Training recurrent spiking neural networks in strong coupling regime

Christopher M. Kim and Carson C. Chow

Laboratory of Biological Modeling, NIDDK/NIH, Bethesda, MD

Abstract

Recurrent neural networks trained to perform complex tasks can provide insights into the dynamic mechanism that underlies computations performed by cortical circuits. However, due to a large number of unconstrained synaptic connections, the recurrent connectivity that emerges from network training may not be biologically plausible. Therefore, it remains unknown if and how biological neural circuits implement dynamic mechanisms proposed by the models. To narrow this gap, we developed a training scheme that, in addition to achieving learning goals, respects the structural and dynamic properties of a standard cortical circuit model, i.e., strongly coupled excitatory-inhibitory spiking neural networks. By preserving the strong mean excitatory and inhibitory coupling of initial networks, we found that most of trained synapses obeyed Dale's law without additional constraints, exhibited large trial-to-trial spiking variability, and operated in inhibition-stabilized regime. We derived analytical estimates on how training and network parameters constrained the changes in mean synaptic strength during training. Our results demonstrate that training recurrent neural networks subject to strong coupling constraints can result in connectivity structure and dynamic regime relevant to cortical circuits.

Introduction

With recent advances in machine learning, recurrent neural networks are being applied to a wide range of problems in neuroscience as a tool for discovering dynamic mechanisms underlying biological neural circuits. The activity of network models trained either on tasks that animals perform or to directly generate neural activity have been shown to be consistent with neural recordings. Examples include context-dependent computation (Mante et al., 2013), neural responses in motor cortex (Sussillo et al., 2015), decision-making with robust transient activities (Chaisangmongkon et al., 2017), and flexible timing by temporal scaling (Wang et al., 2018). However, due to the large number of parameters, the network connectivities are often not constrained to be biologically plausible, thus it remains unclear if and how biological neural circuits could implement the proposed dynamic mechanisms.

One approach for imposing biological constraints is to develop model architectures based on detailed connectivity structure of the neural system. The complete connectome is available for certain invertebrates such as Drosophila, and it is possible to build recurrent neural network models constrained by the detailed structure of the connectome (Litwin-Kumar and Turaga, 2019; Eschbach et al., 2020; Zarin et al., 2019). In cortical regions of higher animals, however, obtaining a connectome and applying it for computational modeling purposes pose significant technical and conceptual challenges. Therefore, it is important to develop alternative methods for constraining the connectivity in a biologically plausible manner independently of the detailed connectome structure.

Recent studies have applied training methods to rate-based and spiking cortical circuits of excitatory and inhibitory neurons (Song et al., 2016; Nicola and Clopath, 2017; DePasquale et al., 2016; Kim et al., 2019) but the focus was to preserve the excitatory-inhibitory structure (i.e. Dale's law) without constraining the synaptic strength to a strong coupling regime (but see Ingrosso and Abbott (2019); Baker et al. (2019)) that exhibits chaotic dynamics (Sompolinsky et al., 1988), asynchronous spiking (van Vreeswijk and Sompolinsky, 1996; Renart et al., 2010; Rosenbaum et al., 2017), spiking rate fluctuations (Ostojic, 2014), and inhibition-stabilization (Tsodyks et al., 1997) as observed in neural recordings. Here, we show that by preserving the mean excitatory and inhibitory synaptic weights of a strongly coupled network, trained networks can exhibit large trial-to-trial variability in spiking activities, respect Dale's law in most synapses without imposing strict constraints, and show a paradoxical phenomenon found in inhibition-stabilized networks. We provided analytical estimates on how the network and training parameters constrain the size of synaptic updates during network training. Our findings demonstrate that including synaptic constraints that maintain strong coupling strength allows spiking networks to respect dynamic properties of standard cortical circuit models in addition to achieving the learning goal.

Results

1 Learning under excitatory and inhibitory synaptic constraints

The initial network consisted of an equal number of excitatory and inhibitory leaky integrate-and-fire neurons ($N_E, N_I = 500$) randomly connected with connection probability $p = 0.1$. The recurrent synaptic weights W_{ij} and the external inputs were sufficiently strong such that the total excitatory inputs to neurons, i.e. the sum of recurrent excitatory current and external input, exceeded spike threshold, and inhibitory feedback was necessary to prevent run-away excitation (Fig. 1A.i). In this parameter regime, neurons emitted spikes asynchronously and irregularly (Fig. 1A.ii), and had large trial-to-trial variability as measured by the Fano factor of spike counts in a 500 ms time window across trials starting at random initial conditions (Fig. 1A.iii). The synaptic connections to every neuron were statistically identical in that the presynaptic neurons were randomly selected, but the number of excitatory (pN_E) and inhibitory (pN_I) synaptic connections were fixed and their weights were constant depending only the the connections type, $W_{\alpha\beta}, \alpha, \beta \in \{E, I\}$ (i.e., fixed indegree and constant weights). This network con-

figuration resulted in a homogeneous firing rate distribution (Fig. 1A.iv). Initial networks with non-identical connectivity statistics that generate wide rate distributions were also investigated in Section. 6.

In the present study, spiking neural networks were trained to generate desired recurrent activity patterns. In related work, Rajan et al. (2016) trained a recurrent rate network to generate sequential activity observed in the posterior parietal cortex in mice, Laje and Buonomano (2013) stabilized inherent chaotic trajectories of the initial rate network, and Kim and Chow (2018) generated arbitrarily complex activity patterns in recurrent spiking networks. We point out that these tasks are different from the standard machine learning tasks, such as image classification, in that individual neurons within the network learn to generate certain activity patterns. Other studies have investigated performing machine learning tasks in spiking networks (Nicola and Clopath, 2017; Huh and Sejnowski, 2018).

Specifically, our goal was to train the synaptic current $u_i(t)$ to each neuron $i = 1, ..., N$ such that it followed target activity pattern $f_i(t)$ defined on time interval $t \in [0, T]$. To trigger the target response, each neuron was stimulated by a constant input with random amplitude for 200 ms (Fig. 1B). We treated every neuron's synaptic current as a read-out, which made our task equivalent to training N recurrently connected read-outs. For spiking network models with current-based synapses, neuron i's synaptic current u_i can be expressed in terms of the spiking activities of other neurons $r_j, j = 1, ..., N$ through the exact relationship $u_i = \sum_j W_{ij} r_j$ (see Eqs. 11 and 14 for details). Therefore, we adjusted the incoming synaptic connections $W_{ij}, j = 1, ..., N$ to neuron i by the recursive least squares algorithm (Haykin, 1996) in order to generate the target activity. This training scheme allowed us to set up independent objective functions for each neuron and potentially update them in parallel.

The objective function included the error between the synaptic currents and target patterns, and, importantly, we considered two types of synaptic regularization The first type regularized the standard L_2-norm of the incoming synaptic weights to each neuron i,

$$C^i_{FORCE} = \int_0^T \big(f_i(t) - u_i(t)\big)^2 dt + \lambda \sum_j W_{ij}^2, \tag{1}$$

and $C_{FORCE} = \sum_{i=1}^N C^i_{FORCE}$ was the full objective function. The second type regularized the sums of excitatory and inhibitory synaptic weights, respectively, to each neuron i, in addition to the L_2-norm,

$$C^i_{ROWSUM} = \int_0^T \big(f_i(t) - u_i(t)\big)^2 dt + \lambda \sum_j W_{ij}^2 + \mu \sum_{\alpha \in \{E,I\}} \big(\sum_{j \in \alpha} W_{ij}\big)^2, \tag{2}$$

and $C_{ROWSUM} = \sum_{i=1}^N C^i_{ROWSUM}$ was the full objective function. The first type, which is equivalent to the algorithm of Sussillo and Abbott (Sussillo and Abbott, 2009) and used previously in spiking networks (Kim and Chow, 2018; Nicola and Clopath, 2017), will be referred to as FORCE training. The second type will be referred to as ROWSUM training since it regularizes the sum of excitatory and inhibitory weights, respectively, in each row of the connectivity matrix.

We note how the hyperparameters λ and μ can be interpreted. At first glance, it appears that large positive values of λ and μ should reduce the magnitude of synaptic weights in trained networks. However, the recursive least squares is different from the standard gradient descent algorithms in that the synaptic update rule is not derived from the gradient of cost function. Instead, it finds a critical point to the difference of gradients of cost function at two consecutive time points (see Methods / Section 2). This implies that large positive values of λ and μ penalize the size of synaptic updates, resulting in small synaptic updates or slow learning rate (see Sussillo and Abbott (2009) for further discussion). Therefore, the synaptic weights trained with large λ and μ stay close to their initial weights if trained by the recursive least squares algorithm.

We first considered simple target patterns consisting of sinusoidal waves $f_i(t) = A\sin(2\pi t/T + \phi_i) + \bar{u}_i$ defined on $t \in [0, T]$ with duration $T = 1000$ ms, random phase ϕ_i, fixed amplitude A and bias $\bar{u}_i$. The amplitude A was the standard deviation of synaptic currents in time averaged over all neurons, and the bias $\bar{u}_i$ for neuron i was the mean synaptic current to neuron i in the initial network. After adjusting the recurrent weights with two training schemes, we examined the differences in trained network activities. Notably, the large fluctuations of excitatory and inhibitory synaptic currents in the initial network were not observed in FORCE-trained networks. Instead, the total synaptic currents closely followed the target patterns with little error (Fig. 1C.i). Such synaptic currents led to highly regular spikes with little variability across trials starting at random initial conditions (Fig. 1C.ii), resulting in a small Fano factor (Fig. 1C.iii). In contrast, ROWSUM-trained networks, which were encouraged to preserve the initial excitatory and inhibitory synaptic weights, exhibited strong synaptic fluctuations, irregular spike trains and large trial-to-trial variability (Fig. 1D.i,ii,iii). Also, the firing rate distribution of trained neurons closely followed the firing rate distribution of untrained neurons, showing that sinusoidal targets minimally altered the firing rate distribution (Fig. 1A.iv).

2 ROWSUM constraint can preserve the mean excitatory and inhibitory synaptic strength

We examined the trained synaptic weights to better understand the underlying network structure that led to different levels of variability in neuron activities in the FORCE- and ROWSUM-trained networks. We found that the excitatory and inhibitory synaptic weights of the initial network were both redistributed around zero by the FORCE learning (Fig. 2A.i), and the fraction of synapses that violated Dale's law increased gradually as the number of training iteration increased (Fig. 2A.ii). The mean strength of excitatory and inhibitory synaptic connections to each neuron also moved towards zero (Fig. 2A.iii). This shift in synaptic weight distributions can be explained by the L_2-regularization term in Eq. 1 that allowed small synaptic weights, and as a consequence, the mean excitatory and inhibitory synaptic connections to each neuron were significantly smaller than those of the initial network. After removing synaptic connections that violated Dale's law, the FORCE-trained network was able to generate the target patterns with high accuracy (Fig. 2A.iv). These findings suggested that the network connectivity resulting from FORCE training was degenerating and highly tuned to produce the target

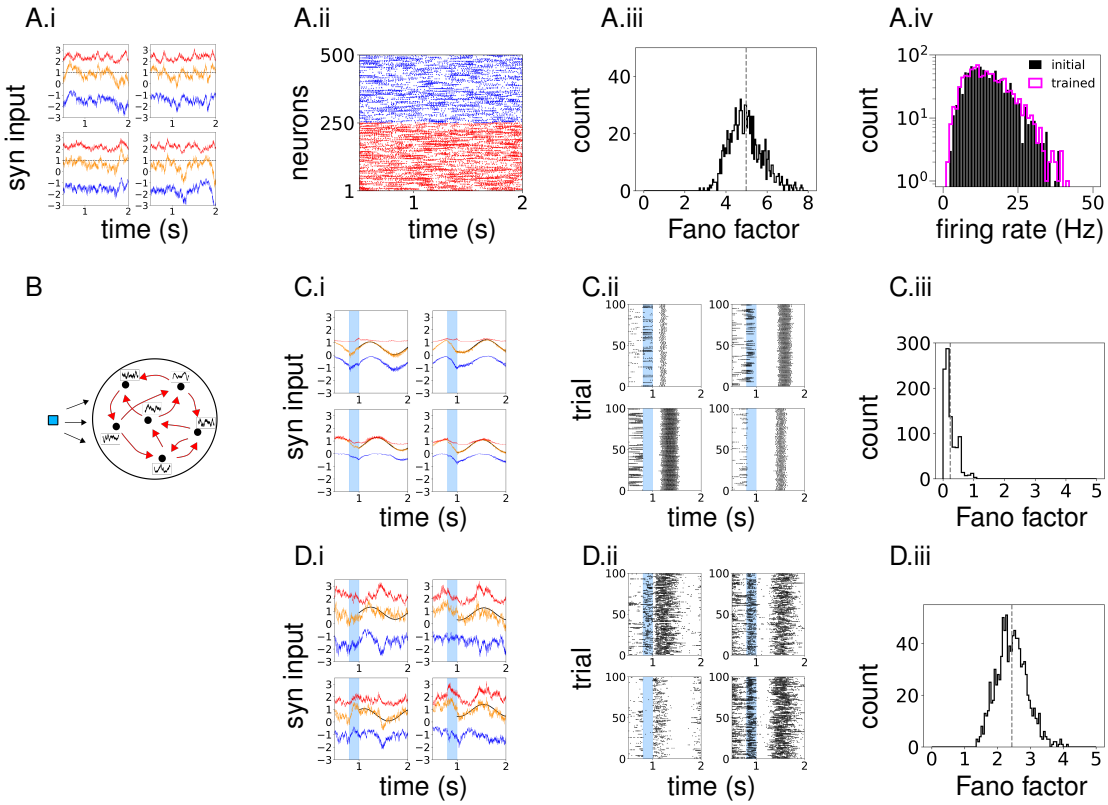

Figure 1: Comparison of FORCE- and ROWSUM-trained network activities. **(A)** Population activity of the initial network. (A.i) Synaptic inputs to example neurons; excitatory (red), inhibitory (blue) and the total (orange) inputs; external input is added to the excitatory and total inputs; dashed line indicates the spike threshold. (A.ii) Spike raster of sample neurons. (A.iii) Fano factor of all neurons; spike counts across trials in a 500 ms time window are used to compute the Fano factor and averaged over the training window to obtain a time-averaged Fano factor for each neuron. (A.iv) Firing rate distribution of neurons; initial network (black), ROWSUM-trained network (magenta). **(B)** Schematic of network training. Brief external input (blue) is applied to trigger learned response; individual neurons learn their own target patterns (black); recurrent synaptic connections (red) are modified to generate the target patterns. **(C)** Population activity of a FORCE-trained network. (C.i) Synaptic inputs to example neurons. (C.ii) Spike trains of the same neurons in (C.i) across trials starting at random initial conditions. (C.iii) Fano factor of trained neurons. **(D)** Population activity of a ROWSUM-trained network; Same color codes as in (C).

dynamics with minimal error.

For the ROWSUM training, we used the same network and training parameters as the FORCE training but included the additional penalty that constrained the sums of excitatory and inhibitory synaptic weights, respectively, to each neuron. We found that the excitatory and inhibitory weights of the trained synapses were distributed around their initial values (Fig. 2B.i), and the fraction synapses that violated Dale's law increased in the early stage of training but gradually decreased and stabilized as the training continued (Fig. 2B.ii). The mean strength of excitatory and inhibitory synaptic connections to each neuron after training deviated little from the initial mean synaptic weights (Fig. 2B.iii), demonstrating that the

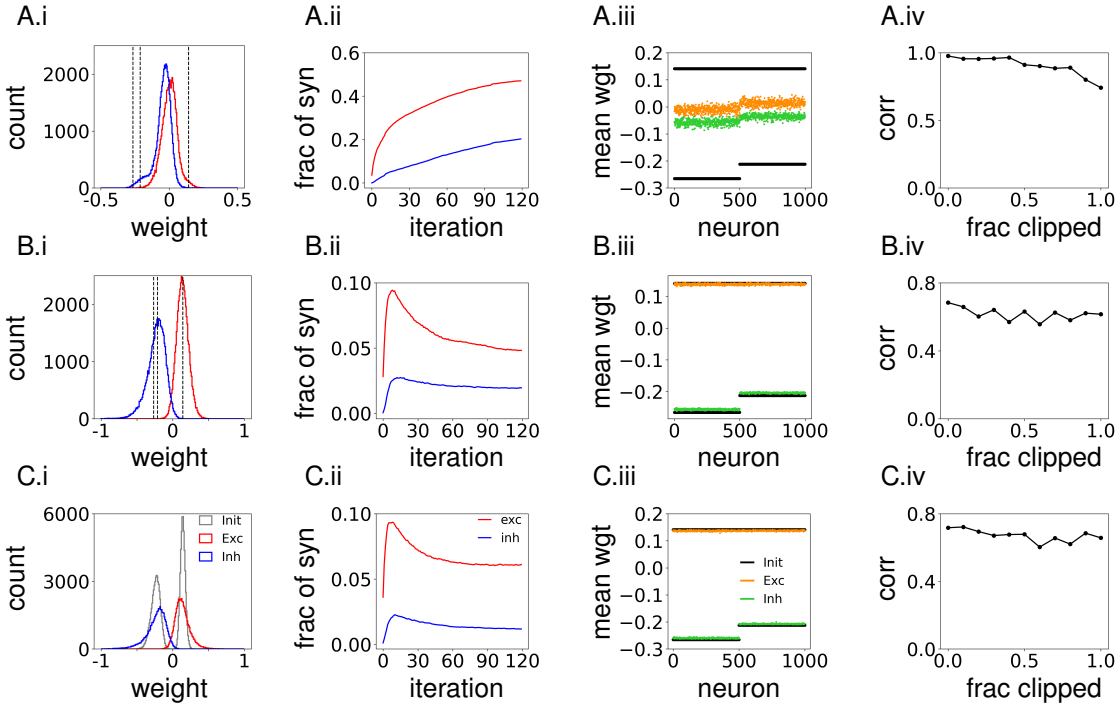

Figure 2: Network connectivity of trained networks. **(A)** FORCE-trained network. (A.i) Distribution of synaptic weights of the FORCE-trained network (120 training iterations); red (blue) shows synapses that are excitatory (inhibitory) in the initial network; dotted line shows the initial synaptic strength. (A.ii) Fraction of excitatory (red) and inhibitory (blue) synapses that violate Dale's law after each training iteration. (A.iii) Mean strength of synaptic connections to individual neurons after FORCE-training; orange (green) dots show connections that are excitatory (inhibitory) in the initial network; black dots show the initial mean synaptic weights. Mean excitatory and inhibitory weights to neurons were identical in the initial network since indegrees were fixed and synaptic weights were constant. (A.iv) Performance of trained networks, measured as the correlation between target patterns and synaptic activities, when synapses violating Dale's law are removed; fraction of clipped synapses equals 1 if all synapses violating Dale's are removed. **(B)** Same as in (A), but for a ROWSUM-trained network with fixed indegree and constant weights. **(C)** Same as in (A), but for a ROWSUM-trained network with random indegrees and Gaussian weights; correction terms are added to reconcile differences in synaptic weights each neuron receive.

ROWSUM-regularization allowed the trained network to preserve the mean excitatory and inhibitory strength of the strongly coupled initial network. The fraction of synapses that violated Dale's law after training (approximately 5% and 2% of excitatory and inhibitory connections, respectively) was significantly smaller than a FORCE-trained network (40% and 20%, respectively), and removing them had only a small impact on the network performance (Fig. 2B.iv).

Next, we investigated if ROWSUM training was applicable to initial networks with random indegrees and weights. To this end, we constructed an Erdos-Renyi graph where each synaptic connection was generated independently with probability p, i.e., $\Pr(W_{ij}) = p$ if neuron j connects to neuron i, resulting in a connectivity matrix with a non-constant indegree distribution. The synaptic weight of each connection was sampled from Gaussian distribution with mean $W_{\alpha\beta}$ and standard deviation $W_{\alpha\beta}/5$ for $\alpha, \beta \in \{E, I\}$ to ensure the initial network respected Dale's

law. In order to reconcile differences in synaptic weights neurons received, we added a correction term to nonzero elements in each row of the connectivity matrix, such that the sum of excitatory and inhibitory synaptic weights to neuron $i \in \alpha$ was equal to $pN_E w_{\alpha E}$ and $pN_I w_{\alpha I}$, respectively. Including such correction terms resulted in a homogeneous rate distribution, as opposed to a wide rate distribution, and trained networks qualitatively agreed with those initialized with fixed indegrees and constant weights (Fig. 2C.i-iv).

These findings demonstrated that constraining the sums of excitatory and inhibitory synaptic weights, respectively, allowed the individual synapses to learn target patterns while keeping the mean synaptic strength close to the initial values. Since the excitatory and inhibitory synaptic weights of the initial network were strong enough to generate variable spiking activity, the trained network was able to inherit the dynamic property of the initial network.

Before turning to limitations of training complex target patterns and initial networks without the correction terms in Sections. 4 and 6, we examined if other aspects of the initial network dynamics were inherited by the trained network when the mean excitatory and inhibitory synaptic weights were preserved.

3 ROWSUM constrained networks inherit dynamic features of inhibition-stabilized network

The initial network was set up to be an inhibition-stabilized network (ISN), where strong recurrent excitatory connections were stabilized by inhibitory feedback without which the network would exhibit run-away excitation (Tsodyks et al., 1997; Sanzeni et al., 2019; Sadeh and Clopath, 2020). An ISN exhibits a paradoxical phenomenon where an external stimulus that excites inhibitory neurons decreases their firing rates. Although the stimulus can transiently increase the firing rates of inhibitory neurons, this activity inhibits excitatory neurons that in turn reduce recurrent excitatory input to inhibitory neurons. The reduction in recurrent excitation is larger than the external stimulus resulting in a net decrease in excitatory input to inhibitory neurons. Such large reduction in recurrent excitation occurs because strongly connected excitatory neurons respond to changes in input with large gain (Tsodyks et al., 1997; Sanzeni et al., 2019; Sadeh and Clopath, 2020).

To verify that the initial network operated in the inhibition-stabilized regime, we injected additional excitatory stimulus to inhibitory neurons (Fig. 3A) and found that both excitatory and inhibitory population firing rates decreased (shaded region in Fig. 3B.i). As the external stimulus to inhibitory neurons was further increased, the excitatory population rate decreased gradually, and once the excitatory population activity was fully suppressed the firing rate of inhibitory neurons started increasing with the external stimulus (Fig. 3B.ii). These dynamic features are consistent with theoretical analysis of ISNs that the paradoxical phenomenon occurs only within a limited range of stimulus intensity. In a strong stimulus regime, the network is effectively reduced to a single inhibitory population, therefore the paradoxical phenomenon can no longer exist.

To better understand how individual neurons responded to the external stimulus, we stimulated inhibitory neurons over multiple trials and examined the trial-averaged firing rates of individual neurons with and without the stimulus. The rela-

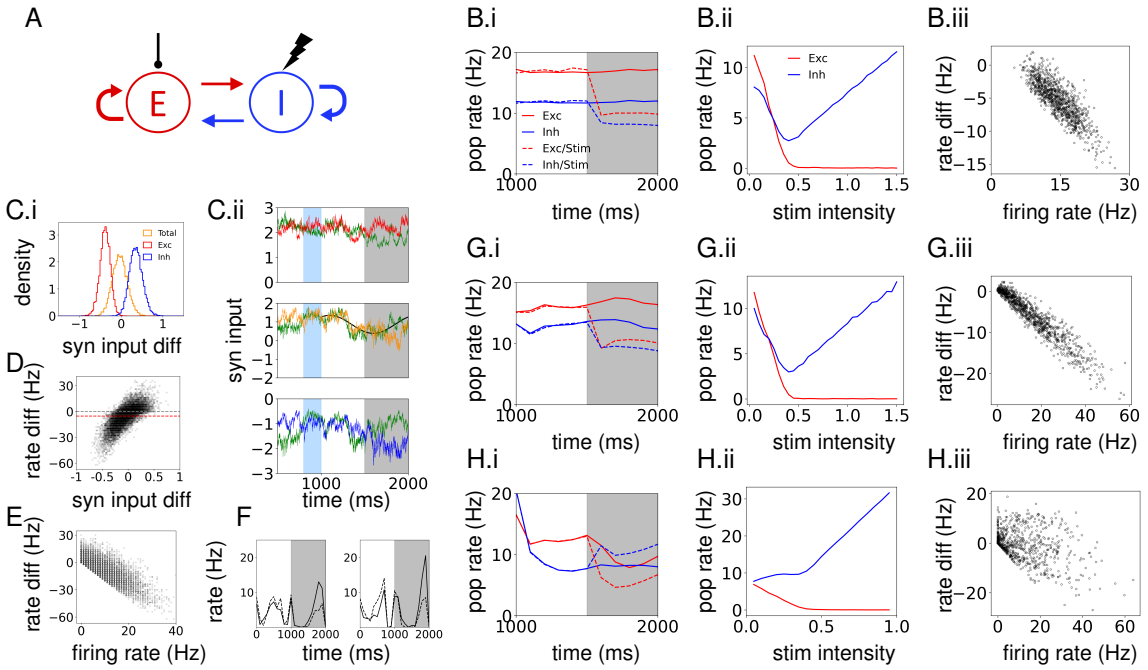

Figure 3: ROWSUM-trained networks exhibit dynamic features of inhibition-stabilization. Shaded region shows the time period during which inhibitory neurons are excited by additional external stimulus. **(A)** Schematic of network structure where inhibitory neurons are stimulated. **(B)** Initial network activities in response to stimulating inhibitory neurons. (B.i) Excitatory (red) and inhibitory (blue) population firing rates with (dotted) and without (solid) the additional stimulus to inhibitory neurons. (B.ii) Population firing rates as a function of stimulus intensity. (B.iii) Trial-averaged firing rates of individual neurons without stimulus vs. changes in neurons firing rates due to the stimulus. **(C)** Synaptic inputs to individual neurons. (C.i) Distribution of changes in time-averaged synaptic inputs to individual neurons due to the stimulus. For excitatory (inhibitory) inputs, negative (positive) value implies reduction in the average input; excitatory (red), inhibitory (blue) and total (orange). (C.ii) Synaptic inputs to a neuron in an unstimulated network is compared to those in a stimulated network (green). **(D)** Changes in total synaptic input vs. changes in the number of emitted spikes due to the stimulus; dots show individual neurons. **(E)** The number of emitted spikes in an unstimulated network vs. changes in the number of emitted spikes due to the stimulus; dots show individual neurons. **(F)** Trial-averaged firing rate of sample neurons with (dotted) and without (solid) the stimulus. **(G)** Same as in (B), but for a ROWSUM-trained network. **(H)** Same as in (B), but for a FORCE-trained network.

tionship between the trial-averaged firing rate of unstimulated neurons and changes in their firing rates due to the stimulus revealed that neurons with high baseline firing rates resulted in large reduction in their spiking activities (Fig. 3B.iii).

Since the ROWSUM training preserved the mean synaptic strength of the initial network, we investigated if, in addition to learning target activity patterns, the trained network preserved unique features of initial network's population dynamics characterized by the strong excitation and feedback inhibition. To test if the ROWSUM-trained network exhibited the dynamic features of ISN, we evoked the learned activity patterns in the trained network, and then injected external stimulus that excited all inhibitory neurons during the second half of the target patterns. We

found that, although the temporal average of recurrent excitatory and inhibitory inputs to individual neurons were both reduced in their magnitudes due to the stimulus (Fig. 4Ci, red and blue), the distribution of total synaptic inputs to neurons were centered around zero and did not exhibit apparent hyperpolarizing effects at the population level (Fig. 4Ci, orange). When we examined single trial responses with and without the stimulus, neurons whose total synaptic inputs were hyperpolarized by the stimulus (negative side of the orange curve in Fig. 4C.i) had large reduction in their spiking activities (negative side in Fig. 4D) in comparison to the increased spiking activities (positive side in Fig. 4D) in other neurons whose total synaptic inputs were depolarized by the stimulus (positive side of orange curve in Fig. 4C.i). As a result, the population firing rate decreased by 5 Hz (red line in Fig. 4D) consistently with the firing rate reduction seen in the stimulated initial network (Fig. 4B.i). In addition, neurons with high spiking rate in the unstimulated network had large reduction in their emitted spikes (Fig. 3E) in agreement with the trial-averaged response in the initial network (Fig. 3B.iii).

The effects of exciting inhibitory neurons were observed more clearly in trial-averaged firing rates in individual neurons and population rates. Without additional stimulus to inhibitory neurons, the trial-averaged firing rates followed the target rate patterns as trained (Fig. 3F). However, when the external stimulus was injected to excite the inhibitory neurons, their firing rates decreased compared to those of unstimulated neurons and similar effects was observed in both the excitatory and inhibitory population firing rates (Fig. 3G.i). We also found that the paradoxical effect was observed in a limited stimulus range for the ROWSUM-trained network (Fig. 3G.ii), and the neurons with high trial-averaged firing rates in an unstimulated network had larger reduction in their spiking activities due to the stimulus (Fig. 3G.iii). These findings show that the ROWSUM-trained network inherited the dynamic features of the initial network.

To test if the FORCE-trained network exhibited similar dynamic features, we trained an inhibition-stabilized network without the ROWSUM regularization term. Injecting excitatory inputs to inhibitory neurons in the FORCE-trained network did not elicit the paradoxical effect: the inhibitory population rate increased and the excitatory population rate decreased (Fig. 3H.i). Moreover, the paradoxical effect was not observed over a wide range of stimulus intensity (Fig. 3H.ii). Also, a large number of neurons increased their trial-averaged firing rates due to the stimulus (Fig. 3H.iii). These findings suggest that the FORCE-trained network did not operate in the inhibition-stabilized regime.

4 Limitations of constraining excitatory and inhibitory synaptic weights

So far, (1) the initial network connectivity was adjusted such that the sum of excitatory and inhibitory synaptic weights to neurons were uniform, and (2) the target patterns were simple sinusoidal perturbations of the initial synaptic activities. Such initial network setup and target patterns led to relatively homogeneous firing rate distribution. We asked if the ROWSUM training was still effective when either the initial network connectivity or the set of target patterns was no longer fine-tuned. To this end, we investigated if both the initial network setup and the choice of target functions can limit the effectiveness of ROWSUM training. In this section, we

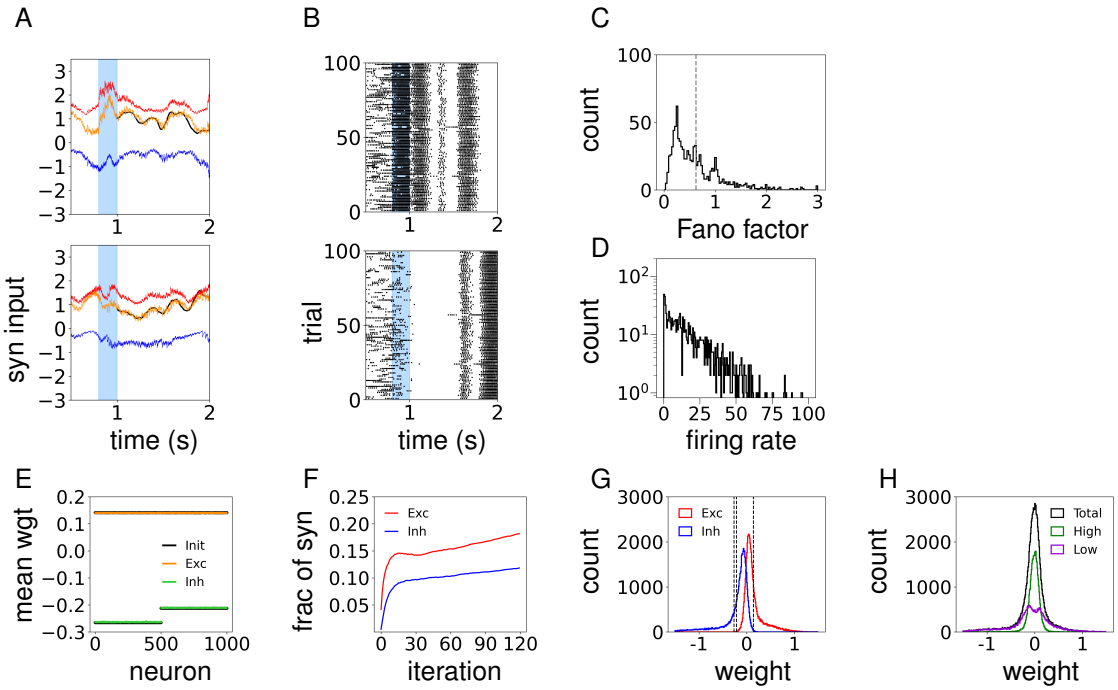

Figure 4: Learning complex activity patterns under the ROWSUM constraint on excitatory and inhibitory synapses. Target patterns consist of trajectories generated from Ornstein-Ulenbeck process. **(A)** Synaptic inputs to trained neurons; excitatory (red), inhibitory (blue) and the total (orange) inputs; external input is added to the excitatory and total inputs. **(B)** Spike trains of the same neurons in (A) across trials. **(C)** Fano factor of trained neurons; dashed line indicates the mean. **(D)** Distribution of neurons firing rates during the trained time window. **(E)** Mean strength of trained synaptic weights to each neuron; orange (green) dots show the mean strength of synaptic weights that are excitatory (inhibitory) prior to training; black dots show the mean strength of initial excitatory (positive) and inhibitory (negative) synaptic weights. **(F)** Fraction of excitatory (red) and inhibitory (blue) synapses that violate Dale's law after each training iteration. **(G)** Distribution of trained synaptic weights; red (blue) shows synapses that are excitatory (inhibitory) in the initial network; dotted line shows the initial synaptic strength. **(H)** Distribution of trained synaptic weights; all the synapses (black); outgoing synapses from top (green) and bottom (purple) 40% of neurons ordered by their firing rates.

considered initial networks whose total excitatory and inhibitory synaptic weights across neurons were uniform as before, but neurons were trained on complex target patterns to encourage a wide range of activity patterns. Initial networks with non-identical synaptic weights across neurons is discussed in Section. 6.

To generate complex target patterns we sampled trajectories from Ornstein-Ulenbeck process with decay time constant 200 ms and applied the moving-average with 100 ms time window to smooth out fast fluctuations. The amplitude of trajectories was scaled so that their standard deviation matched with that of synaptic currents of initial network. Each neuron was trained on independently generated trajectories with the ROWSUM constraints on the sum of excitatory and inhibitory synaptic weights, respectively. We found that, although the trained neurons successfully tracked the complex trajectory patterns (Fig. 4A), the trial-to-trial variability was relatively low (Figs. 4B, C) and the firing rate distribution was wide including many low firing rate neurons (Fig. 4D).

Examining the trained weights showed that the mean strength of excitatory and inhibitory synaptic connections to each neuron did not deviate from their initial values, showing that the ROWSUM constraints worked effectively (Fig. 4E). However, the fraction of synapses that violated Dale's law increased gradually with training iterations (Fig. 4F), and the distribution of synaptic weights of the trained network showed that a large number of synapses were degenerating (Fig. 4G). In particular, the synapses outgoing from neurons with high firing rate were concentrated near zero, in contrast to those from neurons with low firing rates (Fig. 4H).

To better understand how the trained synaptic weights might depend on the firing rate distribution of presynaptic neurons, we analytically estimated the size of synaptic weight updates assuming presynaptic neurons followed simplified rate distributions (see Methods / Section 4 for details). The analytical expressions suggested that, for a homogeneous rate distribution, i.e., all neurons have the same rate, the size of synaptic update is small, i.e., $O(1/(\mu K))$, due to the cancellation effect by the ROWSUM-regularization (Eq. 31). On the other hand, for an inhomogeneous rate distribution consisting of high and low firing rate neurons, the size of synaptic update is large, i.e., $O(1)$, and the synaptic connections from two neuron types are updated in opposite directions (Eq. 32). The analytical results from the inhomogeneous rate distribution were consistent with the full ROWSUM-training in that the synapses from high firing rate neurons concentrated near zero while those from low firing rate neurons had a wide distribution including strong synapses (Fig. 4H). Moreover, analysis of the homogeneous rate distribution suggested that applying the ROWSUM constraints to subpopulations that share similar activity levels can reduce deviations of individual synaptic weights from their mean values.

5 ROWSUM constraints on multiple subpopulations

In order to overcome the limitations of applying ROWSUM constraints to a population of neurons with a wide firing rate distribution, we further divided the neurons into multiple subpopulations where neurons within a subpopulatioin shared similar level of spiking activities. First, to estimate individual neuron's firing rate after training, external inputs that followed the target patterns were injected to each neuron. Then, the excitatory and inhibitory neurons were respectively ordered by their time-averaged firing rates and divided into M subpopulations of equal sizes by assigning the αth N_E/M excitatory and N_I/M inhibitory neurons to E_α and I_α, respectively. Finally, the ROWSUM constraint was applied to synapses from each subpopulation $E_\alpha, I_\alpha, \alpha = 1, ..., M$, separately, resulting in the following synaptic constraint for training neuron i:

$$\sum_{\alpha=1}^{M} \Big[\big(\sum_{j \in E_\alpha} W_{ij} \big)^2 + \big(\sum_{j \in I_\alpha} W_{ij} \big)^2 \Big]. \tag{3}$$

See Figure 5A for a schematic of constraining multiple subpopulations.

Overall, we found that the synaptic fluctuations (Fig. 5B) and trial-to-trial spiking variability (Figs. 5C, D; average Fano factor 1.5) increased when multiple subpopulations were constrained, in contrast to the network constrained by the excitatory and inhibitory connections only (Figs. 4B, C).

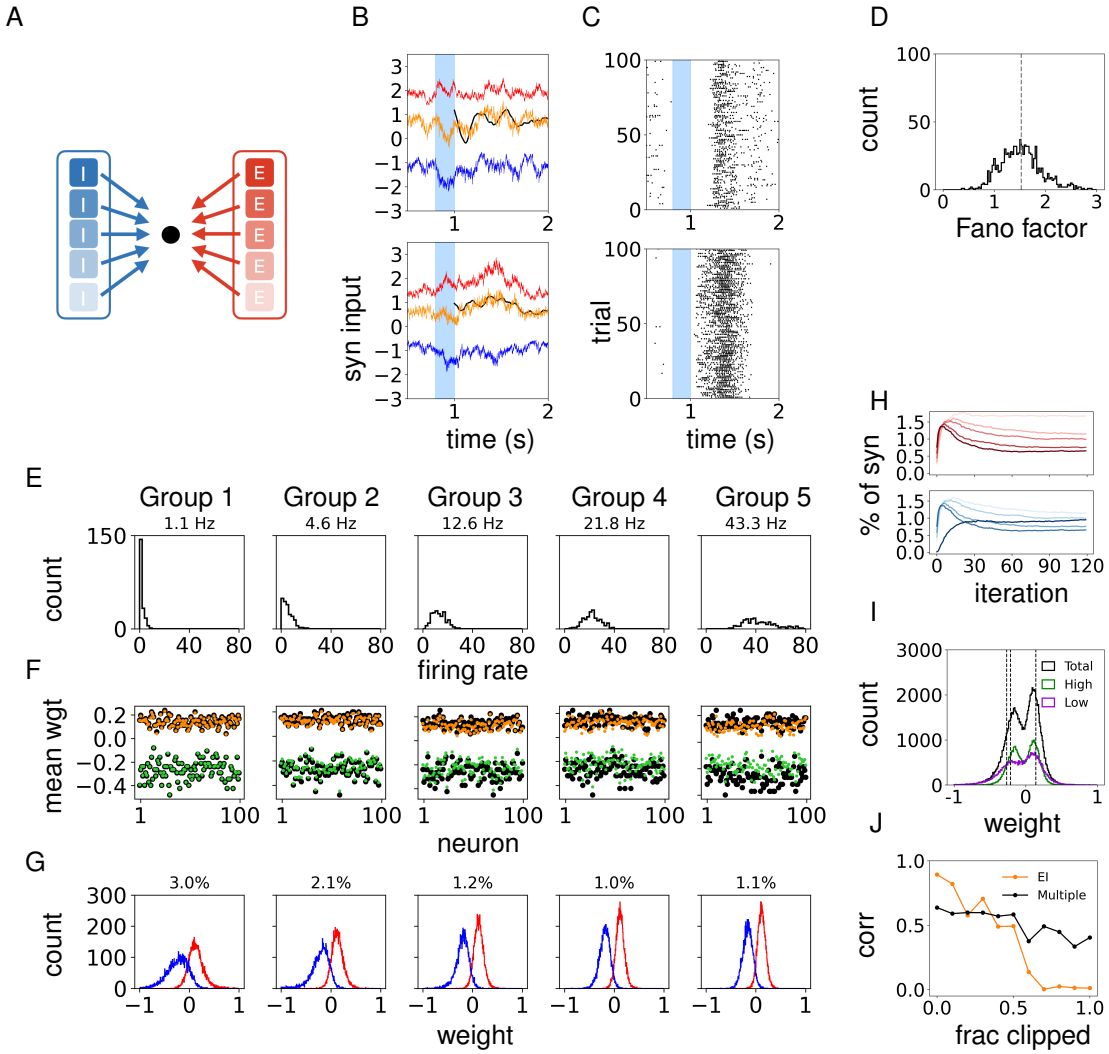

Figure 5: Learning complex activity patterns under the ROWSUM constraint on multiple subpopulations. Target patterns consist of trajectories generated from Ornstein-Ulenbeck process. **(A)** Schematic of constraining multiple subpopulations; black dot indicates one neuron; transparency indicates the within-group firing rate. **(B)** Excitatory (red), inhibitory (blue) and the total (orange) synaptic inputs to trained neurons. **(C)** Spike trains of the same neurons in (a) across multiple trials. **(D)** Fano factor of trained neurons; dashed line indicates the mean. **(E)-(G)** Excitatory and inhibitory neurons are ordered by firing rates and assigned to five subpopulations (i.e., 100 neurons in each group); low to high rate groups are shown from left to right; neurons in the αth excitatory and inhibitory subpopulations are shown in the same panel, Group α. (E) Distribution of firing rates of neurons in Group α; the within-group average firing rates shown. (F) Mean strength of synaptic connections that each neuron in the αth excitatory (orange) and inhibitory (green) subpopulations receive from other neurons in the same subpopulation; black dots show the mean strength of initial connections. (G) Synaptic weight distribution of outgoing connections from neurons in the αth excitatory (red) and inhibitory (blue) subpopulations. The percentage of synapses (with respect to the total number of synapses in the network) that violates Dale's law by the neurons in each group are shown. **(H)** Fraction of excitatory (red) and inhibitory (blue) synapses that violates Dale's law after each training iteration; each line represents outgoing synapses from neurons in the αth subpopulation; light to dark color shows subpopulation 1 to 5. **(I)** Distribution of trained synaptic weights (black); outgoing synapses from top (green) and bottom (purple) 40% of neurons ordered by firing rates; dashed lines indicate initial synaptic weights. **(J)** Performance of trained networks after removing synapses that violate Dale's law; constraints on multiple (black) and excitatory-inhibitory (orange) populations.

Next, we combined the αth excitatory and inhibitory subpopulations into one group and examined their features. The within-group average firing rates increased sequentially in agreement with how neurons were assigned to each group (Fig. 5E; α^{th} excitatory and inhibitory groups shown in the same panel). The mean strength of synaptic connections from low rate neurons showed little deviations from their initial mean values (Fig. 5F, Group 1), while synaptic connections from high rate neurons (Group 5) showed larger deviations. As will be further analyzed in Results / Section 7, the firing rates of presynaptic neurons and the inverse of the number synapses constrained under the ROWSUM regularization are among the factors that determine how much recursive least squares algorithm can change the mean synaptic weights. In light of this analysis, the high firing rates and the reduced number of neurons in a subpopulation provide an explanation why the mean synaptic weights from neurons in high rate groups deviated from the initial mean weights.

Importantly, the fraction of synapses that violated Dale's law was reduced (8%; Figs. 5G, H) compared to that found in the network trained under constrained excitatory and inhibitory connections (14%; Fig 4F). Also, the distribution of synaptic weights from the low and high firing rate neurons no longer exhibited drastic differences (Fig. 5I). Interestingly, the fraction of synapses that violated Dale's law was smaller in groups consisting of high firing rate neurons (see the percentages shown on the panels of Fig. 5G). Tracking the synapses that violated Dale's law during training revealed that, although such synapses increased rapidly during the early phase of training, they diminished quickly and stabilized to lower values, particularly in the high firing rate groups (Fig. 5H). Moreover, removing the synapses that violated Dale's law did not significantly affect the performance in networks trained under multiple subpopulation constraints; however, a network trained under the excitatory and inhibitory constraints suffered from reduced performance (Fig. 5J).

6 Training initial networks with random connectivity

In this section, we turn to initial networks for which the sums of excitatory and inhibitory synaptic weights to each neuron, respectively, are no longer identical across neurons. We constructed an Erdos-Renyi graph with connection probability p, which resulted in a non-constant indegree distribution, and assigned constant weights to each connection type $W_{\alpha\beta}$ for $\alpha, \beta \in \{E, I\}$. To revive the inactive neurons present in such a network, we selected bottom half of the neurons from their mean synaptic weight distribution and added a correction term to the incoming synaptic weights such that the sum matched the expected value $pN_\beta W_{\alpha\beta}$ (See Methods / Section 1 for details). The resulting network exhibited a wide firing rate distribution including many low firing rate neurons (Fig. 6A.i) and large trial-to-trial spiking variability (Figs. 6A.ii-iii).

To avoid compounding the effects of network initialization and target choices, we chose sinusoidal functions as the target patterns since they minimally altered the initial firing rate distribution. Then, we trained the initial network with two different ROWSUM constraints developed in Sections 4 and 5.

First, when only the excitatory and inhibitory synaptic connections to neurons were constrained, the fraction of synapses that violated Dale's law increased with training iterations (Fig. 6B.i) and the spiking activities of the trained network

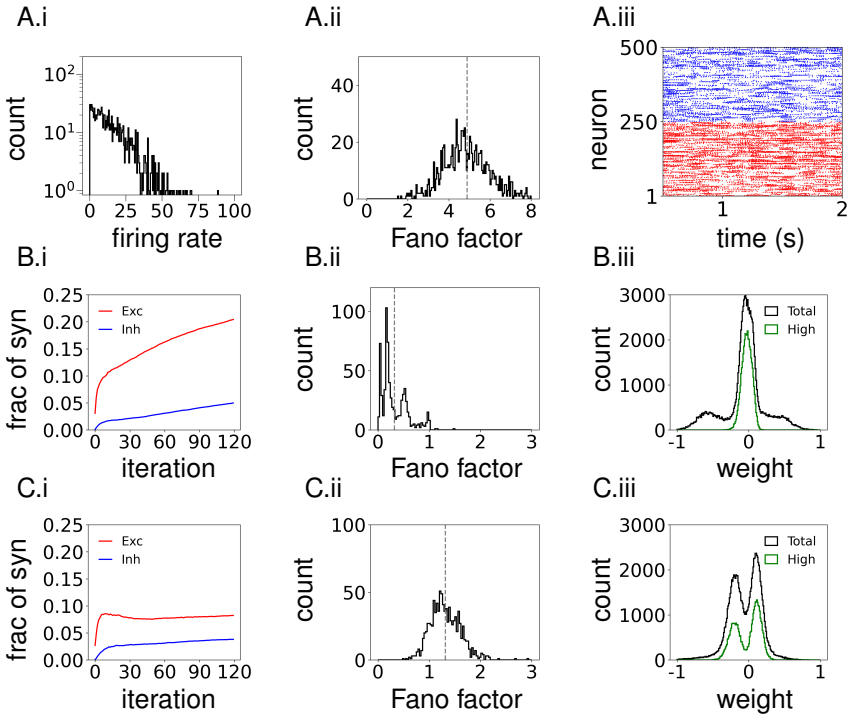

Figure 6: Training initial networks with random connectivity. Synaptic weights to neurons in the bottom half of mean synaptic weight distribution are corrected to their expected values $pN_\beta W_{\alpha\beta}$. Target patterns consist of sinusoidal waves with random phases and fixed amplitude. **(A)** Population activity of an initial network. (A.i) Distribution of neurons' firing rates. (A.ii) Distribution of neurons' Fano factor. (A.iii) Spike raster of sample neurons. **(B)** Network trained with the ROWSUM constraint on excitatory and inhibitory synapses. (B.i) Fraction of synapses that violate Dale's law after each training iteration. (B.ii) Fano factor of trained neurons. (B.iii) Synaptic weight distribution of a trained network (black); Outgoing synapses from top 40% of neurons ordered by firing rates (green). **(C)** Same as in (B), but for a network trained with the ROWSUM constraint on multiple subpopulations.

was highly regular across trials (Fig. 6B.ii). We found that, similarly to networks trained to learn complex target patterns (Fig. 4H), the synaptic connections from high firing rate neurons were degenerating (Fig. 6B.iii).

On the other hand, when the ROWSUM constraint was applied to multiple excitatory and inhibitory subpopulations as discussed in Section. 5 (i.e., five subpopulations for the excitatory and inhibitory neurons, respectively), the fraction of synapses violating Dale's law remained low (Fig. 6C.i) and the spiking variability was relatively large with the mean Fano factor close to 1.5 (Fig. 6C.ii). Moreover, the synaptic weight distribution, particularly the synaptic connections from high rate neurons, did not converge towards zero (Fig. 6C.iii). These training results demonstrated that applying the ROWSUM constraint on multiple subpopulations can be an effective training scheme even when the the initial network included random connectivity that generated a wide firing rate distribution.

7 Analytical upper bound on synaptic weight updates

Since retaining the strong coupling strength of the initial network was critical for generating large synaptic fluctuations and inhibition-dominant dynamics, we further inquired how the ROWSUM-training was able to tightly constrain the mean synaptic weights, such that the mean excitatory and inhibitory synaptic weights to each neuron stayed close to their initial values (Figs. 2B,C, 4E and 5F). In particular, we sought to derive an analytical upper bound on the changes in the sums of excitatory and inhibitory synaptic weights, respectively, in terms of the network and training parameters.

For the analysis, we considered a vector of synaptic connections $\mathbf{w}$ to an arbitrary neuron i in the network whose weights were modified by the RLS algorithm under the ROWSUM-regularization. Elements of $\mathbf{w}$ are denoted by $w_j = W_{ij}$ for $j = 1, ..., K$ where K is the total number of synapses modified under ROWSUM-regularization. Then, the sums of excitatory and inhibitory synaptic weights are defined, respectively, by

$$z_\alpha = \mathbf{w} \cdot \mathbf{1}_\alpha \quad \text{for} \quad \alpha = E, I. \tag{4}$$

Here, $\mathbf{1}_E$ denotes a vector $(1, ..., 1, 0, ..., 0)$ where the number of 1's is equal to the number of excitatory synapses K_E modified under ROWSUM-regularization and the rest of $K_I = K - K_E$ elements is 0. $\mathbf{1}_I = (0, ..., 0, 1, ..., 1)$ is defined similarly for the inhibitory synapses modified under the ROWSUM-regularization.

From the RLS algorithm (Eq. 19 in Methods), the changes in the sums of excitatory and inhibitory synaptic weights at each synaptic update obey

$$\delta z_\alpha^n = z_\alpha^n - z_\alpha^{n-1} = e^n \boldsymbol{\sigma}_\alpha^n \cdot \mathbf{r}^n \quad \text{for} \quad \alpha = E, I \tag{5}$$

where

$$\boldsymbol{\sigma}_\alpha^n = P^n \mathbf{1}_\alpha \tag{6}$$

and

$$P^n = \left[\sum_{t=1}^{n} \mathbf{r}^t (\mathbf{r}^t)' + \lambda I + \mu \mathbf{1}_E \mathbf{1}_E' + \mu \mathbf{1}_I \mathbf{1}_I' \right]^{-1}. \tag{7}$$

Since $\boldsymbol{\sigma}_\alpha^n$ was the only term in the synaptic update rule (Eq. 5) that depended on training and network parameters, it constrained the upper bound on synaptic weight changes. In simulations, the elements of $\boldsymbol{\sigma}_\alpha^n$ fluctuated in the early phase of training but stabilized as the training iteration n increased (Fig. 6A). Based on this observation, we derived an approximation of $\boldsymbol{\sigma}_\alpha^n$ for large n and obtained an estimate on its upper bound.

The analysis of $\boldsymbol{\sigma}_\alpha^n$ presented below revealed that (1) the initial vector $\boldsymbol{\sigma}_\alpha^0 = (\lambda + \mu K_\alpha)^{-1} \mathbf{1}_\alpha$ for $\alpha = E, I$ can be made arbitrarily small if the product μK_α of ROWSUM-penalty μ and the number of modified synapses K_α is large (see Methods / Section 5.1), (2) $\boldsymbol{\sigma}_\alpha^n$ is bounded by $\boldsymbol{\sigma}_\alpha^0$ uniformly in $n \geq 1$ up to a uniform constant (see Methods / Section 5.2), and (3) the update size of mean synaptic weights $\overline{\delta z_\alpha^n} = K_\alpha^{-1} \delta z_\alpha^n$ has an upper bound with a scaling factor $O(1/(\mu K_\alpha))$ thanks to (1) and (2) (see Methods / Section 5.3). This analysis result showed that the product of μ and K_α tightly constrained the mean synaptic updates because

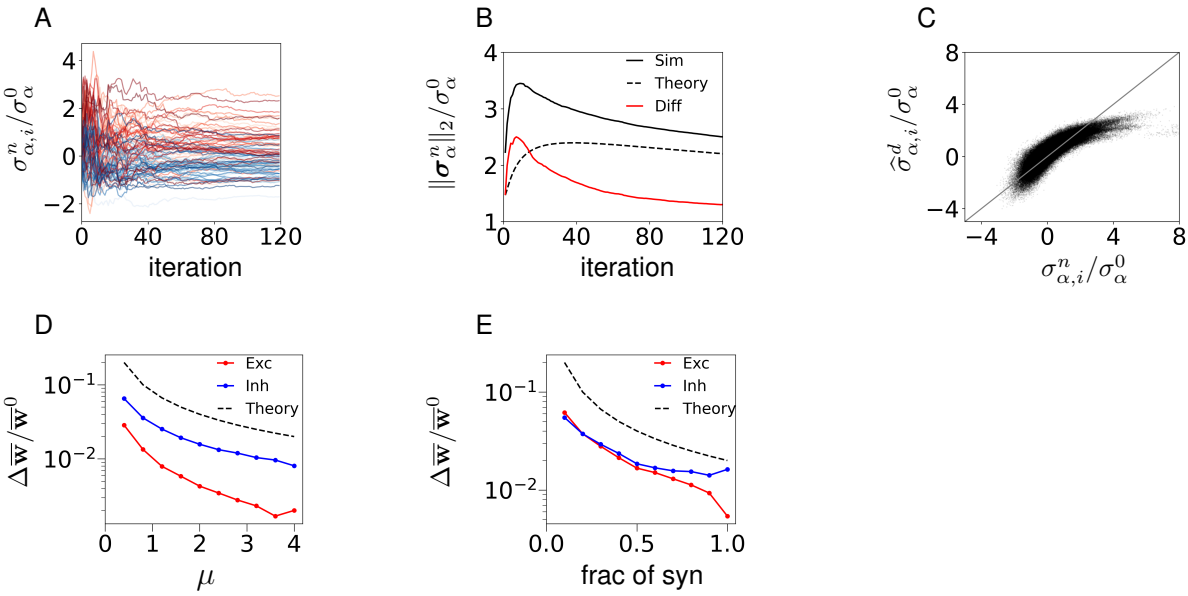

Figure 7: Analytical upper bound on synaptic updates. **(A)** Elements in $\boldsymbol{\sigma}_\alpha^n$ stabilizes if the training iteration d becomes large. $\boldsymbol{\sigma}_\alpha^n$ of sample neurons are shown; excitatory (red) and inhibitory (blue). $\sigma_\alpha^0 = (\lambda + \mu K_\alpha)^{-1}$ denotes the initial value. **(B)** Comparison of the actual $\boldsymbol{\sigma}_\alpha^n$ and the theoretical estimate $\widehat{\boldsymbol{\sigma}}_\alpha^d$ as a function of training iteration d; actual (solid), theory (dotted), difference of actual and theoretical (red). **(C)** Comparison of the actual values $\boldsymbol{\sigma}_\alpha^n$ and the theoretical estimates $\widehat{\boldsymbol{\sigma}}_\alpha^d$ at training iteration $d = 120$. Scatter plot shows the vector elements of all neurons; identity line (gray). **(D)** Total changes in the mean synaptic weights $\Delta\overline{\mathbf{w}} = \sum_{t=1}^n \overline{\delta z_\alpha^t}$ due to training, normalized by the mean initial weight $\overline{\mathbf{w}}^0$, are shown as a function of μ; theoretical estimate of the scaling factor $O(\mu^{-1})$ (black dotted line); mean excitatory (red) and inhibitory (blue) weights. **(E)** Same as in (D) but varying the number of synapses constrained by ROWSUM regularization; fraction of synapses equals 1 if all synapses are regularized by the ROWSUM constraint. Color code same as (D); theoretical estimate of scaling factor is $O(K_\alpha^{-1})$.

$\boldsymbol{\sigma}_\alpha^0$ effectively served as a uniform upper bound on $\boldsymbol{\sigma}_\alpha^n$ that persisted throughout the training iterations n.

An approximation of $\boldsymbol{\sigma}_\alpha^n$ was derived in a few steps. First, the spiking activity $\mathbf{r}^t$ during training was approximated by the firing rates $\bar{\mathbf{r}}^t$ induced by the target patterns and a noise term $\bar{\boldsymbol{\varepsilon}}^t$ that modeled the spiking fluctuations around the firing rates (see Methods / Section 5.2 for details). Such an approximation was based on the observation that spiking activity stayed close to the target patterns during training due to the fast weight updates (e.g., every 10 ms) by the RLS algorithm.

Next, since the target patterns were learned repeatedly, firing rates $\bar{\mathbf{r}}^t$ were assumed to have a period equal to the target length (m-steps) and repeated d-times to cover the entire training duration, i.e., $n = d \cdot m$. Assuming such firing rate patterns, the correlation matrix was approximated by $\sum_{t=1}^n \mathbf{r}^t(\mathbf{r}^t)' \approx d \sum_{t=1}^m (\bar{\mathbf{r}}^t + \sqrt{d}\,\bar{\boldsymbol{\varepsilon}}^t)(\bar{\mathbf{r}}^t + \sqrt{d}\,\bar{\boldsymbol{\varepsilon}}^t)'$ and substituted into the definition of P^n (Eq. 7) to obtain

$$\widehat{\boldsymbol{\sigma}}_\alpha^d = [I_K - P^0\bar{R}(d^{-1}I_m + \bar{R}'P^0\bar{R} + d\,\mathcal{E}'P^0\mathcal{E})^{-1}\bar{R}']\boldsymbol{\sigma}_\alpha^0 \tag{8}$$

where $\bar{R} = [\bar{\mathbf{r}}^1, ..., \bar{\mathbf{r}}^m]$ and $\bar{\mathcal{E}} = [\bar{\boldsymbol{\varepsilon}}^1, ..., \bar{\boldsymbol{\varepsilon}}^m]$. We verified numerically that $\boldsymbol{\sigma}_\alpha^n$ and $\widehat{\boldsymbol{\sigma}}_\alpha^d$ decrease monotonically for large n or d (Fig. 6B), and $\boldsymbol{\sigma}_\alpha^n$ can be estimated by $\widehat{\boldsymbol{\sigma}}_\alpha^d$ (Fig. 6C). Then, by monotonicity, we can find $\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}$ for some constant d_0 that

serves as a uniform upper bound on $\boldsymbol{\sigma}_\alpha^n$:

$$\|\boldsymbol{\sigma}_\alpha^n\|_2 \approx \|\widehat{\boldsymbol{\sigma}}_\alpha^d\|_2 \leq \|\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}\|_2 \leq C_\alpha \sigma_\alpha^0 \quad \text{for} \quad d \gg d_0 \gg 1 \quad \text{and} \quad \alpha = E, I. \quad (9)$$

Moreover, the constant C_α is independent of not only n but also μ and K_α because P^0 converges elementwise when μ or K_α becomes large (see Methods Section 5.3 for details).

Finally, we applied the upper bound on $\boldsymbol{\sigma}_\alpha^n$ (Eq. 9) to the synaptic learning rule (Eq. 5) to estimate the size of mean synaptic weight updates:

$$|\overline{\delta z_\alpha^n}| = K_\alpha^{-1} |\delta z_\alpha^n| \leq |e^n| B_\alpha (\mu K_\alpha)^{-1} \quad \text{for} \quad n \gg 1 \quad \text{and} \quad \alpha = E, I, \quad (10)$$

where $B_\alpha = C_\alpha r_{max} q_\alpha^{-1}$, $r_{\max}$ is the maximal firing rate induced by the target patterns, $q_\alpha = K_\alpha / K$ is the proportion of $\alpha = E, I$ synapses relative to the total number of synapses K modified under ROWSUM-regularization, and e^n is the learning error.

The analytical bound obtained in Eq. 10 suggested two factors that determine the size of synaptic updates. First, the ROWSUM-penalty μ constrained the upper bound of $\overline{\delta z_\alpha^n}$ with a scaling factor μ^{-1}. To verify this analytical estimate, we set up an initial network and trained all the synaptic weights with different μ-values across training sessions. We evaluated the mean excitatory and inhibitory synaptic weights each neuron received, and then averaged over all neurons in the network. We found that the total changes in the mean excitatory and inhibitory weights after training $\sum_{t=1}^n \overline{\delta z_\alpha^t}$ decreased with μ consistently to the analytical estimate $O(\mu^{-1})$ (Fig. 6D).

Second, the upper bound on the size of mean synaptic weight updates $|\overline{\delta z_\alpha^n}|$ was inversely proportional to K_α if it was sufficiently large and $q_\alpha = K_\alpha / K$ remained constant. In other words, increasing the number of $\alpha = E, I$ synapses modified under the ROWSUM-regularization effectively strengthened the ROWSUM-penalty. To verify this analytical result, among pN synapses each neuron received, a fraction $f = K/(Np)$ of synapses was trained under the ROWSUM-regularization while the rest $1 - f$ of the synapses were trained under the standard L_2-regularization. The fraction f of ROWSUM-regularized synapses was increased systematically, while the proportion $q_\alpha = K_\alpha / K$ of ROWSUM-modified $\alpha = E, I$ synapses were fixed to 0.5 for both $\alpha = E, I$. After training, we evaluated the mean excitatory and inhibitory synaptic weights trained under ROWSUM-regularization and confirmed that the total changes in average synaptic strength decreased consistently with the analytical estimate $O(K_\alpha^{-1})$ (Fig. 6E).

Discussion

Recurrent neural networks can learn to perform cognitive tasks (Mante et al., 2013; Sussillo et al., 2015; Chaisangmongkon et al., 2017; Wang et al., 2018) and generate complex activity patterns (Kim and Chow, 2018; Rajan et al., 2016; Laje and Buonomano, 2013). However, due to the large number of connections, the connectivity structure that achieves the learning goal is nonunique (Song et al., 2016; Sussillo et al., 2015). Therefore, it is important to further constrain the recurrent connectivity structure in order to identify solutions relevant to biological neural

circuits. In this study, we showed that trained networks can maintain strong excitatory and inhibitory synaptic strength and, consequently, exhibit dynamic features of strongly coupled cortical circuit models, such as stochastic spiking and inhibition-stabilization, if the mean synaptic strength of the strongly coupled initial network is preserved throughout training.

In a related work, Ingrosso and Abbott (2019) developed a training scheme that maintained strong synaptic strength in trained networks by setting up a strongly coupled excitatory-inhibitory network and penalizing the deviation of synaptic weights from their initial values. Specifically, they considered a regularizer $R_i^{\text{syn}} = \sum_j (W_{ij} - W_{ij}^0)^2$ for each neuron i that constrained the L_2-distance of individual synaptic weights between the initial and trained networks. The main difference to our work is that the regularizer $R_i^{\text{sum}} = (\sum_{j \in E} W_{ij})^2 + (\sum_{j \in I} W_{ij})^2$ considered in our study constrained the sums of excitatory and inhibitory weights, respectively, but not the individual synaptic weights. In fact, including R_i^{syn} to the cost function is equivalent to adding the standard L_2-regularizer $R_i^{L_2} = \sum_j W_{ij}^2$ if the synaptic update rule is derived through the RLS algorithm. This is because the effects of subtracting constants W_{ij}^0 from W_{ij} are nullified when the gradients at two time points are compared to derive a synaptic update rule (see Methods / Section 3). Bounded coordinate descent algorithm implemented by Ingrosso and Abbott did not suffer from this issue, and they showed that regularizing the weights with R_i^{syn}, but not $R_i^{L_2}$, allowed them to preserve the strong synaptic strength of initial networks. For the RLS algorithm, strengthening the L_2-regularizer $R_i^{L_2}$ will keep the weights close to their initial values. Our study demonstrated that constraining the mean synaptic strength through R_i^{sum} imparts the flexibility for individual synapses to learn the task while maintaining strong coupling of the averaged synaptic weights.

Separation of excitatory and inhibitory neurons is an important feature of biological neural circuits that has implications to the dynamical state of the network. In particular, strong recurrent connections within and between the excitatory and inhibitory neurons give rise to dynamic properties such as asynchronous spiking (van Vreeswijk and Sompolinsky, 1996; Renart et al., 2010; Rosenbaum et al., 2017), balanced amplification (Murphy and Miller, 2009), and inhibition-stabilization (Tsodyks et al., 1997). In recent studies, training schemes were developed to preserve the excitatory-inhibitory structure of recurrent connections. For instance, "clipping" methods skipped synaptic updates that violated the sign constraints (Ingrosso and Abbott, 2019; Kim and Chow, 2018); back-propagation-through-time (BPTT) was adapted to rectify the synaptic weights in order to produce zero gradients in the opposite sign regime (Song et al., 2016); in cases where only the read-out and/or feedback weights were trained, the strong recurrent synapses of the initial network remained unchanged while other weighs were constrained to respect the overall excitatory-inhibitory structure (Nicola and Clopath, 2017). However, these studies focused on respecting the excitatory-inhibitory structure without promoting the strong recurrent connections essential for generating novel network dynamics. Our study suggests that regularizing by imposing additional constraints on the mean synaptic weights can lead to trained networks that inherit dynamic properties of a strongly coupled excitatory-inhibitory network.

In a recent work, Baker et al. (2019) considered network and stimulus configurations that deviated from the standard balanced state in that excessive inhibition

present in such a state resulted in threshold-linear stationary rates as a function of external stimulus, as opposed to the linear relationship between positive stationary rates and external inputs in the balanced state. Using the nonlinearity present in threshold-linear activation, they showed that nonlinear computations was possible in the inhibition-excessive excitatory-inhibitory network, pointing out the importance of dynamic regime determined by the synaptic connectivity.

Developing general purpose machine learning algorithms for training spiking neural networks is an active area of research that has made significant progress in the last several years. Recent work has demonstrated that including biological mechanisms, such as spike frequency adaptation, can enhance a spiking neural network's capability (Bellec et al., 2018), and biologically inspired alternatives to BPTT can be applied to perform supervised and reinforcement learning tasks (Bellec et al., 2019). This progress provides a fertile ground to discover the potential role of connectivity constraints present in biological neural circuits. Spiking neural networks trained with relevant biological features and connectivity constraints may serve as a biologically interpretable alternative to artificial neural networks that can provide further insights into the inner working of biological neural circuits.

Methods

1 Spiking network model

The spiking network model consisted of an equal number of N_E excitatory and N_I inhibitory leaky integrate-and-fire neurons connected randomly with connection probability p. The network activity was governed by a set of equations

$$\begin{aligned} \tau_m \dot{v}_i &= -v_i + X_i + \tau_m u_i \\ \tau_s \dot{u}_i &= -u_i + \sum_j W_{ij} \sum_s \delta(t - t_j^s), \end{aligned} \tag{11}$$

where v_i is membrane potential of neuron i, τ_m is membrane time constant, X_i is external input, u_i is synaptic current to neuron i, τ_s is synaptic time constant, W_{ij} is recurrent synaptic weights from neuron j to i and t_j^s is spike times of neuron j. Neurons emitted spikes when the membrane potential reached the spike threshold $v_{thr} = 1$, after which the membrane potential was immediately reset to $v_{reset} = 0$. Specific parameter values used in simulations are summarized in Table 1.

The initial recurrent synaptic weights W_{ij} and the external inputs X_i were sufficiently strong such that the total excitatory inputs to a neuron, i.e. the sum of excitatory current and external input, exceeded the spike threshold; therefore, feedback inhibition was necessary to stabilize the network activity. Each synaptic weight was determined by its connection type, i.e., $W_{ij} = W_{\alpha\beta}$ for neuron j in population β to neuron i in population α, and neurons in the same population received identical inputs, i.e. $X_i = X_\alpha$ for $i \in \alpha$ where $\alpha, \beta \in \{E, I\}$. The relative strength of recurrent excitation, recurrent inhibition and external input to the excitatory population, i.e., $\gamma_E = W_{EE}/W_{IE}, \gamma_I = W_{EI}/W_{II}, \gamma_X = X_E/X_I$, were chosen to be consistent with the balanced condition (van Vreeswijk and Sompolin-

sky, 1998; Renart et al., 2010; Rosenbaum and Doiron, 2014):

$$\gamma_E < \gamma_I < \gamma_X. \tag{12}$$

Finally, the recurrent synaptic weights and the external inputs were defined with reference to $W_E = 1/\sqrt{K_E}$, $W_I = -1.5/\sqrt{K_I}$ and $X = 0.1\sqrt{K_I}$:

$$\begin{aligned} W_{EE} &= \gamma_E W_E, \quad W_{EI} = \gamma_I W_I, \quad X_E = \gamma_X X \\ W_{IE} &= W_E, \qquad W_{II} = W_I, \qquad X_I = X, \end{aligned} \tag{13}$$

where $K_E = pN_E$ and $K_I = pN_I$ are the mean excitatory and inhibitory indegrees, respectively.

In case of an instantaneous synapse (i.e. $\tau_s = 0$), one spike with weights W_E and W_I would deflect the membrane potential by 0.14 and 0.21, respectively, with the chosen network size $N_E, N_I = 500$ and connection probability $p = 0.1$. Since the spike threshold was $v_{thr} = 1$, approximately 8 incoming spikes were enough to trigger a spike. However, due to the synaptic filtering, the actual deflection in membrane potential by a spike was scaled by a factor $1/\tau_s$.

We investigated several types of random connectivity structures for the initial network. First, we considered a network in which the number of synaptic connections to each neuron was fixed to the mean indegree pN, but the pN_E excitatory and pN_I inhibitory presynaptic neurons were selected randomly. The synaptic weights in this network structure were set to constants as described in Eq. 13. Since the mean synaptic weights were identical across neurons, we observed homogeneous firing rate distribution in this connectivity type (Fig. 1A.iv). Previous studies considered spiking networks with fixed indegree and constant weights (Brunel, 2000; Ostojic, 2014; Nicola and Clopath, 2017). This network structure was trained on sinusoidal functions in Sections. 1, 2 and 3 and on complex target functions in Sections. 4 and 5.

Second, we constructed an Erdos-Renyi graph where each synaptic connection was generated independently with probability p, resulting in a connectivity matrix with a non-constant indegree distribution. The synaptic weights were assigned in two different ways. They were either constants as described in Eq. 13 or were sampled from Gaussian distributions with mean $W_{\alpha\beta}$ and standard deviation $W_{\alpha\beta}/5$ to ensure Dale's law was not violated. In these networks, the sum of excitatory and inhibitory synaptic weights to each neuron, i.e., $\sum_{j\in E} W_{ij}$ and $\sum_{j\in I} W_{ij}$, respectively, were not identical across neurons, therefore resulted in a wide firing rate distribution including many low firing rate neurons (Fig. 6A.i). To reconcile the differences in synaptic weights across neurons, we added a correction term $\delta w_i = (pN_\beta W_{\alpha\beta} - \sum_{j\in\beta} W_{ij})/N_\beta^{(i)}$ to all nonzero synaptic connections from population β to neuron $i \in \alpha$, where $N_\beta^{(i)}$ is the number of nonzero connections, such that the total synaptic weight became $pN_\beta W_{\alpha\beta}$. In Section. 2, we considered an Erdos-Renyi connectivity with Gaussian weights in which the synaptic connections to all the neurons were corrected (see Fig. 2C). On the other hand, in Section. 6, we discussed a random connectivity with constant weights in which the synaptic connections were corrected in only a fraction of neurons.

	Neuron parameters	Values
τ_m	membrane time constant	10 ms
v_{thr}	spike threshold	1
v_{reset}	voltage reset	0
	Network parameters	
N	number of neurons	1000
N_E	number of excitatory neurons	500
N_I	number of inhibitory neurons	500
p	connection probability	0.1
	Synaptic parameters	
τ_s	synaptic time constant	20 ms
K	mean indegree	pN
K_E	mean excitatory indegree	pN_E
K_I	mean inhibitory indegree	pN_I
W_E	excitatory synaptic weight	$1/\sqrt{K_E}$
W_I	inhibitory synaptic weight	$-1.5/\sqrt{K_I}$
X	external input	$0.1\sqrt{K_I}$
γ_E	relative strength of recurrent excitation to excitatory population	1.0
γ_I	relative strength of recurrent inhibition to excitatory population	1.25
γ_X	relative strength of external input to excitatory population	1.5
W_{EE}	E to E synaptic weight	$\gamma_E W_E$
W_{IE}	E to I synaptic weight	W_E
W_{EI}	I to E synaptic weight	$\gamma_I W_I$
W_{II}	I to I synaptic weight	W_I
X_E	external input to excitatory neurons	$\gamma_X X$
X_I	external input to inhibitory neurons	X
	Training parameters	
λ	penalty for L_2-regularization	0.1
μ	penalty for ROWSUM-regularization	2.0
N_{loop}	number of training iterations	120
T	length of target patterns	1000 ms

Table 1: Default simulation and training parameters. Any differences from these parameters are mentioned in the main text.

2 Training a network of excitatory and inhibitory neurons

Previous studies have shown that the recurrent synaptic connections can be trained using recursive least squares (RLS) to generate target activity patterns (Kim and Chow, 2018; Laje and Buonomano, 2013; DePasquale et al., 2018; Rajan et al., 2016). Specifically, the goal was to train the recurrent connections W_{ij} such that the synaptic drive $u_i(t)$ to all neurons $i = 1, ..., N$ in the network learned to follow their own target pattern $f_i(t)$ defined on time interval $[0, T]$ when neurons were stimulated by brief external input. In this study, we included additional penalty terms to the cost function that penalized the sums of excitatory and inhibitory synaptic weights to each neuron, respectively, and then derived the RLS algorithm that included the effects of the additional penalty terms.

The learning rule described below trains the activity of one neuron in the recurrent network. However, the same scheme can be applied to all neurons synchronously (i.e., modify synaptic connections to all neuron at the same time) to update the full connectivity matrix. To simplify notation, we dropped neuron index i in the following derivation; $\mathbf{w} = (W_{i1}, ..., W_{iK})$ denotes a vector of synaptic weights that neuron i receives from presynaptic neurons, $\mathbf{r}(t) = (r_1(t), ..., r_K(t))$ denotes the synaptic filtering of spike trains of neurons presynaptic to neuron i, and $f(t), t \in [0, T]$ is the target pattern of neuron i. Neurons were labeled so that the first K_E are excitatory and the remaining $K_I = K - K_E$ are inhibitory.

Using linearity in W_{ij}, the equation for synaptic drive u_i was rewritten in the form,

$$\begin{aligned} u_i(t) &= \sum_j W_{ij} r_j(t) \\ \tau_s \dot{r}_i &= -r_i + \sum_s \delta(t - t_i^s). \end{aligned} \tag{14}$$

With this formulation, the synaptic drive was expressed as a linear function of connectivity W_{ij}, which allowed us to derive the RLS update rule when the synaptic drive activities were trained.

Here we derive an RLS training algorithm for a network consisting of excitatory and inhibitory neurons applied in Results / Sections 1 to 4. Extension of the ROWSUM constraints to multiple subpopulations are discussed at the end of this section. Since maintaining strong excitatory and inhibitory synaptic strength was critical for generating dynamic features of excitatory-inhibitory network, we constrained the sums of excitatory and inhibitory synaptic weights to each neuron, respectively. The cost function was

$$C = \frac{1}{2} \sum_{t=0}^{T} (f^t - \mathbf{w} \cdot \mathbf{r}^t)^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + \frac{\mu}{2} (\mathbf{w} \cdot \mathbf{1}_E)^2 + \frac{\mu}{2} (\mathbf{w} \cdot \mathbf{1}_I)^2 \tag{15}$$

where the first term is the learning error, the second is the L_2-regularization and the remaining two terms comprise the ROWSUM-regularization that penalizes the sums of excitatory and inhibitory synaptic connections to the neuron, respectively. Here $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{K} x_i^2}$ is the standard L_2 norm, $\mathbf{1}_E = (1, .., 1, 0, ..., 0)$ and $\mathbf{1}_I = (0, ..., 0, 1, ..., 1)$ where the number of 1's in $\mathbf{1}_E$ and $\mathbf{1}_I$ is equal to K_E and K_I, respectively.

The gradient of the cost function with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}} C = \sum_t (-f^t \mathbf{r}^t + \mathbf{r}^t (\mathbf{r}^t)' \mathbf{w}) + \lambda \mathbf{w} + \mu \mathbf{1}_E \mathbf{1}_E^T \mathbf{w} + \mu \mathbf{1}_I \mathbf{1}_I^T \mathbf{w}. \quad (16)$$

To derive the RLS algorithm, we computed the gradients at two consecutive time points

$$\mathbf{0} = \nabla_{\mathbf{w}^n} C = \sum_{t=1}^{n} (-f^t \mathbf{r}^t + \mathbf{r}^t (\mathbf{r}^t)' \mathbf{w}^n) + \lambda \mathbf{w}^n + \mu \mathbf{1}_E \mathbf{1}_E' \mathbf{w}^n + \mu \mathbf{1}_I \mathbf{1}_I' \mathbf{w}^n \quad (17)$$

and

$$\mathbf{0} = \nabla_{\mathbf{w}^{n-1}} C = \sum_{t=1}^{n-1} (-f^t \mathbf{r}^t + \mathbf{r}^t (\mathbf{r}^t)' \mathbf{w}^{n-1}) + \lambda \mathbf{w}^{n-1} + \mu \mathbf{1}_E \mathbf{1}_E' \mathbf{w}^{n-1} + \mu \mathbf{1}_I \mathbf{1}_I' \mathbf{w}^{n-1} \quad (18)$$

and subtracted Eq. 18 from Eq. 17 to obtain

$$\begin{aligned} \mathbf{w}^n &= \mathbf{w}^{n-1} + e^n P^n \mathbf{r}^n \\ e^n &= f^n - \mathbf{w}^{n-1} \cdot \mathbf{r}^n \end{aligned} \quad (19)$$

where

$$P^n = \left[\sum_{t=1}^{n} \mathbf{r}^t (\mathbf{r}^t)' + \lambda I + \mu \mathbf{1}_E \mathbf{1}_E' + \mu \mathbf{1}_I \mathbf{1}_I' \right]^{-1} \quad \text{for} \quad n \geq 1 \quad (20)$$

with the initial value

$$P^0 = [\lambda I + \mu \mathbf{1}_E \mathbf{1}_E' + \mu \mathbf{1}_I \mathbf{1}_I']^{-1}. \quad (21)$$

To update P^n iteratively, we used the Woodbury matrix identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U (C^{-1} + V A^{-1} U)^{-1} V A^{-1} \quad (22)$$

where A is invertible and N-by-N, U is N-by-T, C is invertible and T-by-T and V is T-by-N matrices. Then P^n can be calculated iteratively

$$P^n = P^{n-1} - \frac{P^{n-1} \mathbf{r}^n (\mathbf{r}^n)' P^{n-1}}{1 + (\mathbf{r}^n)' P^{n-1} \mathbf{r}^n}.$$

For ROWSUM training on multiple subpopulations, the cost function constrains the sum of synaptic connections from each subpopulation

$$C = \frac{1}{2} \sum_{t=0}^{T} (f^t - \mathbf{w} \cdot \mathbf{r}^t)^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + \frac{\mu}{2} \sum_{\alpha=1}^{M} \left[(\mathbf{w} \cdot \mathbf{1}_{E_\alpha})^2 + (\mathbf{w} \cdot \mathbf{1}_{I_\alpha})^2 \right], \quad (23)$$

where $(\mathbf{1}_{E_\alpha})_i = 1$ if $i \in E_\alpha$ and zero otherwise. A sum of multiple block matrices is introduced into the inverse correlation matrix

$$P^n = \left[\sum_{t=1}^{n} \mathbf{r}^t (\mathbf{r}^t)' + \lambda I + \mu \sum_{\alpha=1}^{M} (\mathbf{1}_{E_\alpha} \mathbf{1}_{E_\alpha}' + \mathbf{1}_{I_\alpha} \mathbf{1}_{I_\alpha}') \right]^{-1} \quad \text{for} \quad n \geq 1 \quad (24)$$

with the initial condition

$$P^n = \left[\lambda I + \mu \sum_{\alpha=1}^{M} (\mathbf{1}_{E_\alpha} \mathbf{1}_{E_\alpha}' + \mathbf{1}_{I_\alpha} \mathbf{1}_{I_\alpha}') \right]^{-1}. \quad (25)$$

3 Translation invariance of RLS algorithm

Here we show that the RLS algorithm is invariant to translation in the sense that adding arbitrary constants to the quadratic penalty terms does not change the synaptic update rule derived from the RLS. Translation invariance arises because of the recursive nature of RLS, i.e., the synaptic update rule uses the difference of the gradient at two time points and thus constants are removed . This justifies the form of the penalty terms in the cost function Eq. 15 since adding a constant to $\mathbf{w}$ in the regularization terms (e.g., $\mathbf{w} - \mathbf{m}$ in the example below) does not affect the synaptic update rule (Eq. 19).

To demonstrate translation invariance, we shifted w_i in the penalty terms by arbitrary number m_i and rewrote the cost function in the following form:

$$C = \frac{1}{2}\sum_{t=0}^{n}(f^t - \mathbf{w}\cdot\mathbf{r}^t)^2 + \frac{\lambda}{2}\|\mathbf{w}-\mathbf{m}\|_2^2 + \frac{\mu}{2}\big((\mathbf{w}-\mathbf{m})\cdot\mathbf{1}_E\big)^2 + \frac{\mu}{2}\big((\mathbf{w}-\mathbf{m})\cdot\mathbf{1}_I\big)^2. \tag{26}$$

Although this cost function is different from the one considered in the previous section, we now show that the synaptic update rule derived from this alternative form is identical to the previously derived Eq. 19. Differentiating the cost function Eq. 26 with respect to $\mathbf{w}^n$ and $\mathbf{w}^{n-1}$ yields

$$\begin{aligned}\mathbf{0} = \nabla_{\mathbf{w}^n} C = \sum_{t=0}^{n}(-f^t\mathbf{r}^t + \mathbf{r}^t(\mathbf{r}^t)'\mathbf{w}^n) \\ + \lambda(\mathbf{w}^n - \mathbf{m}) + \mu\mathbf{1}_E\mathbf{1}_E'(\mathbf{w}^n - \mathbf{m}) + \mu\mathbf{1}_I\mathbf{1}_I'(\mathbf{w}^n - \mathbf{m})\end{aligned} \tag{27}$$

and

$$\begin{aligned}\mathbf{0} = \nabla_{\mathbf{w}^{n-1}} C = \sum_{t=0}^{n-1}(-f^t\mathbf{r}^t + \mathbf{r}^t(\mathbf{r}^t)'\mathbf{w}^{n-1}) \\ + \lambda(\mathbf{w}^{n-1} - \mathbf{m}) + \mu\mathbf{1}_E\mathbf{1}_E'(\mathbf{w}^{n-1} - \mathbf{m}) + \mu\mathbf{1}_I\mathbf{1}_I'(\mathbf{w}^{n-1} - \mathbf{m}).\end{aligned} \tag{28}$$

The $\mathbf{m}$ terms drop out in the final synaptic update rule when we subtract Eq. 28 from Eq. 27. Therefore, the synaptic update rule we obtain from the alternative cost function (Eq. 26) is identical to one (Eq. 19) derived previously from the original cost function (Eq. 15).

4 Effects of rate distribution on synaptic weight updates

To gain insights into how the firing rate distribution of presynaptic neurons affects the ROWSUM-training, we derived an analytical expression for the learning rule at the first synaptic update assuming simplified rate distributions. For a homogeneous rate distribution, we considered a rate vector $\mathbf{r} = r_0\mathbf{1}$ where all K presynaptic neurons have the same rate r_0. For an inhomogeneous rate distribution with high- and low-rate neurons, we considered a rate vector $\mathbf{r} = (r_0, ..., r_0, \epsilon, ..., \epsilon)$ where the first L neurons fire at rate r_0 and the rest of $K - L$ neurons have a small firing rate ϵ.

We focused on a ROWSUM constraint applied to connections from all the presynaptic neurons, instead of considering excitatory and inhibitory neurons separately. The first synaptic update can be written as

$$\Delta \mathbf{w}^1 = \mathbf{w}^1 - \mathbf{w}^0 = e^1 P^1 \mathbf{r} \approx e^1 P^0 \mathbf{r} \tag{29}$$

where

$$P^0 = [\lambda I + \mu \mathbf{1}\mathbf{1}']^{-1} = \frac{1}{\lambda} I - \frac{\mu}{\lambda(\lambda + \mu K)} \mathbf{1}\mathbf{1}' \tag{30}$$

and $P^1 = [\mathbf{r}(\mathbf{r})' + \lambda I + \mu \mathbf{1}\mathbf{1}']^{-1}$ was approximated by P^0 for simplicity.

For the homogeneous rate distribution, the size of synaptic updates is small, i.e. order $O(1/(\mu K))$, for all neurons, thanks to the cancellation of two terms in $P^0 \mathbf{1}$:

$$\Delta \mathbf{w}^1 = \frac{e r_0}{\lambda + \mu K} \mathbf{1}. \tag{31}$$

However, for the inhomogeneous rate distribution with high and low rate neurons, the size of the synaptic update can be of order $O(1)$ (i.e., note the factors $K - L$ and L in Eq. 32), and the direction of synaptic updates for the high and low rate neurons can be opposite to each other (i.e., note the difference in their signs):

$$\Delta w_i^1 = \begin{cases} \frac{e r_0}{\lambda + \mu K}(1 + \lambda^{-1} \mu (K - L)) > 0 & \text{if} \quad 1 \le i \le L, \\ \frac{e r_0}{\lambda + \mu K}(-\lambda^{-1} \mu L) < 0 & \text{if} \quad L + 1 \le i \le K \end{cases} \tag{32}$$

in the limit $\epsilon \to 0$.

The analysis presented here addresses only the first synaptic update. The full training results can be found in Results / Section 4

5 Analytical upper bound on mean synaptic weight updates

In this section, we derived analytical estimates on how the training and network parameters constrained the trained synaptic weights. The sums of excitatory and inhibitory synaptic weights to a neuron were defined, respectively, as

$$z_\alpha^n = \sum_{i \in \alpha} w_i^n = \mathbf{w}^n \cdot \mathbf{1}_\alpha \quad \text{for} \quad \alpha = E, I. \tag{33}$$

We rearranged the RLS learning rule (Eq. 19) to find an expression for changes in the sums of excitatory and inhibitory synaptic weights:

$$\delta z_\alpha^n = z_\alpha^n - z_\alpha^{n-1} = e^n P^n \mathbf{r}^n \cdot \mathbf{1}_\alpha = e^n \boldsymbol{\sigma}_\alpha^n \cdot \mathbf{r}^n \tag{34}$$

where

$$\boldsymbol{\sigma}_\alpha^n = P^n \mathbf{1}_\alpha \quad \text{for} \quad \alpha = E, I. \tag{35}$$

Here we give an overview of the full analysis. We first obtained analytical expressions for the initial value $\sigma_\alpha^0 = (\lambda + \mu K_\alpha)^{-1}$ and $\boldsymbol{\sigma}_\alpha^n, n \ge 1$ to be used in the subsequent analysis. We found that the initial value σ_α^0 is of order $O((\mu K_\alpha)^{-1})$

and used it as the main quantity to bound the rest of $\boldsymbol{\sigma}_\alpha^n, n \geq 1$ (Section 5.1). To show that $\boldsymbol{\sigma}_\alpha^n$ is uniformly bounded in n, we decomposed the spiking activity as a sum of spiking rate and noise, and then obtained an approximation $\widehat{\boldsymbol{\sigma}}_\alpha^d$ of $\boldsymbol{\sigma}_\alpha^n$ which decreases monotonically and thus has a uniform upper bound $\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}$ for some $d \gg d_0$ (Section 5.2). We numerically verified that $\boldsymbol{\sigma}_\alpha^n \approx \widehat{\boldsymbol{\sigma}}_\alpha^d$, and they both decrease monotonically which led to the uniform upper bound on $\|\boldsymbol{\sigma}_\alpha^n\|_2 \approx \|\widehat{\boldsymbol{\sigma}}_\alpha^d\|_2 \leq C_\alpha \sigma_\alpha^0$ and mean synaptic weights $|\overline{\delta z_\alpha^n}| = K_\alpha^{-1} \delta z_\alpha^n \sim O\big((\mu K_\alpha)^{-1}\big)$ (Section 5.3).

The main result deduced from the analysis is that the changes in mean synaptic weights $|\overline{\delta z_\alpha^n}|$ is constrained by the product μK_α. By either increasing the μ-penalty or the number of synapses K_α modified by the ROWSUM-regularization, the changes in mean synaptic weights can be constrained tightly.

5.1 Analytical expressions for $\boldsymbol{\sigma}_\alpha^n$

To find the initial $\boldsymbol{\sigma}_\alpha^0$ for $\alpha = E, I$, we substituted $A = \lambda I_K$, $U = [\mathbf{1_E}, \mathbf{1_I}]$, $V = U'$ and $C = I_2$ to the Woodbury matrix identity (Eq. 22) to obtain

$$P^0 = \frac{1}{\lambda} I_K - \frac{\mu}{\lambda(\lambda + \mu K_E)} \mathbf{1}_E \mathbf{1}_E' - \frac{\mu}{\lambda(\lambda + \mu K_I)} \mathbf{1}_I \mathbf{1}_I'. \tag{36}$$

Here the subscript in I_K denotes the dimensionality of the $K \times K$ identity matrix. Then,

$$\begin{aligned} \boldsymbol{\sigma}_\alpha^0 &= \sigma_\alpha^0 \mathbf{1}_\alpha \\ \text{where} \quad \sigma_\alpha^0 &= \frac{1}{\lambda + \mu K_\alpha} \quad \text{for} \quad \alpha = E, I. \end{aligned} \tag{37}$$

We can see immediately that the initial value σ_α^0 is of order $O(1/(\mu K_\alpha))$. To obtain estimates for $\boldsymbol{\sigma}_\alpha^n$ with $n \geq 1$, we used the Woodbury matrix identity (Eq. 22) to express P^n in the form

$$P^n = P^0 - P^0 R^n [I + (R^n)' P^0 R^n]^{-1} (R^n)' P^0 \tag{38}$$

where $R^n = [\mathbf{r}^1, ..., \mathbf{r}^n]$ and obtained

$$\boldsymbol{\sigma}_\alpha^n = [I - P^0 R^n [I + (R^n)' P^0 R^n]^{-1} (R^n)'] \boldsymbol{\sigma}_\alpha^0. \tag{39}$$

5.2 Asymptotic approximation of $\boldsymbol{\sigma}_\alpha^n$

To show that $\boldsymbol{\sigma}_\alpha^n$ can be bounded uniformly in n, it suffices to show that the right hand side of Eq. 39 does not depend on n. To this end, we analyzed its asymptotic behavior as the training iteration n grew arbitrarily large. We used the fact that (1) the same target pattern was learned repeatedly during training and (2) fast synaptic weight updates (e.g., every 10 ms) by the RLS algorithm kept the spiking activity close to the target pattern. Then, the actual spiking activity $\mathbf{r}^t$ was approximated with a firing rate function $\bar{\mathbf{r}}^t$ induced by the target patterns where $\bar{\mathbf{r}}^t$ is a periodic function with a period equal to the target length (m-steps) and repeated d-times to cover the entire training duration, i.e., $n = d \cdot m$.

To make it clear that $\bar{\mathbf{r}}$ is periodic in time, we used a time index $\bar{t}$ that denotes the time point within the target pattern that have period m. Specifically, the spiking activity is expressed as

$$\mathbf{r}^t = \bar{\mathbf{r}}^{\bar{t}} + \boldsymbol{\varepsilon}_s^{\bar{t}} \quad \text{for} \quad t = 1, ..., n \tag{40}$$

where $t = sm + \bar{t}$ and $s = 1, ..., d$ indicate the training iteration number. We introduced a noise term $\boldsymbol{\varepsilon}_s^{\bar{t}}$ whose i^{th} component is the fluctuation of neuron i's spiking activity around the firing rate $\bar{\mathbf{r}}^{\bar{t}}$ at the sth training session. For the following analysis, we assumed a simple noise model that $\boldsymbol{\varepsilon}_s^{\bar{t}}$ is a multivariate random variable at each $\bar{t}$ with no correlations between neurons within and across training sessions

$$\mathrm{Cov}[\boldsymbol{\varepsilon}_s^{\bar{t}}, \boldsymbol{\varepsilon}_s^{\bar{t}}] = \mathrm{E}[\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_s^{\bar{t}})'] = \alpha I_K \quad \text{for} \quad 1 \leq s \leq d \tag{41}$$

$$\mathrm{Cov}[\boldsymbol{\varepsilon}_s^{\bar{t}}, \boldsymbol{\varepsilon}_q^{\bar{t}}] = \mathrm{E}[\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_q^{\bar{t}})'] = 0 \quad \text{for} \quad s \neq q \text{ and } 1 \leq s, q \leq d. \tag{42}$$

To derive an alternative expression for $\boldsymbol{\sigma}_\alpha^n$ amenable to asymptotic analysis, we expressed the correlation of spiking activities in terms of the firing rates and fluctuations around them:

$$\sum_{t=1}^{n} \mathbf{r}^t(\mathbf{r}^t)' = \sum_{\bar{t}=1}^{m}\sum_{s=1}^{d}(\bar{\mathbf{r}}^{\bar{t}} + \boldsymbol{\varepsilon}_s^{\bar{t}})(\bar{\mathbf{r}}^{\bar{t}} + \boldsymbol{\varepsilon}_s^{\bar{t}})' \tag{43}$$

$$= \sum_{\bar{t}=1}^{m}\left(d\,\bar{\mathbf{r}}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})' + d\,\bar{\mathbf{r}}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})' + d\,\boldsymbol{\varepsilon}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})' + \sum_{s=1}^{d}\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_s^{\bar{t}})'\right) \tag{44}$$

$$\approx d\sum_{\bar{t}=1}^{m}\left(\bar{\mathbf{r}}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})' + \bar{\mathbf{r}}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})' + \boldsymbol{\varepsilon}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})' + d\,\boldsymbol{\varepsilon}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})'\right) \tag{45}$$

$$\approx d\sum_{\bar{t}=1}^{m}(\bar{\mathbf{r}}^{\bar{t}} + \sqrt{d}\,\bar{\boldsymbol{\varepsilon}}^{\bar{t}})(\bar{\mathbf{r}}^{\bar{t}} + \sqrt{d}\,\bar{\boldsymbol{\varepsilon}}^{\bar{t}})'. \tag{46}$$

The noise approximation in Eq. 45 is equivalent to following relationship

$$d\,\bar{\boldsymbol{\varepsilon}}^{\bar{t}}(\bar{\boldsymbol{\varepsilon}}^{\bar{t}})' = d^{-1}\sum_{s=1}^{d}\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_s^{\bar{t}})' + d^{-1}\sum_{s=1}^{d}\sum_{q\neq s}\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_q^{\bar{t}})' \approx d^{-1}\sum_{s=1}^{d}\boldsymbol{\varepsilon}_s^{\bar{t}}(\boldsymbol{\varepsilon}_s^{\bar{t}})' \tag{47}$$

where

$$\bar{\boldsymbol{\varepsilon}}^{\bar{t}} = d^{-1}\sum_{s=1}^{d}\boldsymbol{\varepsilon}_s^{\bar{t}} \tag{48}$$

and the cross-training correlations in Eq. 47 was ignored by the noise assumption given in Eq. 42. In Eqs. 45 and 46, the dominant term is $d\,\boldsymbol{\varepsilon}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})'$, so $\bar{\mathbf{r}}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})'$ and $\boldsymbol{\varepsilon}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})'$ in Eq. 45 were replaced by $\sqrt{d}\,\bar{\mathbf{r}}^{\bar{t}}(\boldsymbol{\varepsilon}^{\bar{t}})'$ and $\sqrt{d}\,\boldsymbol{\varepsilon}^{\bar{t}}(\bar{\mathbf{r}}^{\bar{t}})'$ in Eq. 46.

Next, we substituted the approximated $\sum_t \mathbf{r}^t(\mathbf{r})'$ (Eq. 46) into the definition of P^n (Eq. 38) and suppressed $O(\sqrt{d})$-terms to obtain

$$\widehat{\boldsymbol{\sigma}}_\alpha^d = [I_K - P^0\bar{R}(d^{-1}I_m + \bar{R}'P^0\bar{R} + d\,\mathcal{E}'P^0\mathcal{E})^{-1}\bar{R}']\boldsymbol{\sigma}_\alpha^0 \tag{49}$$

where

$$\bar{R} = [\bar{\mathbf{r}}^1, ..., \bar{\mathbf{r}}^m] \quad \text{and} \quad \bar{\mathcal{E}} = [\bar{\boldsymbol{\varepsilon}}^1, ..., \bar{\boldsymbol{\varepsilon}}^m]. \tag{50}$$

Note that for uncorrelated noise, the term $\mathcal{E}'P^0\mathcal{E}$ yields an identity matrix which effectively serves as a regularizer that makes it possible to invert $\bar{R}'P^0\bar{R} + d\,\mathcal{E}'P^0\mathcal{E}$.

We verified numerically that the actual $\boldsymbol{\sigma}_\alpha^n$ can be well approximated by $\widehat{\boldsymbol{\sigma}}_\alpha^d$ (Fig. 7B). Moreover, since $\|\widehat{\boldsymbol{\sigma}}_\alpha^d\|_2$ decreases monotonically with d, there exists a constant d_0 such that

$$\|\widehat{\boldsymbol{\sigma}}_\alpha^d\|_2 \leq \|\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}\|_2 \quad \text{for} \quad d \gg d_0 \gg 1. \tag{51}$$

5.3 Upper bound on the size of synaptic updates

By its definition in Eq. 49, $\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}$ can be expressed in terms of the initial value σ_α^0

$$\widehat{\boldsymbol{\sigma}}_\alpha^{d_0} = \sigma_\alpha^0 \mathbf{v}_\alpha \tag{52}$$

$$\text{where} \quad \mathbf{v}_\alpha = [I_K - P^0 \bar{R}(d_0^{-1} I_m + \bar{R}' P^0 \bar{R} + d_0\, \mathcal{E}' P^0 \mathcal{E})^{-1} \bar{R}'] \mathbf{1}_\alpha. \tag{53}$$

Next we made two observations regarding the upper bound on $\mathbf{v}_\alpha$. First, it can be bounded by a constant C_α that does not depend on d or n since $P^0 = P^0(\mu, \lambda, K)$ is a function of the penalty terms μ, λ and the number of synapses K, and $\bar{R}$ and $\bar{\mathcal{E}}$ are functions of spiking rate $\bar{\mathbf{r}}$ and noise $\bar{\boldsymbol{\varepsilon}}$, respectively, thus

$$\|\mathbf{v}_\alpha\|_2 \leq C_\alpha. \tag{54}$$

Second, the constant C_α does not depend on μ and K_α if they become sufficiently large. This can be seen from the fact that P^0 converged elementwise as μ or K_α become arbitrarily large. Specifically,

$$P^0 \approx \frac{1}{\lambda} I_K - \frac{1}{\lambda K_E} \mathbf{1}_E \mathbf{1}_E' - \frac{1}{\lambda K_I} \mathbf{1}_I \mathbf{1}_I' \quad \text{if} \quad \mu \gg 1 \tag{55}$$

$$P^0 \approx \frac{1}{\lambda} I_K \quad \text{if} \quad K_E, K_I \gg 1. \tag{56}$$

These observations led to the uniform upper bound

$$\|\widehat{\boldsymbol{\sigma}}_\alpha^d\|_2 \leq \|\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}\|_2 \leq C_\alpha \sigma_\alpha^0 \quad \text{for} \quad d \gg d_0 \gg 1, \tag{57}$$

where C_α does not depend on d, μ and K_α.

Applying this upper bound to the synaptic learning rule (Eq. 34), we obtained that the update size of the sums of excitatory and inhibitory synaptic weights, respectively, at each update are constrained by

$$|\delta z_\alpha^n| \approx |e^n \widehat{\boldsymbol{\sigma}}_\alpha^d \cdot \boldsymbol{r}^n| \tag{58}$$

$$\leq |e^n| \cdot \|\widehat{\boldsymbol{\sigma}}_\alpha^{d_0}\|_2 \cdot \|\mathbf{r}^n\|_2 \tag{59}$$

$$\leq |e^n| \cdot C_\alpha \sigma_\alpha^0 \cdot K r_{max} \tag{60}$$

$$\approx |e^n| B_\alpha \mu^{-1} \quad \text{for all} \quad n \tag{61}$$

where $B_\alpha = C_\alpha r_{max} K / K_\alpha$. We used $\boldsymbol{\sigma}_\alpha^n \approx \widehat{\boldsymbol{\sigma}}_\alpha^d$ to approximate $|\delta z_\alpha^n|$ in the first line, the uniform upper bound on $\widehat{\boldsymbol{\sigma}}_\alpha^d$ (Eq. 57) in the second and third lines, and $\sigma_\alpha^0 = O(1/(\mu K_\alpha))$ (Eq. 37) in the last line.

This estimate showed that the upper bound on the sum of synaptic weights (1) is inversely proportional to μ and (2) does not depend on the number of synaptic connections K_α modified under the ROWSUM-regularization as long as the proportion $q_\alpha = K_\alpha / K$ remains constant. Finally, we concluded that the upper bound on the average synaptic weights was inversely proportional to the product μK_α:

$$|\overline{\delta z_\alpha^n}| = K_\alpha^{-1} |\delta z_\alpha^n| \leq |e^n| B_\alpha (\mu K_\alpha)^{-1}. \tag{62}$$

Acknowledgments

This work was supported by the Intramural Program of the NIH, NIDDK.

References

Baker, C., Zhu, V., and Rosenbaum, R. (2019). Nonlinear stimulus representations in neural circuits with approximate excitatory-inhibitory balance. *BioRxiv*, page 841684.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., and Maass, W. (2018). Long short-term memory and learning-to-learn in networks of spiking neurons. In *Advances in Neural Information Processing Systems*, pages 787–797.

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. (2019). A solution to the learning dilemma for recurrent networks of spiking neurons. *bioRxiv*, page 738385.

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of computational neuroscience*, 8(3):183–208.

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., and Wang, X.-J. (2017). Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504–1517.

DePasquale, B., Churchland, M. M., and Abbott, L. (2016). Using firing-rate dynamics to train recurrent networks of spiking model neurons. *arXiv preprint arXiv:1601.07620*.

DePasquale, B., Cueva, C. J., Rajan, K., Abbott, L., et al. (2018). full-force: A target-based method for training recurrent networks. *PloS one*, 13(2):e0191527.

Eschbach, C., Fushiki, A., Winding, M., Schneider-Mizell, C. M., Shao, M., Arruda, R., Eichler, K., Valdes-Aleman, J., Ohyama, T., Thum, A. S., Gerber, B., Fetter, R. D., Truman, J. W., Litwin-Kumar, A., Cardona, A., and Zlatic, M. (2020). Recurrent architecture for adaptive regulation of learning in the insect brain. *Nature Neuroscience*, 23(4):544–555. Number: 4 Publisher: Nature Publishing Group.

Haykin, S. (1996). *Adaptive Filter Theory (3rd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Huh, D. and Sejnowski, T. J. (2018). Gradient descent for spiking neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 1433–1443. Curran Associates, Inc.

Ingrosso, A. and Abbott, L. (2019). Training dynamically balanced excitatory-inhibitory networks. *PloS one*, 14(8).

Kim, C. M. and Chow, C. C. (2018). Learning recurrent dynamics in spiking networks. *eLife*, 7:e37124.

Kim, R., Li, Y., and Sejnowski, T. J. (2019). Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the National Academy of Sciences*, 116(45):22811–22820.

Laje, R. and Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, 16(7):925–933.

Litwin-Kumar, A. and Turaga, S. C. (2019). Constraining computational models using electron microscopy wiring diagrams. *Current Opinion in Neurobiology*, 58:94–100.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84.

Murphy, B. K. and Miller, K. D. (2009). Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648.

Nicola, W. and Clopath, C. (2017). Supervised learning in spiking neural networks with force training. *Nature communications*, 8(1):2208.

Ostojic, S. (2014). Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons. *Nature neuroscience*, 17(4):594–600.

Rajan, K., Harvey, C. D., and Tank, D. W. (2016). Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142.

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. D. (2010). The asynchronous state in cortical circuits. *science*, 327(5965):587–590.

Rosenbaum, R. and Doiron, B. (2014). Balanced networks of spiking neurons with spatially dependent recurrent connections. *Physical Review X*, 4(2):021039.

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature neuroscience*, 20(1):107.

Sadeh, S. and Clopath, C. (2020). Patterned perturbation of inhibition can reveal the dynamical structure of neural processing. *eLife*, 9:e52757.

Sanzeni, A., Akitake, B., Goldbach, H. C., Leedy, C. E., Brunel, N., and Histed, M. H. (2019). Inhibition stabilization is a widespread property of cortical networks. *bioRxiv*, page 656710.

Sompolinsky, Crisanti, and Sommers (1988). Chaos in random neural networks. *Physical review letters*, 61(3):259–262.

Song, H. F., Yang, G. R., and Wang, X.-J. (2016). Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2).

Sussillo, D. and Abbott, L. (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron*, 63(4):544–557.

Sussillo, D., Churchland, M. M., Kaufman, M. T., and Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7):1025–1033.

Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., and McNaughton, B. L. (1997). Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388.

van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724.

van Vreeswijk, C. and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371.

Wang, J., Narain, D., Hosseini, E. A., and Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature neuroscience*, 21(1):102–110.

Zarin, A. A., Mark, B., Cardona, A., Litwin-Kumar, A., and Doe, C. Q. (2019). A multilayer circuit architecture for the generation of distinct locomotor behaviors in Drosophila. *eLife*, 8:e51781. Publisher: eLife Sciences Publications, Ltd.