populations in sensory cortex, Nature **521**, 511 (2015).

[29] D. Hansel and C. van Vreeswijk, The mechanism of orientation selectivity in primary visual cortex without a functional map, Journal of Neuroscience **32**, 4049 (2012).

[30] J. J. Pattadkal, G. Mato, C. van Vreeswijk, N. J. Priebe, and D. Hansel, Emergent orientation selectivity from random networks in mouse visual cortex, Cell reports **24**, 2042 (2018. PMC6179374).

[31] M. N. Shadlen and W. T. Newsome, The variable discharge of cortical neurons: implications for connectivity, computation, and information coding, Journal of neuroscience **18**, 3870 (1998).

[32] Y. Chen, W. S. Geisler, and E. Seidemann, Optimal decoding of correlated neural population responses in the primate visual cortex, Nature neuroscience **9**, 1412 (2006).

[33] A. Polk, A. Litwin-Kumar, and B. Doiron, Correlated neural variability in persistent state networks, Proceedings of the National Academy of Sciences 10.1073/pnas.1121274109 (2012).

[34] J. Yu and D. Ferster, Membrane potential synchrony in primary visual cortex during sensory stimulation, Neuron **68**, 1187 (2010).

[35] S. Arroyo, C. Bennett, and S. Hestrin, Correlation of synaptic inputs in the visual cortex of awake, behaving mice, Neuron **99**, 1289 (2018).

[36] M. Okun and I. Lampl, Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities, Nature neuroscience **11**, 535 (2008).

[37] Y. Zerlaut, S. Zucca, S. Panzeri, and T. Fellin, The spectrum of asynchronous dynamics in spiking networks as a model for the diversity of non-rhythmic waking states in the neocortex, Cell reports **27**, 1119 (2019).

[38] A. Sanzeni, M. H. Histed, and N. Brunel, Emergence of irregular activity in networks of strongly coupled conductance-based neurons, Physical Review X **12**, 011044 (2022).

[39] R. B. Stein, A theoretical analysis of neuronal variability, Biophysical Journal **5**, 173 (1965).

[40] H. C. Tuckwell, *Introduction to theoretical neurobiology: linear cable theory and dendritic structure*, Vol. 1 (Cambridge University Press, 1988).

[41] M. J. Richardson, Effects of synaptic conductance on the voltage distribution and firing rate of spiking neurons, Physical Review E **69**, 051918 (2004).

[42] M. J. Richardson and W. Gerstner, Synaptic shot noise

and conductance fluctuations affect the membrane voltage with equal significance, Neural computation **17**, 923 (2005).

[43] M. J. Richardson and W. Gerstner, Statistics of subthreshold neuronal voltage fluctuations due to conductance-based synaptic shot noise, Chaos: An Interdisciplinary Journal of Nonlinear Science **16**, 026106 (2006).

[44] S. Marcus, Modeling and analysis of stochastic differential equations driven by point processes, IEEE Transactions on Information theory **24**, 164 (1978).

[45] S. I. Marcus, Modeling and approximation of stochastic differential equations driven by semimartingales, Stochastics: An International Journal of Probability and Stochastic Processes **4**, 223 (1981).

[46] A. Destexhe, M. Rudolph, J.-M. Fellous, and T. Sejnowski, Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons, Neuroscience **107**, 13 (2001).

[47] H. Meffin, A. N. Burkitt, and D. B. Grayden, An analytical model for the 'large, fluctuating synaptic conductance state'typical of neocortical neurons in vivo, Journal of computational neuroscience **16**, 159 (2004).

[48] B. Doiron, A. Litwin-Kumar, R. Rosenbaum, G. K. Ocker, and K. Josić, The mechanics of state-dependent neural correlations, Nature neuroscience **19**, 383 (2016).

[49] G. K. Ocker, Y. Hu, M. A. Buice, B. Doiron, K. Josić, R. Rosenbaum, and E. Shea-Brown, From the statistics of connectivity to the statistics of spike times in neuronal networks, Current opinion in neurobiology **46**, 109 (2017).

[50] W. Rall, Time constants and electrotonic length of membrane cylinders and neurons, Biophysical Journal **9**, 1483 (1969).

[51] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes. Vol. I. Probability and its Applications* (New York). Springer-Verlag, New York,, 2003).

[52] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure* (Springer Science & Business Media, 2007).

[53] B. W. Knight, The relationship between the firing rate of a single neuron and the level of activity in a population of neurons: Experimental evidence for resonant enhancement in the population response, The Journal of general physiology **59**, 767 (1972).

[54] B. Knight, Dynamics of encoding in a population of neurons., The Journal of general physiology **59**, 734 (1972).

[55] J. F. Kingman, Uses of exchangeability, The Annals of Probability **6**, 183 (1978).

[56] D. J. Aldous, Exchangeability and related topics, in *École d'Été de Probabilités de Saint-Flour XIII—1983* (Springer, 1985) pp. 1–198.

[57] B. De Finetti, Funzione caratteristica di un fenomeno aleatorio, in *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928* (1929) pp. 179–190.

[58] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications* (CRC press, 2004).

[59] J. H. Macke, P. Berens, A. S. Ecker, A. S. Tolias, and M. Bethge, Generating spike trains with specified correlation coefficients, Neural computation **21**, 397 (2009).

[60] N. L. Hjort, Nonparametric bayes estimators based on beta processes in models for life history data, the Annals of Statistics , 1259 (1990).

[61] R. Thibaux and M. I. Jordan, Hierarchical beta processes and the indian buffet process, in *Artificial intelligence and statistics* (PMLR, 2007) pp. 564–571.

[62] T. Broderick, M. I. Jordan, and J. Pitman, Beta processes, stick-breaking and power laws, Bayesian analysis **7**, 439 (2012).

[63] P. Berkes, F. Wood, and J. Pillow, Characterizing neural dependencies with copula models, Advances in neural information processing systems **21** (2008).

[64] N. Balakrishnan and C. D. Lai, *Continuous bivariate distributions* (Springer Science & Business Media, 2009).

[65] R. Stratonovich, A new representation for stochastic integrals and equations, SIAM Journal on Control **4**, 362 (1966).

[66] A. Chechkin and I. Pavlyukevich, Marcus versus stratonovich for systems with jump noise, Journal of Physics A: Mathematical and Theoretical **47**, 342001 (2014).

[67] K. Matthes, Zur Theorie der Bedienungsprozesse, in *Trans. Third Prague Conf. Information Theory, Statist. Decision Functions, Random Processes (Liblice, 1962)* (Publ. House Czech. Acad. Sci., Prague, 1964) pp. 513–528.

[68] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen, Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses, Science **273**, 1868 (1996).

[69] Z. Mainen and T. Sejnowski, Reliability of spike timing in neocortical neurons, Science **268**, 1503 (1995).

[70] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes*, A Bradford Book (MIT Press, Cambridge, MA, 1999) pp. xviii+395, exploring the neural code, Computational Neuroscience.

[71] W. R. Softky and C. Koch, The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps, Journal of neuroscience **13**, 334 (1993).

[72] R. M. Bruno and B. Sakmann, Cortex is driven by weak but synchronously active thalamocortical synapses, Science **312**, 1622 (2006).

[73] J.-S. Jouhanneau, J. Kremkow, A. L. Dorrn, and J. F. Poulet, In vivo monosynaptic excitatory transmission between layer 2 cortical pyramidal neurons, Cell reports **13**, 2098 (2015).

[74] A. Pala and C. C. Petersen, In vivo measurement of cell-type-specific synaptic connectivity and synaptic transmission in layer 2/3 mouse barrel cortex, Neuron **85**, 68 (2015).

[75] S. C. Seeman, L. Campagnola, P. A. Davoudian, A. Hoggarth, T. A. Hage, A. Bosma-Moody, C. A. Baker, J. H. Lee, S. Mihalas, C. Teeter, *et al.*, Sparse recurrent excitatory connectivity in the microcircuit of the adult mouse and human cortex, elife **7**, e37349 (2018).

[76] L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli, M.-H. Kim, T. Hage, T. Braun, L. Alfiler, J. Andrade, P. Bohn, R. Dalley, A. Henry, S. Kebede, A. Mukora, D. Sandman, G. Williams, R. Larsen, C. Teeter, T. L. Daigle, K. Berry, N. Dotson, R. Enstrom, M. Gorham, M. Hupp, S. D. Lee, K. Ngo, P. R. Nicovich, L. Potekhina, S. Ransford, A. Gary, J. Goldy, D. McMillen, T. Pham, M. Tieu, L. Siverts, M. Walker,

C. Farrell, M. Schroedter, C. Slaughterbeck, C. Cobb, R. Ellenbogen, R. P. Gwinn, C. D. Keene, A. L. Ko, J. G. Ojemann, D. L. Silbergeld, D. Carey, T. Casper, K. Crichton, M. Clark, N. Dee, L. Ellingwood, J. Gloe, M. Kroll, J. Sulc, H. Tung, K. Wadhwani, K. Brouner, T. Egdorf, M. Maxwell, M. McGraw, C. A. Pom, A. Ruiz, J. Bomben, D. Feng, N. Hejazinia, S. Shi, A. Szafer, W. Wakeman, J. Phillips, A. Bernard, L. Esposito, F. D. D'Orazi, S. Sunkin, K. Smith, B. Tasic, A. Arkhipov, S. Sorensen, E. Lein, C. Koch, G. Murphy, H. Zeng, and T. Jarsky, Local connectivity and synaptic dynamics in mouse and human neocortex, Science **375**, eabj5861 (2022), https://www.science.org/doi/pdf/10.1126/science.abj5861.

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

[78] A. Khintchine, Korrelationstheorie der stationären stochastischen prozesse, Mathematische Annalen **109**, 604 (1934).

[79] P. Lévy and P. Lévy, *Théorie de l'addition des variables aléatoires* (Gauthier-Villars, 1954).

[80] B. F. Qaqish, A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations, Biometrika **90**, 455 (2003).

[81] E. Niebur, Generation of synthetic spike trains with defined pairwise correlations, Neural Computation **19**, 1720 (2007).

[82] J. H. Macke, L. Buesing, J. P. Cunningham, B. M. Yu, K. V. Shenoy, and M. Sahani, Empirical models of spiking in neural populations, Advances in neural information processing systems **24** (2011).

[83] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, Spatio-temporal correlations and visual signalling in a complete neuronal population, Nature **454**, 995 (2008).

[84] I. M. Park, E. W. Archer, K. Latimer, and J. W. Pillow, Universal models for binary spike patterns using centered dirichlet processes, Advances in neural information processing systems **26** (2013).

[85] L. Theis, A. M. Chagas, D. Arnstein, C. Schwarz, and M. Bethge, Beyond glms: a generative mixture modeling approach to neural system identification, PLoS computational biology **9**, e1003356 (2013).

[86] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, Nature **440**, 1007 (2006).

[87] E. Granot-Atedgi, G. Tkačik, R. Segev, and E. Schneidman, Stimulus-dependent maximum entropy models of neural population codes, PLoS computational biology **9**, e1002922 (2013).

[88] S. M. Bohté, H. Spekreijse, and P. R. Roelfsema, The effects of pair-wise and higher-order correlations on the firing rate of a postsynaptic neuron, Neural Computation **12**, 153 (2000).

[89] S.-i. Amari, H. Nakahara, S. Wu, and Y. Sakai, Synchronous firing and higher-order interactions in neuron pool, Neural computation **15**, 127 (2003).

[90] A. Kuhn, A. Aertsen, and S. Rotter, Higher-order statistics of input ensembles and the response of simple model neurons, Neural computation **15**, 67 (2003).

[91] B. Staude, S. Rotter, and S. Grün, Cubic: cumulant based inference of higher-order correlations in massively parallel spike trains, Journal of computational neuroscience **29**, 327 (2010).

[92] B. Staude, S. Rotter, *et al.*, Higher-order correlations in non-stationary parallel spike trains: statistical modeling and inference, Frontiers in computational neuroscience , 16 (2010).

[93] N. Bäuerle and R. Grübel, Multivariate counting processes: copulas and beyond, ASTIN Bulletin: The Journal of the IAA **35**, 379 (2005).

[94] J. Trousdale, Y. Hu, E. Shea-Brown, and K. Josić, A generative spike train model with time-structured higher order correlations, Frontiers in computational neuroscience **7**, 84 (2013).

[95] M. Shelley, D. McLaughlin, R. Shapley, and J. Wielaard, States of high conductance in a large-scale model of the visual cortex, Journal of computational neuroscience **13**, 93 (2002).

[96] M. Rudolph and A. Destexhe, Characterization of subthreshold voltage fluctuations in neuronal membranes, Neural Computation **15**, 2577 (2003).

[97] A. Kumar, S. Schrader, A. Aertsen, and S. Rotter, The high-conductance state of cortical networks, Neural computation **20**, 1 (2008).

[98] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, Vol. 1 (Elsevier, 1992).

[99] H. Risken, Fokker-planck equation, in *The Fokker-Planck Equation* (Springer, 1996) pp. 63–95.

[100] K. Itô, 109. stochastic integral, Proceedings of the Imperial Academy **20**, 519 (1944).

[101] F. Baccelli and T. Taillefumier, Replica-mean-field limits for intensity-based neural networks, SIAM Journal on Applied Dynamical Systems **18**, 1756 (2019).

[102] F. Baccelli and T. Taillefumier, The pair-replica-mean-field limit for intensity-based neural networks, SIAM Journal on Applied Dynamical Systems **20**, 165 (2021).

[103] L. Yu and T. O. Taillefumier, Metastable spiking networks in the replica-mean-field limit, PLoS Computational Biology **18**, e1010215 (2022).

[104] F. Baccelli and P. Brémaud, The palm calculus of point processes, in *Elements of Queueing Theory* (Springer, 2003) pp. 1–74.

[105] M. London, A. Roth, L. Beeren, M. Häusser, and P. E. Latham, Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex, Nature **466**, 123 EP (2010).

[106] L. F. Abbott and C. van Vreeswijk, Asynchronous states in networks of pulse-coupled oscillators, Phys. Rev. E **48**, 1483 (1993).

[107] J. Baladron, D. Fasoli, O. Faugeras, and J. Touboul, Mean-field description and propagation of chaos in networks of hodgkin-huxley and fitzhugh-nagumo neurons, The Journal of Mathematical Neuroscience **2**, 10 (2012).

[108] J. Touboul, G. Hermann, and O. Faugeras, Noise-induced behaviors in neural mean field dynamics, SIAM Journal on Applied Dynamical Systems **11**, 49 (2012).

[109] P. Robert and J. Touboul, On the dynamics of random neuronal networks, J. Stat. Phys. **165**, 545 (2016).

[110] B. B. Averbeck, P. E. Latham, and A. Pouget, Neural correlations, population coding and computation, Nature Reviews Neuroscience **7**, 358 EP (2006).

[111] A. S. Ecker, P. Berens, A. S. Tolias, and M. Bethge, The effect of noise correlations in populations of diversely tuned neurons,

Journal of Neuroscience **31**, 14272 (2011), http://www.jneurosci.org/content/31/40/14272.full.pdf.

[112] Y. Hu, J. Zylberberg, and E. Shea-Brown, The sign rule and beyond: Boundary effects, flexibility, and noise correlations in neural population codes, PLOS Computational Biology **10**, 1 (2014).

[113] G. Buzsáki and K. Mizuseki, The log-dynamic brain: how skewed distributions affect network operations, Nature Reviews Neuroscience **15**, 264 (2014).

[114] R. Iyer, V. Menon, M. Buice, C. Koch, and S. Mihalas, The influence of synaptic weight distribution on neuronal population dynamics, PLoS computational biology **9**, e1003248 (2013).

[115] A. Amarasingham, S. Geman, and M. T. Harrison, Ambiguity and nonidentifiability in the statistical analysis of neural codes, Proceedings of the National Academy of Sciences **112**, 6455 (2015).

[116] B. Scholl, C. I. Thomas, M. A. Ryan, N. Kamasawa, and D. Fitzpatrick, Cortical response selectivity derives from strength in numbers of synapses, Nature **590**, 111 (2021).

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

[118] J. de la Rocha, B. Doiron, E. Shea-Brown, K. Josić, and A. Reyes, Correlation between neural spike trains increases with firing rate, Nature **448**, 802 EP (2007).

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

[120] A.-S. Sznitman, Topics in propagation of chaos, in *École d'Été de Probabilités de Saint-Flour XIX—1989*, Lecture Notes in Math., Vol. 1464 (Springer, Berlin, 1991) pp. 165–251.

[121] X. Erny, E. Löcherbach, and D. Loukianova, Conditional propagation of chaos for mean field systems of interacting neurons, Electronic Journal of Probability **26**, 1 (2021).