bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Synapse-type-specific competitive Hebbian learning forms functional recurrent networks

Samuel Eckmann[*,1] **and Julijana Gjorgjieva**[1,2]

[1]Max Planck Institute for Brain Research, Frankfurt am Main, Germany
[2]TUM School of Life Sciences, Technical University Munich, Germany
[*]Corresponding author. Email: ec.sam@outlook.com

Cortical networks exhibit complex stimulus-response patterns. Previous work has identified the balance between excitatory and inhibitory currents as a central component of cortical computations, but has not considered how the required synaptic connectivity emerges from biologically plausible plasticity rules. Using theory and modeling, we demonstrate how a wide range of cortical response properties can arise from Hebbian learning that is stabilized by the synapse-type-specific competition for synaptic resources. In fully plastic recurrent circuits, this competition enables the development and decorrelation of inhibition-balanced receptive fields. Networks develop an assembly structure with stronger connections between similarly tuned neurons and exhibit response normalization and surround suppression. These results demonstrate how neurons can self-organize into functional circuits and provide a foundational understanding of plasticity in recurrent networks.

Computation in neural circuits is based on the interactions between recurrently connected excitatory (E) and inhibitory (I) neurons.[1–4] In sensory cortices, response normalization, surround and gain modulation, predictive processing, and attention all critically involve inhibitory neurons.[5–10] Theoretical work has highlighted the experimentally observed balance of stimulus selective excitatory and inhibitory input currents as a critical requirement for many neural computations.[11–16] For example, recent models based on balanced E-I networks have explained a wide range of cortical phenomena, such as cross-orientation and surround suppression,[17,18] as well as stimulus-induced neural variability.[19,20] A major caveat of these models is that the network connectivity is usually static and designed by hand. How recurrent synaptic weights self-organize in a biologically plausible manner to generate the non-linear response properties observed experimentally is unknown. Earlier theoretical work on inhibitory plasticity has focused on the balance of excitation and inhibition in single neurons,[21–23] but has not been able to explain the development of inhibition balanced receptive fields when excitatory and inhibitory inputs are both plastic. In more recent recurrent network models, only a fraction of excitatory and inhibitory synapse-types are modeled as plastic, and neural responses exhibit a narrow subset of the different response patterns recorded in experiments.[14,24–29]

Here we present a Hebbian learning framework with minimal assumptions that explains a wide range of experimental observations. In our framework, synaptic strengths evolve according to a Hebbian plasticity rule that is stabilized by the competition for a limited supply of synaptic resources.[30–32] Motivated by the unique protein composition of excitatory and inhibitory synapses, our key assumption is that different synapse-types compete for separate resource pools, which enables the self-organization into functional recurrent networks.

To understand plasticity in recurrently connected E-I networks, we considered simplified circuits of increasing complexity. We first asked how E-I balance and stimulus selectivity can simultaneously develop in a single neuron. The neuron receives input from an upstream population of excitatory neurons, and disynaptic inhibitory input from a population of laterally connected inhibitory neurons that themselves receive input from the same upstream population (Fig. 1A). We studied the self-organization of excitatory and inhibitory synapses that project onto the single post-synaptic neuron (Fig. 1B), assuming that input synapses that project onto inhibitory neurons remained fixed (Fig. 1A). Following experimental results,[33–36] we assumed that inhibitory and excitatory input neurons are equally selective for the orientation of a stimulus grating (Fig. 1C, bottom). We presented uniformly distributed oriented stimuli to the network in random order. Stimuli elicited a Gaussian-shaped response in the population of input neurons (Fig. 1C, top) and thus drove the post-synaptic neuron (see Supplementary Material (SM) Sec. 1 for method details). Synapses are plastic according to a basic Hebbian rule:

$$\Delta \mathbf{w}_A \propto \mathbf{y}_A r, \quad A \in \{E, I\}, \tag{1}$$

where r is the post-synaptic firing rate, and $\mathbf{y}_A$ is a vector that holds the pre-synaptic firing rates of excitatory ($A = E$) and inhibitory ($A = I$) neurons. Experimental results have shown that after the induction of long-term plasticity neither the total excitatory nor the total inhibitory synaptic area change.[31] This suggests that a synapse can only grow at the expense of another synapse – a competitive mechanism potentially mediated by the limited supply of synaptic proteins (Fig. 1D).[32] Motivated by these results, we adopted a competitive normalization rule for both excitatory and inhibitory synapses:

$$\mathbf{w}_A \leftarrow W_A \frac{\mathbf{w}_A + \Delta \mathbf{w}_A}{\|\mathbf{w}_A + \Delta \mathbf{w}_A\|}, \tag{2}$$

where $A \in \{E, I\}$, and W_E, W_I are the maintained total excitatory and inhibitory synaptic weight, respectively. Shortly after random initialization, excitatory and inhibitory weights stabilize (Fig. 1E) and form aligned, Gaussian-shaped tuning curves (Fig. 1F) that reflect the shape of the input stimuli (Fig. 1C). As a result, neural responses become orientation selective while inhibitory and excitatory inputs are equally tuned, which demonstrates the joint development of stimulus selectivity and E-I balance.

To uncover the principles of synapse-type-specific competitive Hebbian learning, we analyzed the feedforward model analytically. It is well established that in the absence of inhibition, competitive Hebbian learning rules generate stimulus selective excitatory receptive fields.[30,37] In the case of a linear activation function, $f(u) \propto u$, the expected total synaptic efficacy changes can be expressed as[30] (SM Sec. 2):

$$\langle \dot{\mathbf{w}}_E \rangle \propto \mathbf{C}\mathbf{w}_E - \gamma \mathbf{w}_E, \tag{3}$$

were $\mathbf{C} = \langle \mathbf{y}_E \mathbf{y}_E^\mathsf{T} \rangle$ is the input covariance matrix, with $\langle \cdot \rangle$ being the temporal average, and γ is a scalar normalization factor that regulates Hebbian growth. Then, fixed points, for which $\langle \dot{\mathbf{w}}_E \rangle = 0$, are eigenvectors of the covariance matrix, i.e., the neuron becomes selective to the first principal component of its input data.[30,37] For a non-linear activation function $f(\mathbf{u})$, neurons become selective for higher-order correlations in their inputs.[38–40] In the following, we call the fixed points of such pure feedforward circuits 'input modes'.

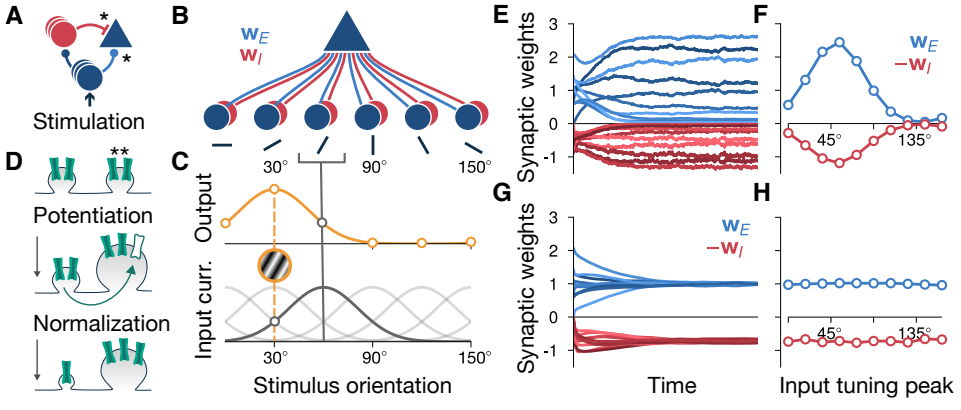

Figure 1: Synapse-type-specific competitive Hebbian learning enables the development of stimulus selectivity and inhibitory balance. (**A**) Feedforward input to a pyramidal neuron consists of direct excitation and disynaptic inhibition. Plastic synapses are marked by *. (**B**) A single post-synaptic pyramidal neuron, targeted by excitatory (E) and inhibitory (I) synapses. (**C**) Excitatory and inhibitory input neurons are equally tuned to the orientation of a stimulus grating (bottom, tuning curve of neurons tuned to 60° highlighted in dark gray) and exhibit a Gaussian-shaped population response when a single grating of 30° is presented (orange plate, dashed line). (**D**) Hebbian potentiation of a synapse (**) is bounded due to a limited amount of synaptic resources, here reflected by a fixed number of synaptic channels. (**E**) Weight convergence of synapses where inhibitory weights are plastic according to synapse-type-specific competitive Hebbian learning. (**F**) Final synaptic weight strength, after training, as a function of the tuning peak of the pre-synaptic neurons. (**G**) & (**H**) Same as (E) and (F), but for classic inhibitory plasticity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

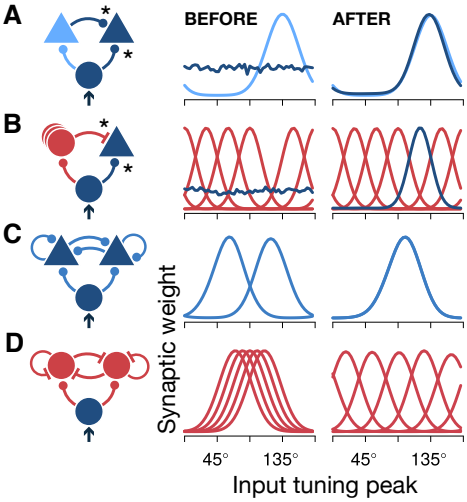

Figure 2: Feedforward tunings are affected by lateral input in microcircuit motifs. (**A**) In addition to feedforward input from a population of orientation tuned excitatory cells (blue circle), a neuron receives lateral input from an excitatory neuron with fixed feedforward tuning (light blue). * indicates plastic synapses. Feedforward tuning curves of the two neurons are shown before (center) and after (right) training. (**B**) Same as in (A), for lateral input from multiple inhibitory neurons. (**C**) Same as in (A), including recurrent excitation and self-excitation, and all synaptic connections being plastic. (**D**) Same as in (C), for inhibitory neurons.

We next examined how inhibitory plasticity affects the development of stimulus selectivity. Previous work has suggested that inhibitory synaptic plasticity in the cortex is Hebbian[41,42] and imposes a target firing rate r_0 on the post-synaptic neuron:[22]

$$\dot{\mathbf{w}}_I \propto \mathbf{y}_I\,(r - r_0)\,. \tag{4}$$

When excitatory synaptic weights remain fixed, this 'classic' inhibitory plasticity leads to balanced excitatory and inhibitory input currents.[22] However, when excitatory synaptic weights are also plastic, neurons develop no stimulus selectivity:[23] Classic inhibitory plasticity must act on a faster timescale than excitatory plasticity to maintain stability.[23] Then the post-synaptic target firing rate is consistently met and average excitatory synaptic weight changes only differ amongst each other due to different average pre-synaptic firing rates, which prevents the development of stimulus selectivity (Fig. 1G, and H) (SM Sec. 2.2).

Synapse-type-specific competitive Hebbian learning (Eq. 1, and 2) can solve this problem. As in Eq. 3, we incorporated the normalization step (Eq. 2) into the update rule (Eq. 1) and considered the simpler case of a linear activation function $f(u) \propto u$ (SM Sec. 3):

$$\langle \dot{\mathbf{w}} \rangle \propto \bar{\mathbf{C}}\mathbf{w} - \gamma \begin{pmatrix} \mathbf{w}_E \\ \mathbf{0} \end{pmatrix} - \rho \begin{pmatrix} \mathbf{0} \\ \mathbf{w}_I \end{pmatrix}, \tag{5}$$

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_E \\ \mathbf{w}_I \end{pmatrix}, \quad \bar{\mathbf{C}} \equiv \left\langle \begin{pmatrix} \mathbf{y}_E\mathbf{y}_E{}^{\mathsf{T}} & -\mathbf{y}_E\mathbf{y}_I{}^{\mathsf{T}} \\ \mathbf{y}_I\mathbf{y}_E{}^{\mathsf{T}} & -\mathbf{y}_I\mathbf{y}_I{}^{\mathsf{T}} \end{pmatrix} \right\rangle, \tag{6}$$

where γ and ρ are scalars that ensure normalization, and we defined the modified covariance matrix $\bar{\mathbf{C}}$. Now the fixed points of the weight dynamics are multiples of the excitatory and the inhibitory part of the eigenvectors of the modified covariance matrix $\bar{\mathbf{C}}$. When excitatory and inhibitory inputs are equally stimulus selective, such that one can approximate $\mathbf{y}_E \propto \mathbf{y}_I$, the modified covariance matrix $\bar{\mathbf{C}}$ is composed of multiples of the original covariance matrix C. This implies that excitatory and inhibitory synaptic weights eventually have identical shape, $\mathbf{w}_E \propto \mathbf{w}_I$, equal to a multiple of an eigenvector of C. Neurons become selective for activity along one particular input direction, while excitatory and inhibitory neural inputs are balanced, which explains the joint development of stimulus selectivity and E-I balance in feedforward circuits, in agreement with our numerical simulations (Fig. 1).

We next investigated the effect of synapse-type-specific competitive Hebbian learning in recurrent networks. In a first step, we considered how lateral input from an excitatory neuron with fixed selectivity for a specific feedforward input mode affects synaptic weight dynamics in a microcircuit motif (Fig. 2A, left). We observed that a downstream neuron becomes preferentially tuned to the feedforward input mode of the lateral projecting neuron (Fig. 2A, right)(compare SM Sec. 4). Similarly, laterally projecting inhibitory neurons repel downstream neurons from their input modes (Fig. 2B). However, when two excitatory neurons are reciprocally connected, they pull each other towards their respective input modes, and their tuning curves and activities become correlated (Fig. 2C). This contradicts experimental observations that brain activity decorrelates over development.[43,44] Recent experimental results have suggested that inhibitory neurons drive decorrelation of neural activities.[45,46] In line with these results, in our model, interconnected inhibitory neurons repel each other and their tuning curves decorrelate (Fig. 2D).

Hence, we asked whether the interaction between excitatory and inhibitory neurons can also decorrelate excitatory neural activities. To address this question we explored the consequences of synapse-type-specific competitive Hebbian learning in a network of recurrently connected excitatory and inhibitory neurons (Fig. 3A, top). We presented different oriented gratings in random order in a network where all feedforward and recurrent synapses are plastic. We observed a sharp increase in response selectivity (Fig. 3A, bottom) that is reflected in the reconfiguration of feedforward synaptic weights. Shortly after random initialization (Fig. 3B, top), excitatory neurons predominantly connect to a subset of input neurons with similar stimulus selectivities (Fig. 3B, center). Different from circuits without inhibition (compare Fig. 2C), tuning curves of excitatory as well as inhibitory neurons decorrelate and cover the whole stimulus space with minimal overlap (Fig. 3B, bottom). After training, synaptic connections become organized in an assembly-like structure, according to their tuning similarity (Fig. 3C, and D) as is observed experimentally.[47–57] We found that inhibitory neurons become as selective for stimulus orientations as excitatory neurons[33–36] (Fig. 3E), while wider inhibitory input (Fig. 3F) from multiple overlapping inhibitory neurons (Fig. 3G) increases neurons' tuning selectivities, in agreement with experimental results.[12,58–61] In summary, synapse-type-specific competitive Hebbian learning in fully plastic recurrent networks can decorrelate neural activities and leads to preferential connectivity between similarly tuned neurons, as observed in cortical circuits.

To uncover how recurrent inhibition can prevent all neurons to become selective for a single input mode, we investigated the fundamental principles of synapse-type-specific competitive Hebbian learning in recurrent networks analytically (SM Sec. 6). To prevent the collapse of all tuning curves onto the same dominant input mode (compare Fig. 2A, and C), we find that its effective attraction has to decrease with the number of neurons that are tuned to that mode (SM Sec. 6.3). In the simplified case of linear activation

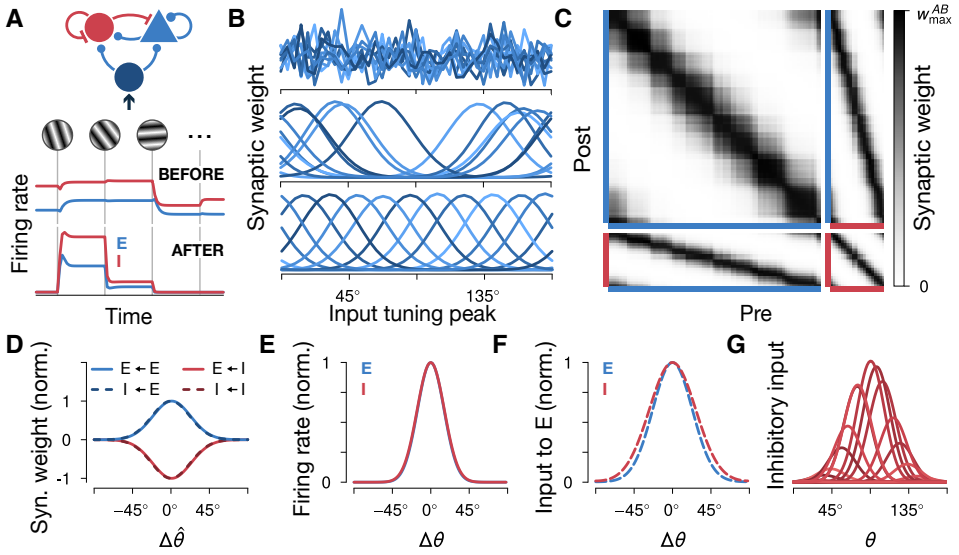

Figure 3: Tuning curve decorrelation in plastic recurrent networks. (**A**) Top: A population of recurrently connected excitatory and inhibitory neurons receives input from a set of input neurons that are tuned to different stimulus orientations (compare Fig. 1B, bottom). Every 200ms a different orientation is presented to the network (vertical gray lines). At the same time, all synapses exhibit plasticity according to a synapse-type-specific Hebbian rule (see Appendix 1 for details). Bottom: Exemplary firing rate activity of one excitatory and one inhibitory neuron before and after training. (**B**) Feedforward tuning curves of $N_E = 10$ excitatory neurons before (top), during (center), and after (bottom) training. (**C**) Connectivity matrices for $N_E = 80$ excitatory (blue) and $N_I = 20$ inhibitory (red) neurons after training. Neurons are sorted according to their preferred orientation $\hat{\theta}$. $w^{AB}_{\max}$ is the largest synaptic weight between population A and B; $A, B \in \{E, I\}$. (**D**) Normalized (norm.) recurrent weight strengths as a function of the difference between preferred orientations of pre- and post-synaptic neurons, $\Delta\hat{\theta} = \hat{\theta}_{\text{post}} - \hat{\theta}_{\text{pre}}$, averaged over all neuron pairs. (**E**) Average firing rate response of inhibitory and excitatory neurons to a stimulus orientation θ, relative to their preferred orientation, $\Delta\theta = \hat{\theta} - \theta$, averaged over all neurons. (**F**) Same as (E) for excitatory and inhibitory inputs to excitatory neurons. (**G**) Inhibitory input to an excitatory neuron with preferred orientation close to 90°. Each curve corresponds to the input from one pre-synaptic inhibitory neuron.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

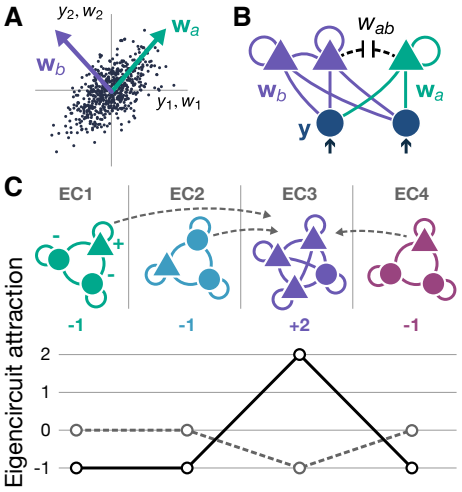

Figure 4: Eigencircuit decomposition and attraction. (**A**) Synaptic weight vectors $\mathbf{w}_a$, $\mathbf{w}_b$ of two neurons that are tuned to two different principle components (top, purple and green) of the input data (each dark blue dot represents one pre-synaptic firing pattern). (**B**) Synaptic weights between differently tuned neurons w_{ab} decay to zero, while neurons tuned to the same eigenvector form an eigencircuit with recurrent connectivity (purple). (**C**) A recurrent network with four eigencircuits (EC). Each excitatory neuron contributes plus one (+), each inhibitory neuron minus one (-) to the total eigencircuit attraction. Due to synaptic plasticity, neurons are pulled towards the most attractive eigencircuit (gray dashed arrows). When all neurons are part of the same eigencircuit (EC3), its attraction becomes negative (bottom).

functions, input modes are eigenvectors of the input covariance matrix (compare Eq.3). Since these eigenvectors are orthogonal by definition, the activities of neurons that are tuned to different eigenvectors are uncorrelated, and their reciprocal connections decay to zero under Hebbian plasticity (Fig. 4A). Then, neurons that are tuned to the same eigenvector form recurrent 'eigencircuits' that are otherwise separated from the rest of the network (SM Sec. 5). Crucially, this decomposition of the network into eigencircuits allows us to write the effective attraction $\hat{\lambda}$ of an input mode as the sum of a feedforward component λ and the variances of the neurons that reside in the eigencircuit (Fig. 4B):

$$\hat{\lambda} = \lambda + \lambda_{\text{eig}} = \lambda + \sum_i \sigma^2_{E,i} - \sum_j \sigma^2_{I,j}, \qquad (7)$$

where we defined the eigencircuit attraction λ_{eig}, and variances σ^2 depend on the total synaptic weights, and the number of excitatory neurons and inhibitory neurons in the eigencircuit. This reveals that the attractive and repulsive effects of excitatory and inhibitory neurons balance each other. In a simplified example, we assumed that all feedforward input modes have equal attraction λ, while each excitatory neuron contributes plus one and each inhibitory neuron minus one to the total attraction (Fig. 4C). Then the eigencircuit attraction becomes $\lambda_{\text{eig}} = n_E - n_I$. All neurons become attracted to the same eigencircuit, which suggests that all tuning curves will eventually collapse onto the same input mode. However, when all neurons become selective to the most attractive input mode, that mode becomes repulsive (Fig. 4C, bottom, dashed line), as each increase in attraction due to additional excitatory neurons is balanced by a decrease in attraction due to additional inhibitory neurons. Consequently, the resulting eigencircuit becomes unstable and neurons become attracted to non-repulsive input modes. This prevents the collapse of tuning curves onto a single input mode, and demonstrates how neurons decorrelate due to recurrent inhibitory interactions.

Our results thus far reveal how synapse-type-specific competitive Hebbian learning can explain the development of structured recurrent connectivity. We next asked whether synapse-type-specific competitive Hebbian learning can also explain the emergence of network computations. For example, neurons in the visual cortex respond to multiple overlayed oriented gratings in a non-linear fashion.[62,63] After training our network with single grating stimuli (Fig. 5A), we found that neural responses to a cross-oriented mask grating that is presented in addition to a regular test grating are normalized, i.e., the response to the combined stimulus is weaker than the sum of the responses to the individual gratings (Fig. 5B, left). When the contrast of the mask grating is lower than the test grating's, the network responds in a winner-takes-all fashion: The higher-contrast test grating dominates activities while the lower-contrast mask grating is suppressed (Fig. 5B, right). As observed experimentally,[62,63] we found that suppression is divisive and shifts the log-scale contrast-response function to the right (Fig. 5C).

We next investigated how the stimulus statistics during training affect receptive field properties. We considered a plastic network where neurons receive tuned input from either a center or a surround region of the visual field (Fig. 5D). During training, we presented either the same oriented grating in both regions (Fig. 5E, top), or a single grating in just one region (Fig. 5E, bottom). We found that after training, the response of a center-tuned neuron exhibits feature-specific surround suppression, reflecting the stimulus statistics during training. When the center and the surround are stimulated separately during training, iso- and cross-oriented stimuli in the surround elicit minimal suppression (Fig. 5F, pink). In case of correlated stimulation of center and surround, suppression is stronger for iso- compared to cross-orientations (Fig. 5F, purple), as has been reported experimentally.[64–66] Co-tuned excitatory and inhibitory inputs mediate suppression from the surround to the center region (Fig. 5G). Such a balance of excitatory and inhibitory lateral inputs has previously been observed in experiments.[67] Together this demonstrates that synapse-type-specific Hebbian learning produces extra-classical receptive fields that modulate feedforward responses via recurrent interactions in accordance with experimental results.

Competitive interactions between synapses have been observed in many different preparations and have been attributed to various mechanisms.[31,68–76] The local competition for a limited supply of synaptic building blocks is a biologically plausible normalization mechanism.[32,77,78] Many synaptic proteins are specific to inhibitory or excitatory synapses and reside in one synapse-type, but not the other.[79] Therefore, in this work we assumed a synapse-type specific competition for different synaptic resource pools and implemented separate normalization constants for inhibitory and excitatory synapses. On a finer scale, synapses of different excitatory and inhibitory neuron subtypes also differ in their protein composition.[80–82] In principle, this allows for the precise regulation of different input pathways via the adjustment of subtype-specific resource pools.[83–89] We anticipate such pathway-specific competition to be crucial for the functional development of any network with multiple neuron subtypes.

In the brain, total synaptic strengths are dynamic and homeostatically regulated on a timescale of hours to days.[90–92] In addition to maintaining average firing rates in response to network-scale perturbations, a prominent framework puts forward homeostatic scaling of synaptic strengths as a stabilizing mechanism of Hebbian growth.[93] However, theoretical models suggest that homeostatic scaling is too slow to balance rapid synaptic plasticity.[94] In our networks, Hebbian growth is instead thought to be stabilized by the rapid competition for a limited pool of synapse-type specific resources, while total synaptic strengths remain fixed. In line with

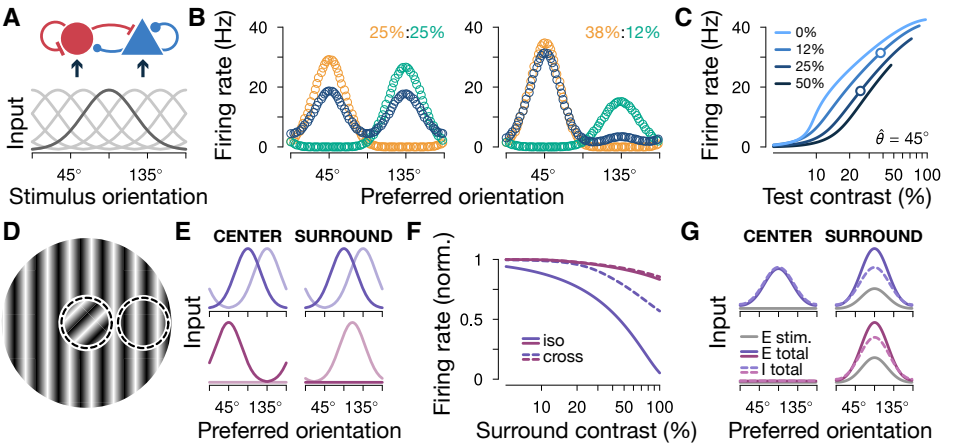

Figure 5: Cross-orientation and surround suppression in a trained neural network. (**A**) A plastic network of excitatory and inhibitory neurons (top) receives input according to fixed tuning curves (bottom). Amplitude corresponds to stimulus contrast. Tuning curve of neurons with preferred orientation of 90° highlighted in dark gray. (**B**) Response of 80 excitatory neurons to a test grating (orange, 45°) and a mask grating (green, 135°) of different contrast levels (inset). Gratings are presented separately (orange & green) or together (dark blue). Each circle corresponds to the response of one excitatory neuron. (**C**) Contrast response curve of a single excitatory neuron (preferred orientation $\hat{\theta} = 45°$) to the test and mask gratings in (B). Different mask contrasts are indicated by color. Circles correspond to contrast levels in (B). (**D**) Center and surround receptive fields with different oriented stimuli. (**E**) Feedforward inputs for two example stimuli (one solid, one transparent) when center and surround are correlated (top) or stimulated separately (bottom). (**F**) Response of one excitatory neuron to center and surround stimulation after training. A center stimulus of preferred orientation was presented at constant contrast while the contrast of a cross- or iso-oriented surround stimulus changed. Colors as in (E). (**G**) Input to excitatory neurons during stimulation (stim.) of the surround region.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

these results, we suggest that scaling of synaptic strengths may not be required for immediate network stability but instead controls the operating regime of the network.[16,95,96]

Our results suggest that synapse-type-specific competitive Hebbian learning is the key mechanism that enables the formation of functional recurrent networks. Rather than hand-tuning connectivity to selectively explain experimental data, our networks emerge from unsupervised, biologically plausible learning rules. In a single framework, they readily explain multiple experimental observations, including the development of stimulus selectivity, excitation-inhibition balance, decorrelated neural activity, assembly structures, response normalization, and orientation-specific surround suppression. Our results demonstrate how the connectivity of inhibition balanced networks is shaped by their input statistics and explain the experience-dependent formation of extra-classical receptive fields.[97–101] In our model, circuit formation depends only on the statistical regularities between input streams and is agnostic to their origin. Therefore, we expect our approach to extend beyond sensory cortices and to provide a fundamental framework for plasticity in recurrent neural networks.

REFERENCES AND NOTES

1. R. Tremblay, S. Lee, B. Rudy, *Neuron* **91**, 260–292 (2016).
2. R. Hattori, K. V. Kuchibhotla, R. C. Froemke, T. Komiyama, *Nature neuroscience* **20**, 1199–1208 (2017).
3. K. A. Pelkey, R. Chittajallu, M. T. Craig, L. Tricoire, J. C. Wester, C. J. McBain, *Physiological reviews* **97**, 1619–1747 (2017).
4. A. Kepecs, G. Fishell, *Nature* **505**, 318–326 (2014).
5. M. Carandini, D. J. Heeger, *Nature Reviews Neuroscience* **13**, 51 (2012).
6. A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, P. C. Bressloff, *Annual review of neuroscience* **40**, 425–451 (2017).
7. K. C. Wood, J. M. Blackwell, M. N. Geffen, *Current opinion in neurobiology* **46**, 200–207 (2017).
8. G. B. Keller, T. D. Mrsic-Flogel, *Neuron* **100**, 424–435 (2018).
9. O. K. Swanson, A. Maffei, *Frontiers in molecular neuroscience* **12**, 168 (2019).
10. K. A. Ferguson, J. A. Cardin, *Nature Reviews Neuroscience* **21**, 80–92 (2020).
11. C. Van Vreeswijk, H. Sompolinsky, *Science* **274**, 1724–1726 (1996).
12. J. S. Isaacson, M. Scanziani, *Neuron* **72**, 231–243 (2011).
13. S. Denève, C. K. Machens, *Nature neuroscience* **19**, 375–382 (2016).
14. G. Hennequin, E. J. Agnes, T. P. Vogels, *Annual review of neuroscience* **40**, 557–579 (2017).
15. S. Sadeh, C. Clopath, *Nature Reviews Neuroscience* **22**, 21–37 (2021).
16. Y. Ahmadian, K. D. Miller, *Neuron* (2021).
17. H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, D. Ferster, *Neuron* **62**, 578–592 (2009).
18. D. B. Rubin, S. D. Van Hooser, K. D. Miller, *Neuron* **85**, 402–417 (2015).
19. G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, K. D. Miller, *Neuron* **98**, 846–860 (2018).
20. R. Echeveste, L. Aitchison, G. Hennequin, M. Lengyel, *Nature neuroscience* **23**, 1138–1149 (2020).
21. Y. Luz, M. Shamir, *PLoS computational biology* **8**, e1002334 (2012).
22. T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, W. Gerstner, *Science* **334**, 1569–1573 (2011).
23. C. Clopath, T. P. Vogels, R. C. Froemke, H. Sprekeler, *BioRxiv*, 066589 (2016).
24. P. D. King, J. Zylberberg, M. R. DeWeese, *Journal of Neuroscience* **33**, 5475–5485 (2013).
25. A. Litwin-Kumar, B. Doiron, *Nature communications* **5**, 1–12 (2014).
26. F. Zenke, E. J. Agnes, W. Gerstner, *Nature communications* **6**, 1–13 (2015).
27. O. Mackwood, L. B. Naumann, H. Sprekeler, *Elife* **10**, e59715 (2021).
28. E. J. Agnes, T. P. Vogels, *bioRxiv* (2021).
29. S. Soldado-Magraner, H. Motanis, R. Laje, D. V. Buonomano, *bioRxiv*, 2020–12 (2021).
30. K. D. Miller, D. J. MacKay, *Neural computation* **6**, 100–126 (1994).
31. J. N. Bourne, K. M. Harris, *Hippocampus* **21**, 354–373 (2011).
32. J. Triesch, A. D. Vo, A.-S. Hafner, *Elife* **7**, e37836 (2018).
33. J. A. Hirsch, L. M. Martinez, C. Pillai, J.-M. Alonso, Q. Wang, F. T. Sommer, *Nature neuroscience* **6**, 1300–1308 (2003).
34. J. A. Cardin, L. A. Palmer, D. Contreras, *Journal of Neuroscience* **27**, 10333–10344 (2007).
35. C. A. Runyan, J. Schummers, A. Van Wart, S. J. Kuhlman, N. R. Wilson, Z. J. Huang, M. Sur, *Neuron* **67**, 847–857 (2010).
36. A. K. Moore, M. Wehr, *Journal of neuroscience* **33**, 13713–13723 (2013).
37. E. Oja, *Journal of mathematical biology* **15**, 267–273 (1982).
38. E. Oja, *Proceedings of the ICANN'91, 1991*, 385–390 (1991).
39. C. S. Brito, W. Gerstner, *PLoS computational biology* **12**, e1005070 (2016).
40. G. Ocker, M. Buice, *Advances in Neural Information Processing Systems* **34** (2021).
41. J. A. D'Amour, R. C. Froemke, *Neuron* **86**, 514–528 (2015).
42. F. Lagzi, M. C. Bustos, A.-M. Oswald, B. Doiron, *bioRxiv* (2021).
43. P. Golshani, J. T. Gonçalves, S. Khoshkhoo, R. Mostany, S. Smirnakis, C. Portera-Cailliau, *Journal of Neuroscience* **29**, 10890–10899 (2009).
44. N. L. Rochefort, O. Garaschuk, R.-I. Milos, M. Narushima, N. Marandi, B. Pichler, Y. Kovalchuk, A. Konnerth, *Proceedings of the National Academy of Sciences* **106**, 15049–15054 (2009).
45. H. N. Mulholland, B. Hein, M. Kaschube, G. B. Smith, *Elife* **10**, e72456 (2021).
46. M. Chini, T. Pfeffer, I. L. Hanganu-Opatz, *bioRxiv* (2021).
47. C. D. Gilbert, T. N. Wiesel, *Journal of Neuroscience* **9**, 2432–2442 (1989).
48. Y. Yoshimura, J. L. Dantzker, E. M. Callaway, *Nature* **433**, 868–873 (2005).
49. Y. Yoshimura, E. M. Callaway, *Nature neuroscience* **8**, 1552–1559 (2005).
50. D. E. Wilson, G. B. Smith, A. L. Jacob, T. Walker, J. Dimidschstein, G. Fishell, D. Fitzpatrick, *Neuron* **93**, 1058–1065 (2017).
51. H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, T. D. Mrsic-Flogel, *Nature* **473**, 87–91 (2011).
52. A. D. Lien, M. Scanziani, *Nature neuroscience* **16**, 1315–1323 (2013).
53. Y.-t. Li, L. A. Ibrahim, B.-h. Liu, L. I. Zhang, H. W. Tao, *Nature neuroscience* **16**, 1324–1330 (2013).
54. L.-y. Li, Y.-t. Li, M. Zhou, H. W. Tao, L. I. Zhang, *Nature neuroscience* **16**, 1179–1181 (2013).
55. L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, T. D. Mrsic-Flogel, *Nature* **518**, 399–403 (2015).
56. M. F. Iacaruso, I. T. Gasler, S. B. Hofer, *Nature* **547**, 449–452, issn: 14764687 (2017).
57. P. Znamenskiy, M.-H. Kim, D. R. Muir, M. F. Iacaruso, S. B. Hofer, T. D. Mrsic-Flogel, *Biorxiv*, 294835 (2018).
58. D. Rose, C. Blakemore, *Nature* **249**, 375–377 (1974).
59. X. Pei, T. Vidyasagar, M. Volgushev, O. Creutzfeldt, *Journal of Neuroscience* **14**, 7130–7140 (1994).
60. G. K. Wu, R. Arbuckle, B.-h. Liu, H. W. Tao, L. I. Zhang, *Neuron* **58**, 132–143 (2008).
61. D. E. Wilson, B. Scholl, D. Fitzpatrick, *Nature* **560**, 97–101 (2018).
62. L. Busse, A. R. Wade, M. Carandini, *Neuron* **64**, 931–942 (2009).
63. S. P. MacEvoy, T. R. Tucker, D. Fitzpatrick, *Nature Neuro.* **12**, 637–645 (2009).
64. C. Blakemore, E. A. Tobin, *Experimental brain research* **15**, 439–440 (1972).
65. J. J. Knierim, D. C. Van Essen, *Journal of neurophysiology* **67**, 961–980 (1992).
66. J. R. Cavanaugh, W. Bair, J. A. Movshon, *Journal of neurophysiology* (2002).
67. H. Adesnik, M. Scanziani, *Nature* **464**, 1155–1160 (2010).
68. Y.-J. Lo, M.-m. Poo, *Science* **254**, 1019–1022 (1991).
69. M. Scanziani, R. C. Malenka, R. A. Nicoll, *Nature* **380**, 446–450 (1996).
70. S. Royer, D. Paré, *Nature* **422**, 518–522 (2003).
71. A. Govindarajan, I. Israely, S.-Y. Huang, S. Tonegawa, *Neuron* **69**, 132–146 (2011).
72. W. C. Oh, L. K. Parajuli, K. Zito, *Cell reports* **10**, 162–169 (2015).
73. S. El-Boustani, J. P. Ip, V. Breton-Provencher, G. W. Knott, H. Okuno, H. Bito, M. Sur, *Science* **360**, 1349–1354 (2018).
74. G. Antunes, F. Simoes-de-Souza, *Scientific reports* **8**, 1–14 (2018).
75. A. Perez-Alvarez, S. Yin, C. Schulze, J. A. Hammer, W. Wagner, T. G. Oertner, *Nature communications* **11**, 1–10 (2020).
76. T. Ravasenga, M. Ruben, A. I. Polenghi, E. M. Petrini, A. Barberis, *bioRxiv* (2021).
77. N. W. Gray, R. M. Weimer, I. Bureau, K. Svoboda, *PLoS biology* **4**, e370 (2006).
78. S. H. Lee, C. Jin, E. Cai, P. Ge, Y. Ishitsuka, K. W. Teng, A. A. De Thomaz, D. Nall, M. Baday, O. Jeyifous, *et al.*, *Elife* **6**, e27744 (2017).
79. M. Sheng, E. Kim, *Cold Spring Harbor perspectives in biology* **3**, a005678 (2011).
80. A. Gupta, Y. Wang, H. Markram, *Science* **287**, 273–278 (2000).
81. A. M. Craig, H. Boudin, *Nature neuroscience* **4**, 569–578 (2001).
82. G. H. Diering, R. L. Huganir, *Neuron* **100**, 314–329 (2018).
83. J. J. Zhu, *Journal of Neuroscience* **29**, 6320–6335 (2009).
84. J. A. Wen, A. L. Barth, *Journal of Neuroscience* **31**, 4456–4465 (2011).
85. J. N. Levinson, A. El-Husseini, *Neuron* **48**, 171–174 (2005).
86. A. A. Chubykin, D. Atasoy, M. R. Etherton, N. Brose, E. T. Kavalali, J. R. Gibson, T. C. Südhof, *Neuron* **54**, 919–931 (2007).
87. R. S. Larsen, P. J. Sjöström, *Current opinion in neurobiology* **35**, 127–135 (2015).
88. M. E. Horn, R. A. Nicoll, *Proceedings of the National Academy of Sciences* **115**, 589–594 (2018).
89. K. D. Fox, A. Pandey, N. R. Hardingham, *bioRxiv* (2022).
90. G. G. Turrigiano, K. R. Leslie, N. S. Desai, L. C. Rutherford, S. B. Nelson, *Nature* **391**, 892–896 (1998).
91. G. G. Turrigiano, S. B. Nelson, *Nature reviews neuroscience* **5**, 97–107 (2004).
92. P. Wenner, *Neural plasticity* **2011** (2011).
93. G. G. Turrigiano, *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160258 (2017).
94. F. Zenke, W. Gerstner, S. Ganguli, *Current opinion in neurobiology* **43**, 166–176 (2017).
95. Y. Ahmadian, D. B. Rubin, K. D. Miller, *Neural Comp.* **25**, 1994–2037 (2013).
96. N. Kraynyukova, T. Tchumatchenko, *Proceedings of the National Academy of Sciences* **115**, 3464–3469 (2018).
97. M. Pecka, Y. Han, E. Sader, T. D. Mrsic-Flogel, *Neuron* **84**, 457–469 (2014).
98. S. V. David, W. E. Vinje, J. L. Gallant, *Journal of Neuroscience* **24**, 6991–7006 (2004).
99. G. Felsen, J. Touryan, F. Han, Y. Dan, *PLoS biology* **3**, e342 (2005).
100. G. Felsen, J. Touryan, Y. Dan, *Network: Computation in Neural Systems* **16**, 139–149 (2005).
101. E. Froudarakis, P. Berens, A. S. Ecker, R. J. Cotton, F. H. Sinz, D. Yatsenko, P. Saggau, M. Bethge, A. S. Tolias, *Nature neuroscience* **17**, 851–857 (2014).

ACKNOWLEDGEMENTS

We thank all members of the Gjorgjieva group for useful feedback throughout the project. This work was supported by the Max Planck Society and the European Research Council (StG 804824 to JG).

AUTHOR CONTRIBUTIONS

SE conceived research with input from GJ. SE performed simulations, derived mathematical results, and prepared figures. SE and JG wrote the manuscript.

SUPPLEMENTARY MATERIAL

SM Text Sections 1 to 6,
Figures S1 to S4,
Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Material for

Synapse-type-specific competitive Hebbian learning forms functional recurrent networks

Samuel Eckmann[*,1] **and Julijana Gjorgjieva**[1,2]

[1]Max Planck Institute for Brain Research, Frankfurt am Main, Germany
[2]TUM School of Life Sciences, Technical University Munich, Germany
[*]Corresponding author. Email: ec.sam@outlook.com

Contents

1 Methods & simulation parameters

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We consider networks of rate coding excitatory (E) and inhibitory (I) neurons that receive input from themselves and a population of feedforward input neurons (F). Membrane potential vectors $\mathbf{u}$ evolve according to

$$\tau_A \dot{\mathbf{u}}_A = -\mathbf{u}_A + \sum_B \mathbf{W}_{AB}\mathbf{r}_B - \mathbf{W}_{AI}\mathbf{r}_I, \quad A \in \{E, I\}, \quad B \in \{E, I, F\} \tag{1}$$

where τ_A is the activity timescale, $\mathbf{W}_{AB}$ are matrices that hold synaptic weights between the pre-synaptic population B and the post-synaptic population A. All dynamics were numerically integrated using the Euler method in timesteps of Δt. Entries of weight matrices were drawn from a half-normal distribution with its peak at zero, and normalized before the start of a simulation (see below). Firing rate vectors $\mathbf{r}_A$ are given as a function $f(\mathbf{u}_A)$ of the membrane potential $\mathbf{u}_A$:

$$\mathbf{r}_A = f(\mathbf{u}_A), \quad f(\mathbf{u}_A) = a[\mathbf{u}_A - b]_+^n, \quad A \in \{E, I\} \tag{2}$$

with $[\cdot]_+ = \max(0, \cdot)$ and scalar constants a, b, and n.

Plasticity and Normalization

Plastic weights evolve according to a Hebbian plasticity rule

$$\dot{\mathbf{W}}_{AB} = \epsilon_{AB}\mathbf{r}_A \odot \mathbf{r}_B, \quad A \in \{E, I\}, \quad B \in \{E, I, F\} \tag{3}$$

where ϵ_{AB} is a scalar learning rate, and $\odot$ indicates the outer product. After each plasticity step, synaptic weights are normalized such that the total excitatory and inhibitory post-synaptic weights are maintainted:

$$w_{AB}^{(ij)} \leftarrow W_{AE}\frac{w_{AB}^{(ij)}}{\sum_j w_{AE}^{(ij)} + \sum_k w_{AF}^{(ik)}}, \quad w_{AI}^{(ij)} \leftarrow W_{AI}\frac{w_{AI}^{(ij)}}{\sum_j w_{AI}^{(ij)}}, \quad A \in \{E, I\}, \quad B \in \{E, F\}, \tag{4}$$

where W_{AE}, W_{AI} are the total excitatory and inhibitory synaptic weights.

In Fig. 1, we set the activity of the inhibitory input neurons equal to the activity of the excitatory input neurons, i.e., $\mathbf{r}_I = \mathbf{r}_F$. For panels E & F of Fig. 1, inhibitory weights evolved according to the classic inhibitory plasticity rule[1] without normalization:

$$\dot{\mathbf{w}}_{EI} = \epsilon_{EI}(r_E - r_0)\mathbf{r}_I, \tag{5}$$

where r_0 is a target firing rate.

Input model

The activity of feedforward input neurons depend on the orientation θ and contrast c of an input grating:

$$\mathbf{r}_F = cA_F \exp\left(\frac{|\theta, \boldsymbol{\theta}_F|^2}{2\sigma_F^2}\right), \tag{6}$$

where the vector $\boldsymbol{\theta}_F$ holds the preferred orientations of the input neurons that are evenly distributioned between 0 and 180°, $|\cdot, \cdot|$ is the angular distance, σ_F is the tuning width, and A_F the maximum firing rate. During training, single gratings, sampled from a uniform distribution between 0° and 180°, were presented to the network for 200ms, before the next stimulus was selected.

In Fig. 5 network stimulation is realized via static feedforward weights. Neuron were assigned a preferred orientation $\hat{\theta}$, evenly distributed between 0° and 180°. Static feedforward weights were initialized as

$$\mathbf{W}_{AF} = \exp\left(\frac{|\hat{\theta}, \boldsymbol{\theta}_F|^2}{2\sigma_W^2}\right). \tag{7}$$

Feedforward weights are normalized to W_{AF} and are not taken into account when normalizing recurrent weights. The activity of input neurons $\mathbf{r}_F$ was determined as described above. Parameters were selected to result in stimulation patterns as inRubin et al..[2] Weight norms W_{AB} were also adapted from Rubin et al.[2]

See Table S1 for an overview of used simulation parameters. Python code will be made available after journal publication.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure	1 E,F	1 G,H	4 A,B	4 C,D,E,F,G	5 A,B,C	5 D,E,F,G
N_E	1	1	10	80	80	80×2
N_I	10	10	10	20	20	20×2
N_F	10	10	40	80	80	80×2
a	1	1	0.04	0.04	0.04	0.04
b	0.25	0.25	0	0	0	0
n	2	2	2	2	2	2
W_{EE}	10	10	2	0.6	3.51	3.51
W_{IE}	-	-	2	0.85	3.35	3.35
W_{EI}	-	5	0.8	0.3	1.84	1.84
W_{II}	-	-	0.5	0.35	1.44	1.44
W_{EF}	-	-	-	-	1.4	1.4
W_{IF}	-	-	-	-	1.4	1.4
c	1	1	1	1	0.5	0.5
A_F	1	1	35	140	80	80
σ_F	20°	20°	12°	12°	$30°/\sqrt{2}$	$30°/\sqrt{2}$
σ_A	-	-	-	-	$30°/\sqrt{2}$	$30°/\sqrt{2}$
Δt	200ms	200ms	10ms	10ms	10ms	10ms
τ_E	200ms	200ms	20ms	20ms	25ms	25ms
τ_I	-	-	17ms	17ms	12.5ms	12.5ms
ϵ_{EE}	-	-	2×10^{-9}ms^{-1}	1.0×10^{-10}ms^{-1}	1.0×10^{-9}ms^{-1}	1.0×10^{-9}ms^{-1}
ϵ_{IE}	-	-	3×10^{-9}ms^{-1}	1.5×10^{-10}ms^{-1}	1.5×10^{-9}ms^{-1}	1.5×10^{-9}ms^{-1}
ϵ_{EI}	4×10^{-4}ms^{-1}	4×10^{-4}ms^{-1}	4×10^{-9}ms^{-1}	2.0×10^{-10}ms^{-1}	2.0×10^{-9}ms^{-1}	2.0×10^{-9}ms^{-1}
ϵ_{II}	-	-	5×10^{-9}ms^{-1}	2.5×10^{-10}ms^{-1}	2.5×10^{-9}ms^{-1}	2.5×10^{-9}ms^{-1}
ϵ_{EF}	2×10^{-4}ms^{-1}	2×10^{-4}ms^{-1}	ϵ_{EE}	ϵ_{EE}	-	-
ϵ_{IF}	-	-	ϵ_{IE}	ϵ_{IE}	-	-
r_0	0.25	-	-	-	-	-

Table S1: Simulation parameters

2 Linear competitive Hebbian learning finds principal components

Before considering inhibitory plasticity, we recapitulate how linear Hebbian learning finds the principal eigenvector of a neuron's inputs. Although first described by Oja,[3] we will mostly follow the derivation by Miller and MacKay[4] that we will later extend to inhibitory neurons.

2.1 Hebbian plasticity without normalization is unstable

We consider a single neuron that receives input from a set of excitatory neurons (Fig. S1A). Its output firing rate r is a weighted sum of the firing rates of its presynaptic inputs $\mathbf{y}$. One can conveniently write this as a dot product:

$$\tau \dot{r} = -r + \sum_i w_i y_i = -r + \mathbf{w}^\mathsf{T} \mathbf{y}, \tag{8}$$

where $\mathbf{w}$ is a vector that holds the synaptic weights, and τ defines the timescale at which the activity changes. The transpose is denoted by $^\mathsf{T}$. In the following, lowercase letters in bold indicate vectors and uppercase letters in bold

matrices or tensors. Following Hebb's principle, synaptic weight changes depend on the pre- and postsynaptic firing rates. In vector notation,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\tau_w \dot{\mathbf{w}} = \mathbf{y} r \tag{9}$$

where the constant τ_w sets the timescale. Assuming that synaptic weights change on a much slower timescale than firing rates, $\tau \ll \tau_w$, we make the simplifying assumption that r reaches its fixed point instantaneously, $r = \mathbf{w}^\mathsf{T}\mathbf{y}$. Then, the average change of the synaptic weights can be expressed as a linear transformation of the original weight vector:

$$\langle \dot{\mathbf{w}} \rangle = \langle \mathbf{y} r \rangle = \langle \mathbf{y}\mathbf{y}^T \mathbf{w} \rangle = \mathbf{C}w, \quad C \equiv \langle \mathbf{y}\mathbf{y}^\mathsf{T} \rangle, \tag{10}$$

where $\langle \cdot \rangle$ is a temporal average and $\mathbf{C}$ is the covariance matrix of the synaptic inputs $\mathbf{y}$, assuming inputs have zero mean, $\langle \mathbf{y} \rangle = \mathbf{0}$. In the following, we only consider the average weight changes and omit the angled notation for convenience. For better readability, we set $\tau_w = 1$. To solve this differential equation, one can express the weight change in the eigenvector basis of the covariance matrix $\mathbf{C}$, which is symmetric and, therefore, has a complete set of orthonormal eigenvectors.

$$\dot{\mathbf{w}}_v \equiv \mathbf{V}^\mathsf{T}\dot{\mathbf{w}} = \mathbf{V}^\mathsf{T}\mathbf{C}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{w} = \Lambda \mathbf{w}_v, \tag{11}$$

$$\Rightarrow \mathbf{w}_v = \mathbf{w}_v(t_0)\exp(\Lambda t). \tag{12}$$

Here, each column of $\mathbf{V}$ holds an eigenvector, and Λ is the diagonal eigenvalue matrix. Each eigenvector component grows exponentially with the rate given by the respective eigenvalue. To constrain this unlimited growth, one can modify the Hebbian learning rule such that it maintains the total synaptic weight.

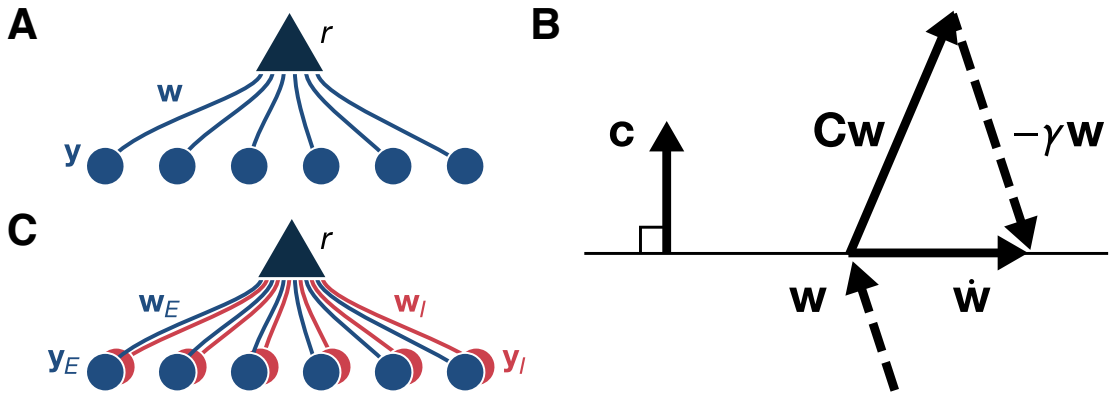

Figure S1: (**A**) A postsynaptic neuron with output firing rate r receives synapses $\mathbf{w}$ from a set excitatory neurons with firing rates y_E. (**B**) The normalization operation constrains synaptic weight changes $\dot{\mathbf{w}}$ to a hyperplane that is perpendicular to the constraint vector $\mathbf{c}$ by substracting a multiple γ of the weightvector $\mathbf{w}$. See text for details. Figure adapted from Miller and MacKay.[4] (**C**) A postsynaptic neuron with output firing rate r receives excitatory synapses $\mathbf{w}_E$ from a set of excitatory neurons with firing rates $\mathbf{y}_E$, and inhibitory synapses $\mathbf{w}_I$ from a set of inhibitory neurons with firing rates $\mathbf{y}_I$.

2.2 Weight constraints stabilize unlimited Hebbian growth

Hebbian plasticity and weight normalization can be considered as two discrete steps. First, growing weights according to the Hebbian rule. Second, normalizing to maintain the total synaptic weight. In this section, we will follow Miller and MacKay[4] and show how one can integrate these two discrete steps into one and derive the effective weight change $\dot{\mathbf{w}}$. One can write the two steps as

$$\tilde{\mathbf{w}} = \mathbf{w}(t) + \mathbf{C}\mathbf{w}\Delta t, \quad \mathbf{w}(t+\Delta t) = \frac{W}{\mathbf{c}^\mathsf{T}\tilde{\mathbf{w}}}\tilde{\mathbf{w}}, \quad W \equiv \mathbf{c}^\mathsf{T}\mathbf{w}(t). \tag{13}$$

This update rule maintains the projection of $\mathbf{w}$ onto the constraint vector $\mathbf{c}$ by multiplicatively scaling the weight vector after the Hebbian learning step, i.e., $\tilde{\mathbf{w}}$. Alternatively, if we let W be a constant, the projection onto $\mathbf{c}$ would be constrained to be equal to that constant. In the following, we instead assume that the weights are already properly normalized and set the projection value as it was before the plasticity timestep, i.e., equal to W as defined above.

$$\mathbf{w}(t+\Delta t) = \beta\left[\mathbf{w}(t) + \mathbf{C}\mathbf{w}(t)\Delta t\right], \quad \beta(\mathbf{w}(t), \Delta t) = \frac{\mathbf{c}^\mathsf{T}\mathbf{w}(t)}{\mathbf{c}^\mathsf{T}\left[\mathbf{C}\mathbf{w}(t)\Delta t + \mathbf{w}(t)\right]}, \tag{14}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483999; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where β describes the multiplicative normalization that depends on the size of the timestep Δt and the previous weight $\mathbf{w}(t)$. It is straightforward to check that the projection of the weight vector onto the constraint vector $\mathbf{c}$ does not change, i.e.,

$$\mathbf{c}^\mathsf{T}\mathbf{w}(t+\Delta t) = \mathbf{c}^\mathsf{T}\mathbf{w}(t). \tag{15}$$

Then, the effective weight change $\dot{\mathbf{w}}$ is given as

$$\dot{\mathbf{w}} = \lim_{\Delta t\to 0}\frac{\mathbf{w}(t+\Delta t)-\mathbf{w}(t)}{\Delta t} = \lim_{\Delta t\to 0}\left[\mathbf{C}\mathbf{w}(t) - \frac{1-\beta}{\beta\Delta t}\mathbf{w}(t+\Delta t)\right], \tag{16}$$

After taking the limit, one gets

$$\lim_{\Delta t\to 0}\frac{1-\beta}{\beta\Delta t} = \frac{\mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{w}}{\mathbf{c}^\mathsf{T}\mathbf{w}}. \tag{17}$$

And finally (compare Fig. S1B)

$$\boxed{\Rightarrow \dot{\mathbf{w}} = \mathbf{C}\mathbf{w} - \gamma\mathbf{w}, \quad \gamma \equiv \frac{\mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{w}}{\mathbf{c}^\mathsf{T}\mathbf{w}}.} \tag{18}$$

Here, γ is a scalar normalization factor that depends on the current weight $\mathbf{w}$. An alternative way to derive $\dot{\mathbf{w}}$ is to guess the shape of the multiplicative normalization term in Eq. 18 and require that the change along the constraint vector is zero, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{c}^\mathsf{T}\mathbf{w}\right) = \mathbf{c}^\mathsf{T}\dot{\mathbf{w}} = \mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{w} - \gamma\mathbf{c}^\mathsf{T}\mathbf{w} \stackrel{!}{=} 0, \quad \Rightarrow \gamma = \frac{\mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{w}}{\mathbf{c}^\mathsf{T}\mathbf{w}}. \tag{19}$$

Note that for $\mathbf{c}$ being a constant vector of ones, the L1-norm of the weight vector is maintained. However, $\mathbf{c}$ does not have to be constant. For example, for $\mathbf{c} = \mathbf{w}$ the L2-norm is maintained. Also note that one can analogously derive effective plasticitiy rules when weights are constrained via subtractive normalization with the ansatz $\dot{\mathbf{w}} = \mathbf{C}\mathbf{w} - \zeta\mathbf{k}$, where $\mathbf{k}$ is a vector of ones.[4]

Fixed points

From Eq. 18 it is clear that multiples of eigenvectors $\mathbf{v}$ of $\mathbf{C}$ are fixed points $\mathbf{w}^*$, for which $\dot{\mathbf{w}}^* = 0$. Explicitely, for a scalar constant a and $\mathbf{w}^* = a\mathbf{v}$ one gets:

$$\dot{\mathbf{w}}^* = a\mathbf{C}\mathbf{v} - \frac{\mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{v}}{\mathbf{c}^\mathsf{T}\mathbf{v}}a\mathbf{v} = a\lambda\mathbf{v} - \frac{\mathbf{c}^\mathsf{T}\lambda\mathbf{v}}{\mathbf{c}^\mathsf{T}\mathbf{v}}a\mathbf{v} = 0. \tag{20}$$

Note that this is independent of the choice of the constraint vector $\mathbf{c}$.

Stability analysis

Multiplicative normalization constrains the norm of the weight vector and therefore prevents the otherwise unlimited growth of Hebbian plasticity. However, in theory, it is still possible that the system is unstable and never settles into a fixed point. Following Miller and MacKay,[4] we will now explorer under what conditions fixed points are stable.

Formally, a fixed point in a linear system is stable when the largest eigenvalue of the Jacobian is negative, or marginally stable when it is equal to zero.[5] The weight dynamics around a fixed point $\mathbf{w}^*$ can be approximated with its Taylor expansion

$$\dot{\mathbf{w}} \approx \dot{\mathbf{w}}^* + \sum_i \left.\frac{\mathrm{d}\dot{\mathbf{w}}}{\mathrm{d}w_i}\right|_* (w_i - w_i^*), \tag{21}$$

$$= \left.\frac{\mathrm{d}\dot{\mathbf{w}}}{\mathrm{d}\mathbf{w}}\right|_* (\mathbf{w} - \mathbf{w}^*), \tag{22}$$

$$= \mathbf{J}^*(\mathbf{w} - \mathbf{w}^*). \tag{23}$$

where $\dot{\mathbf{w}}^*$ is zero, by definition, and $\mathbf{J}^*$ is the Jacobian evaluated at the fixed point. The Jacobian is defined as

$$\mathbf{J}^* \equiv \begin{pmatrix} \left.\frac{\mathrm{d}\dot{w}_1}{\mathrm{d}w_1}\right|_* & \cdots & \left.\frac{\mathrm{d}\dot{w}_1}{\mathrm{d}w_N}\right|_* \\ \vdots & & \vdots \\ \left.\frac{\mathrm{d}\dot{w}_N}{\mathrm{d}w_1}\right|_* & \cdots & \left.\frac{\mathrm{d}\dot{w}_N}{\mathrm{d}w_N}\right|_* \end{pmatrix} \equiv \left.\frac{\mathrm{d}\dot{\mathbf{w}}}{\mathrm{d}\mathbf{w}}\right|_* . \tag{24}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-4.0 International license.

A fixed point is stable if small perturbations away from the fixed point, $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^*$, decay to zero, i.e.,

$$\frac{d}{dt}\Delta\mathbf{w} = \dot{\mathbf{w}} - \dot{\mathbf{w}}^* = \dot{\mathbf{w}} = \mathbf{J}^*\Delta\mathbf{w}, \tag{25}$$

where we approximated $\dot{\mathbf{w}}$ with its Taylor expansion, since the perturbation is small, i.e., $\mathbf{w}$ is close to the fixed point. The result is a linear differential equation that one can solve as

$$\Delta\mathbf{w}(t) = \Delta\mathbf{w}(t_0)\exp(\mathbf{J}^* t), \tag{26}$$

where all vector components decay to zero if all eigenvalues of $\mathbf{J}^*$ are negative[1]. As we will see later, it is useful to rewrite the weight dynamics (Eq. 18) as

$$\dot{\mathbf{w}} = \mathbf{C}\mathbf{w} - \mathbf{w}\gamma, \tag{27}$$

$$= \mathbf{C}\mathbf{w} - \frac{\mathbf{w}\mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{w}}{\mathbf{c}^\mathsf{T}\mathbf{w}}, \tag{28}$$

$$= \left[\mathbb{1} - \frac{\mathbf{w}\mathbf{c}^\mathsf{T}}{\mathbf{c}^\mathsf{T}\mathbf{w}}\right]\mathbf{C}\mathbf{w}. \tag{29}$$

It follows[2]:

$$\left.\frac{d\dot{\mathbf{w}}}{d\mathbf{w}}\right|_* = \left[\mathbb{1} - \frac{\mathbf{v}^*\mathbf{c}^\mathsf{T}}{\mathbf{c}^\mathsf{T}\mathbf{v}^*}\right]\mathbf{C} + \left[-\frac{\mathbb{1}\mathbf{c}^\mathsf{T}}{\mathbf{c}^\mathsf{T}\mathbf{v}^*} + \frac{\mathbf{v}^*\mathbf{c}^\mathsf{T}\mathbf{c}^\mathsf{T}}{(\mathbf{c}^\mathsf{T}\mathbf{v}^*)^2}\right]\mathbf{C}\mathbf{v}^*, \tag{30}$$

$$= \left[\mathbb{1} - \frac{\mathbf{v}^*\mathbf{c}^\mathsf{T}}{\mathbf{c}^\mathsf{T}\mathbf{v}^*}\right][\mathbf{C} - \lambda^*\mathbb{1}], \tag{31}$$

where $\mathbf{w}\big|_* = \mathbf{w}^* = a\mathbf{v}^*$ is the fixed point with $\mathbf{v}^*$ being an eigenvector of $\mathbf{C}$. The scalar a is the length of the fixed point weight vector $\mathbf{w}^*$ (which cancels) and λ^* is the eigenvalue to $\mathbf{v}^*$. To find the eigenvalues of the Jacobian, $\lambda_\mathbf{J}$, we diagonalize $\mathbf{J}$ by switching to the eigenbasis of $\mathbf{C}$. When $\mathbf{V}$ is the matrix that holds the eigenvectors of $\mathbf{C}$ as columns one gets

$$\mathbf{V}^\mathsf{T}\left.\frac{d\dot{\mathbf{w}}}{d\mathbf{w}}\right|_*\mathbf{V} = \left[\mathbb{1} - \mathbf{V}^\mathsf{T}\frac{\mathbf{v}^*\mathbf{c}^\mathsf{T}}{\mathbf{c}^\mathsf{T}\mathbf{v}^*}\mathbf{V}\right]\left[\mathbf{V}^\mathsf{T}\mathbf{C}\mathbf{V} - \lambda^*\mathbb{1}\right], \tag{32}$$

$$= \left[\mathbb{1} - \mathbf{e}^*\frac{\mathbf{c}^\mathsf{T}\mathbf{V}}{\mathbf{c}^\mathsf{T}\mathbf{v}^*}\right][\Lambda - \lambda^*\mathbb{1}], \tag{33}$$

where Λ is a diagonal matrix that holds the eigenvalues of $\mathbf{C}$. Without loss of generality, we can assume that the first column of $\mathbf{V}$ is equal to $\mathbf{v}^*$. Then $\mathbf{e}^* = \mathbf{V}^\mathsf{T}\mathbf{v}^*$ is a column vector of zeros, except for the first entry, that is equal to one. Then, the first bracket becomes an upper triangular matrix with ones on the diagonal, except for the first diagonal entry, which is zero. From this, it follows[3] that the eigenvalues of the Jacobian are

$$\boxed{\lambda_\mathbf{J} = \lambda - \lambda^*.} \tag{34}$$

If λ^* is the largest eigenvalue, i.e., $\mathbf{w}^*$ is a multiple of the principal eigenvector of $\mathbf{C}$, then all $\lambda_\mathbf{J}$ are negative, or zero and the fixed point is marginally stable. If there exists a $\lambda > \lambda^*$, the corresponding $\lambda_\mathbf{J}$ is negative and the fixed point is unstable. In summary, linear Hebbian learning combined with multiplicative normalization finds the principal eigenvector of the input covariance matrix and thus performs principal component analysis (PCA).

Classic Inhibitory plasticity prevents stimulus selectivity

Previous work suggested a homeostatic inhibitory synaptic plasticity rule[1] that enforced a post-synaptic target firing rate r_0:

$$\dot{\mathbf{w}}_I \propto \mathbf{y}_I\,(r - r_0)\,. \tag{35}$$

However, when combined with excitatory plasticity, this classic rule prevents the development of stimulus selectivity.[6] For completeness, we briefly recapitulate this result presented in Clopath et al.:[6] Classic inhibitory plasticity is

[1] This can be seen by formulating the system in the eigenbasis of $\mathbf{J}^*$. Then, the matrix exponential becomes: $\mathbf{V}^{-1}\exp(\mathbf{J}^*)\mathbf{V} = \exp(\Lambda_\mathbf{J})$, where $\mathbf{V}$ holds eigenvectors and $\Lambda_\mathbf{J}$ is a diagonal matrix that holds the eigenvalues of $\mathbf{J}^*$.

[2] To make sense of the vector notation, it helps to first consider the b'th column of $\frac{d\dot{\mathbf{w}}}{d\mathbf{w}}$ which is equal to $\frac{d\dot{\mathbf{w}}}{dw_b}$, where w_b is the b'th vector component of $\mathbf{w}$.

[3] Because the eigenvalues of a product of two triangular matrices is equal to the product of their eigenvalues.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483999; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

required to act faster than excitatory plasticity to enable stable weight dynamics.[6] For much faster inhibitory plasticity, the dynamics of excitatory and inhibitory weights decouples and fixed points of the inhibitory weights $\mathbf{w}_I^*$ can be considered separately from the fixed points of excitatory weights. When excitatory and inhibitory inputs are equally stimulus selective, the fast dynamics of inhibitory weights ensures that the target firing rate is consistently met, i.e., the post-synaptic neuron always responds with the same firing rate r^*.

$$\dot{\mathbf{w}}_I^* = 0 \quad \Rightarrow \quad r^* = r_0 \quad \Rightarrow \quad \dot{\mathbf{w}}_E \propto \mathbf{y}_E r_0 - \text{normalization} \tag{36}$$

When all pre-synaptic neurons have similar average firing rates, $\langle y_E \rangle_i \approx y_0$, and weights change on a slower timescale than activities as is the case biologically, the average excitatory synaptic weight change becomes

$$\langle \dot{\mathbf{w}}_E \rangle \propto \mathbf{c} y_0 r_0 - \text{normalization}, \tag{37}$$

where $\mathbf{c}$ is a vector of ones. The average synaptic weight change is identical across synapses, which prevents the development of stimulus selectivity (Fig. 1E & F). Therefore, classic inhibitory plasticity that enforces a target firing rate cannot explain the joint development of stimulus selectivity and inhibitory balance.

3 Subtype-specific normalization balances E-I receptive fields

We generalize the approach outlined above to the case of simultaneous excitatory and inhibitory normalization. We consider a simplified case of one postsynaptic neuron with firing rate r that receives input from a set of excitatory and inhibitory neurons with firing rates $\mathbf{y_E}$ and $\mathbf{y_I}$, respectively (Fig. S1C). Again, we assume fast activity dynamics and write the activity fixed point as

$$r = \mathbf{y}_E{}^\mathsf{T} \mathbf{w}_E - \mathbf{y}_I{}^\mathsf{T} \mathbf{w}_I{}^\mathsf{T} \equiv \mathbf{y}^\mathsf{T} \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix} \mathbf{w}, \tag{38}$$

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_E \\ \mathbf{w}_I \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_E \\ \mathbf{y}_I \end{pmatrix}, \tag{39}$$

where $\mathbb{1}$ is the unit matrix with appropriate dimension, and we defined the modified weight and input vectors, $\mathbf{w}$ and $\mathbf{y}$. Then one can write the Hebbian part of the time-averaged weight dynamics as

$$\tau \langle \dot{\mathbf{w}} \rangle = \langle \mathbf{y} r \rangle = \left\langle \mathbf{y}\mathbf{y}^\mathsf{T} \right\rangle \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix} \mathbf{w}, \tag{40}$$

$$= \left\langle \begin{pmatrix} \mathbf{y}_E \mathbf{y}_E{}^\mathsf{T} & -\mathbf{y}_E \mathbf{y}_I{}^\mathsf{T} \\ \mathbf{y}_I \mathbf{y}_E{}^\mathsf{T} & -\mathbf{y}_I \mathbf{y}_I{}^\mathsf{T} \end{pmatrix} \right\rangle \mathbf{w} \equiv \bar{\mathbf{C}} \mathbf{w}, \tag{41}$$

where we defined the modified covariance matrix $\bar{\mathbf{C}}$. The matrix τ holds the timescales of excitatory plasticity, τ_E, and inhibitory plasticity, τ_I, as diagonal entries and is zero otherwise. In the following, we drop the bracket notation $\langle \cdot \rangle$ for better readability. As in the case of only excitatory input, we can implement multiplicative normalization by additional constraint terms. Now also for inhibitory weights (compare Eq. 18):

$$\boxed{\tau \dot{\mathbf{w}} = \bar{\mathbf{C}} \mathbf{w} - \gamma \mathbf{w}_E^\circ - \rho \mathbf{w}_I^\circ}, \tag{42}$$

$$\mathbf{w}_E^\circ = \begin{pmatrix} \mathbf{w}_E \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{w}_I^\circ = \begin{pmatrix} \mathbf{0} \\ \mathbf{w}_I \end{pmatrix}, \tag{43}$$

where $\mathbf{0}$ are vectors of zeros of appropriate dimension. The constraint factors γ and ρ follow from the requirement that the weight vector does not grow along the direction of the constraint vectors $\mathbf{c}_E^\circ$ and $\mathbf{c}_I^\circ$. Here we choose them such that the sums over the excitatory and inhibitory weights remain constant, i.e., the L1-norm of the excitatory and inhibitory part of the weight vector is maintained[1]:

$$\mathbf{c}_E^{\circ\mathsf{T}} \dot{\mathbf{w}} \stackrel{!}{=} 0, \quad \mathbf{c}_I^{\circ\mathsf{T}} \dot{\mathbf{w}} \stackrel{!}{=} 0, \tag{44}$$

$$\mathbf{c}_E^{\circ\mathsf{T}} \equiv \Big(1, \dots, 1, 0, \dots, 0\Big), \quad \mathbf{c}_I^{\circ\mathsf{T}} \equiv \Big(0, \dots, 0, 1, \dots, 1\Big). \tag{45}$$

$$\Rightarrow \quad \gamma = \frac{\mathbf{c}_E^{\circ\mathsf{T}} \bar{\mathbf{C}} \mathbf{w}}{\mathbf{c}_E^{\circ\mathsf{T}} \mathbf{w}_E^\circ}, \quad \rho = \frac{\mathbf{c}_I^{\circ\mathsf{T}} \bar{\mathbf{C}} \mathbf{w}}{\mathbf{c}_I^{\circ\mathsf{T}} \mathbf{w}_I^\circ}. \tag{46}$$

[1] The choice of the L1-norm is motivated by the synaptic competition for a fixed amount of resources, where, in the simplest case, each unit of resource linearly increases synaptic strengths. Higher-order L-norms do not affect results in the feedforward learning. However in recurrent networks they can lead to instabilities (see Section 5.3).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483999; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where the number of non-zero entries in $\mathbf{c}_E^\circ$ and $\mathbf{c}_I^\circ$ is equal to the number of excitatory and inhibitory neurons, respectively. Finally, we can write the weight dynamics as

$$\Rightarrow \quad \boxed{\tau\dot{\mathbf{w}} = \left[\mathbb{1} - \frac{\mathbf{w}_E^\circ \mathbf{c}_E^{\circ\mathsf{T}}}{\mathbf{c}_E^{\circ\mathsf{T}}\mathbf{w}_E^\circ} - \frac{\mathbf{w}_I^\circ \mathbf{c}_I^{\circ\mathsf{T}}}{\mathbf{c}_I^{\circ\mathsf{T}}\mathbf{w}_I^\circ}\right]\bar{\mathbf{C}}\mathbf{w}}\,. \tag{47}$$

3.1 Fixed points are multiples of eigenvector components

For the fixpoints we have to find weight vectors $\mathbf{w}^*$ for which the time derivative $\dot{\mathbf{w}}^*$ is equal to zero:

$$\dot{\mathbf{w}}^* = \bar{\mathbf{C}}\mathbf{w} - \gamma\mathbf{w}_E^{*\circ} - \rho\mathbf{w}_I^{*\circ} \tag{48}$$

$$= \bar{\mathbf{C}}\mathbf{w} - \frac{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{C}}\mathbf{w}^*}{\mathbf{c}_E^{\circ\mathsf{T}}\mathbf{w}_E^{*\circ}}\mathbf{w}_E^{*\circ} - \frac{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{C}}\mathbf{w}^*}{\mathbf{c}_I^{\circ\mathsf{T}}\mathbf{w}_I^{*\circ}}\mathbf{w}_I^{*\circ} \stackrel{!}{=} \mathbf{0}. \tag{49}$$

This equation holds if $\mathbf{w}^*$ solves

$$\boxed{\bar{\mathbf{C}}\mathbf{w}^* \stackrel{!}{=} \bar{\lambda}_E\mathbf{w}_E^{*\circ} + \bar{\lambda}_I\mathbf{w}_I^{*\circ}}\,, \tag{50}$$

where $\bar{\lambda}_E$ and $\bar{\lambda}_I$ are any scalars. It is straight forward to check that multiples of eigenvectors $\bar{\mathbf{v}}$ of the modified covariance matrix $\bar{\mathbf{C}}$ are fixed points.

$$\bar{\mathbf{C}}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}_E^\circ + \bar{\lambda}\bar{\mathbf{v}}_I^\circ \stackrel{!}{=} \bar{\lambda}_E\bar{\mathbf{v}}_E^\circ + \bar{\lambda}_I\bar{\mathbf{v}}_I^\circ \quad \Rightarrow \quad \bar{\lambda}_E = \bar{\lambda}_I = \bar{\lambda}. \tag{51}$$

In general, the fixed points depend non-trivially on the tuning of the two populations (compare Section 4, Eq. 99). However, when the inhibitory neurons are tuned to multiples of eigenvectors of the excitatory population's covariance matrix, multiples of the excitatory and inhibitory part of the eigenvectors of the modified covariance matrix $\bar{\mathbf{C}}$ are fixed points. This is what one would expect when the postsynaptic excitatory neuron and the inhibitory population both receive excitatory input from the same external brain region (compare Fig. 1A) and synapses from the external population onto inhibitory neurons are plastic according to a Hebbian rule with multiplicative normalization. In that sense, we assume that inhibitory neurons are as sharply tuned as excitatory neurons. In this scenario, we can express the firing rates of inhibitory neurons as

$$\mathbf{y}_I = \mathbf{Q}^\mathsf{T}\mathbf{y}_E = \mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{y}_E, \tag{52}$$

where each column of $\mathbf{Q} = \mathbf{VA}$ can be thought of as the feedforward weight vector of an inhibitory neuron which is equal to a multiple of an eigenvector $\mathbf{v}$ of the excitatory covariance matrix $\mathbf{C} = \langle \mathbf{y}_E\mathbf{y}_E{}^\mathsf{T}\rangle$. Then $\mathbf{V}$ holds all eigenvectors as columns, and $\mathbf{A}$ is a matrix where each multiple is the only non-zero element per column, such that $\mathbf{AA}^\mathsf{T}$ is a diagonal matrix.

Eigenvectors and eigenvalues of the modified covariance matrix

When inhibitory neurons are tuned to eigenvectors of the excitatory covariance matrix, the modified covariance matrix becomes

$$\bar{\mathbf{C}} = \left\langle \begin{pmatrix} \mathbf{y}_E\mathbf{y}_E{}^\mathsf{T} & -\mathbf{y}_E\mathbf{y}_I{}^\mathsf{T} \\ \mathbf{y}_I\mathbf{y}_E{}^\mathsf{T} & -\mathbf{y}_I\mathbf{y}_I{}^\mathsf{T} \end{pmatrix} \right\rangle = \begin{pmatrix} \mathbf{C} & -\mathbf{CVA} \\ \mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{C} & -\mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{CVA} \end{pmatrix} = \begin{pmatrix} \mathbf{C} & -\mathbf{V}\Lambda\mathbf{A} \\ \mathbf{A}^\mathsf{T}\Lambda\mathbf{V}^\mathsf{T} & -\mathbf{A}^\mathsf{T}\Lambda\mathbf{A} \end{pmatrix}. \tag{53}$$

Then a full set of linearly independent eigenvectors $\bar{\mathbf{V}}$ and their inverse $\bar{\mathbf{V}}^{-1}$ is given as

$$\bar{\mathbf{V}} = \begin{pmatrix} \mathbf{V} & \mathbf{V} \\ \mathbf{A}^\mathsf{T} & \mathbf{A}^{-1} \end{pmatrix}, \quad \bar{\mathbf{V}}^{-1} = \begin{pmatrix} (\mathbb{1} - \mathbf{AA}^\mathsf{T})^{-1} & 0 \\ 0 & (\mathbb{1} - \mathbf{AA}^\mathsf{T})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{V}^\mathsf{T} & -\mathbf{A} \\ -\mathbf{AA}^\mathsf{T}\mathbf{V}^\mathsf{T} & \mathbf{A} \end{pmatrix}, \tag{54}$$

where each column of $\bar{\mathbf{V}}$ is an non-normalized eigenvector and $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T}(\mathbf{AA}^\mathsf{T})^{-1}$. For weight vectors in the right matrix column of $\bar{\mathbf{V}}$ in Eq. 54, the excitatory and inhibitory components of the membrane potential exactly cancel, and no plasticity is induced. For multiple postsynaptic neurons with firing rates $\mathbf{r}$, where each neuron is tuned to one of these eigenvectors, one gets

$$\mathbf{r} = \mathbf{y}^\mathsf{T}\begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix}\mathbf{W}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{V} \\ \mathbf{A}^{-1} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_E \\ \mathbf{y}_I \end{pmatrix}, \quad \mathbf{y}_I = \mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{y}_E, \tag{55}$$

$$\Rightarrow \quad \mathbf{r} = \left(\mathbf{y}_E{}^\mathsf{T}, \mathbf{y}_E{}^\mathsf{T}\mathbf{VA}\right)\begin{pmatrix} \mathbf{V} \\ \mathbf{A}^{-1} \end{pmatrix} = \mathbf{0}. \tag{56}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483889; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Since the postsynaptic firing rate is zero for these eigenvectors, we call them "null eigenvectors" or "null fixed points". For usual choices for the excitatory and inhibitory weight norms where excitation is larger than inhibition, this set of fixed point solutions is irrelevant. Finally, the eigenvalue spectrum is

$$\bar{\mathbf{C}}\bar{\mathbf{V}} \equiv \bar{\mathbf{V}}\bar{\Lambda}, \quad \bar{\Lambda} = \begin{pmatrix} \Lambda(\mathbb{1} - \mathbf{A}\mathbf{A}^\mathsf{T}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{57}$$

Note that null-eigenvectors define the null-space of the $\bar{\mathbf{C}}$ matrix.

Non-eigenvector fixed points

We first consider eigenvectors that result in non-zero postsynaptic activity and show that they meet the condition. We make the ansatz that the matrix of fixed point vectors $\mathbf{W}^*$ has the shape

$$\mathbf{W}^* = \begin{pmatrix} \mathbf{V}\mathbf{K}_E \\ \mathbf{A}^\mathsf{T}\mathbf{K}_I \end{pmatrix}, \quad \bar{\mathbf{C}}\mathbf{W}^* = \begin{pmatrix} \mathbf{W}_E^* \bar{\Lambda}_E \\ \mathbf{W}_I^* \bar{\Lambda}_I \end{pmatrix}, \tag{58}$$

where $\mathbf{K}_E$ and $\mathbf{K}_I$ and are diagonal scaling matrices of arbitrary constants and the second equation follows from the fixed point condition in Eq. 50. We now show that for any $\mathbf{K}_E$, $\mathbf{K}_I$ we can find diagonal matrices $\bar{\Lambda}_E$, $\bar{\Lambda}_I$ that fulfill this condition. Note that for $\mathbf{K}_E = \mathbb{1}$ and $\mathbf{K}_I = (\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}$ columns of $\mathbf{W}^*$ are null-eigenvectors. Therefore, if true, the theorem holds for regular as well as null-eigenvectors. We write explicitly

$$\Rightarrow \quad \bar{\mathbf{C}}\mathbf{W}^* = \begin{pmatrix} \mathbf{C} & -\mathbf{V}\Lambda\mathbf{A} \\ \mathbf{A}^\mathsf{T}\Lambda\mathbf{V}^\mathsf{T} & -\mathbf{A}^\mathsf{T}\Lambda\mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{V}\mathbf{K}_E \\ \mathbf{A}^\mathsf{T}\mathbf{K}_I \end{pmatrix} = \begin{pmatrix} \mathbf{V}\mathbf{K}_E\bar{\Lambda}_E \\ \mathbf{A}^\mathsf{T}\mathbf{K}_I\bar{\Lambda}_I \end{pmatrix} \tag{59}$$

$$\mathbf{C}\mathbf{V}\mathbf{K}_E - \mathbf{V}\Lambda\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{K}_I = \mathbf{V}\mathbf{K}_E\bar{\Lambda}_E, \tag{60}$$

$$\mathbf{A}^\mathsf{T}\Lambda\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{K}_E - \mathbf{A}^\mathsf{T}\Lambda\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{K}_I = \mathbf{A}^\mathsf{T}\mathbf{K}_I\bar{\Lambda}_I. \tag{61}$$

$$\mathbf{V}\mathbf{K}_E\Lambda - \mathbf{V}\mathbf{K}_I\mathbf{A}\mathbf{A}^\mathsf{T}\Lambda = \mathbf{V}\mathbf{K}_E\bar{\Lambda}_E, \tag{62}$$

$$\mathbf{A}^\mathsf{T}\mathbf{K}_E\Lambda - \mathbf{A}^\mathsf{T}\mathbf{K}_I\Lambda\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{A}^\mathsf{T}\mathbf{K}_I\bar{\Lambda}_I, \tag{63}$$

where we made use of the fact that independent of their subscript, the $\mathbf{K}$, Λ, and $\mathbf{A}\mathbf{A}^\mathsf{T}$ matrices are diagonal and commute. By comparing the left and right sides of the equations, we find

$$\bar{\Lambda}_E = \Lambda\left(\mathbb{1} - \mathbf{K}_E^{-1}\mathbf{K}_I\mathbf{A}\mathbf{A}^\mathsf{T}\right), \tag{64}$$

$$\bar{\Lambda}_I = \Lambda\left(\mathbf{K}_I^{-1}\mathbf{K}_E - \mathbf{A}\mathbf{A}^\mathsf{T}\right), \tag{65}$$

which are diagonal matrices, as required.

3.2 Stability analysis

We first consider the stability of fixed point eigenvectors of the modified covariance matrix and discuss the general case afterwards. With Eq. 47, for the Jacobian $\mathbf{J}$ it follows (compare Eq. 31)

$$\tau\,\mathbf{J}\Big|_* = \tau\,\frac{\mathrm{d}\dot{\mathbf{w}}}{\mathrm{d}\mathbf{w}}\Big|_* = \left[\mathbb{1} - \frac{\bar{\mathbf{v}}_E^{*\circ}\mathbf{c}_E^{\circ\mathsf{T}}}{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{v}}_E^{*\circ}} - \frac{\bar{\mathbf{v}}_I^{*\circ}\mathbf{c}_I^{\circ\mathsf{T}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}}\right]\left[\bar{\mathbf{C}} - \bar{\lambda}^*\mathbb{1}\right], \tag{66}$$

where $\bar{\mathbf{v}}_E^*$ and $\bar{\mathbf{v}}_E^*$ are the excitatory and the inhibitory part of the eigenvector fixed point $\bar{\mathbf{v}}$ with eigenvalue $\bar{\lambda}^*$ and the superscript "$\circ$" indicates an additional set of zeros to reach the correct dimensionality of the vector (compare Eq. 43). To find the eigenvalues $\bar{\lambda}_\mathbf{J}$ of the Jacobian, we switch to the eigenbasis of the modified covariance matrix[1]:

$$\Rightarrow \quad \bar{\mathbf{V}}^{-1}\,\mathbf{J}\Big|_*\,\bar{\mathbf{V}} = \bar{\mathbf{V}}^{-1}\tau^{-1}\bar{\mathbf{V}}\bar{\mathbf{V}}^{-1}\left[\mathbb{1} - \frac{\bar{\mathbf{v}}_E^{*\circ}\mathbf{c}_E^{\circ\mathsf{T}}}{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{v}}_E^{*\circ}} - \frac{\bar{\mathbf{v}}_I^{*\circ}\mathbf{c}_I^{\circ\mathsf{T}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}}\right]\bar{\mathbf{V}}\left[\bar{\Lambda} - \bar{\lambda}^*\mathbb{1}\right], \tag{67}$$

where we inserted $\bar{\mathbf{V}}\bar{\mathbf{V}}^{-1} \equiv \mathbb{1}$. The result is a block diagonal matrix where each block corresponds to one regular eigenvector and its null-eigenvector, i.e., all eigenvectors with the same excitatory component. To better see this,

[1] Note that we must make use of the inverse instead of the transpose since, in general, the eigenvector matrix $\bar{\mathbf{V}}$ is not orthonormal.

we define $\epsilon \equiv \tau^{-1}$ and write.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\bar{\mathbf{V}}^{-1}\tau^{-1}\bar{\mathbf{V}} = \begin{pmatrix} (\mathbb{1} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} & 0 \\ 0 & (\mathbb{1} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{V}^\mathsf{T} & -\mathbf{A} \\ -\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \epsilon_E & 0 \\ 0 & \epsilon_I \end{pmatrix} \begin{pmatrix} \mathbf{V} & \mathbf{V} \\ \mathbf{A}^\mathsf{T} & \mathbf{A}^{-1} \end{pmatrix} \tag{68}$$

$$= \begin{pmatrix} (\mathbb{1} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} & 0 \\ 0 & (\mathbb{1} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} \end{pmatrix} \begin{pmatrix} \epsilon_E - \epsilon_I\mathbf{A}\mathbf{A}^\mathsf{T} & \epsilon_E - \epsilon_I \\ (-\epsilon_E + \epsilon_I)\mathbf{A}\mathbf{A}^\mathsf{T} & -\epsilon_E\mathbf{A}\mathbf{A}^\mathsf{T} + \epsilon_I \end{pmatrix}. \tag{69}$$

As one would expect, for $\epsilon_E = \epsilon_I$, this is equal to the identity matrix. When we switch columns and rows such that pairs of regular and corresponding null-eigenvectors form blocks, this becomes a block diagonal matrix. Note that this does not change the determinant or the eigenvalues of the matrix as for each row switch, there is a corresponding column switch that maintains the characteristic polynomial. Alternatively, we can assume that the matrix of eigenvectors $\mathbf{V}$ and its inverse $\mathbf{V}^{-1}$ are appropriately sorted. Without loss of generality, we assume that the first two columns of $\mathbf{V}$ are the fixed point's eigenvector $\mathbf{v}^*$ and its corresponding null-eigenvector and write

$$\bar{\mathbf{V}}^{-1}\tau^{-1}\bar{\mathbf{V}} = \begin{pmatrix} (1-a^{*2})^{-1} & 0 & \mathbf{0} \\ 0 & (1-a^{*2})^{-1} & \\ \mathbf{0} & & \ddots \end{pmatrix} \begin{pmatrix} \epsilon_E - \epsilon_I a^{*2} & \epsilon_E - \epsilon_I & \mathbf{0} \\ (-\epsilon_E + \epsilon_I)a^{*2} & -\epsilon_E a^{*2} + \epsilon_I & \\ \mathbf{0} & & \ddots \end{pmatrix}, \tag{70}$$

where a^{*2} is the first element on the diagonal of $\mathbf{A}\mathbf{A}^\mathsf{T}$ which corresponds to the fixed point eigenvector and $\mathbf{0}$ are matrices of zeros and appropriate dimensionality. Similarly, we can write the second part of Eq. 67 as a block triangular matrix. Before sorting, we write

$$\bar{\mathbf{V}}^{-1}\left[\frac{\bar{\mathbf{v}}_E^{*\circ}\mathbf{c}_E^{\circ\mathsf{T}}}{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{v}}_E^{*\circ}} + \frac{\bar{\mathbf{v}}_I^{*\circ}\mathbf{c}_I^{\circ\mathsf{T}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}}\right]\bar{\mathbf{V}} \equiv \bar{\mathbf{V}}^{-1}\left[\bar{\mathbf{v}}_E^{*\circ}\mathbf{d}_E{}^\mathsf{T} + \bar{\mathbf{v}}_I^{*\circ}\mathbf{d}_I{}^\mathsf{T}\right] = \bar{\mathbf{V}}^{-1}\begin{pmatrix} \bar{\mathbf{v}}_E^*\mathbf{d}_E{}^\mathsf{T} \\ \bar{\mathbf{v}}_I^*\mathbf{d}_I{}^\mathsf{T} \end{pmatrix}, \tag{71}$$

$$\mathbf{d}_E{}^\mathsf{T} = \frac{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{V}}}{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{v}}_E^{*\circ}} = \mathbf{c}^\mathsf{T}, \quad \mathbf{d}_I{}^\mathsf{T} = \frac{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{V}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}}, \tag{72}$$

where $\mathbf{d}_I{}^\mathsf{T}$ holds the L1-norm of all eigenvectors' inhibitory part as a fraction of the L1-norm of the fixed point eigenvector's inhibitory part. For the excitatory part, this fraction is always one, and $\mathbf{d}_E{}^\mathsf{T}$ is equal to $\mathbf{c}^\mathsf{T}$, a column vector of ones.

$$\bar{\mathbf{V}}^{-1}\begin{pmatrix} \bar{\mathbf{v}}_E^*\mathbf{d}_E{}^\mathsf{T} \\ \bar{\mathbf{v}}_I^*\mathbf{d}_I{}^\mathsf{T} \end{pmatrix} = \mathbf{N}\begin{pmatrix} \mathbf{V}^\mathsf{T} & -\mathbf{A} \\ -\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T} & \mathbf{A} \end{pmatrix}\begin{pmatrix} \bar{\mathbf{v}}_E^*\mathbf{c}^\mathsf{T} \\ \bar{\mathbf{v}}_I^*\mathbf{d}_I{}^\mathsf{T} \end{pmatrix}, \quad \bar{\mathbf{v}}_E^* = \mathbf{V}\mathbf{e}^*, \quad \bar{\mathbf{v}}_I^* = \mathbf{A}^\mathsf{T}\mathbf{e}^*, \tag{73}$$

$$= \mathbf{N}\begin{pmatrix} \mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{e}^*\mathbf{c}^\mathsf{T} - \mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{e}^*\mathbf{d}_I{}^\mathsf{T} \\ -\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{e}^*\mathbf{c}^\mathsf{T} + \mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{e}^*\mathbf{d}_I{}^\mathsf{T} \end{pmatrix}, \tag{74}$$

$$= \mathbf{N}\begin{pmatrix} \mathbf{e}^*\mathbf{c}^\mathsf{T} - a^{*2}\mathbf{e}^*\mathbf{d}_I{}^\mathsf{T} \\ -a^{*2}\mathbf{e}^*\mathbf{c}^\mathsf{T} + a^{*2}\mathbf{e}^*\mathbf{d}_I{}^\mathsf{T} \end{pmatrix}, \tag{75}$$

where we defined the normalization matrix $\mathbf{N}$ of the inverse eigenvector matrix $\bar{\mathbf{V}}^{-1}$ (compare Eq. 54). The vector $\mathbf{e}^*$ is zero except for one entry, equal to one, which corresponds to the fixed point eigenvector, such that the equations above hold. Note that the matrix above holds non-zero values in only two columns corresponding to the fixed point eigenvector and its null-eigenvector. After rearranging, we get

$$\bar{\mathbf{V}}^{-1}\left[\frac{\bar{\mathbf{v}}_E^{*\circ}\mathbf{c}_E^{\circ\mathsf{T}}}{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{v}}_E^{*\circ}} + \frac{\bar{\mathbf{v}}_I^{*\circ}\mathbf{c}_I^{\circ\mathsf{T}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}}\right]\bar{\mathbf{V}} = \mathbf{N}\begin{pmatrix} 1 - a^{*2}d_I^* & 1 - a^{*2}d_I^{\circledast} & \cdots \\ -a^{*2} + a^{*2}d_I^* & -a^{*2} + a^{*2}d_I^{\circledast} & \\ \mathbf{0} & & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbb{1} & \cdots \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{76}$$

where d_I^* and $d_I^{\circledast}$ are the first and second entries of $\mathbf{d}_I{}^\mathsf{T}$ and correspond to the fixed point eigenvector and its null-eigenvector. The last equality holds because $d_I^* = 1$ and $d_I^{\circledast} = 1/a^{*2}$: Before rearranging rows and columns, we can write

$$\mathbf{d}_I{}^\mathsf{T} = \frac{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{V}}}{\mathbf{c}_I^{\circ\mathsf{T}}\bar{\mathbf{v}}_I^{*\circ}} = \frac{1}{a^*}\left(\mathbf{c}_I{}^\mathsf{T}\mathbf{A}^\mathsf{T}, \mathbf{c}_I{}^\mathsf{T}\mathbf{A}^{-1}\right), \tag{77}$$

Remembering that $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}$ we rearrange and get

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

$$\mathbf{d}_I^\mathsf{T} = \left(d_I, d_I, \dots\right) = \frac{1}{a^*}\left(a^*, \frac{a^*}{a^{*2}}, \dots\right) = \left(1, \frac{1}{a^{*2}}, \dots\right). \tag{78}$$

from which the equality follows.

In summary, we find that after rearrangement Eq. 67 is a block triangular matrix.

$$\Rightarrow \quad \bar{\mathbf{V}}^{-1}\,\mathbf{J}\Big|_*\,\bar{\mathbf{V}} = \bar{\mathbf{V}}^{-1}\tau^{-1}\bar{\mathbf{V}}\begin{pmatrix}\mathbf{0} & \cdots\\ \mathbf{0} & \mathbb{1}\end{pmatrix}\left[\bar{\Lambda} - \bar{\lambda}^*\mathbb{1}\right], \tag{79}$$

Therefore, to find the eigenvalues, we consider each diagonal block separately. The first block corresponds to disturbances in the direction of the fixed point eigenvector or its null-eigenvector. From the matrix product above, we see immediately that their corresponding eigenvalues must be zero. For disturbances in the direction of a non-fixed point eigenvector $\bar{\mathbf{v}}^\dagger$ or its null-eigenvector we consider an exemplary block matrix:

$$\mathbf{J}_*^\dagger \equiv \bar{\mathbf{V}}^{\dagger-1}\,\mathbf{J}\Big|_*\,\bar{\mathbf{V}}^\dagger = \frac{1}{1-a^{\dagger 2}}\begin{pmatrix}\epsilon_E - \epsilon_I a^{\dagger 2} & \epsilon_E - \epsilon_I\\ (-\epsilon_E + \epsilon_I)a^{\dagger 2} & -\epsilon_E a^{\dagger 2} + \epsilon_I\end{pmatrix}\begin{pmatrix}\bar{\lambda}^\dagger - \bar{\lambda}^* & 0\\ 0 & -\bar{\lambda}^*\end{pmatrix}, \tag{80}$$

where $\bar{\mathbf{V}}^\dagger$ is a two-column matrix that holds $\bar{\mathbf{v}}^\dagger$ and its null-eigenvector. The eigenvalues of this matrix are negative under two conditions. First, its determinant must be positive, and second, its trace must be negative. After some algebra, these two conditions read

$$\det(\mathbf{J}_*^\dagger) \overset{!}{>} 0 \quad \Rightarrow \quad -\left(\bar{\lambda}^\dagger - \bar{\lambda}^*\right)\bar{\lambda}^*\epsilon_E\epsilon_I \overset{!}{>} 0, \tag{81}$$

$$\mathrm{tr}(\mathbf{J}_*^\dagger) \overset{!}{<} 0 \quad \Rightarrow \quad (\lambda^\dagger - \bar{\lambda}^*) - \frac{\epsilon_I}{\epsilon_E}\left(a^{\dagger 2}\lambda^\dagger + \bar{\lambda}^*\right) \overset{!}{<} 0. \tag{82}$$

Principal component analysis in inhibitory modified input spaces

First, we assume that excitatory and inhibitory plasticity are equally fast, i.e., $\epsilon_I = \epsilon_E$. Then the first stability condition above states that the fixed point $\bar{\mathbf{v}}^*$ with the largest eigenvalue, $\bar{\lambda}^* > \bar{\lambda}^\dagger, \forall\bar{\lambda}^\dagger$, is stable, provided that it is not repulsive[1], i.e., provided that its eigenvalue is larger than zero: $\bar{\lambda}^* > 0$. For $\epsilon_I = \epsilon_E$, the second condition reduces to $\bar{\lambda}^\dagger - 2\bar{\lambda}^* < 0$ which holds if the first condition is met. Therefore, the neuron finds the principal component of the modified input space while it takes the tuning of the inhibitory input population into account.

Fast inhibition increases stability

Unlike many other inhibitory plasticity rules, we do not require that inhibitory plasticity is faster than excitatory plasticity. In the extreme case of static inhibition, $\epsilon_E = 0$, the second condition is still satisfied if the fixed point attraction $\bar{\lambda}^*$ is larger than the excitatory attraction $\lambda^\dagger$ of any other eigenvector alone[2]. For growing $\epsilon_I > 0$, the repulsive influence of the inhibitory part of competing eigenvectors increases until they become overamplified, and the second condition is always fulfilled. In practice, fast inhibition helps to stabilize the system and otherwise does not crucially affect the dynamics. Therefore, we consider slightly faster inhibitory than excitatory plasticity for numerical simulations.

Stability of non-eigenvector fixed points

In principle, we can choose the total synaptic excitatory and inhibitory weights maintained during plasticity. Until now, we assumed that fixed points $\mathbf{w}^*$ are multiples of eigenvectors $\bar{\mathbf{v}}$ of the modified covariance matrix $\bar{\mathbf{C}}$, which puts a strong constraint on our choice for the weight norms[3]. As shown before, the general shape of a fixed point $\mathbf{w}^*$ is

$$\mathbf{w}^* = \begin{pmatrix}k_E\bar{\mathbf{v}}_E\\ k_I\bar{\mathbf{v}}_I\end{pmatrix} = \begin{pmatrix}\mathbb{1}k_E & \mathbf{0}\\ \mathbf{0} & \mathbb{1}k_I\end{pmatrix}\bar{\mathbf{v}} \equiv \begin{pmatrix}\mathbf{k}_E & \mathbf{0}\\ \mathbf{0} & \mathbf{k}_I\end{pmatrix}\bar{\mathbf{v}} \equiv \mathbf{K}\bar{\mathbf{v}}, \tag{83}$$

where k_E and k_I are scalar constants. We recapitulate the general weight dynamics as given in Eq. 42:

$$\tau\dot{\mathbf{w}} = \bar{\mathbf{C}}\mathbf{w} - \gamma\mathbf{w}_E^\circ - \rho\mathbf{w}_I^\circ. \tag{84}$$

[1] An eigenvector can become repulsive if inhibition is sufficiently strong, i.e., if $\bar{\lambda}^* = \lambda^*(1-a^{*2}) < 0 \;\Rightarrow\; a^{*2} > 1$.

[2] Remember that the total attraction is a combination of the excitatory attraction λ^* minus the inhibitory repulsion $a^{*2}\lambda^*$. When inhibitory weights are static, they remain tuned to the fixed point. Then, only the excitatory attractions of competing eigenvectors $\lambda^\dagger$ are relevant for stability.

[3] More precisely this constraints the ratio between the weight norms, as an additional scalar factor does not change the dynamics.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Instead of evaluating the Jacobian in a general fixed point, we switch to a new coordinate system in which the fixed point is an eigenvector of the modified covariance matrix. We define

$$\hat{\mathbf{w}} \equiv \mathbf{K}^{1/2}\mathbf{w}, \tag{85}$$

from which the weight dynamics becomes

$$\hat{\tau}\dot{\hat{\mathbf{w}}} = \hat{\mathbf{C}}\hat{\mathbf{w}} - \hat{\gamma}\hat{\mathbf{w}}_E^\circ - \hat{\rho}\hat{\mathbf{w}}_I^\circ, \tag{86}$$

$$\hat{\tau} = \tau\mathbf{K}, \quad \hat{\mathbf{C}} = \mathbf{K}^{1/2}\bar{\mathbf{C}}\mathbf{K}^{1/2} = \left\langle \begin{pmatrix} k_E^{1/2}\mathbf{y}_E \\ k_I^{1/2}\mathbf{y}_I \end{pmatrix} \left(k_E^{1/2}\mathbf{y}_E^\mathsf{T}, k_I^{1/2}\mathbf{y}_I^\mathsf{T} \right)^\mathsf{T} \right\rangle, \tag{87}$$

$$\hat{\gamma} = \mathbf{K}^{1/2}\frac{\mathbf{c}_E^{\circ\mathsf{T}}\bar{\mathbf{C}}\mathbf{K}^{1/2}\hat{\mathbf{w}}}{\mathbf{c}_E^{\circ\mathsf{T}}\mathbf{K}^{1/2}\hat{\mathbf{w}}_E^\circ}\mathbf{K}^{1/2} = \frac{\mathbf{c}_E^{\circ\mathsf{T}}\hat{\mathbf{C}}\hat{\mathbf{w}}}{\mathbf{c}_E^{\circ\mathsf{T}}\hat{\mathbf{w}}_E^\circ}, \quad \hat{\rho} = \frac{\mathbf{c}_I^{\circ\mathsf{T}}\hat{\mathbf{C}}\hat{\mathbf{w}}}{\mathbf{c}_I^{\circ\mathsf{T}}\hat{\mathbf{w}}_I^\circ}. \tag{88}$$

In this coordinate system, we are interested in the general fixed point, which becomes

$$\hat{\mathbf{w}}^* = \mathbf{K}^{-1/2}\mathbf{w}^* = \mathbf{K}^{-1/2}\mathbf{K}\bar{\mathbf{v}} = \mathbf{K}^{1/2}\bar{\mathbf{v}}. \tag{89}$$

It is straight forward to proof that $\hat{\mathbf{w}}^*$ is an eigenvector of $\hat{\mathbf{C}}$ with eigenvalue $\hat{\lambda}^* = (k_E - k_I a^{*2})\lambda^*$. In principle, we can now proceed in finding the eigenvalues of the Jacobian, as explained before. However, we would have to employ the eigenvector basis in the new coordinates $\hat{\mathbf{V}} = \mathbf{K}^{-1/2}\bar{\mathbf{V}}$ for triangularization. As before, one finds that stability is largely determined by the eigenvalues $\hat{\lambda}$.

Effective timescales and attraction landscapes

Apart from providing a way to determine if a general fixed point is stable, the change of variables approach provides additional insight: Let's assume the total synaptic inhibitory weight of a neuron is very small, much smaller than any eigenvector of $\bar{\mathbf{C}}$ would suggest, i.e., $k_I \ll 1$, while the excitatory weight norm is equal to one, which implies $k_E = 1$. As one would expect intuitively, the neuron does not exhibit much of the inhibitory attraction landscape (compare $\hat{\mathbf{C}}$ in Eq. 87), and its stability would be primarily determined by the excitatory attraction of the different eigenvector modes, i.e., $\hat{\lambda} \approx \lambda$. In the extreme case, when the inhibitory weight norm is zero, i.e., $k_I = 0$, only the activity of the excitatory population is relevant. Another important aspect is that the effective timescale $\hat{\tau} = \tau\mathbf{K}$ depends on the magnitude of the weight norms. For example, when the inhibitory weight norm k_I decreases while the excitatory weight norm k_E remains fixed, the effective inhibitory plasticity becomes faster, since $\hat{\tau}_I = \tau_I k_I$. This can be interpreted as a fiercer competition of the same number of presynaptic neurons for fewer synaptic resources, which leads to faster dynamics. Note that one can achieve the same effect of faster effective plasticity by increasing the number of competitors while maintaining the number of available resources.

4 Lateral input warps attraction landscape

As a first step towards fully recurrent plastic networks, we consider a model of two neurons that receive feedforward input from a population of input neurons. Additionally, the first neuron projects laterally onto the second neuron without receiving any lateral input itself (Fig. S2A & B). Since there is no recurrence, this is effectively still a feedforward circuit, similar to the case considered in Section 3, where instead of an additional excitatory neuron, we considered additional inhibitory neurons that are tuned to eigenvectors of the excitatory population.

Let the first neuron have a fixed set of feedforward weights $\mathbf{q}$. We are interested in the system's fixed points, i.e., how the second neuron adapts its feedforward weights and its lateral weight under a linear Hebbian plasticity rule. From the perspective of the second neuron, the input space is increased by one dimension due to the additional lateral input. We denote the new input vector as

$$\bar{\mathbf{y}} = \left(\mathbf{y}^\mathsf{T}, \mathbf{q}^\mathsf{T}\mathbf{y}\right)^\mathsf{T}, \tag{90}$$

and the input weights onto the second neuron as

$$\bar{\mathbf{w}} = \left(\mathbf{w}^\mathsf{T}, w_q\right)^\mathsf{T}. \tag{91}$$

Effectively this is still a feedforward network without feedback, and the covariance matrix $\bar{\mathbf{C}}$ of the new inputs $\bar{\mathbf{y}}$ fully determines the synaptic weight dynamics:

$$\bar{\mathbf{C}} = \left\langle \bar{\mathbf{y}}\bar{\mathbf{y}}^\mathsf{T} \right\rangle = \left\langle \begin{pmatrix} \mathbf{y}\mathbf{y}^\mathsf{T} & \mathbf{y}\mathbf{y}^\mathsf{T}\mathbf{q} \\ \mathbf{q}^\mathsf{T}\mathbf{y}\mathbf{y}^\mathsf{T} & \mathbf{q}^\mathsf{T}\mathbf{y}\mathbf{y}^\mathsf{T}\mathbf{q} \end{pmatrix} \right\rangle = \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{q} \\ \mathbf{q}^\mathsf{T}\mathbf{C} & \mathbf{q}^\mathsf{T}\mathbf{C}\mathbf{q} \end{pmatrix}, \tag{92}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

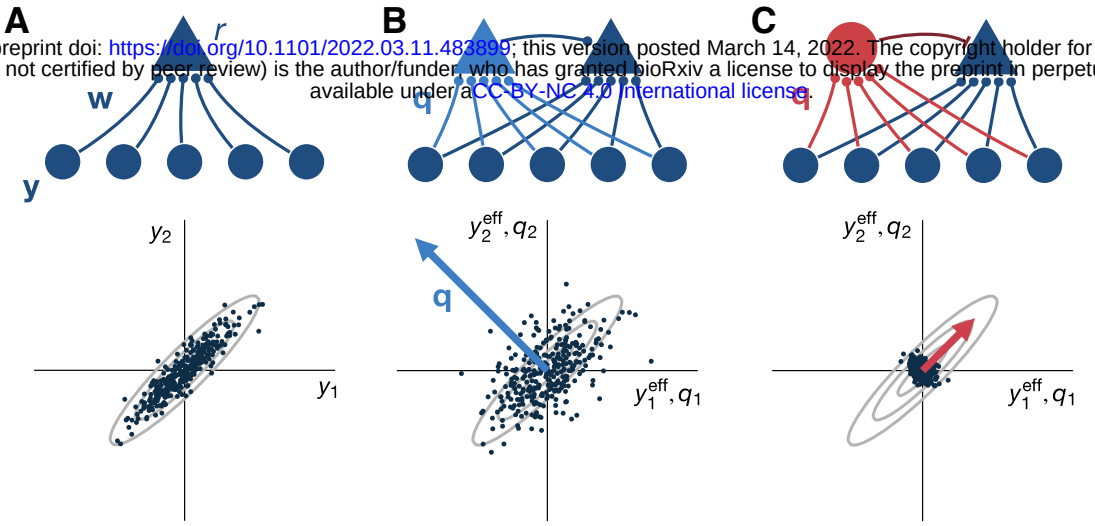

Figure S2: Input space warping due to lateral connectivity. (**A**) Top: a single neuron with firing rate r receives synaptic inputs $\mathbf{w}$ from a population of excitatory neurons $\mathbf{y}$. Bottom: input distribution projected onto the first two input dimensions. Each dot represents the firing rates of the first two neurons during one input pattern. (Contour lines in light gray). Under a linear Hebbian learning rule, the neuron becomes selective for the direction of maximum variance, the first principal component (see Section 2). (**B**) Top: Same as (A) for a neuron that receives input from a laterally projecting excitatory neuron which is tuned to an eigenvector $\mathbf{q}$ of the original input covariance matrix. Bottom: the effective input space $\mathbf{y}^{\text{eff}}$ of the target neuron (dark blue triangle) is warped such that the variance along the eigenvector $\mathbf{q}$ (blue arrow) is stretched in proportion to the absolute value of the weight vector $\mathbf{q}$. The contour lines of the original input distribution from (A) are shown in light gray for reference. (**C**) Top: Same as (B) for a laterally projecting inhibitory neuron. Bottom: Now, the effective input space is compressed. See text for details.

where $\mathbf{C}$ is the covariance matrix of the original input $\mathbf{y}$. As before, assuming multiplicative normalization, the eigenvectors $\bar{\mathbf{v}}$ of the modified covariance matrix $\bar{\mathbf{C}}$ are fixed points $\bar{\mathbf{w}}^*$ (compare Section 3.1):

$$\bar{\mathbf{w}}^* = \bar{\mathbf{v}} \equiv \left(\mathbf{w}^{*\mathsf{T}}, w_q^*\right)^{\mathsf{T}}, \quad \bar{\mathbf{C}}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}, \tag{93}$$

where w_q^* and $\mathbf{w}^*$ are the lateral input weight and the feedforward input weight in the fixed point, respectively. We solve for eigenvalues $\bar{\lambda}$ and eigenvectors $\bar{\mathbf{v}}$:

$$\bar{\mathbf{C}}\bar{\mathbf{v}} = \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{q} \\ \mathbf{q}^{\mathsf{T}}\mathbf{C} & \mathbf{q}^{\mathsf{T}}\mathbf{C}\mathbf{q} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ w_q \end{pmatrix} = \bar{\lambda} \begin{pmatrix} \mathbf{w} \\ w_q \end{pmatrix}. \tag{94}$$

$$\mathbf{C}\mathbf{w} + \mathbf{C}\mathbf{q}w_q = \bar{\lambda}\mathbf{w}, \tag{95}$$
$$\mathbf{q}^{\mathsf{T}}\mathbf{C}\mathbf{w} + \mathbf{q}^{\mathsf{T}}\mathbf{C}\mathbf{q}w_q = \bar{\lambda}w_q. \tag{96}$$

$$\mathbf{C}\left(\mathbf{w} + \mathbf{q}w_q\right) = \bar{\lambda}\mathbf{w}, \tag{97}$$
$$\mathbf{q}^{\mathsf{T}}\mathbf{C}\left(\mathbf{w} + \mathbf{q}w_q\right) = \bar{\lambda}w_q. \tag{98}$$

Inserting the first into the second expression gives $w_q = \mathbf{q}^{\mathsf{T}}\mathbf{w}$ which, when inserted into the first expression, results in:

$$\boxed{\mathbf{C}\left(\mathbb{1} + \mathbf{q}\mathbf{q}^{\mathsf{T}}\right)\mathbf{w} = \bar{\lambda}\mathbf{w}.} \tag{99}$$

The solution to this equation gives the feedforward weight vector. For general $\mathbf{q}$, the solution is not straightforward: When we consider the equation in the input eigenspace, where Eq. 99 becomes

$$\mathbf{\Lambda}\left(\mathbb{1} + \mathbf{q}_v\mathbf{q}_v^{\mathsf{T}}\right)\mathbf{w}_v = \bar{\lambda}\mathbf{w}_v, \tag{100}$$

with $\mathbf{\Lambda}$ being the diagonal matrix of eigenvalues and the subscript $(\cdot)_v$ indicates a vector in the eigenbasis of $\mathbf{C}$. In this basis, eigenvectors of $\mathbf{C}$ are unit vectors, i.e., $\mathbf{v}_v = \mathbf{e}_v$, where $\mathbf{e}_v$ is a vector of zeros with one entry equal to one that corresponds to the respective eigenvector. When $\mathbf{q}$ contains components of more than one eigenvector, the matrix $\mathbf{q}_v\mathbf{q}_v^{\mathsf{T}}$ is no longer diagonal and eigenvectors of $\mathbf{C}$, $\mathbf{w}_v = \mathbf{e}_v$, do not solve the equation. However, when we assume that the first neuron has plastic feedforward input, we know that it will converge to a multiple of an eigenvector of the feedforward input $\mathbf{q} \propto \mathbf{v}_i$, where $\mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Then multiples of eigenvectors of $\mathbf{C}$ solve Eq. 99. To find the eigenvalues

$\bar{\lambda}$, which determine fixed point stability (compare Section 3.2), we distinguish two cases: First, $\mathbf{w}$ may be proportional to a different eigenvector than $\mathbf{q}$:

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\mathbf{w} \propto \mathbf{v}_j \neq \mathbf{v}_i \propto \mathbf{q} \quad \Rightarrow \quad \mathbf{q}^\mathsf{T}\mathbf{w} \propto \mathbf{v}_i^\mathsf{T}\mathbf{v}_j = 0, \quad \Rightarrow \quad \bar{\lambda}_j = \lambda_j = \sigma_j^2. \tag{101}$$

Therefore, input modes that are orthogonal to the tuning of a laterally projecting neuron maintain their attraction. Second, $\mathbf{w}$ may be proportional to the same eigenvector as $\mathbf{q}$:

$$\mathbf{w} \propto \mathbf{v} \propto \mathbf{q} \quad \Rightarrow \quad \mathbf{q}^\mathsf{T}\mathbf{w} \propto \mathbf{v}^\mathsf{T}\mathbf{v} = 1, \quad \Rightarrow \quad \bar{\lambda} = \lambda + \lambda a_q^2 = \sigma^2 + \sigma_q^2, \tag{102}$$

where a_q is equal to $\|\mathbf{q}\|$, the L2-norm of $\mathbf{q}$. Input modes aligned with the tuning of the laterally projecting neuron increase their attraction by the variance of that neuron. Together this can be interpreted as a stretching of the original input space in the direction of the laterally projecting neuron's feedforward weight vector $\mathbf{q}$ (Fig. S2B). Then fixed points are of the following general shape:

$$\mathbf{w}^* = \mathbf{v}, \quad \mathbf{q} = a_q\mathbf{v}, \tag{103}$$

$$\Rightarrow \quad \bar{\mathbf{w}}^* = \begin{pmatrix} \mathbf{w}^* \\ w_q^* \end{pmatrix} = k \begin{pmatrix} \mathbf{w}^* \\ \mathbf{q}^\mathsf{T}\mathbf{w}^* \end{pmatrix} = k \begin{pmatrix} \mathbf{v} \\ a_q \end{pmatrix}, \tag{104}$$

where k is a scalar constant. Similarly, when the laterally projecting neuron is inhibitory, the modified covariance matrix becomes $\bar{\mathbf{C}}' = \begin{pmatrix} \mathbf{C} & -\mathbf{Cq} \\ \mathbf{q}^\mathsf{T}\mathbf{C} & -\mathbf{q}^\mathsf{T}\mathbf{Cq} \end{pmatrix}$ (compare Eq. 53) and it follows that the effective input space is compressed[1] (Fig. S2C):

$$\bar{\lambda} = \lambda - \lambda a_q^2 = \sigma^2 - \sigma_q^2, \tag{105}$$

This can be generalized to multiple excitatory and inhibitory neurons such that the total attraction towards a feedforward eigenvector becomes

$$\bar{\lambda} = \lambda \left(1 + \|\mathbf{a}_E\|^2 - \|\mathbf{a}_I\|^2\right), \tag{106}$$

$$\Rightarrow \quad \boxed{\bar{\lambda} = \sigma^2 + \|\boldsymbol{\sigma}_E\|^2 - \|\boldsymbol{\sigma}_I\|^2}, \tag{107}$$

where $\mathbf{a}_E$, $\mathbf{a}_I$ hold the vector norms of the laterally projecting neurons and $\boldsymbol{\sigma}_E$, $\boldsymbol{\sigma}_I$ hold their standard deviations. In this general case, one can write the fixed points as

$$\boxed{\bar{\mathbf{w}}^* = \mathbf{K} \begin{pmatrix} \mathbf{v} \\ \mathbf{a}_E \\ \mathbf{a}_I \end{pmatrix} = \mathbf{K}' \begin{pmatrix} \sigma\mathbf{v} \\ \boldsymbol{\sigma}_E \\ \boldsymbol{\sigma}_I \end{pmatrix}}, \tag{108}$$

where $\mathbf{K}$ and $\mathbf{K}'$ are diagonal matrices that scale the inhibitory and excitatory part of the vector (compare Eq. 83) and the second equality holds for $\mathbf{K} = \sigma\mathbf{K}'$. This means that the total synaptic weight distributes among synapses in proportion to the standard deviation of their postsynaptic activities.

Note that these results are independent of what causes the laterally projecting neurons' tuning. For example, in addition to feedforward input, a neuron can be integrated into a recurrent circuit of neurons that are all tuned to the same eigenvector. Then σ_E^2 results from recurrent interaction in addition to the norm of the feedforward weight vector $\|\mathbf{q}\|$. We will consider such circuits in the following sections.[2]

5 Eigencircuits

We assume that the activity in a recurrent network with linear activation functions is dominated by feedforward activity such that neurons become selective for different eigenvectors of the feedforward input covariance matrix $\mathbf{C} = \langle yy^T \rangle$. Then the average Hebbian growth of a synapse that connects two neurons that are tuned to different

[1] For sufficiently large vector norms $\|\mathbf{q}\|$ the eigenvector mode becomes repulsive. Then the transformation of the input space can no longer be visualized as intuitively as in Fig. S2.

[2] Another example is neurons that project from outside the local circuit, e.g., from another brain area that is higher up in the processing hierarchy.

eigenvectors is:

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\langle W_{ij} \rangle \propto \langle r_i r_j \rangle \tag{109}$$
$$= \langle \mathbf{v}_i^\mathsf{T} \mathbf{y}\mathbf{y}^\mathsf{T} \mathbf{v}_j \rangle \tag{110}$$
$$= \mathbf{v}_i^\mathsf{T} \mathbf{C}\mathbf{v}_j \tag{111}$$
$$= \mathbf{v}_i^\mathsf{T} \lambda_j \mathbf{v}_j = \lambda_j v_i^\mathsf{T} v_j \tag{112}$$
$$= 0. \tag{113}$$

Due to the competition for synaptic resources, the synapse loses out to the non-zero growth of other synapses and decays to zero over time. In its steady state, the circuit is separated into sub-circuits with recurrent connections within but not between them. Since there is one sub-circuit per eigenvector of the covariance matrix, we call these decoupled circuits 'eigencircuits'.

5.1 Variance propagation

To compute the effective attraction of an input mode when synaptic weights have converged, it is sufficient to know the variances of all neurons that are tuned to that mode (compare section 4). As a first step, we investigate how variances propagate through the network, i.e., our goal is to express the standard deviation σ_r of a neuron as a function of the standard deviations of its presynaptic inputs. In general, the postsynaptic firing rate r is given as

$$r = \mathbf{w}^\mathsf{T}\mathbf{y} + \mathbf{w}_E{}^\mathsf{T}\mathbf{y}_E - \mathbf{w}_I{}^\mathsf{T}\mathbf{y}_I. \tag{114}$$

In an eigencircuit, all presynaptic inputs with non-zero synaptic weight are tuned to the same eigenvector $\mathbf{v}$. We only consider these non-zero entries and write

$$\mathbf{y}_E = \mathbf{a}_E(\mathbf{v}^\mathsf{T}\mathbf{y}), \quad \mathbf{y}_I = \mathbf{a}_I(\mathbf{v}^\mathsf{T}\mathbf{y}), \tag{115}$$

where the vectors $\mathbf{a}_E$ and $\mathbf{a}_I$ set the vectors' response magnitudes which are proportional to their standard deviations. For the weight vector, we require that the excitatory and inhibitory parts are normalized to maintain the total amount of inhibitory and excitatory synaptic resources.

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{w}_E \end{pmatrix} = W_E \frac{\bar{\mathbf{v}}_E}{\|\bar{\mathbf{v}}_E\|_p}, \quad \mathbf{w}_I = W_I \frac{\bar{\mathbf{v}}_I}{\|\bar{\mathbf{v}}_I\|_p}, \tag{116}$$

$$\bar{\mathbf{v}}_E = \begin{pmatrix} \mathbf{v} \\ \mathbf{a}_E \end{pmatrix}, \quad \bar{\mathbf{v}}_I = \mathbf{a}_E, \tag{117}$$

where W_E, W_I are scalar weight norms and $\bar{\mathbf{v}}_E$, $\bar{\mathbf{v}}_E$ are the excitatory and inhibtory part of the modified covariance matrix that we obtain from Eq. 108. The p-norm, $\|\cdot\|_p$, is maintained due to competition for synaptic resources. For the postsynaptic firing rate, it follows

$$r = \left(\frac{1 + \|\mathbf{a}_E\|^2}{\|\bar{\mathbf{v}}_E\|_p} W_E - \frac{\|\mathbf{a}_I\|^2}{\|\bar{\mathbf{v}}_I\|_p} W_I \right) (\mathbf{v}^\mathsf{T}\mathbf{y}). \tag{118}$$

The first bracket is a scalar pre-factor which makes it straightforward to compute the standard deviation:

$$\sigma_r = \left(\frac{1 + \|\mathbf{a}_E\|^2}{\|\bar{\mathbf{v}}_E\|_p} W_E - \frac{\|\mathbf{a}_I\|^2}{\|\bar{\mathbf{v}}_I\|_p} W_I \right) \sigma = \frac{1 + \|\mathbf{a}_E\|^2 \sigma^2}{\|\bar{\mathbf{v}}_E\|_p \sigma} W_E - \frac{\|\mathbf{a}_I\|^2 \sigma^2}{\|\bar{\mathbf{v}}_I\|_p \sigma} W_I, \tag{119}$$

$$\Rightarrow \quad \boxed{\sigma_r = \frac{\left\|\boldsymbol{\sigma}^E\right\|^2}{\left\|\boldsymbol{\sigma}^E\right\|_p} W_E - \frac{\left\|\boldsymbol{\sigma}^I\right\|^2}{\left\|\boldsymbol{\sigma}^I\right\|_p} W_I}, \tag{120}$$

$$\boldsymbol{\sigma}^E = \left(\sigma, \boldsymbol{\sigma}_E{}^\mathsf{T}\right)^\mathsf{T}, \quad \boldsymbol{\sigma}^I = \boldsymbol{\sigma}_I, \tag{121}$$

where we no longer distinguish between feedforward and recurrent input. For a network in the steady state, i.e., when synaptic weights converged, this provides the standard deviation of the postsynaptic neuron as a function of the standard deviations of its inputs. In the next section, we will use this relation to find the total attraction of an eigencircuit.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

5.2 Consitency conditions provide eigencircuit firing rate variances

In its steady state, an eigencircuit with n_E excitatory and n_I inhibitory neurons has to fulfil the variance propagation equation (Eq. 120). In the fully connected eigencircuit, each condition depends on the variances of all neurons, and all neurons have the same presynaptic inputs. This provides $N = n_E + n_I$ consistency conditions for the N unknown standard deviations. The condition for a single excitatory neuron i reads

$$\sigma_E^i = W_{EE}^i \left(\frac{\sigma^2 + \|\boldsymbol{\sigma}_E\|^2}{\sigma + \|\boldsymbol{\sigma}_E\|_1} \right) - W_{EI}^i \left(\frac{\|\boldsymbol{\sigma}_I\|^2}{\|\boldsymbol{\sigma}_I\|_1} \right), \tag{122}$$

where we chose the L1-norm, $p = 1$, for normalization (but see section 5.3). We make the simplifying assumption that all neurons have similar weight norms, i.e., $W_{AB}^i \approx W_{AB},\ \forall i$ and $A, B \in \{E, I\}$. Then, also their standard deviations are similar, and we can approximate:

$$\|\boldsymbol{\sigma}\|^2 = \sum_i (\sigma^i)^2 \approx n\sigma^2, \quad \text{if} \quad \sigma^i \approx \sigma\ \forall i. \tag{123}$$

$$\Rightarrow \quad \frac{\sigma^2 + \|\boldsymbol{\sigma}_E\|^2}{\sigma + \|\boldsymbol{\sigma}_E\|_1} \approx \frac{\sigma^2 + n_E\sigma_E^2}{\sigma + n_E\sigma_E}, \quad \text{if} \quad \sigma_E^i \approx \sigma_E\ \forall i. \tag{124}$$

The standard deviations of excitatory and inhibitory neurons become

$$\sigma_E = W_{EE} \left(\frac{\sigma^2 + n_E\sigma_E^2}{\sigma + n_E\sigma_E} \right) - W_{EI} \left(\frac{n_I\sigma_I^2}{n_I\sigma_I} \right), \tag{125}$$

$$\sigma_I = W_{IE} \left(\frac{\sigma^2 + n_E\sigma_E^2}{\sigma + n_E\sigma_E} \right) - W_{II} \left(\frac{n_I\sigma_I^2}{n_I\sigma_I} \right). \tag{126}$$

After some algebra, this yields the standard deviations of single excitatory and inhibitory neurons as a function of the number of neurons in the eigencircuit, n_E, n_I, their weight norms W_{AB}, and the standard deviation, σ, of the feedforward input along the corresponding eigenvector:

$$\boxed{\sigma_I = \frac{W_{IE}}{1 + W_{II}} \frac{1}{\Phi} \sigma_E}, \quad \Phi \equiv \left[W_{EE} - \frac{W_{EI}W_{IE}}{1 + W_{II}} \right], \tag{127}$$

$$\Rightarrow \quad \boxed{\sigma_E = \frac{1}{2(1-\Phi)n_E} \left(-1 \pm \sqrt{1 + 4\Phi(1-\Phi)n_E} \right) \sigma}. \tag{128}$$

Note that for $\Phi < 1$ there exists a real solution for σ_E, independent of n_E.

5.3 A note on the choice of weight norm

The choice of the weight norm that is maintained via multiplicative normalization is non-trivial. Biologically we motivated normalization by the competition for a limited amount of synaptic resources. We assumed the simplest case, where the L1-norm is maintained, and each resource unit translates to one unit of synaptic strength. An alternative choice would be to maintain the L2-norm. In the variance propagation equation (Eq. 120) this corresponds to $p = 2$ which becomes

$$\sigma_r = \left\| \boldsymbol{\sigma}^E \right\| W_E - \left\| \boldsymbol{\sigma}^I \right\| W_I. \tag{129}$$

For a single inhibitory neuron the eigencircuit consistency condition becomes (compare Eq. 122):

$$\sigma_I = \frac{W_{IE}}{1 + W_{II}} \left(\sigma^2 + \|\boldsymbol{\sigma}_E\|^2 \right)^{\frac{1}{2}}, \tag{130}$$

where we once more assumed that all neurons have similar weight norms, $W_{AB}^i \approx W_{AB},\ \forall i$. For the variance of an excitatory neuron, it follows

$$\sigma_E^2 = \Phi^2 \left(\sigma^2 + \|\boldsymbol{\sigma}_E\|^2 \right) = \Phi^2 \left(\sigma^2 + n_E\sigma_E^2 \right), \tag{131}$$

$$\Rightarrow \quad \boxed{\sigma_E^2 = \frac{\Phi^2}{1 - \Phi^2 n_E} \sigma^2}. \tag{132}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

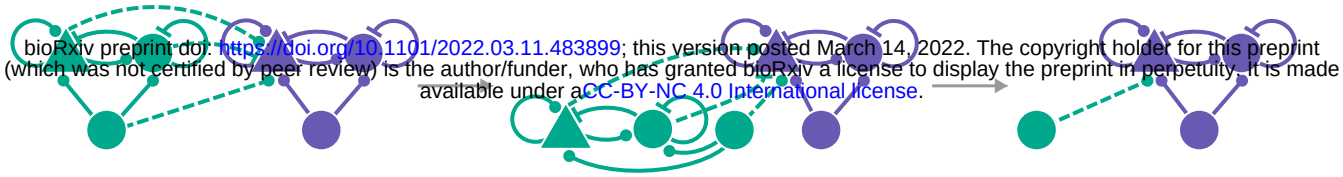

Figure S3: When considering the stability of a neuron in an eigencircuit with respect to a perturbation towards another eigencircuit (left, dashed lines), the other eigencircuit contributes as an effectively feedforward input (center). Then one can consider an equivalent circuit (right) where the attraction of the competing input mode (green) is increased by the attraction of its former eigencircuit (see text for details).

For an increasing number of excitatory neurons n_E, the variance of a single excitatory neuron grows and diverges for $\Phi^2 n_E = 1$. For even larger n_E, variances would have to be negative to fulfill the consistency condition, which is not possible. It follows that for sufficiently large n_E there exist no fixed points. This is not unique to the L2-norm but holds for any $p > 1$. Such norms allow for a larger total synaptic weight (in terms of its L1-norm) when distributed across multiple synapses. Additional neurons provide additional recurrent synapses, which leads to the growth of the effective recurrent excitation until activities can no longer be stabilized by recurrent inhibition. For a suitable choice of the weight norms, Φ can, in principle, become small enough to balance the number of excitatory neurons in any eigencircuit to maintain positive variances. However, this requires additional fine-tuning and fails when n_E becomes unexpectedly large.

6 Fully plastic recurrent E-I networks

In the following, we consider the stability of eigencircuits in a network of recurrently connected excitatory and inhibitory neurons. Specifically, we would like to know when a neuron from one eigencircuit becomes attracted to another eigencircuit. First, we make some simplifying assumptions. Since each neuron can potentially be bidirectionally connected to all other neurons, the system's dimensionality grows quadratically with the number of neurons. We are only interested in the general principles and consider two eigencircuits, each with one excitatory and one inhibitory neuron.

Similar to the feedforward case, we find the following general form for the Jacobian (compare Eq. 66)

$$\tau \left.\frac{\mathrm{d}\dot{\mathbf{w}}}{\mathrm{d}\mathbf{w}}\right|_* = [\mathbb{1} - ...] \left(\hat{\mathbf{C}}^* - \hat{\lambda}^* \mathbb{1} + \left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_* \mathbf{w}^* \right), \tag{133}$$

where the last term takes into account that the modified covariance matrix is no longer fixed but depends on the weights themself: In general, when a synaptic weight is perturbed, the postsynaptic neuron's variance changes, which affects the presynaptic neuron's variance via recurrent connections.

We consider infinitesimal perturbations of fixed points where the circuit is separated into unconnected eigencircuits (see Section 5). When a neuron in eigencircuit A is perturbed towards a different eigencircuit B, it changes its variance (Fig. 3, left). However, neurons in eigencircuit B are unaffected because there are no recurrent connections between eigencircuit A and B. When assessing stability, it is, therefore, sufficient to consider the recurrence within a single eigencircuit. Input from other eigencircuits can be treated as feedforward input to that circuit (Fig. 3, center). For that reason, we consider an equivalent circuit where the original attraction of the feedforward input mode λ is increased by the eigencircuit attraction, $\lambda \leftarrow \hat{\lambda} = \lambda + \lambda_{\text{eig}}$ (Fig. S3, right). In this circuit, firing rates are given by (compare section 3)

$$y_E = \mathbf{w}_{EF}{}^\mathsf{T} \mathbf{y} + w_{EE} y_E - w_{EI} y_I, \tag{134}$$

$$y_I = \mathbf{w}_{IF}{}^\mathsf{T} \mathbf{y} + w_{IE} y_E - w_{II} y_I, \tag{135}$$

$$y_E = \frac{1}{1 - w_{EE} + \dfrac{w_{EI} w_{IE}}{1 + w_{II}}} \left(\mathbf{w}_{EF}{}^\mathsf{T} - \frac{w_{EI} \mathbf{w}_{IF}{}^\mathsf{T}}{1 + w_{II}} \right) \mathbf{y} \equiv \mathbf{a}_E{}^\mathsf{T} \mathbf{y}, \tag{136}$$

$$y_I = \frac{1}{1 + w_{II} + \dfrac{w_{IE} w_{EI}}{1 - w_{EE}}} \left(\mathbf{w}_{IF}{}^\mathsf{T} + \frac{w_{IE} \mathbf{w}_{EF}{}^\mathsf{T}}{1 - w_{EE}} \right) \mathbf{y} \equiv \mathbf{a}_I{}^\mathsf{T} \mathbf{y}. \tag{137}$$

The weight dynamics is

$$\tau \dot{\mathbf{w}} = \begin{pmatrix} \dot{\mathbf{w}}_{EF} \\ \dot{w}_{EE} \\ \dot{w}_{EI} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{y}\mathbf{y}^\mathsf{T} & \mathbf{y} y_E & -\mathbf{y} y_I & \\ y_E \mathbf{y}^\mathsf{T} & y_E y_E & -y_E y_I & \mathbf{0} \\ y_I \mathbf{y}^\mathsf{T} & y_I y_E & -y_I y_I & \\ & \mathbf{0} & & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{w}_{EF} \\ w_{EE} \\ w_{EI} \\ \vdots \end{pmatrix} - \begin{pmatrix} \gamma_E & 0 & 0 & \\ 0 & \gamma_E & 0 & \mathbf{0} \\ 0 & 0 & \rho_E & \\ & \mathbf{0} & & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{w}_{EF} \\ w_{EE} \\ w_{EI} \\ \vdots \end{pmatrix}, \tag{138}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where ellipsis indicate similar terms for afferent weights of the inhibitory neuron. We define the modified covariance matrix

$$\hat{\mathbf{C}} = \begin{pmatrix} \langle \mathbf{y}\mathbf{y}^\mathsf{T} \rangle & \langle \mathbf{y} y_E \rangle & -\langle \mathbf{y} y_I \rangle & \\ \langle y_E \mathbf{y}^\mathsf{T} \rangle & \langle y_E y_E \rangle & -\langle y_E y_I \rangle & \mathbf{0} \\ \langle y_I \mathbf{y}^\mathsf{T} \rangle & \langle y_I y_E \rangle & -\langle y_I y_I \rangle & \\ & \mathbf{0} & & \ddots \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{a}_E & -\mathbf{C}\mathbf{a}_I & \\ \mathbf{a}_E^\mathsf{T}\mathbf{C} & \mathbf{a}_E^\mathsf{T}\mathbf{C}\mathbf{a}_E & -\mathbf{a}_E^\mathsf{T}\mathbf{C}\mathbf{a}_I & \mathbf{0} \\ \mathbf{a}_I^\mathsf{T}\mathbf{C} & \mathbf{a}_I^\mathsf{T}\mathbf{C}\mathbf{a}_E & -\mathbf{a}_I^\mathsf{T}\mathbf{C}\mathbf{a}_I & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \tag{139}$$

and write the average synaptic change as

$$\Rightarrow \quad \tau \langle \dot{\mathbf{w}} \rangle \equiv \hat{\mathbf{C}}\mathbf{w} - \mathbf{\Gamma}\mathbf{w}, \tag{140}$$

where $\mathbf{\Gamma}$ is a diagonal matrix that holds the scalar constraint factors (compare Eq. 42). Note that this is a non-linear dynamical system since the modified covariance matrix depends on plastic synaptic weights. We make the simplifying assumption that the plasticity of excitatory and inhibitory synapses is equally fast, $\tau_E = \tau_I$. Then $\boldsymbol{\tau} = \tau \mathbb{1}$, which does not affect the fixed points or the stability of the system[1]. Therefore, we set $\boldsymbol{\tau} = \mathbb{1}$.

6.1 Fixed points

Fixed points $\mathbf{w}^*$ must fulfill the following condition

$$\hat{\mathbf{C}}^* \mathbf{w}^* - \mathbf{\Gamma}^* \mathbf{w}^* \stackrel{!}{=} 0. \tag{141}$$

where $\hat{\mathbf{C}}^*$ is the modified covariance matrix evaluated in the fixed point. We consider a fixed point where all neurons form a single eigencircuit and are tuned to the same input mode $\mathbf{v}^*$. Then we can write the excitatory and inhibitory firing rates as

$$y_E^* = \mathbf{a}_E^{*\mathsf{T}} \mathbf{y} = \mathbf{y}^\mathsf{T} \mathbf{a}_E^*, \quad \mathbf{a}_E^* = a_E^* \mathbf{v}^*, \tag{142}$$

$$y_I^* = \mathbf{a}_I^{*\mathsf{T}} \mathbf{y} = \mathbf{y}^\mathsf{T} \mathbf{a}_I^*, \quad \mathbf{a}_I^* = a_I^* \mathbf{v}^*, \tag{143}$$

Note that the superscript '*' indicates a value in the fixed point of the weight dynamics and not a fixed point of the firing rate activity. Different input patterns $\mathbf{y}$ still result in different neural activities y_E^*. The modified covariance matrix in the fixed point becomes

$$\hat{\mathbf{C}}^* = \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{v}^* a_E^* & -\mathbf{C}\mathbf{v}^* a_I^* & \\ a_E^* \mathbf{v}^{*\mathsf{T}}\mathbf{C} & a_E^* \mathbf{v}^{*\mathsf{T}}\mathbf{C}\mathbf{v}^* a_E^* & -a_E^* \mathbf{v}^{*\mathsf{T}}\mathbf{C}\mathbf{v}^* a_I^* & \mathbf{0} \\ a_I^* \mathbf{v}^{*\mathsf{T}}\mathbf{C} & a_I^* \mathbf{v}^{*\mathsf{T}}\mathbf{C}\mathbf{v}^* a_E^* & -a_I^* \mathbf{v}^{*\mathsf{T}}\mathbf{C}\mathbf{v}^* a_I^* & \\ & \mathbf{0} & & \ddots \end{pmatrix} \tag{144}$$

$$= \begin{pmatrix} \mathbf{C} & \lambda^* a_E^* \mathbf{v}^* & -\lambda^* a_I^* \mathbf{v}^* & \\ \lambda^* a_E^* \mathbf{v}^{*\mathsf{T}} & \lambda^* a_E^{*2} & -\lambda^* a_E^* a_I^* & \mathbf{0} \\ \lambda^* a_I^* \mathbf{v}^{*\mathsf{T}} & \lambda^* a_I^* a_E^* & -\lambda^* a_I^{*2} & \\ & \mathbf{0} & & \ddots \end{pmatrix}. \tag{145}$$

which can be diagonalized by the eigenvector matrix $\hat{\mathbf{V}}^*$ and its inverse:

$$\hat{\mathbf{V}}^* = \begin{pmatrix} \mathbf{V}_{\setminus *} & \mathbf{v}^* & \mathbf{v}^* a_E^* & \mathbf{v}^* a_I^* & \\ \mathbf{0} & a_E^* & -1 & 0 & \mathbf{0} \\ \mathbf{0} & a_I^* & 0 & 1 & \\ & & \mathbf{0} & & \ddots \end{pmatrix}, \quad \hat{\mathbf{V}}^{*-1} = \mathcal{N}^{-1} \begin{pmatrix} \mathcal{N}\mathbf{V}_{\setminus *}^\mathsf{T} & \mathbf{0} & \mathbf{0} & \\ \mathbf{v}^{*\mathsf{T}} & a_E^* & -a_I^* & \mathbf{0} \\ a_E \mathbf{v}^{*\mathsf{T}} & -(1 - a_I^{*2}) & -a_E^* a_I^* & \\ -a_I \mathbf{v}^{*\mathsf{T}} & -a_I^* a_E^* & 1 + a_E^{*2} & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \tag{146}$$

$$\mathcal{N} \equiv 1 + a_E^{*2} - a_I^{*2}, \tag{147}$$

where the subscript $(\cdot)_{\setminus *}$ indicates that a matrix does not contain an entry that corresponds to the input mode $\mathbf{v}^*$. The first set of eigenvectors has eigenvalues $\mathbf{\Lambda}_{\setminus *}$, which are eigenvalues of the feedforward covariance matrix $\mathbf{C}$, since

[1] It does not affect the sign of the eigenvalues of the Jacobian, since τ is always positive. In principle, however, different timescales for excitatory and inhibitory weights can affect stability (compare Section 3.2)

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483995; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

there are no neurons in the recurrent circuit[1] that are tuned to these input modes. The last two eigenvectors in the first block are null-eigenvectors with eigenvalue zero. Their excitatory feedforward component is balanced by either an inhibitory component or a negative recurrent excitatory component[2]. The remaining eigenvector has eigenvalue $1 + a_E^{*2} - a_I^{*2}$. Similar to the feedforward case (compare section 3.1), arbitrary multiples of the separately normalized parts of eigenvectors of $\hat{\mathbf{C}}^*$ are fixed points[3]. Inserting these fixed points into Eq. 136 and 137, provides conditions to determine a_E^* and a_I^*. We consider the simplest case of an eigenvector of the following shape:

$$\hat{\mathbf{v}}^* = \begin{pmatrix} \mathbf{v}^* \\ a_E^* \\ a_I^* \\ \mathbf{v}^* \\ a_E^* \\ a_I^* \end{pmatrix}, \quad \hat{\lambda}^* = \lambda^* \left(1 + a_E^{*2} - a_I^{*2}\right), \tag{148}$$

6.2 Stability analysis

When are neurons stable, and when do they destabilize and become attracted to a different input mode? To answer this question, we consider a small fixed point perturbations $\Delta\mathbf{w}$, where the excitatory neuron shifts its tuning in the direction of a different input mode $\mathbf{v}^\dagger$:

$$\Delta\mathbf{w} \propto \begin{pmatrix} \mathbf{v}^\dagger \\ \mathbf{0} \end{pmatrix}. \tag{149}$$

Then, the neuron is stable with respect to the perturbation if the perturbation decays to zero. To check this, we solve the following differential equation that holds for small perturbations (compare section 2.2)

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta\mathbf{w}) = \mathbf{J}^*(\Delta\mathbf{w}). \tag{150}$$

We will consider the dynamics in the non-orthogonal eigenbasis $\hat{\mathbf{V}}^*$ of the modified covariance matrix $\hat{\mathbf{C}}$. In this basis, the perturbation is defined as

$$\Delta\mathbf{w} = \hat{\mathbf{V}}^* \Delta\mathbf{w}_v. \tag{151}$$

and its dynamics becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta\mathbf{w}_v) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\hat{\mathbf{V}}^{*-1}\Delta\mathbf{w}\right) = \hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^*\hat{\mathbf{V}}^{*-1}(\Delta\mathbf{w}) = \hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^*\mathbf{e}^\dagger, \tag{152}$$

where $\mathbf{e}^\dagger$ is a vector of zeros with a single non-zero entry that corresponds to the perturbation mode $\mathbf{v}^\dagger$. Without loss of generality, we assume that eigenvectors in $\hat{\mathbf{V}}^*$ are sorted such that the first entry of $\mathbf{e}^\dagger$ is non-zero. Note that for perturbations $\Delta\mathbf{w}'$ that do not shift a neuron away from its fixed point input mode $\mathbf{v}^*$, the first entries of the perturbation vector expressed in the eigenbasis ($\hat{\mathbf{V}}^*\Delta\mathbf{w}'$) are zero. Such perturbations have no component in the direction of other input modes $\mathbf{V}_{\setminus *}$ (compare Eq. 146). In the following, we will derive the transformed Jacobian $\hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^*$.

The transformed Jacobian

First we consider the regular Jacobian $\mathbf{J}^*$. We rewrite the dynamics in Eq. 140 as

$$\dot{\mathbf{w}} = \left[\mathbb{1} - \frac{(\mathbf{w}_{EF}^\circ + \mathbf{w}_{EE}^\circ)\mathbf{c}_{EE}^{\circ\,\mathsf{T}}}{\mathbf{c}_{EE}^{\circ\,\mathsf{T}}(\mathbf{w}_{EF}^\circ + \mathbf{w}_{EE}^\circ)} - \ldots\right]\hat{\mathbf{C}}\mathbf{w} \tag{153}$$

where the second term in the bracket corresponds to the normalization of all excitatory synapses onto the excitatory neuron, and additional normalization terms are indicated by ellipsis (compare Eq. 47). Then the

[1] No neurons that are not input neurons.

[2] In our simulations, we constrain synaptic weights to be positive. When a null-eigenvector is added to a regular eigenvector, the net synaptic inputs remain unchanged. For example, a decrease in recurrent excitation due to a negative excitatory component of the null eigenvector is balanced by an increase in feedforward excitation.

[3] The only exception is the rightmost null-eigenvector. There, the inhibitory and the excitatory weights are aligned such that the postsynaptic activity is zero, which does not allow for arbitrary scaling of the weight norms.

Jacobian has the following shape (compare Eq. 31)

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\mathbf{J}^* = \left.\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}\mathbf{w}}\right|_* = \left[\mathbb{1} - \frac{(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})\mathbf{c}^{\circ\,\mathsf{T}}_{EE}}{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})} - \dots\right]\left(\hat{\mathbf{C}} - \hat{\lambda}^*\mathbb{1} + \left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_* \mathbf{w}^*\right), \tag{154}$$

where we accounted for the weight dependence of the modified covariance matrix $\hat{\mathbf{C}}$ which results in the tensor $\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}$. To find the transformed Jacobian $\hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^*$, we consider the first bracket:

$$\hat{\mathbf{V}}^{*-1}\left[\mathbb{1} - \frac{(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})\mathbf{c}^{\circ\,\mathsf{T}}_{EE}}{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})} - \dots\right]\hat{\mathbf{V}}^* \tag{155}$$

$$= \hat{\mathbf{V}}^{*-1}\left[\mathbb{1} - \frac{(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})\mathbf{c}^{\circ\,\mathsf{T}}_{EE}}{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})} - \dots\right]\begin{pmatrix} \mathbf{V}_{\setminus *} & \mathbf{v}^* & \mathbf{v}^* a^*_E & \mathbf{v}^* a^*_I & \\ \mathbf{0} & a^*_E & -1 & 0 & \mathbf{0} \\ \mathbf{0} & a^*_I & 0 & 1 & \\ & & \mathbf{0} & & \ddots \end{pmatrix} \tag{156}$$

$$= \left[\mathbb{1} - \hat{\mathbf{V}}^{*-1}\mathbf{H}\hat{\mathbf{v}}^*\right], \quad \mathbf{H}_i \equiv \begin{pmatrix} \mathbb{1}h^{(i)}_{EE} & & & & & \\ & h^{(i)}_{EE} & & & & \\ & & h^{(i)}_{EI} & & & \\ & & & \mathbb{1}h^{(i)}_{IE} & & \\ & & & & h^{(i)}_{IE} & \\ & & & & & h^{(i)}_{II} \end{pmatrix}_i, \tag{157}$$

where $\mathbf{H}$ is a tensor such that $\mathbf{H}_1\hat{\mathbf{v}}^*$ is the first column of $\mathbf{H}\hat{\mathbf{v}}^*$ and

$$h^{(i)}_{EE} \equiv \frac{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}}{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})}\hat{\mathbf{V}}^*_i, \tag{158}$$

where $\hat{\mathbf{V}}^*_i$ is the ith column of $\hat{\mathbf{V}}^*$. Then

$$\hat{\mathbf{V}}^{*-1}\mathbf{H}\hat{\mathbf{v}}^* = \mathcal{N}^{-1}\begin{pmatrix} \mathcal{N}\mathbf{V}_{\setminus *}{}^{\mathsf{T}} & \mathbf{0} & \mathbf{0} & \\ \mathbf{v}^{*\mathsf{T}} & a^*_E & -a^*_I & \mathbf{0} \\ a_E\mathbf{v}^{*\mathsf{T}} & -(1-a^{*2}_I) & -a^*_E a^*_I & \\ -a_I\mathbf{v}^{*\mathsf{T}} & -a^*_I a^*_E & 1+a^{*2}_E & \\ & \mathbf{0} & & \ddots \end{pmatrix}\mathbf{H}\hat{\mathbf{v}}^* = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & \vdots & \mathbf{0} \\ & \mathbf{0} & & & \ddots \end{pmatrix} \tag{159}$$

$$\Rightarrow \quad \hat{\mathbf{V}}^{*-1}\left[\mathbb{1} - \frac{(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})\mathbf{c}^{\circ\,\mathsf{T}}_{EE}}{\mathbf{c}^{\circ\,\mathsf{T}}_{EE}(\hat{\mathbf{v}}^{*\circ}_{EF} + \hat{\mathbf{v}}^{*\circ}_{EE})} - \dots\right]\hat{\mathbf{V}}^* = \begin{pmatrix} \mathbb{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & \vdots & \mathbf{0} \\ & \mathbf{0} & & & \ddots \end{pmatrix} \tag{160}$$

where ellipsis indicate potentially non-zero entries. After transformation, the second bracket of Eq. 154 becomes

$$\hat{\mathbf{V}}^{*-1}\left(\hat{\mathbf{C}}^* - \hat{\lambda}^*\mathbb{1} + \left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_* \mathbf{w}^*\right)\hat{\mathbf{V}}^* = \left(\hat{\Lambda}^* - \hat{\lambda}^*\mathbb{1} + \hat{\mathbf{V}}^{*-1}\left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_* \mathbf{w}^*\hat{\mathbf{V}}^*\right). \tag{161}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We next consider the first columns of $\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\Big|_* \mathbf{w}^*$, for which we compute the matrix $\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}w^b_{EF}}$, where w^b_{EF} is the bth feedforward weight onto the excitatory neuron, which has the weight onto the excitatory neuron,

$$\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}w^b_{EF}}\Big|_* = \frac{\mathrm{d}}{\mathrm{d}w^b_{EF}} \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{a}_E & -\mathbf{C}\mathbf{a}_I & \\ \mathbf{a}_E{}^\mathsf{T}\mathbf{C} & \mathbf{a}_E{}^\mathsf{T}\mathbf{C}\mathbf{a}_E & -\mathbf{a}_E{}^\mathsf{T}\mathbf{C}\mathbf{a}_I & \mathbf{0} \\ \mathbf{a}_I{}^\mathsf{T}\mathbf{C} & \mathbf{a}_I{}^\mathsf{T}\mathbf{C}\mathbf{a}_E & -\mathbf{a}_I{}^\mathsf{T}\mathbf{C}\mathbf{a}_I & \\ & \mathbf{0} & & \ddots \end{pmatrix}\Bigg|_* \tag{162}$$

$$= \begin{pmatrix} \mathbf{0} & \mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_E}{\mathrm{d}w^b_{EF}}\Big|_* & -\mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_I}{\mathrm{d}w^b_{EF}}\Big|_* & \\ \frac{\mathrm{d}\mathbf{a}_E{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C} & \left(\frac{\mathrm{d}\mathbf{a}_E{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C}\mathbf{a}^*_E + \mathbf{a}^{*\mathsf{T}}_E \mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_E}{\mathrm{d}w^b_{EF}}\Big|_*\right) & -\left(\frac{\mathrm{d}\mathbf{a}_E{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C}\mathbf{a}^*_I + \mathbf{a}^{*\mathsf{T}}_E \mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_I}{\mathrm{d}w^b_{EF}}\Big|_*\right) & \mathbf{0} \\ \frac{\mathrm{d}\mathbf{a}_I{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C} & \left(\frac{\mathrm{d}\mathbf{a}_I{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C}\mathbf{a}^*_E + \mathbf{a}^{*\mathsf{T}}_I \mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_E}{\mathrm{d}w^b_{EF}}\Big|_*\right) & -\left(\frac{\mathrm{d}\mathbf{a}_I{}^\mathsf{T}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{C}\mathbf{a}^*_I + \mathbf{a}^{*\mathsf{T}}_I \mathbf{C}\,\frac{\mathrm{d}\mathbf{a}_I}{\mathrm{d}w^b_{EF}}\Big|_*\right) & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \tag{163}$$

where we used the definition of $\hat{\mathbf{C}}$ from Eq. 139. The vectors $\mathbf{a}_E$ and $\mathbf{a}_I$ are defined in Eq. 136 & 137. It follows:

$$\frac{\mathrm{d}\mathbf{a}_E}{\mathrm{d}w^b_{EF}}\Big|_* = \frac{1}{1 - w^*_{EE} + \dfrac{w^*_{EI} w^*_{IE}}{1 + w^*_{II}}}\,\mathbf{e}_b \equiv \mu_E \mathbf{e}_b, \tag{164}$$

$$\frac{\mathrm{d}\mathbf{a}_I}{\mathrm{d}w^b_{EF}}\Big|_* = \frac{1}{1 + w^*_{II} + \dfrac{w^*_{IE} w^*_{EI}}{1 - w^*_{EE}}}\,\frac{w_{IE}}{1 - w_{EE}}\,\mathbf{e}_b \equiv \mu_I \mathbf{e}_b, \tag{165}$$

$$\mathbf{C}\mathbf{a}^*_E = \lambda^* a^*_E \mathbf{v}^*, \quad \mathbf{C}\mathbf{a}^*_E = \lambda^* a^*_I \mathbf{v}^*, \tag{166}$$

$$\Rightarrow \quad \frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}w^b_{EF}}\Big|_* = \begin{pmatrix} \mathbf{0} & \mu_E \mathbf{C}\mathbf{e}_b & -\mu_I \mathbf{C}\mathbf{e}_b & \\ \mu_E \mathbf{e}_b{}^\mathsf{T}\mathbf{C} & 2\lambda^* a^*_E \mu_E \mathbf{v}^{*\mathsf{T}}\mathbf{e}_b & -\lambda^*(\mu_E a^*_E + \mu_I a^*_E)\mathbf{v}^{*\mathsf{T}}\mathbf{e}_b & \mathbf{0} \\ \mu_I \mathbf{e}_b{}^\mathsf{T}\mathbf{C} & \lambda^*(\mu_I a^*_E + \mu_E a^*_I)\mathbf{v}^{*\mathsf{T}}\mathbf{e}_b & 2\lambda^* a^*_I \mu_I \mathbf{e}_b \mathbf{v}^{*\mathsf{T}} & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \tag{167}$$

$$\Rightarrow \quad \frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}w^b_{EF}}\Big|_* \mathbf{w}^* = \begin{pmatrix} \beta_E \mathbf{C}\mathbf{e}_b \\ g_1 \mathbf{v}^{*\mathsf{T}}\mathbf{e}_b \\ g_2 \mathbf{v}^{*\mathsf{T}}\mathbf{e}_b \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{w}^* = \begin{pmatrix} \mathbf{v}^* \\ w^*_{EE} \\ w^*_{EI} \\ \vdots \end{pmatrix}, \quad \beta_E = \mu_E w^*_{EE} - \mu_I w^*_{EI}. \tag{168}$$

$$\Rightarrow \quad \frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}_{EF}}\Big|_* \mathbf{w}^* = \begin{pmatrix} \beta_E \mathbf{C} \\ g_1 \mathbf{v}^{*\mathsf{T}} \\ g_2 \mathbf{v}^{*\mathsf{T}} \\ \mathbf{0} \end{pmatrix}. \tag{169}$$

We find other columns in a similar fashion and write

$$\Rightarrow \quad \frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\Big|_* \mathbf{w}^* = \begin{pmatrix} \beta_E \mathbf{C} & g_3 \mathbf{v}^* & g_6 \mathbf{v}^* & \\ g_1 \mathbf{v}^{*\mathsf{T}} & g_4 & g_7 & \mathbf{0} \\ g_2 \mathbf{v}^{*\mathsf{T}} & g_5 & g_8 & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \tag{170}$$

where $g_{(\cdot)}$ are scalars. After applying the transformation, we get

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\hat{\mathbf{V}}^{*-1}\left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_{*}\mathbf{w}^{*}\hat{\mathbf{V}}^{*} \tag{171}$$

$$= \hat{\mathbf{V}}^{*-1}\begin{pmatrix} \beta_E\mathbf{C} & g_3\mathbf{v}^* & g_6\mathbf{v}^* & & \\ g_1\mathbf{v}^{*\mathsf{T}} & g_4 & g_7 & \mathbf{0} & \\ g_2\mathbf{v}^{*\mathsf{T}} & g_5 & g_8 & & \\ & \mathbf{0} & & \ddots & \end{pmatrix}\begin{pmatrix} \mathbf{V}_{\setminus *} & \mathbf{v}^* & \mathbf{v}^*a_E^* & \mathbf{v}^*a_I^* & \\ \mathbf{0} & a_E^* & -1 & 0 & \mathbf{0} \\ \mathbf{0} & a_I^* & 0 & 1 & \\ & \mathbf{0} & & & \ddots \end{pmatrix} \tag{172}$$

$$= \hat{\mathbf{V}}^{*-1}\begin{pmatrix} \beta_E\mathbf{C}\mathbf{V}_{\setminus *} & g_9\mathbf{v}^* & g_{12}\mathbf{v}^* & g_{15}\mathbf{v}^* & \\ \mathbf{0} & g_{10} & g_{13} & g_{16} & \mathbf{0} \\ \mathbf{0} & g_{11} & g_{14} & g_{17} & \\ & \mathbf{0} & & & \ddots \end{pmatrix} \tag{173}$$

$$= \mathcal{N}^{-1}\begin{pmatrix} \mathcal{N}\mathbf{V}_{\setminus *}{}^{\mathsf{T}} & \mathbf{0} & \mathbf{0} & \\ \mathbf{v}^{*\mathsf{T}} & a_E^* & -a_I^* & \\ a_E\mathbf{v}^{*\mathsf{T}} & -(1-a_I^{*2}) & -a_E^*a_I^* & \mathbf{0} \\ -a_I\mathbf{v}^{*\mathsf{T}} & -a_I^*a_E^* & 1+a_E^{*2} & \\ & \mathbf{0} & & \ddots \end{pmatrix}\begin{pmatrix} \beta_E\mathbf{V}_{\setminus *}\mathbf{\Lambda}_{\setminus *} & g_9\mathbf{v}^* & g_{12}\mathbf{v}^* & g_{15}\mathbf{v}^* & \\ \mathbf{0} & g_{10} & g_{13} & g_{16} & \mathbf{0} \\ \mathbf{0} & g_{11} & g_{14} & g_{17} & \\ & \mathbf{0} & & & \ddots \end{pmatrix} \tag{174}$$

$$= \begin{pmatrix} \beta_E\mathbf{\Lambda}_{\setminus *} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \mathbf{0} & g_{18} & g_{21} & g_{24} & \mathbf{0} \\ \mathbf{0} & g_{19} & g_{22} & g_{25} & \\ \mathbf{0} & g_{20} & g_{23} & g_{26} & \\ & \mathbf{0} & & & \ddots \end{pmatrix}. \tag{175}$$

The fully transformed Jacobian becomes (compare Eq. 161)

$$\hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^* = \hat{\mathbf{V}}^{*-1}\left[\mathbb{1} - \frac{(\hat{\mathbf{v}}_{EF}^{*\circ} + \hat{\mathbf{v}}_{EE}^{*\circ})\mathbf{c}_{EE}^{\circ\,\mathsf{T}}}{\mathbf{c}_{EE}^{\circ\,\mathsf{T}}(\hat{\mathbf{v}}_{EF}^{*\circ} + \hat{\mathbf{v}}_{EE}^{*\circ})} - \dots\right]\hat{\mathbf{V}}^{*-1}\left(\hat{\mathbf{\Lambda}}^* - \hat{\lambda}^*\mathbb{1} + \hat{\mathbf{V}}^{*-1}\left.\frac{\mathrm{d}\hat{\mathbf{C}}}{\mathrm{d}\mathbf{w}}\right|_{*}\mathbf{w}^*\hat{\mathbf{V}}^*\right) \tag{176}$$

Finally, by inserting Eq. 160 and 175 we get

$$\hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^* = \begin{pmatrix} \mathbb{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & \vdots & \mathbf{0} \\ & \mathbf{0} & & & \ddots \end{pmatrix}\left(\hat{\mathbf{\Lambda}}^* - \hat{\lambda}^*\mathbb{1} + \begin{pmatrix} \beta_E\mathbf{\Lambda}_{\setminus *} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \mathbf{0} & g_{18} & g_{21} & g_{24} & \mathbf{0} \\ \mathbf{0} & g_{19} & g_{22} & g_{25} & \\ \mathbf{0} & g_{20} & g_{23} & g_{26} & \\ & \mathbf{0} & & & \ddots \end{pmatrix}\right). \tag{177}$$

$$\Rightarrow \quad \hat{\mathbf{V}}^{*-1}\mathbf{J}^*\hat{\mathbf{V}}^* = \begin{pmatrix} \left(\mathbf{\Lambda}_{\setminus *} - \mathbb{1}\hat{\lambda}^* + \beta_E\mathbf{\Lambda}_{\setminus *}\right) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\ \vdots & & \vdots & \vdots & \vdots & \mathbf{0} \\ & \mathbf{0} & & & & \ddots \end{pmatrix}. \tag{178}$$

Stability conditions

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The dynamics of a fixed point perturbation $\Delta\mathbf{w}_v$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta\mathbf{w}_v) = \begin{pmatrix} \left(\mathbf{\Lambda}_{\setminus *} - \mathbb{1}\hat{\lambda}^* + \beta_E \mathbf{\Lambda}_{\setminus *}\right) & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & \mathbf{0} \\ \mathbf{0} & & & \ddots \end{pmatrix} \Delta\mathbf{w}_v. \tag{179}$$

It follows that perturbations $\Delta\mathbf{w}'$ that do not shift a neuron's tuning in the direction of a different input mode do also not induce such a shift in their later dynamics, i.e., the first vector entries of such perturbations remain zero. In contrast, perturbations in the direction of a different input mode $\Delta\mathbf{w}_v = \mathbf{e}^\dagger$ induce perturbations within the original eigencircuit: A decrease in feedforward and recurrent excitatory synaptic weights within the eigencircuit balances the increase of feedforward excitation synaptic weights due to the perturbation to maintain the weight norm[1]. However, as explained above, these second-order perturbations are contained within the eigencircuit, i.e., they can not induce perturbations in the direction of non-eigencircuit input modes. To answer the question when neurons switch eigencircuits, we therefore only consider the dynamics along the direction of the original perturbation by projecting the dynamics onto the perturbation vector at time zero $\mathbf{e}_0^\dagger = (\mathbf{v}^{\dagger\mathsf{T}}, \mathbf{0}^\mathsf{T})^\mathsf{T}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{e}_0^{\dagger\mathsf{T}}\mathbf{e}^\dagger\right) = \left(\lambda^\dagger - \hat{\lambda}^* + \beta_E\lambda^\dagger\right)(\mathbf{e}_0^{\dagger\mathsf{T}}\mathbf{e}^\dagger). \tag{180}$$

which provides the general stability condition for the excitatory neuron

$$\boxed{\left(\hat{\lambda}^\dagger - \hat{\lambda}^* + \beta_E\hat{\lambda}^\dagger\right) < 0}, \tag{181}$$

where $\hat{\lambda^\dagger} = \lambda^\dagger + \lambda^\dagger_{\mathrm{eig}}$. In the circuit considered here, we integrated the eigencircuit attraction into the feedforward input, i.e., $\lambda^\dagger_{\mathrm{eig}}$ is zero. For β_E we find

$$\beta_E = \frac{1}{1 - w^*_{EE} + \dfrac{w^*_{EI}w^*_{IE}}{1+w^*_{II}}} w^*_{EE} - \frac{1}{1 + w^*_{II} + \dfrac{w^*_{IE}w^*_{EI}}{1-w^*_{EE}}}\left(\frac{w^*_{EI}w^*_{IE}}{1-w^*_{EE}}\right), \tag{182}$$

$$= \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\mathbf{w}_{EF}{}^\mathsf{T}\mathbf{y})}\right|_* w^*_{EE} - \left.\frac{\mathrm{d}y_I}{\mathrm{d}(\mathbf{w}_{EF}{}^\mathsf{T}\mathbf{y})}\right|_* w^*_{EI}, \tag{183}$$

$$= \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\mathbf{w}_{EF}{}^\mathsf{T}\mathbf{y})}\right|_* \left[w^*_{EE} - \frac{w^*_{EI}w^*_{IE}}{1+w^*_{II}}\right], \tag{184}$$

$$\Rightarrow \quad \boxed{\beta_E = \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\mathbf{w}_{EF}{}^\mathsf{T}\mathbf{y})}\right|_* - 1}, \tag{185}$$

where Eq. 183 follows from Eq. 164. This can be seen readily from Eq. 136, i.e.,

$$\left.\frac{\mathrm{d}y_E}{\mathrm{d}(\mathbf{w}_{EF}{}^\mathsf{T}\mathbf{y})}\right|_* = \frac{1}{1 - w^*_{EE} + \dfrac{w^*_{EI}w^*_{IE}}{1+w^*_{II}}}, \tag{186}$$

$$\left.\frac{\mathrm{d}y_I}{\mathrm{d}(\mathbf{w}_{IF}{}^\mathsf{T}\mathbf{y})}\right|_* = \frac{1}{1 + w^*_{II} + \dfrac{w^*_{IE}w^*_{EI}}{1-w^*_{EE}}}. \tag{187}$$

Following the same framework, we find the stability condition when perturbing the inhibitory neuron:

$$\boxed{\left(\hat{\lambda}^\dagger - \hat{\lambda}^* + \beta_I\hat{\lambda}^\dagger\right) < 0}, \tag{188}$$

$$\beta_I = \frac{1}{1 - w^*_{EE} + \dfrac{w^*_{EI}w^*_{IE}}{1+w^*_{II}}}\left(\frac{-w^*_{EI}w^*_{IE}}{1+w^*_{II}}\right) - \frac{1}{1 + w^*_{II} + \dfrac{w^*_{IE}w^*_{EI}}{1-w^*_{EE}}} w^*_{II}, \tag{189}$$

$$= \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\mathbf{w}_{IF}{}^\mathsf{T}\mathbf{y})}\right|_* w^*_{IE} - \left.\frac{\mathrm{d}y_I}{\mathrm{d}(\mathbf{w}_{IF}{}^\mathsf{T}\mathbf{y})}\right|_* w^*_{II}, \tag{190}$$

$$= \left.\frac{\mathrm{d}y_I}{\mathrm{d}(\mathbf{w}_{IF}{}^\mathsf{T}\mathbf{y})}\right|_* \left[-\frac{w^*_{EI}w^*_{IE}}{1-w^*_{EE}} - w^*_{II}\right]. \tag{191}$$

[1] Compare Eq. 158: $h^{(i)}_{EE} \neq 0$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$\Rightarrow \boxed{\beta_I = \frac{dy_I}{d(\mathbf{w}_{IE}^{\top}\mathbf{y})}\Big|_{*} - 1}, \tag{192}$$

6.3 Decorrelation condition

How can neurons self-organize to represent all parts of their input space instead of clustering all their tuning curves around a dominant input mode? To answer this question, we consider the stability of eigencircuits (Eq. 181 & 188). In particular, we consider the case when contributions from recurrent excitatory and inhibitory connectivity motives balance each other, such that β_E and β_I are negative but close to zero[1]. This can be achieved by a suitable choice of weight norms (see Section 6.4 for a discussion of the case $\beta_{E/I} > 0$). Then, the network is stable when all input modes are equally attractive, i.e.,

$$\hat{\lambda}^a = \lambda^a + \lambda^a_{\text{eig}} \stackrel{!}{=} \hat{\lambda}^b, \quad \forall a, b. \tag{193}$$

For homogeneous input spaces, where $\lambda^a = \lambda^b = \lambda, \forall a, b$, the only other stable configuration is when all neurons are tuned to the same input mode. Such a configuration does not reflect neural tunings in the healthy brain, where all parts of the stimulus space are represented. To prevent such a global clustering of neural tunings, we require that the corresponding eigencircuit is unstable. This is the case when the total attraction of the eigencircuit's input mode $\hat{\lambda}^*$ is smaller than the attraction of an unoccupied input mode $\lambda^\dagger$:

$$\hat{\lambda}^* < \lambda^\dagger, \tag{194}$$

$$\Rightarrow \quad \sum_i \sigma^2_{E,i} - \sum_I \sigma^2_{I,j} + \lambda < \lambda, \tag{195}$$

$$\Rightarrow \quad N_E\sigma_E^2 - N_I\sigma_I^2 < 0, \tag{196}$$

$$\Rightarrow \quad \boxed{N_E\sigma_E^2 < N_I\sigma_I^2}, \tag{197}$$

where σ_E^2, σ_I^2 are the average variances of the population and N_E, N_I are the total number of inhibitory and excitatory neurons.

In general, we want the total attraction of an input mode to decrease when additional neurons join an eigencircuit. Such a decrease prevents the self-reinforcing feedback loop where an attractive input mode becomes even more attractive when additional neurons become tuned to that mode which attracts even more neurons and so forth. One can not prevent this feedback loop if the network only contains excitatory neurons because the total attraction would strictly increase with the number of neurons in an eigencircuit. However, this is not the case in networks that also contain inhibitory neurons. When the repulsive contribution of inhibitory neurons exceeds the contribution of excitatory neurons, the total attraction is decreased. Then, assuming a sufficient amount of available neurons, differences in input mode feedforward attractions will be equalized, which prevents clustering of tuning curves. For every excitatory neuron, there are N_I/N_E many inhibitory neurons that become attracted to an eigencircuit[2]. Then the change in total attraction is

$$\Delta\hat{\lambda} = \hat{\lambda}(n_E + 1, n_I + \frac{N_I}{N_E}) - \hat{\lambda}(n_E, n_I), \tag{198}$$

where n_E, n_I are the original number of excitatory and inhibitory neurons in the eigencircuit. As before, we assume that neurons have similar weight norms, such that their variances are also similar (compare Eq. 123), and write the total attraction of the input mode as (making use of Eq. 127)

$$\hat{\lambda}(n_E, n_I) = \lambda + n_E\sigma_E(n_E, n_I) - n_I\sigma_I(n_E, n_I), \tag{199}$$

$$= \lambda + \left(n_E - \eta^2 n_I\right)\sigma_E(n_E, n_I), \quad \eta \equiv \frac{W_{IE}}{1 + W_{II}}\frac{1}{\Phi}. \tag{200}$$

The variance σ_E does not depend on n_I and strictly decreases with n_E (compare Eq. 128)[3]. Then, an upper bound for the change in total attraction is[4]

$$\Delta\hat{\lambda} < \left(1 - \eta^2\frac{N_I}{N_E}\right)\sigma_E(n_E + 1, n_I\frac{N_I}{N_E}) \stackrel{!}{<} 0, \tag{201}$$

[1] If $\beta_\cdot < 0$, while $|\beta_\cdot|$ is sufficiently large, eigencircuits that correspond to relatively weakly attractive input modes can nonetheless be stable. Small $|\beta_\cdot|$ ensure that modes with relatively weaker attraction remain unstable. Vice versa, when $\beta_\cdot$ is positive, more attractive eigencircuits are unstable due to their recurrent internal interactions. However, they may appear more attractive for a neuron that is tuned to a different input mode, which can lead to oscillatory dynamics.

[2] We assume that plasticity of synaptic weights onto inhibitory neurons is at least as fast as the plasticity of synaptic weights onto excitatory neurons such that both types of neurons adjust their tunings almost simultaneously.

[3] For $\Phi < 1$, which ensures real solutions to Eq. 128, independent of n_E.

[4] Assuming that the original eigencircuit was attractive: $n_E - \eta^2 n_I > 0$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

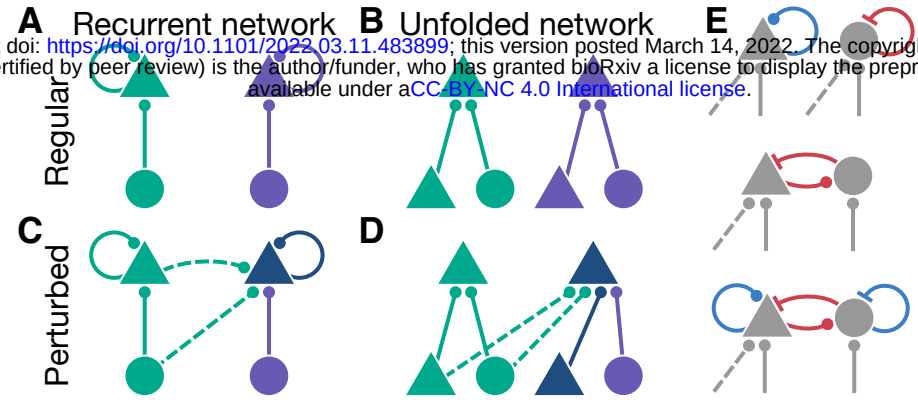

Figure S4: Input space warping due to lateral connectivity. (**A**) Two excitatory neurons (triangles) are tuned to two different but equally attractive input modes (circles, green and purple). Both neurons recurrently connect to themselves but not each other. (**B**) The same circuit as in *A*, unfolded to highlight presynaptic partners. Both input modes are balanced in their attraction. (**C**) Perturbing the purple excitatory neuron towards the green input mode (dashed lines) shifts its tuning (dark blue) such that it now response to both the green and the purple input modes. (**D**) Unfolded circuit shown in *C*. Due to the perturbation, the green input mode is now more attractive, and the previously purple excitatory neuron shifts its tuning. (**E**) Stabilizing (red) and destabilizing (blue) effect of different circuit patterns in case of a perturbation of a neuron in the direction of an equally attractive input mode (dashed line). Recurrent excitation destabilizes (top, left) while monosynaptic recurrent inhibition stabilizes the circuit (top right). Similarly, disynaptic recurrent inhibition has a stabilizing effect (center). In a fully recurrent E-I network (bottom), the net effect results from recurrent excitatory and inhibitory connectivity motifs. The self-inhibition of a recurrently connected inhibitory neuron decreases the stabilizing effect of the disynaptic recurrent inhibition pattern.

$$\Rightarrow \quad \boxed{\eta^2 > \frac{N_E}{N_I}} \quad \Leftrightarrow \quad N_E\sigma_E^2 < N_I\sigma_I^2, \tag{202}$$

which provides a sufficient condition to ensure that an increase of eigencircuit occupancy decreases the total attraction, and neurons become selective to all input modes. If the input space is inhomogeneous, i.e., different input modes have different feedforward attractions, all neurons may still become selective to the most attractive mode when their combined contribution does not sufficiently decrease the mode's total attraction. However, one can always prevent this by increasing the total number of neurons in the network. Investigating this scenario in more detail is left for future work.

6.4 Eigencircuits are stabilized by intra-eigencircuit inhibition and destabilized by intra-eigencircuit excitation

When the number of excitatory and inhibitory neurons in all eigencircuits is the same such that input modes are equally attractive, the network can still be unstable (compare Eq. 181 and Eq. 188 for $\hat{\lambda}^* = \hat{\lambda}^\dagger$, and $\beta_E > 0$, $\beta_I > 0$). This is due to recurrent interactions that destabilize the circuit by amplifying small perturbations. Consider an excitatory network where two neurons with equal weight norms are selective for two non-overlapping input modes of equal attraction and are recurrently connected to themselves but not each other (Fig. S4A). The effective attraction of an input mode is determined by the time-averaged activities of a neuron's presynaptic inputs, where it does not matter whether a connection is recurrent or not. In other words, a neuron can not distinguish if a synapse is recurrent or feedforward. Therefore, we can unfold the recurrent network and observe that the effective mode attraction is a combination of the feedforward input and the recurrent self-excitation (Fig. S4B). When one neuron is perturbed towards the opposing input mode, the tuning of the perturbed neuron changes slightly in the direction of that mode (Fig. S4C, dark blue). This tuning change leads to an attraction increase of the opposite mode, which is now more attractive – the perturbation is unstable (Fig.S4D). Similarly, if the recurrent connection is inhibitory instead, it decreases the attraction to the opposite input mode and thus stabilizes the network. In more complex networks, the combined effects of inhibitory and excitatory recurrence determine the stability of the system (Fig. S4E). These results are supported by our mathematical analysis above, where recurrent synaptic connections affect stability (compare Eq. 181, 185 & 186 for $\hat{\lambda}^\dagger = \hat{\lambda}^*$).

References

[1] Tim P Vogels, Henning Sprekeler, Friedemann Zenke, Claudia Clopath, and Wulfram Gerstner. "Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks". In: *Science* 334.6062 (2011), pp. 1569–1573.

[2] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. "The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex". In: *Neuron* 85.2 (2015), pp. 402–417.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.11.483899; this version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC-BY-NC-ND 4.0 International license.

[3] Erkki Oja. “Simplified neuron model as a principal component analyzer”. In: *Journal of mathematical biology* 15.3 (1982), pp. 267–273.

[4] Kenneth D Miller and David JC MacKay. “The role of constraints in Hebbian learning”. In: *Neural computation* 6.1 (1994), pp. 100–126.

[5] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[6] Claudia Clopath, Tim P Vogels, Robert C Froemke, and Henning Sprekeler. “Receptive field formation by interacting excitatory and inhibitory synaptic plasticity”. In: *BioRxiv* (2016), p. 066589.