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

Samuel Eckmann ✉[1,2], **Edward James Young**[2], **and Julijana Gjorgjieva**[1,3]

[1]Max Planck Institute for Brain Research, Frankfurt am Main, Germany, [2]Computational and Biological Learning Lab, University of Cambridge, United Kingdom, [3]TUM School of Life Sciences, Technical University Munich, Germany, ✉Corresponding author, Email: ec.sam@outlook.com

Cortical networks exhibit complex stimulus-response patterns that are based on specific recurrent interactions between neurons. For example, the balance between excitatory and inhibitory currents has been identified as a central component of cortical computations. However, it remains unclear how the required synaptic connectivity can emerge in developing circuits where synapses between excitatory and inhibitory neurons are simultaneously plastic. Using theory and modeling, we propose that a wide range of cortical response properties can arise from a single plasticity paradigm that acts simultaneously at all excitatory and inhibitory connections – Hebbian learning that is stabilized by the synapse-type-specific competition for a limited supply of synaptic resources. In plastic recurrent circuits, this competition enables the formation and decorrelation of inhibition-balanced receptive fields. Networks develop an assembly structure with stronger synaptic connections between similarly tuned excitatory and inhibitory neurons and exhibit response normalization and orientation-specific center-surround suppression, reflecting the stimulus statistics during training. These results demonstrate how neurons can self-organize into functional networks and suggest an essential role for synapse-type-specific competitive learning in the development of cortical circuits.

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 can explain a wide range of cortical phenomena, such as cross-orientation and surround suppression (*17*, *18*), as well as stimulus-induced neural variability (*19*–*21*). A major caveat of these models is that the network connectivity is usually static and designed by hand, albeit based on experimental measurements. In contrast, in the brain, synapses are plastic and adjust to the statistics of sensory inputs. How synaptic weights self-organize in a biologically plausible manner to generate many of the non-linear response properties observed experimentally is not well understood. Earlier theoretical work on inhibitory plasticity has focused on the balance of excitation and inhibition in single neurons (*22*–*24*), 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*, *25*–*29*).

Here we present a Hebbian learning framework with minimal assumptions that explains a wide range of experimental observations. Our framework is based on two key properties: First, all synaptic strengths evolve according to a Hebbian plasticity rule that is stabilized by the competition for a limited supply of synaptic resources (*30*–*33*). Second, motivated by the unique protein composition of excitatory and inhibitory synapses, different synapse-types compete for separate resource pools. Building on classical work on Hebbian plasticity (*30*, *31*), we develop an analytical framework that provides an intuitive understanding of the weight dynamics in recurrent networks of excitatory and inhibitory neurons. In numerical simulations, we reveal how the synapse-type-specific competition for resources enables the self-organization of neurons into functional networks. Beyond the formation of inhibition-balanced feedforward receptive fields, we demonstrate that emergent recurrent connectivity can generate a wide range of computations observed in cortical circuits.

Results

Synapse-type-specific plasticity enables the joint development of stimulus selectivity and E-I balance. To understand plasticity in recurrently connected E-I networks, we considered simplified circuits of increasing complexity. We first asked how

Significance Statement

Cortical circuits perform diverse computations, primarily determined by highly structured synaptic connectivity patterns that develop during early sensory experience via synaptic plasticity. To understand how these structured connectivity patterns emerge, we introduce a general learning framework for networks of recurrently connected neurons. The framework is rooted in the biologically plausible assumption that synapses compete for limited synaptic resources, which stabilizes synaptic growth. Motivated by the unique protein composition of different synapse types, we assume that different synapse types compete for separate resource pools. Using theory and simulation, we show how this synapse-type-specific competition allows the stable development of structured synaptic connectivity patterns, as well as diverse computations like response normalization and surround suppression.

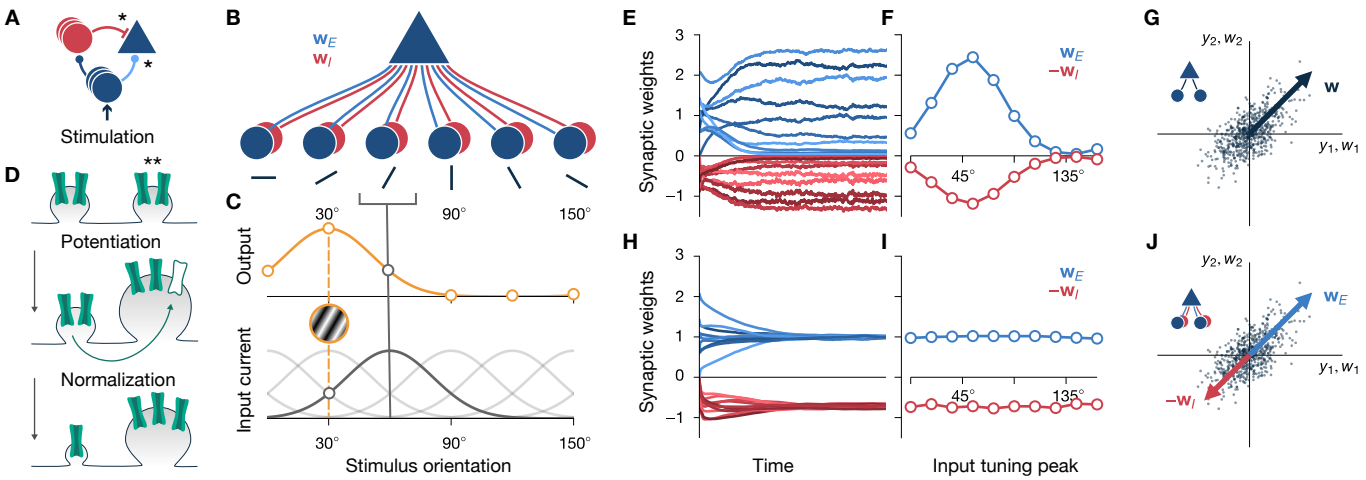

Figure 1: Synapse-type-specific competitive Hebbian learning enables the development of stimulus selectivity and inhibitory balance. (*A*) Feedforward input to a model pyramidal neuron during (blue triangle) stimulation. The neuron receives direct excitation (lightblue) and disynaptic inhibition (red). Plastic synapses are marked by *. (*B*) A single postsynaptic pyramidal neuron receives synaptic input from a population of excitatory ($\mathbf{w}_E$), and a population of inhibitory ($\mathbf{w}_I$) neurons. (*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 (orange, solid line) when a single grating of 30° is presented (orange plate, dashed line). (*D*) Hebbian potentiation of a synapse (**) is normalized due to a limited amount of synaptic resources in the dendritic branch, here reflected by a fixed number of synaptic channels (green). (*E*) Weight convergence of synapses of the feedforward circuit in *B*, where excitatory (blue) and inhibitory (red) weights are plastic according to synapse-type-specific competitive Hebbian learning. All synaptic weights were initialized randomly. (*F*) Final synaptic weight strengths, after training, as a function of the tuning peak of the corresponding presynaptic neurons. (*G*) Excitatory synaptic weight vector (blue arrow) of a single pyramidal neuron with linear activation function. The pyramidal neuron receives input from two excitatory neurons (y_1 and y_2, compare inset). Each dot corresponds to one input pattern. After training, the weight vector aligns with the direction of maximum variance, which corresponds to the principal eigenvector of the input covariance matrix. (*H* & *I*) Same as in *E* and *F*, but for classic inhibitory plasticity. The development of stimulus selectivity is prevented by fast inhibitory plasticity. (*J*) Excitatory (blue) and inhibitory (red) synaptic weight vectors of a single pyramidal neuron with linear activation function. The pyramidal neuron receives input from two pairs of excitatory and inhibitory neurons (y_1 and y_2, compare inset). Each excitatory-inhibitory input pair has identical firing activities y_i. After training via synapse-type-specific competitive Hebbian learning, the excitatory and inhibitory weight vectors both align with the principal component, i.e., excitatory and inhibitory synaptic weights are balanced.

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. 1*A*). We studied the self-organization of excitatory and inhibitory synapses that project onto the single postsynaptic neuron (Fig. 1*B*), assuming that input synapses that project onto inhibitory neurons remained fixed (Fig. 1*A*). Following experimental results (*34*–*37*), we assumed that inhibitory and excitatory input neurons are equally selective for the orientation of a stimulus grating (Fig. 1*C*, 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. 1*C*, top) and thus drove the postsynaptic neuron (see Methods for details). Synapses are plastic according to a basic Hebbian rule:

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

where r is the postsynaptic firing rate, $\mathbf{y}_A$ is a vector that holds the presynaptic firing rates of excitatory ($A = E$) and inhibitory ($A = I$) neurons, and $\Delta \mathbf{w}_A$ are the corresponding synaptic weight changes. Experimental results have shown that after the induction of long-term plasticity neither the total excitatory nor the total inhibitory synaptic area change (*32*). 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. 1*D*) (*33*). Motivated by these results, we adopted a competitive normalization rule for both excitatory and inhibitory synapses:

$$\mathbf{w}_A \leftarrow W_A \frac{\mathbf{w}_A + \Delta \mathbf{w}_A}{\|\mathbf{w}_A + \Delta \mathbf{w}_A\|}, \qquad [2]$$

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. 1*E*) and form aligned, Gaussian-shaped tuning curves (Fig. 1*F*) that reflect the shape of the input stimuli (Fig. 1*C*). 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.

Excitatory plasticity performs principal component analysis. 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*, *31*). In the case of a linear activation function, $r \propto u \equiv \mathbf{w}^\mathsf{T}\mathbf{y}$, the expected total synaptic efficacy changes can be expressed as (*31*):

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

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. The neuron becomes selective to the first principal component of its

input data, i.e., the fixed point input weight vector aligns with the input space direction of maximum variance (*30*, *31*) (Fig. 1*G*; see Supplementary Material (SM) Sec. 1.2 for details). For non-linear activation functions $r = f(u)$, neurons become selective for higher-order correlations, e.g., independent components, in their inputs (*38*, *39*). Such learning rules have been shown to result in feedforward receptive fields that resemble simple cell receptive fields in visual cortex (*40*, *41*). In the following, we call the fixed points of such pure feedforward circuits 'input modes'. This entails principal components, in the case of linear activation functions, and more complex, e.g., simple-cell-like, receptive fields in the case of non-linear activation functions.

Classic inhibitory plasticity prevents stimulus selectivity. 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 (*42*, *43*) and imposes a target firing rate r_0 on the postsynaptic neuron (*23*):

$$\langle \dot{\mathbf{w}}_I \rangle \propto \langle \mathbf{y}_I \, (r - r_0) \rangle, \qquad [4]$$

where synaptic change becomes zero when the postsynaptic firing rate r is equal to the target rate r_0. With this 'classic' inhibitory plasticity rule, inhibitory synaptic weight growth is unbounded. However, since an increase of inhibitory synaptic weights usually entails a decrease in postsynaptic firing rate r the plasticity rule is self-limiting and synaptic weights stop growing once the target firing rate r_0 is reached. When excitatory synaptic weights remain fixed, classic inhibitory plasticity leads to balanced excitatory and inhibitory input currents (*23*). However, when excitatory synaptic weights are also plastic, neurons develop no stimulus selectivity (*24*): Classic inhibitory plasticity must act on a faster timescale than excitatory plasticity to maintain stability (*24*). Then the postsynaptic target firing rate is consistently met and average excitatory synaptic weight changes only differ amongst each other due to different average presynaptic firing rates, which prevents the development of stimulus selectivity (Fig. 1*H* & *I*; see SM Sec. 1.2.3 for details).

Synapse-type-specific competition enables balanced principal component analysis. Synapse-type-specific competitive Hebbian learning (Eq. 1, and 2) enables the joint development of stimulus selectivity and balanced input currents. In contrast to classic inhibitory plasticity, under synapse-type-specific competitive Hebbian learning, inhibitory synaptic growth is not stabilized by a target firing rate. Instead, as excitatory synapses, inhibitory synapses compete for a limited supply of synaptic resources that maintain the total amount of synaptic strength. As we did for excitatory synapses (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 $f(u) \propto u$:

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

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

where γ and ρ are scalars that ensure normalization of excitatory and inhibitory weights, respectively. In addition, we defined the modified covariance matrix $\bar{\mathbf{C}}$. Then multiples of the excitatory and the inhibitory part of the eigenvectors of the modified covariance matrix $\bar{\mathbf{C}}$ are fixed points of the weight dynamics (see SM Sec. 2 for details). 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 $\mathbf{C}$ (cf. Eq. 6). This implies that, if excitatory and inhibitory synaptic weights have identical shape, $\mathbf{w}_E \propto \mathbf{w}_I$, equal to a multiple of an eigenvector of $\mathbf{C}$, the system is in a fixed point (Fig. 1*J*), where $\langle \dot{\bar{\mathbf{w}}} \rangle = 0$ (cf. Eq. 5). Neurons become selective for activity along one particular input direction, while excitatory and inhibitory neural inputs are co-tuned, which explains the joint development of stimulus selectivity and E-I balance in feedforward circuits, in agreement with our numerical simulations with non-linear activation functions (Fig. 1*E* & *F*).

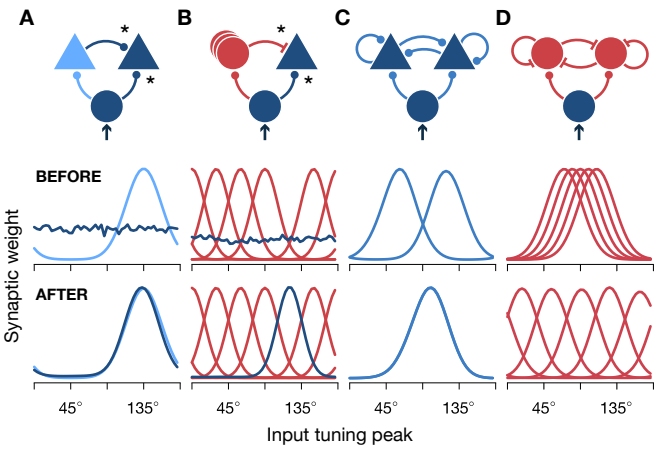

Figure 2: Feedforward tunings are affected by lateral input in microcircuit motifs. (*A*) In addition to feedforward input from a population of orientation tuned excitatory cells (blue circle), a neuron receives lateral input from an excitatory neuron with fixed feedforward tuning (light blue). * indicates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

Lateral inputs shape feedforward weight dynamics. We wanted to understand how fully plastic recurrent networks of excitatory and inhibitory neurons can self-organize into functional circuits. Therefore, 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 simple microcircuit motif (Fig. 2*A*, top). We observed that a downstream neuron becomes preferentially tuned to the feedforward input mode of the lateral projecting neuron (Fig. 2*A*, bottom; cf. SM Sec. 3). Similarly, laterally projecting inhibitory neurons repel downstream neurons from their input modes (Fig. 2*B*). 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. 2*C*). This contradicts experimental observations that brain activity decorrelates over development (*44*, *45*). In line with these results, in our model, interconnected inhibitory neurons repel each other and their tuning curves decorrelate (Fig. 2*D*).

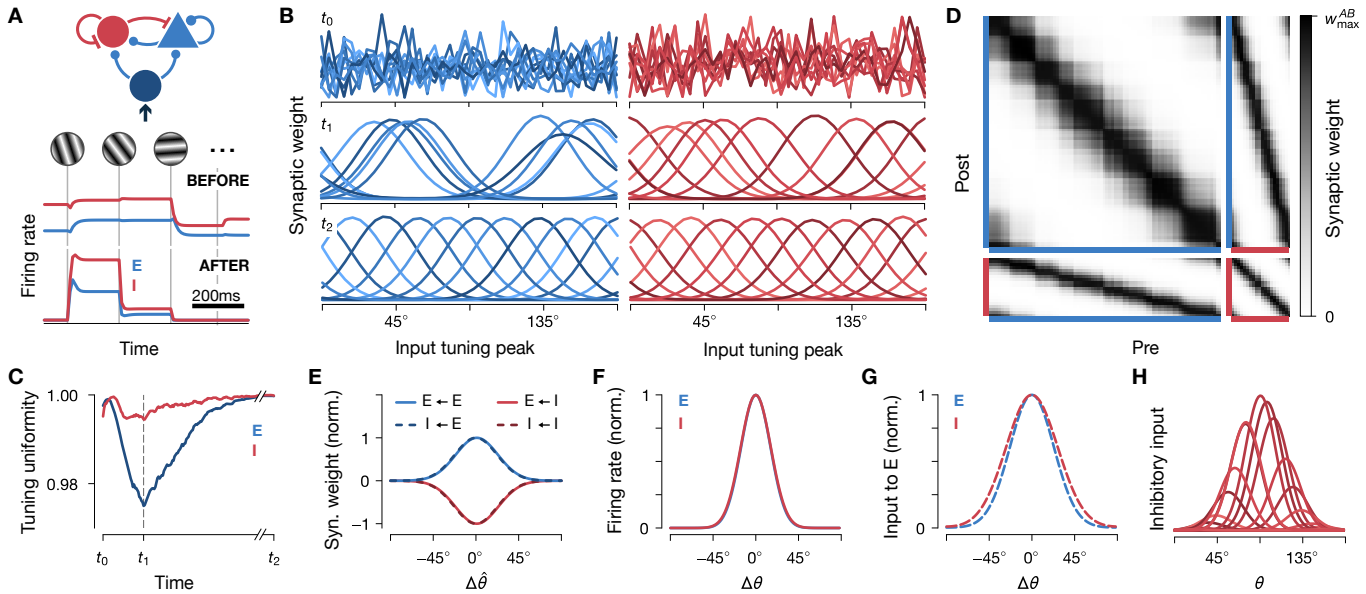

Figure 3: Tuning curve decorrelation in plastic recurrent networks. (*A*) Top: A population of recurrently connected excitatory and inhibitory neurons receives input from a set of input neurons that are tuned to different stimulus orientations (cf. Fig. 1*B*, bottom). Every 200ms a different orientation 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 Methods). Bottom: typical firing rate activity of one excitatory (blue) and one inhibitory (red) neuron before and after training. (*B*) Feedforward tuning curves of $N_E = 10$ excitatory neurons before (t_0, top), during (t_1, center), and after (t_2, bottom) training. Synaptic weights were initialized randomly. Different color shades indicate weights of different postsynaptic neurons. Compare SM Movies M1 & M2. (*C*) Feedforward population tuning uniformity (see Methods) of excitatory and inhibitory neurons in *B*. Time points t_0, t_1, t_2 correspond to time points in *B*. (*D*) Connectivity matrices after training a network of $N_E = 80$ excitatory (blue) and $N_I = 20$ inhibitory (red) neurons. Neurons are sorted according to their preferred orientation $\hat{\theta}$, as measured by their peak response to different oriented gratings. $w^{AB}_{\max}$ is the largest synaptic weight between population *A* and *B*; $A, B \in \{E, I\}$. (*E*) Normalized (norm.) recurrent weight strengths as a function of the difference between preferred orientations of the pre- and postsynaptic neurons, $\Delta\hat{\theta} = \hat{\theta}_{\text{post}} - \hat{\theta}_{\text{pre}}$, averaged over all neuron pairs. Input weights to excitatory (solid) and inhibitory (dashed) neurons overlap. (*F*) 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. Curves for excitatory (blue) and inhibitory (red) neurons overlap. (*G*) Same as in *F*, but for average excitatory and inhibitory inputs to excitatory neurons. (*H*) Inhibitory input to an excitatory neuron with preferred orientation close to 90°. Each curve corresponds to the input from one presynaptic inhibitory neuron for stimuli of different orientations θ.

Tuning curve decorrelation in fully plastic recurrent E-I networks. Recent experimental studies have suggested that inhibitory neurons drive decorrelation of neural activities (*46*, *47*). Hence, we asked whether the interaction between excitatory and inhibitory neurons can serve to decorrelate not only inhibitory but also 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. We presented different oriented gratings in random order to a network where all feedforward and recurrent synapses are plastic (Fig. 3*A*, top). We observed a sharp increase in response selectivity (Fig. 3*A*, bottom) that is reflected in the reconfiguration of feedforward synaptic weights (cf. SM Movies M1 & M2): Shortly after random initialization (Fig. 3*B*, top), excitatory neurons predominantly connect to a subset of input neurons with similar stimulus selectivities (Fig. 3*B*, center left). We quantified the uniformity of the distribution of feedforward tuning curves during training (Fig. 3*C*, see Methods) and found that inhibitory neurons maintained a much wider coverage of the input stimulus space than the excitatory population (cf. Fig. 3*B*, center, t_1). Eventually, tuning curves of excitatory as well as inhibitory neurons decorrelate and cover the whole stimulus space with minimal overlap (Fig. 3*B*, bottom), in sharp contrast to circuits without inhibition, where tuning curves become clustered (cf. Fig. 2*C*). After training, neurons are organized in an assembly-like structure. Neurons that are similarly tuned became more strongly connected (Fig. 3*D* & *E*), as is observed experimentally (*48*–*58*). We found that inhibitory neurons become as selective for stimulus orientations as excitatory neurons (*34*–*37*) (Fig. 3*F*), while excitatory input is balanced by similarly tuned inhibitory input (Fig. 3*G*) from multiple overlapping inhibitory neurons (Fig. 3*H*), in agreement with experimental results (*12*, *59*–*64*); but see (*65*–*70*).

In summary, we find that synapse-type-specific competitive Hebbian learning in fully plastic recurrent networks is sufficient to decorrelate neural activities and leads to preferential connectivity between similarly tuned neurons, as observed in cortical circuits.

Inhibitory neurons balance excitatory attraction and enable decorrelation. To uncover how recurrent inhibition can prevent all neurons from becoming selective for a single input mode, we investigated the fundamental principles of synapse-type-specific competitive Hebbian learning in recurrent networks analytically (SM Sec. 5). In the simplified case of linear activation functions, input modes are eigenvectors of the input covariance matrix (cf. Eq.3). Since these eigenvectors are orthogonal by definition (Fig. 4*A*), the activities of neurons that are tuned to different eigenvectors are uncorrelated, and their reciprocal connections decay to zero under Hebbian plasticity (Fig. 4*B*). Then, neurons that are tuned to the same input mode form recurrent 'eigencircuits' that are otherwise separated from the rest of the network (SM Sec. 4). We characterize

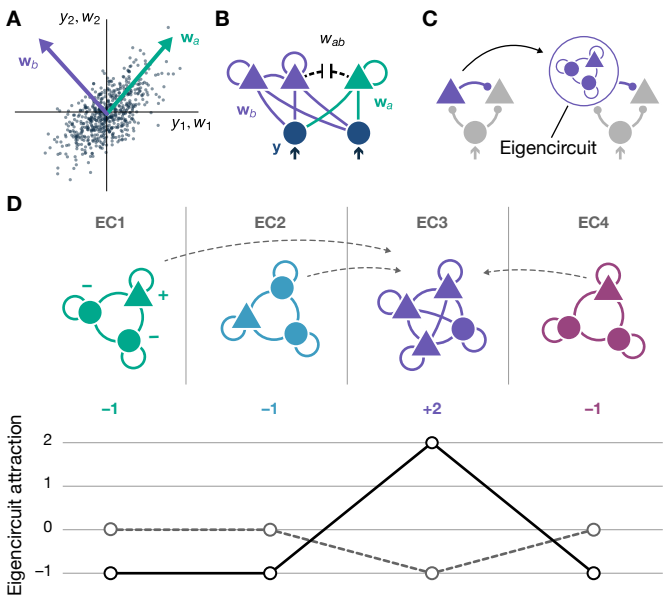

Figure 4: Illustration of eigencircuit decomposition and attraction. (*A*) Feedforward synaptic weight vectors $\mathbf{w}_a$, $\mathbf{w}_b$ of two neurons that are tuned to two different principal components (top, purple and green) of the input data. Each dark blue dot represents one presynaptic firing pattern (cf. Fig. 1*G*). (*B*) Synaptic weights w_{ab} between neurons that are tuned to different eigenvectors decay to zero, while neurons tuned to the same eigenvector form recurrently connected eigencircuits (purple). (*C*) As single, laterally projecting neurons shape the effective attraction of their input mode (left; cf. Fig.2), eigencircuits also increase or decrease the effective attraction of their respective eigenvector direction (right). (*D*) A recurrent network of excitatory (triangles) and inhibitory (circles) neurons that are distributed across four decoupled eigencircuits (EC, top). Each excitatory neuron contributes plus one (+), each inhibitory neuron minus one (-) to the eigencircuit attraction, λ_{eig} (solid line, bottom). Due to synaptic plasticity, neurons are pulled towards the most attractive eigencircuit, EC3 (gray dashed arrows, top). After all neurons integrate into the same eigencircuit (EC3), its attraction becomes negative, while the now unoccupied eigencircuits (EC1, EC2, EC4) are neutral (dashed line, bottom).

a mode's effective attraction as a number such that, if a mode has a higher attraction than a competing mode, then neurons responding to the mode with lower attraction are unstable and shift their tuning towards the mode with higher attraction (see SM for details). Just like single, laterally projecting neurons (SM Sec. 3), eigencircuits also modify the effective attraction of their input mode (Fig. 4*C*). The decomposition of the network into eigencircuits allows to write the effective attraction $\bar{\lambda}$ of each input mode as the sum of a feedforward component λ and the variances of the neurons that reside in the respective eigencircuit (cf. SM Sec. 4.1 & 4.2):

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

where we defined the contribution of recurrently projecting neurons to the effective attraction of an input mode as the eigencircuit attraction, λ_{eig}. Note that, in general, variances $\sigma^2_{E/I}$ depend on the total synaptic weights, and the number of excitatory and inhibitory neurons in the eigencircuit (SM Sec. 4.2). This reveals that the attractive and repulsive effects of excitatory and inhibitory neurons can balance each other. In a simplified example, we assumed that all input modes have equal feedforward attraction, equal to λ, while each excitatory neuron contributes plus one and each inhibitory neuron minus one to the effective attraction (Fig. 4*D*, top). Then the eigencircuit attractions becomes $\lambda_{\text{eig}} = n_E - n_I$ (Fig. 4*D*, bottom, solid line). In this configuration, the network is unstable: All neurons are attracted towards the input mode with the highest effective attraction (EC3), 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 would become repulsive (Fig. 4*D*, bottom, dashed gray line), as each increase in attraction due to an additional excitatory neuron is balanced by a decrease in attraction due to two additional inhibitory neurons. Consequently, the resulting eigencircuit is unstable and neurons are repelled towards non-repulsive, unoccupied input modes; distributed across the stimulus space.

While this example conveys the core principle of how recurrently connected neurons adjust their tunings, the actual dynamics of synaptic weights are more complex (SM Sec. 5). In particular, neurons do not switch their tuning between input modes in discrete steps but shift their tuning gradually. Due to the recurrent nature of the circuit, even small tuning shifts affect the attractions of the respective eigencircuits (cf. SM Sec. 5.2.3). In our simulations, we therefore never observe a full collapse of all tuning curves onto the same input mode before neurons distribute across the stimulus space. Instead, neurons rapidly develop tuned feedforward receptive fields that gradually shift to maximise tuning uniformity, with little to no oscillatory dynamics (Fig. 3*B* & *C* and SM Movies M1 & M2).

In the simplified case of linear activation functions, we derive the following condition that prevents the collapse of all tuning curves onto a single input mode:

$$N_E \sigma_E^2 < N_I \sigma_I^2, \qquad [8]$$

where σ_E^2, σ_I^2 are the average of the variances of the excitatory and inhibitory firing rates, and N_E, N_I are the total number of neurons in the network (cf. SM Sec. 5.2.4). These results show that recruiting recurrent inhibition can prevent tuning curve collapse and enables decorrelation, where a lower number of inhibitory neurons can be compensated by an increase in neural activation.

Plastic recurrent E-I networks perform response normalization and exhibit winner-takes-all dynamics. 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 non-linear network computations. For example, the firing rate response of neurons in the visual cortex to multiple overlayed oriented gratings is normalized in a non-linear fashion (*71*, *72*). While this form of normalization is mostly of thalamic origin (*73*–*75*), there is most likely also a cortical component(*72*, *76*). A recently introduced E-I network model with static, hand-crafted connectivity can explain these modulations (*18*, *77*). We explored if the recurrent connectivity can instead be learned from a network's input stimulus statistics. We consider a circuit with fixed feedforward tuning and plastic recurrent connectivity (Fig. 5*A*). After training the network with single oriented grating stimuli (Fig. 5*A*, bottom), 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. 5*B*,

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. 5*B*, right). As observed experimentally (*71*, *72*, *78*), we found that this orientation-specific response normalization is divisive and shifts the log-scale contrast-response function to the right (Fig. 5*C*).

Sensory input statistics shape computational functions of recurrent circuits. We next investigated how the stimulus statistics during training affect receptive field properties. We considered a plastic network where two neural populations receive tuned input from either a center or a surround region of the visual field (Fig. 5*D*). During training, we presented either the same oriented grating in both regions (Fig. 5*E*, top, purple), or a single grating in just one region (Fig. 5*E*, bottom, red), at 50% contrast (cf. Table 1). These stimulus statistics heavily influenced the recurrent connectivity structure in the network. When identical oriented stimuli are presented to the center and surround regions during training, neurons with similar orientation tuning become most strongly connected, independent from which region the neurons receive their feedforward input (Fig. 5*F*). However, when the center and surround regions are stimulated separately during training, neurons only connect to similarly tuned neurons within the same region and cross-region connectivity decays to zero (Fig. 5*G*). These differences in the recurrent connectivity structure are also reflected in the networks' response properties. We found that after training, the response of center-tuned neurons exhibits orientation-specific surround suppression, reflecting the stimulus statistics during training. When the center and the surround regions are stimulated separately during training, iso- or cross-oriented stimuli in the surround both elicit minimal suppression of the center-tuned population's response to a center stimulus (Fig. 5*H* & *I*, red). In contrast, in the case of correlated stimulation of the center and surround regions during training, the response of the center population is markedly suppressed when an additional surround stimulus is presented (Fig. 5*H* & *I*, purple). Importantly, suppression is stronger for iso- compared to cross-orientations (Fig. 5*I*, solid and dashed lines), as has been reported experimentally (*79*–*82*). We further investigated the lateral interactions between neurons tuned to the center and surround regions by presenting an oriented stimulus only in the surround region, while observing the total excitatory and inhibitory inputs to excitatory neurons (Fig. 5*J*). We found that the total excitatory input to stimulated excitatory neurons in the surround was larger than the total inhibitory input (Fig. 5*J*, right column). When center and surround neurons were stimulated together during training, both center and surround received similar, balanced E and I recurrent input, but the surround cells also received feedforward excitation, yielding more total excitation (Fig. 5*J*, top, purple). When center neurons were not stimulated with the surround neurons during training, they received no input from a surround-only stimulus (Fig. 5*J*, bottom, red). In the case of correlated stimulation of the center and surround regions during training, this lateral input was orientation-specific. Center neurons tuned to the same orientations as stimulated neurons in the surround received stronger input than center neurons tuned to different orientations (Fig. 5*J*, top left), reflecting the input stimulus statistics during training (Fig. 5*E*) and the resulting recurrent connectivity (Fig. 5*F*). A similar balance of excitatory and inhibitory lateral inputs has previously been observed in barrel cortex (*83*). Together, this demonstrates that synapse-type-specific competitive Hebbian learning produces extra-classical receptive fields that modulate feedforward responses via recurrent interactions that reflect the input statistics during training.

Discussion

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 circuits emerge from a single unsupervised, biologically plausible learning paradigm that acts simultaneously at all synapses. In a single framework, our networks 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. These 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 (*84*–*88*). Unlike previous models (*89*–*94*), our networks are composed of excitatory and inhibitory neurons with fully plastic recurrent connectivity.

Early theoretical work on inhibitory plasticity assumed that synapses evolve to maintain the mean firing rate of postsynaptic excitatory neurons (*23*). When excitatory input is static, this leads to neural tunings where inhibition and excitation are balanced. However, when excitatory are simultaneously plastic according to a simple Hebbian rule, the circuit is unstable and can not explain the joint development of feedforward stimulus tuning and inhibitory balance (*24*) (SM Sec. 1.2.3). The system can be stabilized when the Hebbian growth of excitatory synapses is controlled by a BCM-like plasticity threshold. This introduces fierce competition between different input streams in the form of subtractive weight normalization, which leads to winner-takes-all dynamics among synapses that do not allow for the development of extended receptive fields (*24*, *31*, *95*). Later models have proposed more intricate plasticity rules, some of which consider, e.g., voltages or currents, in addition to pre- and postsynaptic action potentials (*25*, *28*, *96*–*102*), as summarized in several recent reviews (*14*, *103*–*106*). In recent years, there has also been a resurgence of interest in normative approaches (*28*, *29*, *107*). In these approaches, it is postulated that synaptic plasticity rules act to optimize an objective function that describes a desirable network property. Motivated by the notorious instability of recurrent networks, one obvious objective is stability, e.g., in the form of firing rate homeostasis. Following early theoretical work that suggested such a homeostatic role for synaptic plasticity of inhibitory synapses onto excitatory neurons(*23*), two recent studies propose a similar role for the plasticity of other recurrent synapse types (*28*, *29*). Indeed, such plasticity rules allow the formation of inhibition balanced receptive fields (*28*), and stabilize network activity, even when faced with strong recurrent connections (*29*). However, none of these rules have been applied in fully plastic recurrent networks with structured feedforward input. Even in complex models that use many different forms of plasticity, some synapse types are kept static after initialization, to maintain stable network activity (*23*, *26*, *27*, *101*). While such networks still show many interesting dynamics, they lack the rich computational functions of circuits with structured connectivity between all neuron types (*18*, *77*). In contrast, our learning

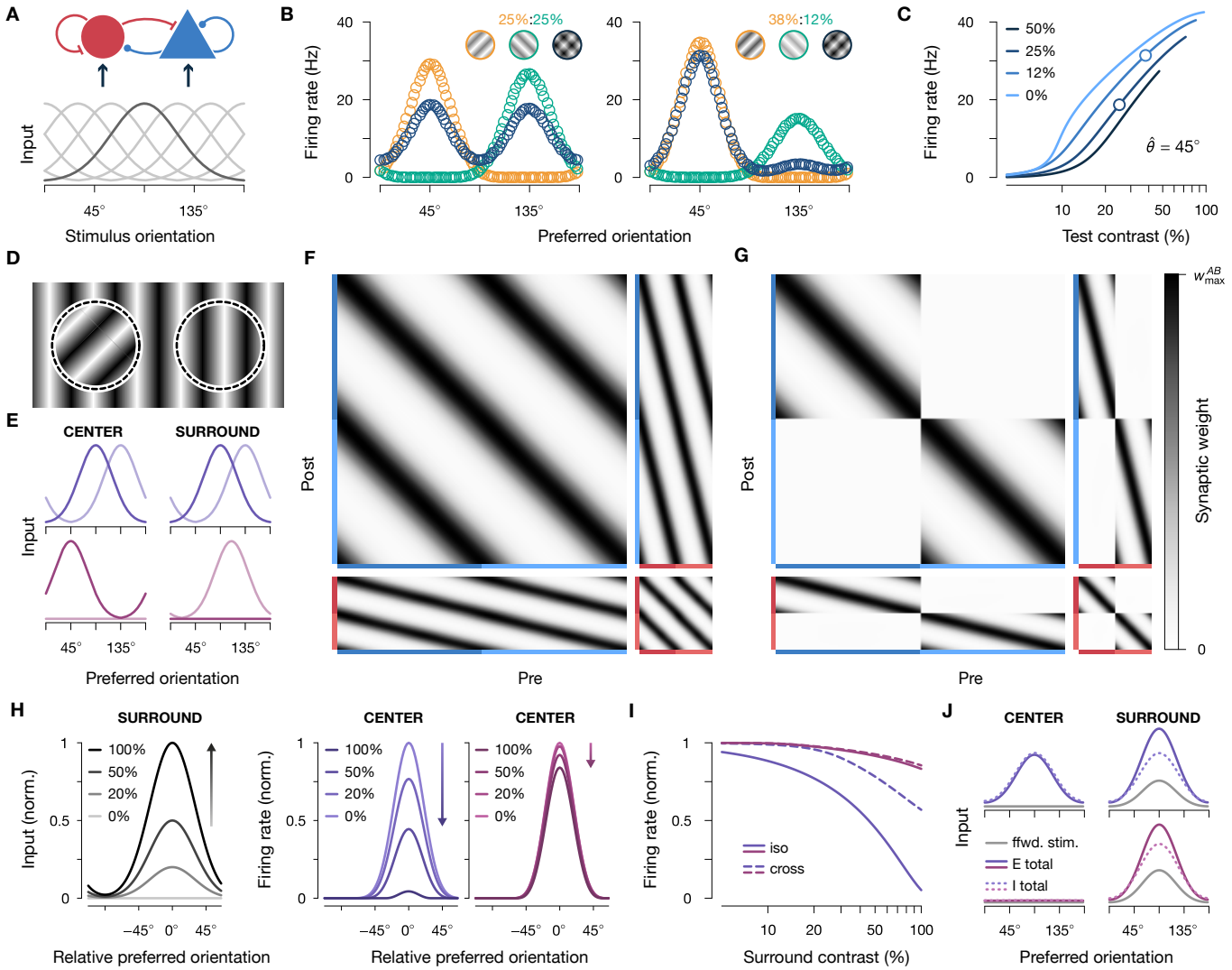

Figure 5: Cross-orientation and surround suppression in trained neural networks. (*A*) A plastic recurrent network of excitatory and inhibitory neurons (top) receives input according to fixed feedforward tuning curves (bottom). Input amplitudes were modulated with stimulus contrast. Tuning curve of neurons 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 (insets, grating contrasts increased for better visibility). Gratings are presented separately (orange & green) or together (dark blue). Each open circle corresponds to the response of one excitatory neuron. (*C*) Contrast response curve of a single excitatory neuron with preferred orientation $\hat{\theta}$ = 45° to the test and mask gratings in *B*. Different mask contrasts are indicated by different color shades. The bottom/top circles correspond to the left/right contrast level configurations in *B*. (*D*) Center (left) and surround region (right) with different oriented stimuli. (*E*) Example stimuli during training with different stimulus statistics. Top: Neurons tuned to the same orientation, but different regions (center region, left; or surround region, right) receive identical input; two example stimuli are shown in solid and transparent purple, respectively. Bottom: Neurons tuned to the center and surround regions are stimulated separately; two example stimuli are shown in solid and transparent red, respectively. Either the surround or the center regions are stimulated, while the other region receives zero input. (*F*) Recurrent connectivity matrix between excitatory (blue) and inhibitory (red) neurons (cf. Fig. 3*D*) after training the network with correlated center and surround stimuli (corresponds to purple color in *E*, top). Neurons are sorted according to their feedforward orientation tuning. Color shades indicate tuning to the center (dark) or surround (light) region. (*G*) Same as in *F*, but for a network trained with single gratings that were presented either in the center or the surround region (corresponds to red color in *E*, bottom). (*H*) Suppression of excitatory population activity in response to increasing surround stimulation for two networks trained under different stimulus statistics. Left: neurons tuned to the center region are stimulated by an oriented grating of constant, 100% contrast (not shown) while neurons tuned to the surround region are stimulated with an oriented grating of increasing contrast (shades; compare insets). Identical stimulation protocol for both training statistics. Center and right: the activity of excitatory neurons that are tuned to the center region is suppressed with increasing surround contrast. The magnitude of suppression depends on the stimulus statistics during training (purple vs. red, colors as in *E*). (*I*) 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- (dashed) or iso-oriented (solid) surround stimulus changed. Colors indicate different stimulus statistics during training (as in *E*). (*J*) Total excitatory (solid) and inhibitory (dotted) input to excitatory neurons during stimulation of only the surround region with an oriented grating of 90°. Excitatory input due to feedforward stimulation (ffwd. stim.) is shown in light gray. Colors (top vs. bottom) indicate different input statistics during training (as in *E*).

rule is minimalistic and only relies on general Hebbian synaptic growth that is stabilized by competitive interactions. Importantly, our theory does not depend on a specific biophysical implementation of the Hebbian plasticity paradigm. We only require that synapses follow the basic Hebbian principle of synaptic strengthening following concurrent pre- and post-

synaptic activity. In the past, competitive Hebbian learning has been investigated theoretically for excitatory synaptic inputs to single neurons (30, 31, 39, 108, 109), but not for inhibitory inputs or in recurrent networks. Our analysis demonstrates that competitive Hebbian plasticity is a suitable learning mechanism for networks of recurrently connected excitatory and inhibitory neurons, while being analytically tractable and biologically plausible.

Competitive interactions between synapses have been observed in many different preparations and have been attributed to various mechanisms (110–121). While previous work has focused on competitive interactions between excitatory synapses, our results support the notion that similar competitive processes are also active at inhibitory synapses (32, 122). The local competition for a limited supply of synaptic building blocks is a biologically plausible normalization mechanism (33, 115, 120, 123, 124). Many synaptic proteins are specific to inhibitory or excitatory synapses and reside in one synapse-type, but not the other (125, 126). Therefore, in this work, we assume a synapse-type-specific competition for different synaptic resource pools and implement 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 (126–129). In principle, this allows for the precise regulation of different input pathways via the adjustment of subtype-specific resource pools (130–136). Furthermore, axons of different neuron subtypes target spatially separated regions on the dendritic tree, allowing for pathway-specific local competition. For example, somatostatin-positive cortical Martinotti cells target the apical dendritic tree of pyramidal cells, while parvalbumin-positive basket cells form synapses closer to the soma (1), which suggests that afferents of these cell types compete for separate resources pools. We anticipate such subtype-specific mechanisms to be crucial for the functional development of any network with multiple neuron subtypes (137, 138).

In the brain, total synaptic strengths are dynamic and homeostatically regulated on a timescale of hours to days (139–142). 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 (143). However, theoretical models suggest that homeostatic scaling is too slow to balance rapid synaptic plasticity (144). In our networks, Hebbian growth is instead thought to be stabilized by the competition for a limited pool of synapse-type-specific resources, while total synaptic strengths remain fixed. This competition is fast due to rapid interactions on a molecular level (33, 120). Compared to Hebbian growth, infinitely fast, as a synapse can only grow at the expense of another. Therefore, we suggest that homeostatic scaling of total synaptic strengths is not required for immediate network stability but instead controls the operating regime of the network (16, 77, 145).

Our results demonstrate how multi-synaptic, inhibitory interactions can decorrelate excitatory neurons. In contrast, inhibitory neurons can inhibit each other mono-synaptically and do not require additional recurrent interactions for decorrelation. Accordingly, we observe that during training, inhibitory neurons are more decorrelated compared to excitatory neurons (Fig. 3*C*). These insights complement recent experimental results that suggest an instrumental role of inhibition in the decorrelation of excitatory networks in mouse prefrontal cortex during early development (47). Recent experimental studies in ferret visual cortex report conflicting evidence — either supporting (46) or contradicting (146) aligned developmental trajectories of excitatory and inhibitory populations. In our simulations, we observe similar developmental trajectories for excitatory and inhibitory populations. However, we focused on synaptic plasticity and did not consider other processes, like critical periods (147, 148), that are known to shape circuit development.

Cortical computations rely on strong recurrent synaptic weights that result in neural activities that can deviate significantly from the input stimulus pattern (15, 16, 18) (cf. Fig. 5*B*, left, combined grating response). Such a decoupling of network activity from feedforward input due to recurrent interactions can lead to neural tunings that do not reflect the input stimulus statistics (cf. SM Sec. 3). In our theory (SM Sec. 4), we assume that neurons are tuned to feedforward modes and thereby implicitly assume that network activity is dominated by feedforward input. In our numerical simulations of fully plastic recurrent networks, we find that for intermediate levels of recurrence (cf. Table 1, Fig. 1, 2 & 3), the network's activities are indeed dominated by feedforward inputs. In case of strong recurrence (Fig. 5), we ensure feedforward dominance by presenting single oriented gratings that match the fixed feedforward tunings of neurons (cf. Fig. 5). Such gratings elicit a Gaussian-shaped response that is sharpened due to the recurrent connectivity, but maintains the general correlation structure compared to purely feedforward-driven networks (compare tuning widths in Fig. 5*A*, bottom, and *B*, single grating response). Biological cortical networks are strongly recurrently connected (149–153). However, neural activity and the induction and polarity of synaptic plasticity are regulated by neuromodulators (154–158), which may control the destabilizing effect of strong recurrent connectivity. In addition, different synapse types do not develop simultaneously but progress through different developmental stages (137, 159, 160). For example, the development of recurrent excitatory connections is delayed compared to that of feedforward synapses (131, 161). Taking these factors into account will be essential for future models of developing recurrent circuits.

In our networks, structured feedforward input is crucial for the development of orientation selective receptive fields. However, already at the time before eye opening cortical neurons exhibit substantial selectivity for stimulus orientation, without having been exposed to the statistical regularities of visual inputs (162–164). One hypothesis is that, instead, spontaneous activity in the retina provides the statistical structure required for the initial development of orientation selectivity (165–167). In our model, circuit formation depends only on the statistical regularities between input streams and is agnostic with respect 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.

Materials and Methods

Computational model. 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

$$\tau_A \dot{\mathbf{u}}_A = -\mathbf{u}_A + \mathbf{W}_{AF}\mathbf{r}_F + \mathbf{W}_{AE}\mathbf{r}_E - \mathbf{W}_{AI}\mathbf{r}_I, \quad A \in \{E, I\}, \qquad [9]$$

where τ_A is the activity timescale. $\mathbf{W}_{AB}$ are matrices that hold synaptic weights between the presynaptic population B and the postsynaptic population A with $B \in \{E, I, F\}$. All differential equations were numer-

ically integrated using the Euler method in timesteps of Δt. Entries of weight matrices were drawn from a normal distribution with mean μ_W equal to two times the standard deviation σ_W, which yields mainly positive entries. Negative entries were set to their absolute value. Before the start of the simulation, excitatory and inhibitory weights were normalized as described below. Unless stated otherwise, prior to normalization, all *recurrent* excitatory and inhibitory weights were set to zero, i.e., initialized networks were purely feedforward. Firing rate vectors $\mathbf{r}_A$ are given as a function $f(\mathbf{u}_A)$ of the membrane potential $\mathbf{u}_A$:

$$\mathbf{r}_A = f(\mathbf{u}_A), \quad f(\mathbf{u}_A) = a[\mathbf{u}_A - b]_+^n, \quad A \in \{E, I\} \qquad [10]$$

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

Plasticity and normalization. Plastic weights evolve according to a Hebbian plasticity rule

$$\dot{\mathbf{W}}_{AB} = \epsilon_{AB}\mathbf{r}_A\mathbf{r}_B^{\mathsf{T}}, \quad A \in \{E, I\}, \quad B \in \{E, I, F\} \qquad [11]$$

where ϵ_{AB} is a scalar learning rate, and ^{T} indicates the transpose. After each plasticity step, synaptic weights are normalized such that the total excitatory and inhibitory postsynaptic weights are maintained:

$$w_{AB}^{(ij)} \leftarrow W_{AE}\frac{w_{AB}^{(ij)}}{\sum_j w_{AE}^{(ij)} + \sum_k w_{AF}^{(ik)}}, \qquad [12]$$

$$w_{AI}^{(ij)} \leftarrow W_{AI}\frac{w_{AI}^{(ij)}}{\sum_j w_{AI}^{(ij)}}, \quad A \in \{E, I\}, \quad B \in \{E, F\}, \qquad [13]$$

where W_{AE}, W_{AI} are the total excitatory and inhibitory synaptic weight norms. Weights are updated and normalized in every integration timestep Δt, in sync with the network dynamics.

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 *H* & *I* of Fig. 1, inhibitory weights evolved according to the classic inhibitory plasticity rule (*23*) without normalization:

$$\dot{\mathbf{w}}_{EI} = \epsilon_{EI}(r_E - r_0)\mathbf{r}_I, \qquad [14]$$

where r_0 is a target firing rate.

Input model. The activity of feedforward input neurons depends on the orientation θ and contrast c of an input grating:

$$\mathbf{r}_F = cA_F \exp\left(\frac{|\theta, \theta_F|^2}{2\sigma_F^2}\right), \qquad [15]$$

where the vector θ_F holds the preferred orientations of the input neurons that are evenly distributed between 0 and 180°, σ_F is the tuning width, A_F the maximum firing rate, and $|\cdot,\cdot|$ is the angular distance , i.e., the shortest distance around a circle of circumference 180°. 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

$$\mathbf{W}_{AF} = \exp\left(\frac{|\hat{\theta}, \theta_F|^2}{2\sigma_\theta^2}\right). \qquad [16]$$

For Fig. 5, feedforward weights are normalized separately to W_{AF} before the start of the simulations (cf. Table 1). In this case, feedforward weights are fixed and are not taken into account when normalizing recurrent weights. Feedforward weights of static neurons in Fig. 2*A* & *B* are processed in the same fashion. For Fig. 5, parameters were selected to result in stimulation patterns as in Rubin *et al.* (*18*). Weight norms W_{AB} were also adapted from Rubin *et al.* (*18*). See Table 1 for an overview of used simulation parameters.

Tuning curve uniformity measure. In Fig. 3*C*, we quantified the uniformity of the distribution of tuning curves during learning and defined:

$$p_j^A = \sum_i w_{AF}^{(ij)} / \sum_{ij} w_{AF}^{(ij)}, \quad A \in \{E, I\}, \qquad [17]$$

where p_j^A is the normalized total synaptic output weight of input neuron j onto the excitatory (E) and inhibitory (I) neural population. Then $\sum_j p_j^A = 1$, and we can define the tuning uniformity U_A as the normalized Shannon entropy H_{p^A}.

$$U_A = \mathrm{H}_{p^A}/\log(N_F) = -\sum_j p_j^A log(p_j^A)/\log(N_F), \quad A \in \{E, I\}. \qquad [18]$$

U_A is maximal, equal to one, if p^A is uniformly distributed, and minimal, equal to zero, if all synaptic weight is concentrated in a single input neuron. Such a concentration is highly unlikely. In our simulations, weight distributions are much closer to a uniform distribution, and the uniformity measure is close to one.

Acknowledgements

We thank the members of the Computation in Neural Circuits group for helpful feedback throughout the project. We thank Dylan Festa for additional comments and for proofreading the manuscript. This work was supported by the Max Planck Society, the Engineering and Physical Sciences Research Council (DTP grant EP/T517847/1 to E.J.Y), and the European Research Council (StG 804824 to J.G.).).

Author contributions

SE conceived research with input from JG. SE performed simulations, derived mathematical results, prepared figures, and wrote the Supplementary Material and the first draft of the manuscript. EJY derived mathematical results and contributed to the Supplementary Material. SE and JG wrote the manuscript.

Supplementary material

Supplementary material includes: SM Text Sections 1 to 6, Figures S1 to S4, Movies M1 and M2. Python code to reproduce the key results of the study is publicly available on GitHub (*168*).

References

1. R. Tremblay *et al.*, GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron* (2016).
2. R. Hattori *et al.*, Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nature neuroscience* (2017).
3. K. A. Pelkey *et al.*, Hippocampal GABAergic inhibitory interneurons. *Physiological reviews* (2017).
4. A. Kepecs, G. Fishell, Interneuron cell types are fit to function. *Nature* (2014).
5. M. Carandini, D. J. Heeger, Normalization as a canonical neural computation. *Nature Reviews Neuroscience* (2012).
6. A. Angelucci *et al.*, Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience* (2017).
7. K. C. Wood *et al.*, Cortical inhibitory interneurons control sensory processing. *Current opinion in neurobiology* (2017).
8. G. B. Keller, T. D. Mrsic-Flogel, Predictive processing: a canonical cortical computation. *Neuron* (2018).
9. O. K. Swanson, A. Maffei, From hiring to firing: activation of inhibitory neurons and their recruitment in behavior. *Frontiers in molecular neuroscience* (2019).
10. K. A. Ferguson, J. A. Cardin, Mechanisms underlying gain modulation in the cortex. *Nature Reviews Neuroscience* (2020).
11. C. Van Vreeswijk, H. Sompolinsky, Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* (1996).
12. J. S. Isaacson, M. Scanziani, How inhibition shapes cortical activity. *Neuron* (2011).
13. S. Denève, C. K. Machens, Efficient codes and balanced networks. *Nature neuroscience* (2016).
14. G. Hennequin *et al.*, Inhibitory plasticity: balance, control, and codependence. *Annual review of neuroscience* (2017).
15. S. Sadeh, C. Clopath, Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience* (2021).
16. Y. Ahmadian, K. D. Miller, What is the dynamical regime of cerebral cortex? *Neuron* (2021).
17. H. Ozeki *et al.*, Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* (2009).
18. D. B. Rubin *et al.*, The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* (2015).
19. G. Hennequin *et al.*, The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron* (2018).

Figure	1*E-F*, *H-I*	2*A*, *B*, *C*, *D* (*A*, *B*)	3*A-C*, *D-H*	5*A-C*, *D-J*
N_E	1	2, 1, 2, 0	10, 80	80, 80×2
N_I	10	0, 5, 0, 5	10, 20	20, 20×2
N_F	10	50	40, 80	80, 80×2
a	1	0.04 (0.2, 0.08)	0.04	0.04
b	0.25	0	0	0
n	2	2	2	2
μ_W	0.1	0.1	0.2	0.2
σ_W	0.05	0.01	0.1	0.1
W_{EE}	10	1, 1, 4, -	2, 0.6	3.51
W_{IE}	-	-	2, 0.85	3.35
W_{EI}	5, -	-, 0.5, -, -	0.8, 0.3	1.84
W_{II}	-	-, -, -, 0.5	0.5, 0.35	1.44
W_{EF}	-	(0.9, -)	-	$1.4^{\ddagger}$
W_{IF}	-	(-, 1)	-	$1.4^{\ddagger}$
c	1	1	1	0.5
A_F	1	1	35, 140	80
σ_F	20°	12°	12°	$30^\circ/\sqrt{2}$
σ_θ	-	-, -, 22°, 22° (22°, 16°)	-	$30^\circ/\sqrt{2}$
Δt	200ms	20ms	10ms	10ms
τ_E	200ms	40ms	20ms	25ms
τ_I	-	28ms	17ms	12.5ms
ϵ_{EE}	-	$0.4 \times 10^{-8}\text{ms}^{-1}$	$2 \times 10^{-9}\text{ms}^{-1}$, $1.0 \times 10^{-10}\text{ms}^{-1}$	$1.0 \times 10^{-9}\text{ms}^{-1}$
ϵ_{IE}	-	$0.8 \times 10^{-8}\text{ms}^{-1}$	$3 \times 10^{-9}\text{ms}^{-1}$, $1.5 \times 10^{-10}\text{ms}^{-1}$	$1.5 \times 10^{-9}\text{ms}^{-1}$
ϵ_{EI}	$2 \times 10^{-4}\text{ms}^{-1}$, $4 \times 10^{-4}\text{ms}^{-1}$	$0.6 \times 10^{-8}\text{ms}^{-1}$	$4 \times 10^{-9}\text{ms}^{-1}$, $2.0 \times 10^{-10}\text{ms}^{-1}$	$2.0 \times 10^{-9}\text{ms}^{-1}$
ϵ_{II}	-	$1.0 \times 10^{-8}\text{ms}^{-1}$	$5 \times 10^{-9}\text{ms}^{-1}$, $2.5 \times 10^{-10}\text{ms}^{-1}$	$2.5 \times 10^{-9}\text{ms}^{-1}$
ϵ_{EF}	$1 \times 10^{-4}\text{ms}^{-1}$, $2 \times 10^{-4}\text{ms}^{-1}$	ϵ_{EE}	ϵ_{EE}	-
ϵ_{IF}	-	ϵ_{IE}	ϵ_{IE}	-
r_0	-, 0.25	-	-	-

Table 1: Simulation parameters. Different parameters for different panels are separated by commas. Dashes indicate that parameters were not used in the simulation. For Fig. 5, weight norms of static synaptic weights are indicated by '‡'. For Fig. 2, parameters of static, laterally projecting neurons are given in brackets (comma separated for panels *A*, *B*).

20. R. Echeveste *et al.*, Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature neuroscience* (2020).
21. W. Soo, M. Lengyel, Training stochastic stabilized supralinear networks by dynamics-neutral growth. *Advances in Neural Information Processing Systems* (2022).
22. Y. Luz, M. Shamir, Balancing feed-forward excitation and inhibition via Hebbian inhibitory synaptic plasticity. *PLoS computational biology* (2012).
23. T. P. Vogels *et al.*, Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* (2011).
24. C. Clopath *et al.*, Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. *BioRxiv* (2016).
25. P. D. King *et al.*, Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *Journal of Neuroscience* (2013).
26. A. Litwin-Kumar, B. Doiron, Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature communications* (2014).
27. F. Zenke *et al.*, Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature communications* (2015).
28. O. Mackwood *et al.*, Learning excitatory-inhibitory neuronal assemblies in recurrent networks. *Elife* (2021).
29. S. Soldado-Magraner *et al.*, Paradoxical self-sustained dynamics emerge from orchestrated excitatory and inhibitory homeostatic plasticity rules. *Proceedings of the National Academy of Sciences* (2022).
30. E. Oja, Simplified neuron model as a principal component analyzer. *Journal of mathematical biology* (1982).
31. K. D. Miller, D. J. MacKay, The role of constraints in Hebbian learning. *Neural computation* (1994).
32. J. N. Bourne, K. M. Harris, Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus* (2011).
33. J. Triesch *et al.*, Competition for synaptic building blocks shapes synaptic plasticity. *Elife* (2018).
34. J. A. Hirsch *et al.*, Functionally distinct inhibitory neurons at the first stage of visual cortical processing. *Nature neuroscience* (2003).
35. J. A. Cardin *et al.*, Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex. *Journal of Neuroscience* (2007).
36. C. A. Runyan *et al.*, Response features of parvalbumin-expressing interneurons suggest precise roles for subtypes of inhibition in visual cortex. *Neuron* (2010).

37. A. K. Moore, M. Wehr, Parvalbumin-expressing inhibitory interneurons in auditory cortex are well-tuned for frequency. *Journal of neuroscience* (2013).
38. E. Oja, Learning in non-linear constrained Hebbian networks. *Proceedings of the ICANN'91, 1991* (1991).
39. G. Ocker, M. Buice, Tensor decompositions of higher-order correlations by nonlinear Hebbian plasticity. *Advances in Neural Information Processing Systems* (2021).
40. A. J. Bell, T. J. Sejnowski, The "independent components" of natural scenes are edge filters. *Vision research* (1997).
41. C. S. Brito, W. Gerstner, Nonlinear Hebbian learning as a unifying principle in receptive field formation. *PLoS computational biology* (2016).
42. J. A. D'Amour, R. C. Froemke, Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* (2015).
43. F. Lagzi *et al.*, Assembly formation is stabilized by Parvalbumin neurons and accelerated by Somatostatin neurons. *bioRxiv* (2021).
44. P. Golshani *et al.*, Internally mediated developmental desynchronization of neocortical network activity. *Journal of Neuroscience* (2009).
45. N. L. Rochefort *et al.*, Sparsification of neuronal activity in the visual cortex at eye-opening. *Proceedings of the National Academy of Sciences* (2009).
46. H. N. Mulholland *et al.*, Tightly coupled inhibitory and excitatory functional networks in the developing primary visual cortex. *Elife* (2021).
47. M. Chini *et al.*, An increase of inhibition drives the developmental decorrelation of neural activity. *ELife* (2022).
48. C. D. Gilbert, T. N. Wiesel, Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience* (1989).
49. Y. Yoshimura *et al.*, Excitatory cortical neurons form fine-scale functional networks. *Nature* (2005).
50. Y. Yoshimura, E. M. Callaway, Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nature neuroscience* (2005).
51. D. E. Wilson *et al.*, GABAergic neurons in ferret visual cortex participate in functionally specific networks. *Neuron* (2017).
52. H. Ko *et al.*, Functional specificity of local synaptic connections in neocortical networks. *Nature* (2011).
53. A. D. Lien, M. Scanziani, Tuned thalamic excitation is amplified by visual cortical circuits. *Nature neuroscience* (2013).
54. Y.-t. Li *et al.*, Linear transformation of thalamocortical input by intracortical excitation. *Nature neuroscience* (2013).
55. L.-y. Li *et al.*, Intracortical multiplication of thalamocortical signals in mouse auditory cortex. *Nature neuroscience* (2013).
56. L. Cossell *et al.*, Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* (2015).
57. M. F. Iacaruso *et al.*, Synaptic organization of visual space in primary visual cortex. *Nature*, issn: 14764687 (2017).
58. P. Znamenskiy *et al.*, Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *Biorxiv* (2018).
59. T. W. Troyer *et al.*, Contrast-invariant orientation tuning in cat visual cortex: thalamocortical input tuning and correlation-based intracortical connectivity. *Journal of Neuroscience* (1998).
60. J. S. Anderson *et al.*, Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of neurophysiology* (2000).
61. L. M. Martinez *et al.*, Laminar processing of stimulus orientation in cat visual cortex. *The Journal of physiology* (2002).
62. J. Mariño *et al.*, Invariant computations in local cortical networks with balanced excitation and inhibition. *Nature neuroscience* (2005).
63. A. Y. Tan *et al.*, Orientation selectivity of synaptic input to neurons in mouse and cat primary visual cortex. *Journal of Neuroscience* (2011).
64. D. E. Wilson *et al.*, Differential tuning of excitation and inhibition shapes direction selectivity in ferret visual cortex. *Nature* (2018).
65. D. Rose, C. Blakemore, Effects of bicuculline on functions of inhibition in visual cortex. *Nature* (1974).
66. X. Pei *et al.*, Receptive field analysis and orientation selectivity of postsynaptic potentials of simple cells in cat visual cortex. *Journal of Neuroscience* (1994).
67. C. Monier *et al.*, Orientation and direction selectivity of synaptic inputs in visual cortical neurons: a diversity of combinations produces spike tuning. *Neuron* (2003).
68. G. K. Wu *et al.*, Lateral sharpening of cortical frequency tuning by approximately balanced inhibition. *Neuron* (2008).
69. B.-h. Liu *et al.*, Broad inhibition sharpens orientation selectivity by expanding input dynamic range in mouse simple cells. *Neuron* (2011).
70. Y.-t. Li *et al.*, Synaptic basis for differential orientation selectivity between complex and simple cells in mouse visual cortex. *Journal of Neuroscience* (2015).
71. L. Busse *et al.*, Representation of concurrent stimuli by population activity in visual cortex. *Neuron* (2009).
72. S. P. MacEvoy *et al.*, A precise form of divisive suppression supports population coding in the primary visual cortex. *Nature Neuro.* (2009).
73. B. Li *et al.*, Origins of cross-orientation suppression in the visual cortex. *Journal of Neurophysiology* (2006).
74. N. J. Priebe, D. Ferster, Mechanisms underlying cross-orientation suppression in cat visual cortex. *Nature neuroscience* (2006).
75. D. Barbera *et al.*, Feedforward mechanisms of cross-orientation interactions in mouse V1. *Neuron* (2022).
76. F. Sengpiel, V. Vorobyov, Intracortical origins of interocular suppression in the visual cortex. *Journal of Neuroscience* (2005).
77. Y. Ahmadian *et al.*, Analysis of the stabilized supralinear network. *Neural Comp.* (2013).
78. T. C. Freeman *et al.*, Suppression without inhibition in visual cortex. *Neuron* (2002).
79. C. Blakemore, E. A. Tobin, Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental brain research* (1972).
80. J. J. Knierim, D. C. Van Essen, Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of neurophysiology* (1992).
81. J. R. Cavanaugh *et al.*, Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of neurophysiology* (2002).
82. B. S. Webb *et al.*, Early and late mechanisms of surround suppression in striate cortex of macaque. *Journal of Neuroscience* (2005).
83. H. Adesnik, M. Scanziani, Lateral competition for cortical space by layer-specific horizontal circuits. *Nature* (2010).
84. M. Pecka *et al.*, Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron* (2014).
85. S. V. David *et al.*, Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience* (2004).
86. G. Felsen *et al.*, Cortical sensitivity to visual features in natural scenes. *PLoS biology* (2005).
87. G. Felsen *et al.*, Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. *Network: Computation in Neural Systems* (2005).
88. E. Froudarakis *et al.*, Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature neuroscience* (2014).
89. O. Schwartz, E. P. Simoncelli, Natural signal statistics and sensory gain control. *Nature neuroscience* (2001).
90. P. Berkes *et al.*, Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* (2011).
91. M. Zhu, C. J. Rozell, Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLoS computational biology* (2013).
92. M. F. Burg *et al.*, Learning divisive normalization in primary visual cortex. *PLOS Computational Biology* (2021).
93. V. Veerabadran *et al.*, presented at the SVRHM 2021 Workshop @ NeurIPS.
94. J. Fu *et al.*, Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv* (2023).
95. E. L. Bienenstock *et al.*, Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* (1982).
96. F. I. Kleberg *et al.*, Excitatory and inhibitory STDP jointly tune feedforward neural circuits to selectively propagate correlated spiking activity. *Frontiers in computational neuroscience* (2014).
97. F. Effenberger *et al.*, Self-organization in balanced state networks by STDP and homeostatic plasticity. *PLoS computational biology* (2015).
98. S. Sadeh *et al.*, Emergence of functional specificity in balanced networks with synaptic plasticity. *PLoS computational biology* (2015).
99. J. Aljadeff *et al.*, Cortical credit assignment by Hebbian, neuromodulatory and inhibitory plasticity. *arXiv preprint arXiv:1911.00307* (2019).
100. V. Pedrosa, C. Clopath, Voltage-based inhibitory synaptic plasticity: network regulation, diversity, and flexibility. *bioRxiv* (2020).
101. E. J. Agnes, T. P. Vogels, Interacting synapses stabilise both learning and neuronal dynamics in biological networks. *BioRxiv* (2021).
102. C. Miehl, J. Gjorgjieva, Stability and learning in excitatory synapses by nonlinear inhibitory plasticity. *PLOS Computational Biology* (2022).
103. T. P. Vogels *et al.*, Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Frontiers in neural circuits* (2013).
104. H. Sprekeler, Functional consequences of inhibitory plasticity: homeostasis, the excitation-inhibition balance and beyond. *Current opinion in neurobiology* (2017).
105. Y. K. Wu *et al.*, Regulation of circuit organization and function through inhibitory synaptic plasticity. *Trends in Neurosciences* (2022).
106. C. Miehl *et al.*, Formation and computational implications of assemblies in neural circuits. *The Journal of Physiology* (2023).
107. C. Pehlevan *et al.*, Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural computation* (2017).
108. C. Von der Malsburg, Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* (1973).
109. V. Delattre *et al.*, Network-timing-dependent plasticity. *Frontiers in cellular neuroscience* (2015).
110. T. Magchielse, E. Meeter, The effect of neuronal activity on the competitive elimination of neuromuscular junctions in tissue culture. *Developmental Brain Research* (1986).
111. P. G. Nelson *et al.*, Synaptic connections in vitro: modulation of number and efficacy by electrical activity. *Science* (1989).
112. Y.-J. Lo, M.-m. Poo, Activity-dependent synaptic competition in vitro: heterosynaptic suppression of developing synapses. *Science* (1991).

113. M. Scanziani *et al.*, Role of intercellular interactions in heterosynaptic long-term depression. *Nature* (1996).
114. S. Royer, D. Paré, Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* (2003).
115. R. Fonseca *et al.*, Competing for memory: hippocampal LTP under regimes of reduced protein synthesis. *Neuron* (2004).
116. I. Rabinowitch, I. Segev, Two opposing plasticity mechanisms pulling a single synapse. *Trends in neurosciences* (2008).
117. A. Govindarajan *et al.*, The dendritic branch is the preferred integrative unit for protein synthesis-dependent LTP. *Neuron* (2011).
118. W. C. Oh *et al.*, Heterosynaptic structural plasticity on local dendritic segments of hippocampal CA1 neurons. *Cell reports* (2015).
119. S. El-Boustani *et al.*, Locally coordinated synaptic plasticity of visual cortex neurons in vivo. *Science* (2018).
120. G. Antunes, F. Simoes-de-Souza, AMPA receptor trafficking and its role in heterosynaptic plasticity. *Scientific reports* (2018).
121. A. Perez-Alvarez *et al.*, Endoplasmic reticulum visits highly active spines and prevents runaway potentiation of synapses. *Nature communications* (2020).
122. T. Ravasenga *et al.*, Spatial regulation of coordinated excitatory and inhibitory synaptic plasticity at dendritic synapses. *Cell Reports* (2022).
123. N. W. Gray *et al.*, Rapid redistribution of synaptic PSD-95 in the neocortex in vivo. *PLoS biology* (2006).
124. S. H. Lee *et al.*, Super-resolution imaging of synaptic and Extra-synaptic AMPA receptors with different-sized fluorescent probes. *Elife* (2017).
125. M. Sheng, E. Kim, The postsynaptic organization of synapses. *Cold Spring Harbor perspectives in biology* (2011).
126. M. van Oostrum *et al.*, The proteomic landscape of synaptic diversity across brain regions and cell types. *bioRxiv* (2023).
127. A. Gupta *et al.*, Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. *Science* (2000).
128. A. M. Craig, H. Boudin, Molecular heterogeneity of central synapses: afferent and target regulation. *Nature neuroscience* (2001).
129. G. H. Diering, R. L. Huganir, The AMPA receptor code of synaptic plasticity. *Neuron* (2018).
130. J. J. Zhu, Activity level-dependent synapse-specific AMPA receptor trafficking regulates transmission kinetics. *Journal of Neuroscience* (2009).
131. J. A. Wen, A. L. Barth, Input-specific critical periods for experience-dependent plasticity in layer 2/3 pyramidal neurons. *Journal of Neuroscience* (2011).
132. J. N. Levinson, A. El-Husseini, Building excitatory and inhibitory synapses: balancing neuroligin partnerships. *Neuron* (2005).
133. A. A. Chubykin *et al.*, Activity-dependent validation of excitatory versus inhibitory synapses by neuroligin-1 versus neuroligin-2. *Neuron* (2007).
134. M. W. Self *et al.*, Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proceedings of the National Academy of Sciences* (2012).
135. M. E. Horn, R. A. Nicoll, Somatostatin and parvalbumin inhibitory synapses onto hippocampal pyramidal neurons are regulated by distinct mechanisms. *Proceedings of the National Academy of Sciences* (2018).
136. C. Bernard *et al.*, Cortical wiring by synapse type-specific control of local protein synthesis. *Science* (2022).
137. R. S. Larsen, P. J. Sjöström, Synapse-type-specific plasticity in local circuits. *Current opinion in neurobiology* (2015).
138. A. R. McFarlan *et al.*, The plasticitome of cortical interneurons. *Nature Reviews Neuroscience* (2022).
139. G. G. Turrigiano *et al.*, Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* (1998).
140. G. G. Turrigiano, S. B. Nelson, Homeostatic plasticity in the developing nervous system. *Nature reviews neuroscience* (2004).
141. P. Wenner, Mechanisms of GABAergic homeostatic plasticity. *Neural plasticity* (2011).
142. I. Spiegel *et al.*, Npas4 regulates excitatory-inhibitory balance within neural circuits through cell-type-specific gene programs. *Cell* (2014).
143. G. G. Turrigiano, The dialectic of Hebb and homeostasis. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2017).
144. F. Zenke *et al.*, The temporal paradox of Hebbian learning and homeostatic plasticity. *Current opinion in neurobiology* (2017).
145. N. Kraynyukova, T. Tchumatchenko, Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of Sciences* (2018).
146. J. T. Chang, D. Fitzpatrick, Development of visual response selectivity in cortical GABAergic interneurons. *Nature Communications* (2022).
147. T. K. Hensch, Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience* (2005).
148. C. N. Levelt, M. Hübener, Critical-period plasticity in the visual cortex. *Annual review of neuroscience* (2012).
149. A. Peters, B. R. Payne, Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. *Cerebral cortex* (1993).
150. A. Peters *et al.*, A numerical analysis of the geniculocortical input to striate cortex in the monkey. *Cerebral Cortex* (1994).
151. R. J. Douglas, K. A. Martin, Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* (2004).
152. S. Lefort *et al.*, The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron* (2009).
153. V. Braitenberg, A. Schüz, *Cortex: statistics and geometry of neuronal connectivity* (Springer Science & Business Media, 2013).
154. G. H. Seol *et al.*, Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. *Neuron* (2007).
155. V. Pawlak *et al.*, Timing is not everything: neuromodulation opens the STDP gate. *Frontiers in synaptic neuroscience* (2010).
156. R. C. Froemke, Plasticity of cortical excitatory-inhibitory balance. *Annual review of neuroscience* (2015).
157. Z. Brzosko *et al.*, Neuromodulation of spike-timing-dependent plasticity: past, present, and future. *Neuron* (2019).
158. A. A. Disney, Neuromodulatory control of early visual processing in macaque. *Annual Review of Vision Science* (2021).
159. A. Maffei, G. Turrigiano, The age of plasticity: developmental regulation of synaptic plasticity in neocortical microcircuits. *Progress in brain research* (2008).
160. A. E. Takesian, T. K. Hensch, Balancing plasticity/stability across brain development. *Progress in brain research* (2013).
161. H. Ko *et al.*, The emergence of functional microcircuits in visual cortex. *Nature* (2013).
162. D. H. Hubel, T. N. Wiesel, Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of neurophysiology* (1963).
163. T. N. Wiesel, D. H. Hubel, Ordered arrangement of orientation columns in monkeys lacking visual experience. *Journal of comparative neurology* (1974).
164. B. Chapman, M. P. Stryker, Development of orientation selectivity in ferret visual cortex and effects of deprivation. *Journal of Neuroscience* (1993).
165. J. B. Ackman *et al.*, Retinal waves coordinate patterned activity throughout the developing visual system. *Nature* (2012).
166. A. Thompson *et al.*, Activity-dependent development of visual receptive fields. *Current opinion in neurobiology* (2017).
167. F. J. Martini *et al.*, Spontaneous activity in developing thalamic and cortical sensory networks. *Neuron* (2021).
168. S. Eckmann, *Synapse-type-specific competitive Hebbian learning*, https://github.com/comp-neural-circuits/Synapse-type-specific-competitive-Hebbian-learning, 2022.

Supplementary Material for

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

Samuel Eckmann✉[,1,2], **Edward James Young**[2], **and Julijana Gjorgjieva**[1,3]

[1]Max Planck Institute for Brain Research, Frankfurt am Main, Germany
[2]Computational and Biological Learning Lab, University of Cambridge, United Kingdom
[3]TUM School of Life Sciences, Technical University Munich, Germany
✉Corresponding author, Email: ec.sam@outlook.com

Contents

1 Linear competitive Hebbian learning finds principal components **2**
1.1 Hebbian plasticity without normalization is unstable . . . 2
1.2 Weight constraints stabilize unlimited Hebbian growth . . . 3
1.2.1 Fixed points . . . 4
1.2.2 Stability analysis . . . 4
1.2.3 Classic Inhibitory plasticity prevents stimulus selectivity . . . 5

2 Synapse-type-specific normalization balances E-I receptive fields **6**
2.1 Fixed points . . . 7
2.1.1 Eigenvectors of the modified covariance matrix are fixed points . . . 7
2.1.2 Eigenvectors and eigenvalues of the modified covariance matrix . . . 7
2.1.3 Non-eigenvector fixed points . . . 8
2.1.4 General fixed points . . . 9
2.2 Stability analysis . . . 10
2.2.1 Principal component analysis in inhibition modified input space . . . 13
2.2.2 Fast inhibition increases stability . . . 13
2.2.3 Stability of non-eigenvector fixed points . . . 13

3 Lateral input stretches and compresses the feedforward input space **15**

4 Eigencircuits **18**
4.1 Variance propagation . . . 19
4.2 Consistency conditions provide eigencircuit firing rate variances . . . 20
4.3 A note on the choice of weight norm . . . 21

5 E-I networks with fully plastic recurrent connectivity **21**
5.1 Fixed points . . . 23
5.2 Stability analysis . . . 24
5.2.1 The transformed Jacobian . . . 25
5.2.2 Stability conditions . . . 29
5.2.3 Eigencircuit stability depends on recurrent connectivity . . . 30
5.2.4 Decorrelation condition . . . 31

6 Movie captions **32**

1 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 (1), we will mostly follow the derivation by Miller and MacKay (2) that we will later extend to inhibitory neurons.

1.1 Hebbian plasticity without normalization is unstable

We consider a single neuron that receives input from a set of excitatory neurons (Fig. S1*A*). Its output firing rate r is a weighted sum of the firing rates of its pre-synaptic inputs $\mathbf{y}$. One can conveniently write this as a dot product:

$$\tau_r \dot{r} = -r + \sum_i w_i y_i = -r + \mathbf{w}^\mathsf{T}\mathbf{y}, \qquad [1]$$

where $\mathbf{w}$ is a vector that holds the synaptic weights, and τ_r defines the timescale at which the activity changes. In the following, lowercase letters in bold indicate vectors, and uppercase letters in bold matrices. Following Hebb's principle, synaptic weight changes depend on the pre- and post-synaptic firing rates. In vector notation:

$$\boldsymbol{\tau}\dot{\mathbf{w}} = \mathbf{y}r \qquad [2]$$

where the constant τ sets the timescale of plasticity. Assuming that synaptic weights change on a much slower timescale than firing rates, $\tau_r \ll \tau$, we make the simplifying assumption that r reaches its fixed point instantaneously, i.e., $\tau_r = 1$ and $r = \mathbf{w}^\mathsf{T}\mathbf{y}$, and consider the same plasticity timescale for all synapses $\boldsymbol{\tau} = \mathbb{1}$. Then, the average change of the synaptic weights can be expressed as a linear transformation of the original weight vector:

$$\langle\dot{\mathbf{w}}\rangle = \langle\mathbf{y}r\rangle = \langle\mathbf{y}\mathbf{y}^T\mathbf{w}\rangle = \mathbf{C}\boldsymbol{w}, \quad \boldsymbol{C} \equiv \langle\boldsymbol{y}\boldsymbol{y}^\mathsf{T}\rangle, \qquad [3]$$

where $\langle\cdot\rangle$ is a temporal average and $\boldsymbol{C}$ 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. To solve this differential equation, we express the weight change in the eigenvector basis of the covariance matrix $\mathbf{C}$, which is symmetric and positive-semidefinite and, therefore, has a complete set of orthonormal eigenvectors with non-negative eigenvalues.

$$\dot{\mathbf{w}}_v \equiv \mathbf{V}^{-1}\dot{\mathbf{w}} = \mathbf{V}^\mathsf{T}\mathbf{C}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{w} = \boldsymbol{\Lambda}\mathbf{w}_v, \qquad [4]$$

$$\Rightarrow \boldsymbol{w}_v = \exp(\boldsymbol{\Lambda}t)\boldsymbol{w}_v(t_0). \qquad [5]$$

Here, $\boldsymbol{\Lambda}$ is the diagonal eigenvalue matrix, and each column of $\mathbf{V}$ holds mutually orthogonal eigenvectors, i.e., $\mathbf{V}\mathbf{V}^\mathsf{T} = \mathbb{1}$, and $\mathbf{V}^{-1} = \mathbf{V}^\mathsf{T}$. Each eigenvector component grows exponentially at a rate given by the respective eigenvalue, which we identify with the *attraction* of the input component. We call eigenvector components with positive eigenvalue *attractive*, and the eigenvector component with the largest eigenvalue the *most attractive* input mode. We will later see that eigenvalues that describe the dynamics of input modes can become negative (Sec. 2). We will call such input modes with negative corresponding eigenvalue *repulsive*.

In summary, we find that unconstrained Hebbian plasticity results in the unlimited growth of synaptic weights and is therefore unstable. One way to constrain this unlimited growth is to modify the Hebbian learning rule such that the total synaptic weight is maintained.

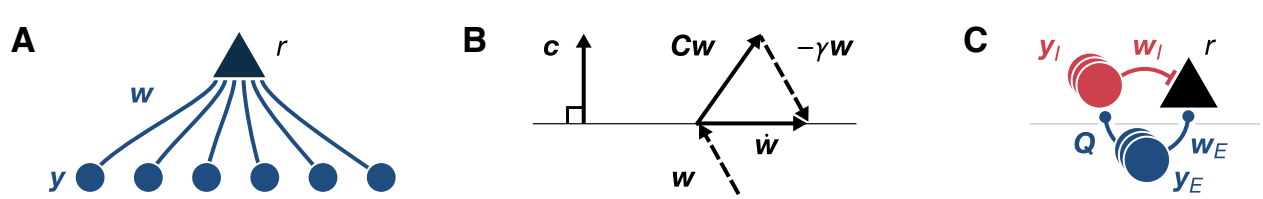

Figure S1: (*A*) Feedforward excitatory circuit. A post-synaptic neuron with output firing rate r receives synapses $\boldsymbol{w}$ from a set of excitatory neurons with firing rates y_E. (*B*) The normalization operation constrains synaptic weight changes $\dot{\boldsymbol{w}}$ to a hyperplane that is perpendicular to the constraint vector $\boldsymbol{c}$ by subtracting a multiple γ of the weight vector $\boldsymbol{w}$. See text for details. Figure adapted from Miller and MacKay (2). (*C*) Feedforward inhibitory circuit. A post-synaptic neuron with output firing rate r receives excitatory synapses $\boldsymbol{w}_E$ from a population of N_E excitatory neurons with firing rates $\boldsymbol{y}_E$, and inhibitory synapses $\boldsymbol{w}_I$ from a population of N_I inhibitory neurons with firing rates $\boldsymbol{y}_I$. The gray horizontal line indicates the separation between two hypothetical brain regions or cortical layers.

1.2 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 (2) and show how one can integrate these two discrete steps into one and derive the effective weight change $\dot{\boldsymbol{w}}$. One can write the two steps as

$$\tilde{\boldsymbol{w}} = \boldsymbol{w}(t) + \boldsymbol{C}\boldsymbol{w}\Delta t, \quad \boldsymbol{w}(t+\Delta t) = \frac{W}{\boldsymbol{c}^\mathsf{T}\tilde{\boldsymbol{w}}}\tilde{\boldsymbol{w}}, \quad W \equiv \boldsymbol{c}^\mathsf{T}\boldsymbol{w}(t). \tag{6}$$

This update rule maintains the projection of $\boldsymbol{w}$ onto the constraint vector $\boldsymbol{c}$ by multiplicatively scaling the weight vector after the Hebbian learning step, i.e., $\tilde{\boldsymbol{w}}$. Alternatively, if we let W be a constant, the projection onto $\boldsymbol{c}$ 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.

$$\boldsymbol{w}(t+\Delta t) = \beta\,[\boldsymbol{w}(t) + \boldsymbol{C}\boldsymbol{w}(t)\Delta t], \quad \beta(\boldsymbol{w}(t), \Delta t) = \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}(t)}{\boldsymbol{c}^\mathsf{T}\,[\boldsymbol{C}\boldsymbol{w}(t)\Delta t + \boldsymbol{w}(t)]}, \tag{7}$$

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

$$\boldsymbol{c}^\mathsf{T}\boldsymbol{w}(t+\Delta t) = \boldsymbol{c}^\mathsf{T}\boldsymbol{w}(t). \tag{8}$$

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

$$\dot{\boldsymbol{w}} = \lim_{\Delta t\to 0}\frac{\boldsymbol{w}(t+\Delta t) - \boldsymbol{w}(t)}{\Delta t} = \lim_{\Delta t\to 0}\left[\frac{\beta-1}{\Delta t}\boldsymbol{w}(t) + \beta\boldsymbol{C}\boldsymbol{w}(t)\right] \tag{9}$$

$$= \lim_{\Delta t\to 0}\left[\frac{\beta-1}{\Delta t}\boldsymbol{w}(t) + \beta\boldsymbol{C}\boldsymbol{w}(t) + \boldsymbol{C}\boldsymbol{w}(t) - \boldsymbol{C}\boldsymbol{w}(t)\right] \tag{10}$$

$$= \lim_{\Delta t\to 0}\left[\boldsymbol{C}\boldsymbol{w}(t) - \frac{1-\beta}{\Delta t}\,[\boldsymbol{w}(t) + \boldsymbol{C}\boldsymbol{w}(t)\Delta t]\right] \tag{11}$$

$$= \lim_{\Delta t\to 0}\left[\boldsymbol{C}\boldsymbol{w}(t) - \frac{1-\beta}{\beta\Delta t}\boldsymbol{w}(t+\Delta t)\right], \tag{12}$$

where, in the first and last steps, we used the definition of $\boldsymbol{w}(t+\Delta t)$ in Eq. 7. Next, we take the limit

$$\lim_{\Delta t\to 0}\frac{1-\beta}{\beta\Delta t} = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\frac{1}{\beta} - 1\right) \tag{13}$$

$$= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\frac{\boldsymbol{c}^\mathsf{T}\,[\boldsymbol{C}\boldsymbol{w}\Delta t + \boldsymbol{w}]}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}} - 1\right) \tag{14}$$

$$= \lim_{\Delta t\to 0}\frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{C}\boldsymbol{w}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}} + \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}\Delta t} - \frac{1}{\Delta t} = \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{C}\boldsymbol{w}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}}. \tag{15}$$

In summary, we get (cf. Fig. S1*B*):

$$\boxed{\Rightarrow \dot{\boldsymbol{w}} = \boldsymbol{C}\boldsymbol{w} - \gamma\boldsymbol{w}, \quad \gamma \equiv \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{C}\boldsymbol{w}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}}.} \tag{16}$$

Here, γ is a scalar normalization factor that depends on the current weight $\boldsymbol{w}$.

An alternative way to derive $\dot{\boldsymbol{w}}$ is to guess the shape of the multiplicative normalization term in Eq. 16 and require that the change along the constraint vector is zero[1], i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{c}^\mathsf{T}\boldsymbol{w}\right) = \boldsymbol{c}^\mathsf{T}\dot{\boldsymbol{w}} = \boldsymbol{c}^\mathsf{T}\boldsymbol{C}\boldsymbol{w} - \gamma\boldsymbol{c}^\mathsf{T}\boldsymbol{w} \overset{!}{=} 0, \quad \Rightarrow \gamma = \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{C}\boldsymbol{w}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}}, \tag{17}$$

. Note that for $\boldsymbol{c}$ being a constant vector of ones, the L1-norm of the weight vector is maintained. However, $\boldsymbol{c}$ does not have to be constant. For example, for $\boldsymbol{c} = \boldsymbol{w}$ the L2-norm is maintained. Also, note that one can analogously derive effective plasticity rules when weights are constrained via subtractive normalization with the ansatz $\dot{\boldsymbol{w}} = \boldsymbol{C}\boldsymbol{w} - \zeta\boldsymbol{k}$, where $\boldsymbol{k}$ is a vector of ones (2).

[1] We indicate an equality or condition that we want to be fulfilled with an exclamation point over the equal sign.

1.2.1 Fixed points

From Eq. 16 it is clear that all fixed points, for which $\dot{\boldsymbol{w}}^* = 0$, are multiples of eigenvectors $\boldsymbol{v}$ of $\boldsymbol{C}$. Explicitly, for a scalar constant a and $\boldsymbol{w}^* = a\boldsymbol{v}$ one gets:

$$\dot{\boldsymbol{w}}^* = a\boldsymbol{C}\boldsymbol{v} - \frac{\boldsymbol{c}^\top \boldsymbol{C}\boldsymbol{v}}{\boldsymbol{c}^\top \boldsymbol{v}}\, a\boldsymbol{v} = a\lambda\boldsymbol{v} - \frac{\boldsymbol{c}^\top \lambda\boldsymbol{v}}{\boldsymbol{c}^\top \boldsymbol{v}} a\boldsymbol{v} = 0. \quad [18]$$

Note that this is independent of the choice of the constraint vector $\boldsymbol{c}$. We next consider the stability of these fixed points.

1.2.2 Stability analysis

In the previous sections, we showed how multiplicative normalization constrains the norm of the weight vector and therefore prevents the otherwise unlimited growth of Hebbian plasticity. However, even when the total synaptic weight is constrained, synaptic weights might still be unstable and never settle into a fixed point, e.g., experiencing oscillatory dynamics and unstable fixed points. Following Miller and MacKay (2), we will now explore 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 (3). The weight dynamics around a fixed point $\boldsymbol{w}^*$ can be approximated with its Taylor expansion:

$$\dot{\boldsymbol{w}} \approx \dot{\boldsymbol{w}}^* + \sum_i \left.\frac{\mathrm{d}\dot{\boldsymbol{w}}}{\mathrm{d}w_i}\right|_* (w_i - w_i^*), \quad [19]$$

$$= \left.\frac{\mathrm{d}\dot{\boldsymbol{w}}}{\mathrm{d}\boldsymbol{w}}\right|_* (\boldsymbol{w} - \boldsymbol{w}^*), \quad [20]$$

$$= \boldsymbol{J}^*(\boldsymbol{w} - \boldsymbol{w}^*), \quad [21]$$

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

$$\boldsymbol{J}^* \equiv \begin{pmatrix} \left.\frac{\mathrm{d}\dot{w}_1}{\mathrm{d}w_1}\right|_* & \cdots & \left.\frac{\mathrm{d}\dot{w}_1}{\mathrm{d}w_N}\right|_* \\ \vdots & & \vdots \\ \left.\frac{\mathrm{d}\dot{w}_N}{\mathrm{d}w_1}\right|_* & \cdots & \left.\frac{\mathrm{d}\dot{w}_N}{\mathrm{d}w_N}\right|_* \end{pmatrix} \equiv \left.\frac{\mathrm{d}\dot{\boldsymbol{w}}}{\mathrm{d}\boldsymbol{w}}\right|_*. \quad [22]$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\boldsymbol{w} = \dot{\boldsymbol{w}} - \dot{\boldsymbol{w}}^* = \dot{\boldsymbol{w}} \approx \boldsymbol{J}^*\Delta\boldsymbol{w}, \quad [23]$$

where we approximated $\dot{\boldsymbol{w}}$ with its Taylor expansion (Eq. 19), since the perturbation is small, i.e., $\boldsymbol{w}$ is close to the fixed point. The result is a linear differential equation that one can solve as

$$\Delta\boldsymbol{w}(t) = \exp\left(\boldsymbol{J}^* t\right)\Delta\boldsymbol{w}(t_0), \quad [24]$$

where all vector components decay to zero if all eigenvalues of $\boldsymbol{J}^*$ are negative[1,2]. As we will see later, it is useful to rewrite the weight dynamics (Eq. 16) as

$$\dot{\boldsymbol{w}} = \boldsymbol{C}\boldsymbol{w} - \boldsymbol{w}\gamma, \quad [25]$$

$$= \boldsymbol{C}\boldsymbol{w} - \frac{\boldsymbol{w}\boldsymbol{c}^\top \boldsymbol{C}\boldsymbol{w}}{\boldsymbol{c}^\top \boldsymbol{w}}, \quad [26]$$

$$= \left[\mathbb{1} - \frac{\boldsymbol{w}\boldsymbol{c}^\top}{\boldsymbol{c}^\top \boldsymbol{w}}\right]\boldsymbol{C}\boldsymbol{w}. \quad [27]$$

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

[2] In general, the real part of the eigenvalues of the Jacobian have to be negative for a fixed point to be stable. However, since C is a covariance matrix, it is positive definite with positive, real eigenvalues. We will see (Eq. 32) that from this it follows that the eigenvalues of the Jacobian are also real.

It follows[1]:

$$\left.\frac{\mathrm{d}\dot{\boldsymbol{w}}}{\mathrm{d}\boldsymbol{w}}\right|_* = \left[\mathbb{1} - \frac{\boldsymbol{v}^*\boldsymbol{c}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*}\right]\boldsymbol{C} + \left[-\frac{\mathbb{1}\boldsymbol{c}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*} + \frac{\boldsymbol{v}^*\boldsymbol{c}^\mathsf{T}\boldsymbol{c}^\mathsf{T}}{(\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*)^2}\right]\boldsymbol{C}\boldsymbol{v}^*, \tag{28}$$

$$= \left[\mathbb{1} - \frac{\boldsymbol{v}^*\boldsymbol{c}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*}\right]\left[\boldsymbol{C} - \lambda^*\mathbb{1}\right], \tag{29}$$

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

$$\boldsymbol{V}^\mathsf{T}\left.\frac{\mathrm{d}\dot{\boldsymbol{w}}}{\mathrm{d}\boldsymbol{w}}\right|_* \boldsymbol{V} = \left[\mathbb{1} - \boldsymbol{V}^\mathsf{T}\frac{\boldsymbol{v}^*\boldsymbol{c}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*}\boldsymbol{V}\right]\left[\boldsymbol{V}^\mathsf{T}\boldsymbol{C}\boldsymbol{V} - \lambda^*\mathbb{1}\right], \tag{30}$$

$$= \left[\mathbb{1} - \boldsymbol{e}^*\frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{V}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{v}^*}\right]\left[\Lambda - \lambda^*\mathbb{1}\right], \tag{31}$$

where Λ is a diagonal matrix that holds the eigenvalues of $\boldsymbol{C}$. Without loss of generality, we can assume that the first column of $\boldsymbol{V}$ is equal to $\boldsymbol{v}^*$. Then $\boldsymbol{e}^* = \boldsymbol{V}^\mathsf{T}\boldsymbol{v}^*$ is a column vector of zeros, except for the first entry, which 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[2] that the eigenvalues of the Jacobian are

$$\boxed{\lambda_{\boldsymbol{J}} = \lambda - \lambda^*.} \tag{32}$$

If λ^* is the largest eigenvalue, i.e., $\boldsymbol{w}^*$ is a multiple of the principal eigenvector of $\boldsymbol{C}$, then all $\lambda_{\boldsymbol{J}}$ are negative or zero, and the fixed point is marginally stable. If there exists a $\lambda > \lambda^*$, the corresponding $\lambda_{\boldsymbol{J}}$ is positive and the fixed point is unstable. Therefore, the eigenvector corresponding to the principal eigenvalue is the only (marginally) stable fixed point. In summary, linear Hebbian learning combined with multiplicative normalization becomes selective for the principal eigenvector of the input covariance matrix and thus performs principal component analysis (PCA). Next, we consider what happens when a neuron also receives inhibitory input.

1.2.3 Classic Inhibitory plasticity prevents stimulus selectivity

Previous work suggested a homeostatic inhibitory synaptic plasticity rule (*4*) that enforced a post-synaptic target firing rate r_0:

$$\dot{\mathbf{w}}_I \propto \mathbf{y}_I\,(r - r_0)\,. \tag{33}$$

However, when combined with excitatory plasticity, this classic rule prevents the development of stimulus selectivity. For completeness, we briefly recapitulate this result presented in Clopath *et al.*(*5*): The authors find that classic inhibitory plasticity is required to act faster than excitatory plasticity to enable stable weight dynamics (*5*). 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^*.

$$\dot{\mathbf{w}}_I^* = 0 \quad\Rightarrow\quad r^* = r_0 \quad\Rightarrow\quad \dot{\mathbf{w}}_E \propto \mathbf{y}_E r_0 - \text{normalization} \tag{34}$$

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

$$\langle\dot{\mathbf{w}}_E\rangle \propto \mathbf{c}y_0 r_0 - \text{normalization}, \tag{35}$$

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. 1*E* & *F*). Therefore, classic inhibitory plasticity that enforces a target firing rate cannot explain the joint development of stimulus selectivity and inhibitory balance. Instead, we propose that, as excitatory weights, also inhibitory weights are constrained via a competitive process that normalizes the total inhibitory input a neuron receives.

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

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

2 Synapse-type-specific normalization balances E-I receptive fields

Different from the normalization of excitatory weights, the normalization of inhibitory weights is not motivated by the requirement for stability. Inhibitory synaptic plasticity that depends on neural activity is self-limiting, since increasing inhibitory weights eventually prevent the neuron from firing, and thus prevent further plasticity. Instead, we motivate the normalization of inhibitory synaptic weights by the competition for a limited amount of synaptic building blocks that may also drive excitatory normalization (see Main text for details).

In the following, we generalize the approach outlined above for excitatory weight normalization to the case of simultaneous excitatory and inhibitory normalization. We consider a simplified circuit of a single post-synaptic neuron with firing rate r that receives lateral input from N_I inhibitory neurons, while all neurons receive feedforward input from a population of N_E excitatory neurons[1]. Then $\boldsymbol{y_E}$ and $\boldsymbol{y_I}$ are vectors that hold the firing rates of the excitatory and inhibitory populations. We now explore the self-organization of excitatory and inhibitory synaptic weights, $\boldsymbol{w}_E$ and $\boldsymbol{w}_I$, that project onto the single post-synaptic neuron, while input synapses $\boldsymbol{Q}$ that project onto inhibitory neurons remain fixed (Fig. S1*C*). Then, the rate dynamics reads

$$\tau_r \dot{r} = -r + \boldsymbol{y}_E{}^\mathsf{T} \boldsymbol{w}_E - \boldsymbol{y}_I{}^\mathsf{T} \boldsymbol{w}_I{}^\mathsf{T} = -r + \bar{\boldsymbol{y}}^\mathsf{T} \begin{pmatrix} \mathbb{1} & \boldsymbol{0} \\ \boldsymbol{0} & -\mathbb{1} \end{pmatrix} \bar{\boldsymbol{w}}, \tag{36}$$

$$\bar{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{w}_E \\ \boldsymbol{w}_I \end{pmatrix}, \quad \bar{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}_E \\ \boldsymbol{y}_I \end{pmatrix}, \quad \boldsymbol{y}_I = \boldsymbol{Q}\boldsymbol{y}_E, \tag{37}$$

where $\mathbb{1}$ is the unit matrix with appropriate dimension, $\boldsymbol{0}$ are matrices of zeros and appropriate dimensionality, and we defined the modified weight and input vectors, $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{y}}$. Similar to before, we assume fast activity dynamics, $\tau_r = 1$, and write the Hebbian part of the time-averaged weight dynamics as

$$\bar{\boldsymbol{\tau}} \left\langle \dot{\bar{\boldsymbol{w}}} \right\rangle = \langle \bar{\boldsymbol{y}} r \rangle = \left\langle \bar{\boldsymbol{y}} \bar{\boldsymbol{y}}^\mathsf{T} \right\rangle \begin{pmatrix} \mathbb{1} & \boldsymbol{0} \\ \boldsymbol{0} & -\mathbb{1} \end{pmatrix} \bar{\boldsymbol{w}}, \tag{38}$$

$$= \left\langle \begin{pmatrix} \boldsymbol{y}_E \boldsymbol{y}_E{}^\mathsf{T} & -\boldsymbol{y}_E \boldsymbol{y}_I{}^\mathsf{T} \\ \boldsymbol{y}_I \boldsymbol{y}_E{}^\mathsf{T} & -\boldsymbol{y}_I \boldsymbol{y}_I{}^\mathsf{T} \end{pmatrix} \right\rangle \bar{\boldsymbol{w}} \equiv \bar{\boldsymbol{C}} \bar{\boldsymbol{w}}, \tag{39}$$

where we defined the modified covariance matrix $\bar{\boldsymbol{C}}$. In general, we assume that all synapses of one type, excitatory or inhibitory, change equally fast (cf. Table 1). Then, the matrix $\bar{\boldsymbol{\tau}}$ holds the timescales of excitatory plasticity, $\boldsymbol{\tau}_E = \mathbb{1}\tau_E$, and inhibitory plasticity, $\boldsymbol{\tau}_I = \mathbb{1}\tau_I$, as matrices on the diagonal, 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 (cf. Eq. 16):

$$\boxed{\bar{\boldsymbol{\tau}} \dot{\bar{\boldsymbol{w}}} = \bar{\boldsymbol{C}} \bar{\boldsymbol{w}} - \gamma \bar{\boldsymbol{w}}_E - \rho \bar{\boldsymbol{w}}_I}, \tag{40}$$

$$\bar{\boldsymbol{w}}_E = \begin{pmatrix} \boldsymbol{w}_E \\ \boldsymbol{0} \end{pmatrix}, \quad \bar{\boldsymbol{w}}_I = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{w}_I \end{pmatrix}, \tag{41}$$

where $\boldsymbol{0}$ indicates vectors of zeros of appropriate dimension (N_I and N_E) that we do not specify for better readability.

The constraint factors γ and ρ follow from the requirement that the weight vector does not grow along the direction of the constraint vectors $\bar{\boldsymbol{c}}_E$ and $\bar{\boldsymbol{c}}_I$. 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[2].

$$\bar{\boldsymbol{c}}_E{}^\mathsf{T} \dot{\bar{\boldsymbol{w}}} \stackrel{!}{=} 0, \quad \bar{\boldsymbol{c}}_I{}^\mathsf{T} \dot{\bar{\boldsymbol{w}}} \stackrel{!}{=} 0, \tag{42}$$

$$\bar{\boldsymbol{c}}_E{}^\mathsf{T} \equiv \Big(1, \dots, 1, 0, \dots, 0\Big), \quad \bar{\boldsymbol{c}}_I{}^\mathsf{T} \equiv \Big(0, \dots, 0, 1, \dots, 1\Big), \tag{43}$$

where the number of non-zero entries in $\bar{\boldsymbol{c}}_E$ and $\bar{\boldsymbol{c}}_I$ is equal to the number of excitatory N_E and inhibitory neurons N_I, respectively. Based on these requirements we derive expressions for the scalar constraint factors γ and ρ:

$$\Rightarrow \quad \gamma = \frac{\bar{\boldsymbol{c}}_E{}^\mathsf{T} \bar{\boldsymbol{C}} \bar{\boldsymbol{w}}}{\bar{\boldsymbol{c}}_E{}^\mathsf{T} \bar{\boldsymbol{w}}_E}, \quad \rho = \frac{\bar{\boldsymbol{c}}_I{}^\mathsf{T} \bar{\boldsymbol{C}} \bar{\boldsymbol{w}}}{\bar{\boldsymbol{c}}_I{}^\mathsf{T} \bar{\boldsymbol{w}}_I}. \tag{44}$$

[1] Note that N_I does not necessarily equal N_E.

[2] 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 the learning of feedforward receptive fields. However, in recurrent networks, they can lead to instabilities (cf. Sec. 4.3).

Finally, we can write the weight dynamics as

$$\Rightarrow \quad \boxed{\bar{\tau}\dot{\bar{\boldsymbol{w}}} = \left[\mathbb{1} - \frac{\bar{\boldsymbol{w}}_E \bar{\boldsymbol{c}}_E{}^\top}{\bar{\boldsymbol{c}}_E{}^\top \bar{\boldsymbol{w}}_E} - \frac{\bar{\boldsymbol{w}}_I \bar{\boldsymbol{c}}_I{}^\top}{\bar{\boldsymbol{c}}_I{}^\top \bar{\boldsymbol{w}}_I}\right] \bar{\boldsymbol{C}}\bar{\boldsymbol{w}}} \,. \tag{45}$$

2.1 Fixed points

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

$$\bar{\tau}\dot{\bar{\boldsymbol{w}}}^* = \bar{\boldsymbol{C}}\bar{\boldsymbol{w}}^* - \gamma\bar{\boldsymbol{w}}_E^* - \rho\bar{\boldsymbol{w}}_I^* \tag{46}$$

$$= \bar{\boldsymbol{C}}\bar{\boldsymbol{w}}^* - \frac{\bar{\boldsymbol{c}}_E{}^\top \bar{\boldsymbol{C}}\bar{\boldsymbol{w}}^*}{\bar{\boldsymbol{c}}_E{}^\top \bar{\boldsymbol{w}}_E^*}\bar{\boldsymbol{w}}_E^* - \frac{\bar{\boldsymbol{c}}_I{}^\top \bar{\boldsymbol{C}}\bar{\boldsymbol{w}}^*}{\bar{\boldsymbol{c}}_I{}^\top \bar{\boldsymbol{w}}_I^*}\bar{\boldsymbol{w}}_I^* \overset{!}{=} \boldsymbol{0}, \tag{47}$$

which is equivalent to

$$\boxed{\bar{\boldsymbol{C}}\bar{\boldsymbol{w}}^* \overset{!}{=} \lambda_E\bar{\boldsymbol{w}}_E^* + \lambda_I\bar{\boldsymbol{w}}_I^*} \,, \tag{48}$$

for λ_E and λ_I being arbitrary scalar.

2.1.1 Eigenvectors of the modified covariance matrix are fixed points

It is straightforward to check that multiples of eigenvectors $\bar{\boldsymbol{v}}$ of the modified covariance matrix $\bar{\boldsymbol{C}}$ with eigenvalue $\bar{\lambda}$ are fixed points:

$$\bar{\boldsymbol{C}}\bar{\boldsymbol{v}} = \bar{\lambda}\bar{\boldsymbol{v}}_E + \bar{\lambda}\bar{\boldsymbol{v}}_I = \lambda_E\bar{\boldsymbol{v}}_E + \lambda_I\bar{\boldsymbol{v}}_I \quad \Rightarrow \quad \lambda_E = \lambda_I = \bar{\lambda}. \tag{49}$$

In the following, we will refer to eigenvectors of the modified covariance matrix as fixed point eigenvectors, and to eigenvectors of the feedforward excitatory covariance matrix $\boldsymbol{C}$ as feedforward eigenvectors. This leaves us with the task of finding the eigenvectors of $\bar{\boldsymbol{C}}$. In general, eigenvectors depend non-trivially on the tuning of the laterally projecting population (cf. Sec. 3, Eq. 121). However, the problem simplifies when the laterally projecting inhibitory neurons are tuned to multiples of eigenvectors of the excitatory population's covariance matrix. This is what one would expect when the post-synaptic excitatory neuron r and the inhibitory population $\boldsymbol{y}_E$ both receive excitatory input from the same external brain region $\boldsymbol{y}_E$ and synapses from the external population onto inhibitory neurons are plastic according to a Hebbian rule with multiplicative normalization (cf. Fig. S1*C*). Although we showed in Section 1.2.2 that without recurrent interactions only the principal eigenvector is a stable fixed point, we will find that with suitable recurrent interactions any feedforward eigenvector can be stable (cf. Sec. 3 & 5.2.3). Formally we set

$$\boldsymbol{y}_I = \boldsymbol{Q}\boldsymbol{y}_E = \boldsymbol{A}^\top \boldsymbol{V}^\top \boldsymbol{y}_E, \tag{50}$$

where each row of $\boldsymbol{Q} = \boldsymbol{A}^\top \boldsymbol{V}^\top$ is the feedforward weight vector of an inhibitory neuron which is equal to a positive multiple, a, of an eigenvector $\boldsymbol{v}$ of the excitatory covariance matrix $\boldsymbol{C} = \langle \boldsymbol{y}_E \boldsymbol{y}_E{}^\top \rangle$. Then $\boldsymbol{V}$ holds all eigenvectors as columns, and $\boldsymbol{A}$ is a matrix where each multiple is the only non-zero element per column, such that $\boldsymbol{A}\boldsymbol{A}^\top$ is a diagonal matrix. We will now show that in this scenario multiples of the excitatory and inhibitory part of the eigenvectors of the modified covariance matrix $\bar{\boldsymbol{C}}$ are fixed points. As a first step, we explicitly calculate the eigenvectors.

2.1.2 Eigenvectors and eigenvalues of the modified covariance matrix

In the previous section, we have seen that eigenvectors $\bar{\boldsymbol{v}}$ of the modified covariance matrix $\bar{\boldsymbol{C}}$ are fixed points. In this section, we will find an explicit expression for these eigenvectors when inhibitory neurons are tuned to feedforward eigenvectors, i.e., inhibitory neurons are tuned to eigenvectors $\boldsymbol{v}$ of the excitatory covariance matrix $\boldsymbol{C}$. Making use of Eq. 50 the modified covariance matrix becomes

$$\bar{\boldsymbol{C}} = \left\langle \begin{pmatrix} \boldsymbol{y}_E\boldsymbol{y}_E{}^\top & -\boldsymbol{y}_E\boldsymbol{y}_I{}^\top \\ \boldsymbol{y}_I\boldsymbol{y}_E{}^\top & -\boldsymbol{y}_I\boldsymbol{y}_I{}^\top \end{pmatrix} \right\rangle = \begin{pmatrix} \boldsymbol{C} & -\boldsymbol{C}\boldsymbol{V}\boldsymbol{A} \\ \boldsymbol{A}^\top\boldsymbol{V}^\top\boldsymbol{C} & -\boldsymbol{A}^\top\boldsymbol{V}^\top\boldsymbol{C}\boldsymbol{V}\boldsymbol{A} \end{pmatrix} = \begin{pmatrix} \boldsymbol{C} & -\boldsymbol{V}\Lambda\boldsymbol{A} \\ \boldsymbol{A}^\top\Lambda\boldsymbol{V}^\top & -\boldsymbol{A}^\top\Lambda\boldsymbol{A} \end{pmatrix}. \tag{51}$$

Then, a full set[1] of linearly independent eigenvectors $\bar{\boldsymbol{V}}$ and their inverse $\bar{\boldsymbol{V}}^{-1}$ is given as[2].

$$\bar{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{VA} \\ \boldsymbol{A}^\mathsf{T} & \mathbb{1} \end{pmatrix}, \quad \bar{\boldsymbol{V}}^{-1} = \begin{pmatrix} (\mathbb{1} - \boldsymbol{AA}^\mathsf{T})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\mathbb{1} - \boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}^\mathsf{T} & -\boldsymbol{A} \\ -\boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T} & \mathbb{1} \end{pmatrix}, \tag{52}$$

where each column of $\bar{\boldsymbol{V}}$ is an non-normalized eigenvector. The eigenvalue spectrum is

$$\bar{\boldsymbol{C}}\bar{\boldsymbol{V}} \overset{!}{=} \bar{\boldsymbol{V}}\bar{\Lambda} \quad \Rightarrow \bar{\Lambda} = \begin{pmatrix} \Lambda(\mathbb{1} - \boldsymbol{AA}^\mathsf{T}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{53}$$

Similar to before, we call eigenvectors of the modified covariance matrix with positive eigenvalue *attractive*. Different from the case of only excitatory feedforward input, eigenvalues of the modified covariance matrix can also be negative. In this case, we call the corresponding eigenvector *repulsive* (cf. Sec. 1.1).

For eigenvectors in the right matrix column of $\bar{\boldsymbol{V}}$ in Eq. 52, the excitatory and inhibitory components of the membrane potential exactly cancel, post-synaptic firing rates are zero, and no plasticity is induced: For multiple post-synaptic neurons with firing rates $\boldsymbol{r}$, where each neuron is tuned to one of these eigenvectors, one gets

$$\boldsymbol{r} = \bar{\boldsymbol{y}}^\mathsf{T} \begin{pmatrix} \mathbb{1} & \boldsymbol{0} \\ \boldsymbol{0} & -\mathbb{1} \end{pmatrix} \bar{\boldsymbol{V}}^\circ, \quad \bar{\boldsymbol{V}}^\circ = \begin{pmatrix} \boldsymbol{VA} \\ \mathbb{1} \end{pmatrix}, \quad \bar{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}_E \\ \boldsymbol{y}_I \end{pmatrix}, \quad \boldsymbol{y}_I = \boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T}\boldsymbol{y}_E, \tag{54}$$

$$\Rightarrow \quad \boldsymbol{r} = \left(\boldsymbol{y}_E{}^\mathsf{T}, -\boldsymbol{y}_E{}^\mathsf{T}\boldsymbol{VA}\right) \begin{pmatrix} \boldsymbol{VA} \\ \mathbb{1} \end{pmatrix} = \boldsymbol{0}. \tag{55}$$

Since these eigenvectors result in post-synaptic firing rates of zero, and they define the null space of the $\bar{\boldsymbol{C}}$ matrix (Eq. 53), we call them 'null eigenvectors' or 'null fixed points', and all eigenvectors that are not null eigenvectors 'regular' eigenvectors or fixed points. Note that for each additional inhibitory neuron that is tuned to a feedforward eigenvector, there is an additional null eigenvector, since inhibitory synaptic weights can now shift between the original, and the additional inhibitory neuron to cancel post-synaptic firing. Overall, there are always N_I null eigenvectors and N_E regular eigenvectors[3].

We have already shown in Section 2.1.1 that eigenvectors of $\bar{\boldsymbol{C}}$ are fixed points. Each eigenvector specifies an exact ratio between the excitatory and inhibitory weight norm. Since our learning rule separately maintains the total excitatory and inhibitory synaptic weights, reaching any of these fixed points would require detailed fine-tuning at the point of initialization. In the next section, we show a more general set of fixed points that does not require any fine tuning of weight norms.

2.1.3 Non-eigenvector fixed points

In this section, we show that there exist fixed points that are not eigenvectors of the modified covariance matrix. In particular, arbitrary multiples of the excitatory and inhibitory parts of regular eigenvectors, i.e., of eigenvectors that result in non-zero post-synaptic activity, are fixed points. We make the ansatz that the matrix $\bar{\boldsymbol{W}}^*$ holds fixed points as columns and has the shape

$$\bar{\boldsymbol{W}}^* = \begin{pmatrix} \boldsymbol{VK}_E \\ \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I \end{pmatrix}, \tag{56}$$

where $\boldsymbol{K}_E$ and $\boldsymbol{K}_I$ and are diagonal scaling matrices of arbitrary constants. The fixed point condition that follows from Eq. 48 is

$$\bar{\boldsymbol{C}}\bar{\boldsymbol{W}}^* \overset{!}{=} \begin{pmatrix} \boldsymbol{W}_E^*\Lambda_E \\ \boldsymbol{W}_I^*\Lambda_I \end{pmatrix}. \tag{57}$$

[1] Note that $\boldsymbol{A}$ and $\boldsymbol{VA}$ are of dimension $N_E \times N_I$, and $\bar{\boldsymbol{V}}$ is of dimension $(N_E + N_I) \times (N_E + N_I)$.

[2] To show that $\bar{\boldsymbol{V}}^{-1}$ is the inverse of $\bar{\boldsymbol{V}}$ it is useful to define the Moore-Penrose inverse $\boldsymbol{A}^{-1} = \boldsymbol{A}^\mathsf{T}(\boldsymbol{AA}^\mathsf{T})^{-1}$ and note that $\boldsymbol{A}^\mathsf{T}(\mathbb{1} - \boldsymbol{AA}^\mathsf{T})^{-1} = (\mathbb{1} - \boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}$.

[3] Similarly, each additional laterally projecting excitatory neuron adds another null eigenvector. In that case, the lateral excitatory weight component and the feedforward excitatory weight component have opposite signs such that they cancel each other (cf. Sec. 5.1).

We now show that for any $\boldsymbol{K}_E$, $\boldsymbol{K}_I$ we can find diagonal matrices Λ_E, Λ_I that fulfil this condition[1]. We write explicitly

$$\Rightarrow \quad \overline{\boldsymbol{C}}\overline{\boldsymbol{W}}^* = \begin{pmatrix} \boldsymbol{C} & -\boldsymbol{V}\Lambda\boldsymbol{A} \\ \boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{V}^\mathsf{T} & -\boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{A} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}\boldsymbol{K}_E \\ \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I \end{pmatrix} \overset{!}{=} \begin{pmatrix} \boldsymbol{V}\boldsymbol{K}_E\Lambda_E \\ \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I\Lambda_I \end{pmatrix} \tag{58}$$

$$\boldsymbol{C}\boldsymbol{V}\boldsymbol{K}_E - \boldsymbol{V}\Lambda\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I \overset{!}{=} \boldsymbol{V}\boldsymbol{K}_E\Lambda_E, \tag{59}$$

$$\boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{K}_E - \boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I \overset{!}{=} \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I\Lambda_I. \tag{60}$$

$$\boldsymbol{V}\boldsymbol{K}_E\Lambda - \boldsymbol{V}\boldsymbol{K}_I\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\Lambda \overset{!}{=} \boldsymbol{V}\boldsymbol{K}_E\Lambda_E, \tag{61}$$

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{K}_E\Lambda - \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I\Lambda\boldsymbol{A}\boldsymbol{A}^\mathsf{T} \overset{!}{=} \boldsymbol{A}^\mathsf{T}\boldsymbol{K}_I\Lambda_I, \tag{62}$$

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

$$\Lambda_E = \Lambda\left(\mathbb{1} - \boldsymbol{K}_E^{-1}\boldsymbol{K}_I\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right), \tag{63}$$

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

which are diagonal matrices, as required[2]. Before we consider the stability of these fixed points in Section 2.2, we first show that there is an additional set of fixed points.

2.1.4 General fixed points

Having covered various special cases of fixed points for the dynamics, we now consider the general problem. Recall that fixed points are defined to satisfy Equation 48:

$$\overline{\boldsymbol{C}}\overline{\boldsymbol{w}}^* = \lambda_E\overline{\boldsymbol{w}}_E^* + \lambda_I\overline{\boldsymbol{w}}_I^* \tag{65}$$

Expanding this using our expression for $\overline{\boldsymbol{C}}$ (Eq. 51), we can see that this is equivalent to:

$$\boldsymbol{V}\Lambda\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* - \boldsymbol{V}\Lambda\boldsymbol{A}\boldsymbol{w}_I^* = \lambda_E\boldsymbol{w}_E^*, \tag{66}$$

$$\boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* - \boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{A}\boldsymbol{w}_I^* = \lambda_I\boldsymbol{w}_I^*, \tag{67}$$

and equivalently

$$\Lambda(\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* - \boldsymbol{A}\boldsymbol{w}_I^*) = \lambda_E\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*, \tag{68}$$

$$\boldsymbol{A}^\mathsf{T}\Lambda(\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* - \boldsymbol{A}\boldsymbol{w}_I^*) = \lambda_I\boldsymbol{w}_I^*. \tag{69}$$

Inserting the first into the second expression, we can conclude that

$$\lambda_E\boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* = \lambda_I\boldsymbol{w}_I^*. \tag{70}$$

If $\lambda_E = \lambda_I \neq 0$, then we know that $\overline{\boldsymbol{w}}^*$ is an eigenvector of the modified covariance matrix, as discussed in Section 2.1.1. In the case that $\lambda_E = \lambda_I = 0$, we have the null eigenvectors discussed in Section 2.1.2. We therefore now address the case that $\lambda_E \neq \lambda_I$.

We begin with the case $\lambda_I \neq 0$. Then we can insert Eq. 70 into Eq. 68 to arrive at

$$\Lambda\left(\mathbb{1} - \frac{\lambda_E}{\lambda_I}\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right)\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^* = \lambda_E\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*, \tag{71}$$

which, together with Eq. 70, gives necessary and sufficient conditions for a fixed point. From Eq. 71, we conclude that $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is an eigenvalue of the diagonal matrix $\Lambda\left(\mathbb{1} - \frac{\lambda_E}{\lambda_I}\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right)$ with eigenvalue λ_E. When $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is one-hot, then the

[1] For $\boldsymbol{K}_E = \mathbb{1}$ and $\boldsymbol{K}_I = (\boldsymbol{A}\boldsymbol{A}^\mathsf{T})^{-1}$ columns of $\boldsymbol{W}^*$ holds null eigenvectors that can be formed by a linear combination of null eigenvectors $\overline{\boldsymbol{V}}^\circ$ given in Eq. 54.

[2] In general, this is not the case for null eigenvectors. Following the same formalism for null eigenvectors $\boldsymbol{W}^{*\mathsf{T}} = (\boldsymbol{K}_E{}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T}, \boldsymbol{K}_I{}^\mathsf{T})^\mathsf{T}$ one finds the condition that $\boldsymbol{A}^\mathsf{T}\Lambda\boldsymbol{A}$ must be diagonal. In general, this is not the case, e.g., when multiple inhibitory neurons are tuned to the same eigenvector, i.e., when multiple columns of $\boldsymbol{A}$ hold the same vector up to a constant factor.

vector $\bar{\boldsymbol{w}}^*$ consists of an arbitrary multiple of the excitatory and inhibitory parts of a regular eigenvector, as covered in Section 2.1.3.

We now turn our attention to the case where $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is not simply one-hot. We can now say that for each component j of $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ which is non-zero, the following equation must hold:

$$\lambda_j\left(1-\frac{\lambda_E}{\lambda_I}(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}\right)=\lambda_E. \tag{72}$$

This is a linear system in the pair of variables λ_E and λ_E/λ_I. We work under the mild assumptions that the eigenvalues λ_j, the diagonal elements $(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}$, and their product $\lambda_j(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}$ are distinct for each j. These conditions will in general hold in the absence of fine tuning. In this case, λ_E and λ_I provide two degrees of freedom and there will only be solutions when $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is (at most) two-hot, having non-zero components, j and k. Such solutions satisfy:

$$\begin{pmatrix}\lambda_j\\ \lambda_k\end{pmatrix}=\begin{pmatrix}\lambda_j(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj} & 1\\ \lambda_k(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{kk} & 1\end{pmatrix}\begin{pmatrix}\lambda_E/\lambda_I\\ \lambda_E\end{pmatrix}, \tag{73}$$

which we can solve to obtain the expressions:

$$\lambda_E=\lambda_j\lambda_k\frac{(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}-(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{kk}}{\lambda_j(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}-\lambda_k(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{kk}},\quad \lambda_I=\lambda_j\lambda_k\frac{(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{jj}-(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{kk}}{\lambda_j-\lambda_k}. \tag{74}$$

The components of the two-hot solution are determined by the initial values of $\boldsymbol{c}_I{}^\mathsf{T}\boldsymbol{w}_I$ and $\boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{w}_E$, which are kept constant throughout training. Although two-hot fixed points do not require fine tuning of excitatory and inhibitory weight norms, we did not observe them in any of our numerical simulations and therefore assume they are unstable.

The final case to be considered is when $\lambda_I=0,\lambda_E\neq 0$. In this situation, Eq. 70 tells us that $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is in the kernel of $\boldsymbol{A}^\mathsf{T}$ and therefore in the kernel of the diagonal matrix $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$[1]. By using Eq. 68, we can therefore conclude that

$$\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\Lambda(\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*-\boldsymbol{A}\boldsymbol{w}_I^*)=0. \tag{75}$$

We work under the assumption that, in the absence of fine tuning, Λ has distinct non-zero eigenvalues. In this case, the first term in Equation 75 is zero, and $\boldsymbol{A}\boldsymbol{w}_I^*$ must also be in the kernel of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ and therefore in the kernel of $\boldsymbol{A}^\mathsf{T}$. So $\boldsymbol{w}_I^*$ is in the kernel of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ and therefore the kernel of $\boldsymbol{A}$. By Equation 68, this tells us that $\Lambda\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*=\lambda_E\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$, and therefore $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is an eigenvector of Λ with eigenvalue λ_E. We therefore arrive at a fixed point for the system in which $\boldsymbol{V}^\mathsf{T}\boldsymbol{w}_E^*$ is one-hot with support on the kernel of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, and $\boldsymbol{w}_I^*$ is in the kernel of $\boldsymbol{A}$. This implies $\boldsymbol{w}_I^{*\mathsf{T}}\boldsymbol{y}_I=0$ (cf. Eq. 50) which is biologically implausible since we constrain synaptic weights $\boldsymbol{w}_I^*$ and firing rates $\boldsymbol{y}_I$ to be positive.

Under mild assumptions regarding Λ and $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, we have thus exhaustively characterized the fixed points of the system.

2.2 Stability analysis

We first consider the stability of fixed points that are regular eigenvectors of the modified covariance matrix and discuss the case of non-eigenvector fixed points afterwards. With Eq. 45, for the Jacobian $\bar{\boldsymbol{J}}$ it follows (cf. Eq. 29)

$$\bar{\boldsymbol{\tau}}\,\bar{\boldsymbol{J}}\Big|_*=\bar{\boldsymbol{\tau}}\,\frac{\mathrm{d}\dot{\bar{\boldsymbol{w}}}}{\mathrm{d}\bar{\boldsymbol{w}}}\Big|_*=\left[\mathbb{1}-\frac{\bar{\boldsymbol{v}}_E^*\bar{\boldsymbol{c}}_E{}^\mathsf{T}}{\bar{\boldsymbol{c}}_E{}^\mathsf{T}\bar{\boldsymbol{v}}_E^*}-\frac{\bar{\boldsymbol{v}}_I^*\bar{\boldsymbol{c}}_I{}^\mathsf{T}}{\bar{\boldsymbol{c}}_I{}^\mathsf{T}\bar{\boldsymbol{v}}_I^*}\right]\left[\bar{\boldsymbol{C}}-\bar{\lambda}^*\mathbb{1}\right], \tag{76}$$

where $\bar{\boldsymbol{v}}_E^*$ and $\bar{\boldsymbol{v}}_I^*$ are the excitatory and the inhibitory part of the eigenvector fixed point $\bar{\boldsymbol{w}}^*=\bar{\boldsymbol{v}}^*$ with eigenvalue $\bar{\lambda}^*$, with an additional set of zeros to reach the correct dimensionality of the vector (cf. Eq. 41). To find the eigenvalues $\bar{\lambda}_{\boldsymbol{J}}$ of the Jacobian, we switch to the eigenbasis of the modified covariance matrix[2]:

$$\Rightarrow\quad \bar{\boldsymbol{V}}^{-1}\,\bar{\boldsymbol{J}}\Big|_*\bar{\boldsymbol{V}}=\bar{\boldsymbol{V}}^{-1}\boldsymbol{\tau}^{-1}\bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^{-1}\left[\mathbb{1}-\frac{\bar{\boldsymbol{v}}_E^*\bar{\boldsymbol{c}}_E{}^\mathsf{T}}{\bar{\boldsymbol{c}}_E{}^\mathsf{T}\bar{\boldsymbol{v}}_E^*}-\frac{\bar{\boldsymbol{v}}_I^*\bar{\boldsymbol{c}}_I{}^\mathsf{T}}{\bar{\boldsymbol{c}}_I{}^\mathsf{T}\bar{\boldsymbol{v}}_I^*}\right]\bar{\boldsymbol{V}}\left[\bar{\Lambda}-\bar{\lambda}^*\mathbb{1}\right], \tag{77}$$

where we inserted $\bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^{-1}\equiv\mathbb{1}$. The result is a block triangular matrix where each block on the diagonal corresponds to one regular eigenvector and its potentially multiple null eigenvectors. To better see this, we consider the first

[1] Note that $\ker(\boldsymbol{A}^\mathsf{T})=\ker(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})$ and $\ker(\boldsymbol{A})=\ker(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})$, for any matrix $\boldsymbol{A}$.

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

and second part of Eq. 77 separately. We define $\bar{\boldsymbol{\epsilon}} \equiv \bar{\boldsymbol{\tau}}^{-1}$, which remains a diagonal matrix with time constants for excitatory and inhibitory synapses on the diagonal, $\boldsymbol{\epsilon}_E = \mathbb{1}\epsilon_E$ and $\boldsymbol{\epsilon}_I = \mathbb{1}\epsilon_I$. Inserting the definition of the eigenvectors matrix and its inverse (Eq. 52) we write

$$\bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{\tau}}^{-1}\bar{\boldsymbol{V}} = \begin{pmatrix} (\mathbb{1} - \boldsymbol{A}\boldsymbol{A}^\mathsf{T})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\mathbb{1} - \boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}^\mathsf{T} & -\boldsymbol{A} \\ -\boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T} & \mathbb{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon}_E & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\epsilon}_I \end{pmatrix} \begin{pmatrix} \boldsymbol{V} & \boldsymbol{V}\boldsymbol{A} \\ \boldsymbol{A}^\mathsf{T} & \mathbb{1} \end{pmatrix} \tag{78}$$

$$= \begin{pmatrix} (\mathbb{1} - \boldsymbol{A}\boldsymbol{A}^\mathsf{T})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\mathbb{1} - \boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon}_E - \boldsymbol{\epsilon}_I\boldsymbol{A}\boldsymbol{A}^\mathsf{T} & (\boldsymbol{\epsilon}_E - \boldsymbol{\epsilon}_I)\boldsymbol{A} \\ (\boldsymbol{\epsilon}_I - \boldsymbol{\epsilon}_E)\boldsymbol{A}^\mathsf{T} & \boldsymbol{\epsilon}_I - \boldsymbol{\epsilon}_E\boldsymbol{A}^\mathsf{T}\boldsymbol{A} \end{pmatrix}. \tag{79}$$

As one would expect, for $\boldsymbol{\epsilon}_E = \boldsymbol{\epsilon}_I$, this is equal to a scalar times 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 $\bar{\boldsymbol{V}}$ and its inverse $\bar{\boldsymbol{V}}^{-1}$ are already appropriately sorted. Without loss of generality, we assume that the first columns of $\bar{\boldsymbol{V}}$ are the fixed point's eigenvector $\bar{\boldsymbol{v}}^*$ and its corresponding null eigenvectors, and write[1]

$$\bar{\boldsymbol{V}}^{-1}\bar{\boldsymbol{\tau}}^{-1}\bar{\boldsymbol{V}} = \begin{pmatrix} (1 - \boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^*)^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (1 - \boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}})^{-1} & \\ & \boldsymbol{0} & \ddots \end{pmatrix} \begin{pmatrix} \epsilon_E - \epsilon_I\boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^* & (\epsilon_E - \epsilon_I)\boldsymbol{a}^{*\mathsf{T}} & \boldsymbol{0} \\ (\epsilon_I - \epsilon_E)\boldsymbol{a}^* & \epsilon_I - \epsilon_E\boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}} & \\ \boldsymbol{0} & & \ddots \end{pmatrix}, \tag{80}$$

where $\boldsymbol{a}^*$ is a column vector that holds the multiples of the inhibitory neurons that are tuned to the feedforward eigenvector $\boldsymbol{v}^*$. As before, $\boldsymbol{0}$ are matrices of zeros and appropriate dimensionality, and ellipsis indicate continuing blocks on the diagonal with similar terms that belong to the non-fixed point eigenvectors and their null eigenvectors[2].

Similarly, we can write the second part of Eq. 77 as a block triangular matrix. Before sorting, we write

$$\bar{\boldsymbol{V}}^{-1}\left[\frac{\bar{\boldsymbol{v}}_E^*\bar{\boldsymbol{c}}_E^\mathsf{T}}{\bar{\boldsymbol{c}}_E^\mathsf{T}\bar{\boldsymbol{v}}_E^*} + \frac{\bar{\boldsymbol{v}}_I^*\bar{\boldsymbol{c}}_I^\mathsf{T}}{\bar{\boldsymbol{c}}_I^\mathsf{T}\bar{\boldsymbol{v}}_I^*}\right]\bar{\boldsymbol{V}} \equiv \bar{\boldsymbol{V}}^{-1}\left[\bar{\boldsymbol{v}}_E^*\boldsymbol{d}_E^\mathsf{T} + \bar{\boldsymbol{v}}_I^*\boldsymbol{d}_I^\mathsf{T}\right] = \bar{\boldsymbol{V}}^{-1}\begin{pmatrix} \boldsymbol{v}_E^*\boldsymbol{d}_E^\mathsf{T} \\ \boldsymbol{v}_I^*\boldsymbol{d}_I^\mathsf{T} \end{pmatrix}, \tag{81}$$

$$\boldsymbol{d}_E^\mathsf{T} = \frac{\bar{\boldsymbol{c}}_E^\mathsf{T}\bar{\boldsymbol{V}}}{\bar{\boldsymbol{c}}_E^\mathsf{T}\bar{\boldsymbol{v}}_E^*}, \quad \boldsymbol{d}_I^\mathsf{T} = \frac{\bar{\boldsymbol{c}}_I^\mathsf{T}\bar{\boldsymbol{V}}}{\bar{\boldsymbol{c}}_I^\mathsf{T}\bar{\boldsymbol{v}}_I^*}, \quad \boldsymbol{v}_E^* = \boldsymbol{V}\boldsymbol{e}^*, \quad \boldsymbol{v}_I^* = \boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*, \tag{82}$$

where $\boldsymbol{d}_E^\mathsf{T}$ and $\boldsymbol{d}_I^\mathsf{T}$ are row vectors that hold the L1-norms of the eigenvectors' excitatory and inhibitory parts as a fraction of the L1-norm of the fixed point eigenvector's excitatory and inhibitory parts. The vector $\boldsymbol{e}^*$ is zero except for one entry, equal to one, which corresponds to the fixed point feedforward eigenvector $\boldsymbol{v}^*$. We continue by multiplying the inverse eigenvector matrix $\bar{\boldsymbol{V}}^{-1}$ from the left:

$$\bar{\boldsymbol{V}}^{-1}\begin{pmatrix} \boldsymbol{v}_E^*\boldsymbol{d}_E^\mathsf{T} \\ \boldsymbol{v}_I^*\boldsymbol{d}_I^\mathsf{T} \end{pmatrix} = \boldsymbol{N}\begin{pmatrix} \boldsymbol{V}^\mathsf{T} & -\boldsymbol{A} \\ -\boldsymbol{A}^\mathsf{T}\boldsymbol{V}^\mathsf{T} & \mathbb{1} \end{pmatrix}\begin{pmatrix} \boldsymbol{V}\boldsymbol{e}^*\boldsymbol{d}_E^\mathsf{T} \\ \boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_I^\mathsf{T} \end{pmatrix} = \boldsymbol{N}\begin{pmatrix} \boldsymbol{e}^*\boldsymbol{d}_E^\mathsf{T} - \boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_I^\mathsf{T} \\ -\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_E^\mathsf{T} + \boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_I^\mathsf{T} \end{pmatrix}, \tag{83}$$

where we defined the normalization matrix $\boldsymbol{N}$ of the inverse eigenvector matrix $\bar{\boldsymbol{V}}^{-1}$ (cf. Eq. 52) to improve readability. It follows that the matrix above holds non-zero values in only a few rows, corresponding to the fixed point eigenvector (top block) and its null eigenvectors (bottom block). After rearranging, we get

$$\boldsymbol{N}\begin{pmatrix} \boldsymbol{e}^*\boldsymbol{d}_E^\mathsf{T} - \boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_I^\mathsf{T} \\ -\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_E^\mathsf{T} + \boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*\boldsymbol{d}_I^\mathsf{T} \end{pmatrix} = \boldsymbol{N}\begin{pmatrix} d_E^* - \boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^*d_I^* & \boldsymbol{d}_E^{\circledast\mathsf{T}} - \boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^*\boldsymbol{d}_I^{\circledast\mathsf{T}} & \cdots \\ -\boldsymbol{a}^*d_E^* + \boldsymbol{a}^*d_I^* & -\boldsymbol{a}^*\boldsymbol{d}_E^{\circledast\mathsf{T}} + \boldsymbol{a}^*\boldsymbol{d}_I^{\circledast\mathsf{T}} & \\ & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \tag{84}$$

where d_E^*, d_I^* and $\boldsymbol{d}_E^{\circledast\mathsf{T}}$, $\boldsymbol{d}_I^{\circledast\mathsf{T}}$ are the entries of $\boldsymbol{d}_E^\mathsf{T}$, $\boldsymbol{d}_I^\mathsf{T}$ that correspond to the fixed point eigenvector and its null eigenvectors, respectively. As before, ellipsis indicate additional non-zero entries. To find the respective entries of $\boldsymbol{d}_E$, $\boldsymbol{d}_I$ we use the definition of $\bar{\boldsymbol{V}}$ (Eq. 52) to write

[1] Note that when sorted, $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ is a block diagonal matrix. Further, as noted before, the matrix $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ is always diagonal.

[2] The dimensionalities of these blocks depend on the number inhibitory neurons tuned to the respective feedforward eigenvector, i.e., if there are $n_I^\dagger$ inhibitory neurons tuned to a specific feedforward eigenvector $\boldsymbol{v}^\dagger$, the dimensionality of the corresponding block is $1+n_I$, due to one regular eigenvector and $n_I^\dagger$ corresponding null eigenvectors.

$$\boldsymbol{d}_E{}^\mathsf{T} = \frac{\bar{\boldsymbol{c}}_E{}^\mathsf{T}\bar{\boldsymbol{V}}}{\bar{\boldsymbol{c}}_E{}^\mathsf{T}\bar{\boldsymbol{v}}_E^*} = \frac{1}{\boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{v}^*}\left(\boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{V},\ \boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{V}\boldsymbol{A}\right), \tag{85}$$

$$\boldsymbol{d}_I{}^\mathsf{T} = \frac{\bar{\boldsymbol{c}}_I{}^\mathsf{T}\bar{\boldsymbol{V}}}{\bar{\boldsymbol{c}}_I{}^\mathsf{T}\bar{\boldsymbol{v}}_I^*} = \frac{1}{\boldsymbol{c}_I{}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*}\left(\boldsymbol{c}_I{}^\mathsf{T}\boldsymbol{A}^\mathsf{T},\ \boldsymbol{c}_I{}^\mathsf{T}\mathbb{1}\right). \tag{86}$$

After rearranging the entries that correspond to the fixed point eigenvector and its null eigenvectors to the front we get

$$\boldsymbol{d}_E{}^\mathsf{T} = \left(d_E^*,\ \boldsymbol{d}_E^{\circledast\mathsf{T}},\ \ldots\right) = \frac{1}{\boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{v}^*}\left(\boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{V}\boldsymbol{e}^*,\ \boldsymbol{c}_E{}^\mathsf{T}\boldsymbol{v}^*\boldsymbol{a}^{*\mathsf{T}}\right) = \left(1,\ \boldsymbol{a}^{*\mathsf{T}},\ \ldots\right), \tag{87}$$

$$\boldsymbol{d}_I{}^\mathsf{T} = \left(d_I^*,\ \boldsymbol{d}_I^{\circledast\mathsf{T}},\ \ldots\right) = \frac{1}{\boldsymbol{c}_I{}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*}\left(\boldsymbol{c}_I{}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{e}^*,\ \boldsymbol{c}_I{}^\mathsf{T}\right) = \left(1,\ \frac{\boldsymbol{c}_I{}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{a}^*},\ \ldots\right), \tag{88}$$

where $\boldsymbol{e}^*$ selects the proper columns and $\boldsymbol{c}^\mathsf{T}$ is a row vector of ones of appropriate dimensionality. We insert Eq. 87 & 88 into Eq. 84 and find

$$\bar{\boldsymbol{V}}^{-1}\left[\frac{\bar{\boldsymbol{v}}_E^*\bar{\boldsymbol{c}}_E{}^\mathsf{T}}{\bar{\boldsymbol{c}}_E{}^\mathsf{T}\bar{\boldsymbol{v}}_E^*} + \frac{\bar{\boldsymbol{v}}_I^*\bar{\boldsymbol{c}}_I{}^\mathsf{T}}{\bar{\boldsymbol{c}}_I{}^\mathsf{T}\bar{\boldsymbol{v}}_I^*}\right]\bar{\boldsymbol{V}} = \boldsymbol{N}\begin{pmatrix} 1-\boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^* & \boldsymbol{0} & \\ \boldsymbol{0} & \frac{\boldsymbol{a}^*\boldsymbol{c}_I{}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{a}^*} - \boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}} & \cdots \\ & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \equiv \begin{pmatrix} 1 & \boldsymbol{0} & \\ \boldsymbol{0} & \boldsymbol{M}^* & \cdots \\ & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \tag{89}$$

$$\boldsymbol{M}^* = \left(\mathbb{1} - \boldsymbol{a}^*\boldsymbol{a}^\mathsf{T}\right)^{-1}\left(\frac{\boldsymbol{a}^*\boldsymbol{c}_I{}^\mathsf{T}}{\boldsymbol{c}^\mathsf{T}\boldsymbol{a}^*} - \boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}}\right), \tag{90}$$

where we defined the matrix $\boldsymbol{M}^*$.

In summary, we find that after rearrangement, Eq. 77 is a block triangular matrix.

$$\Rightarrow \quad \bar{\boldsymbol{V}}^{-1}\,\bar{\boldsymbol{J}}\Big|_*\,\bar{\boldsymbol{V}} = \boldsymbol{N}\begin{pmatrix} \epsilon_E - \epsilon_I\boldsymbol{a}^{*\mathsf{T}}\boldsymbol{a}^* & (\epsilon_E-\epsilon_I)\boldsymbol{a}^{*\mathsf{T}} & \boldsymbol{0} \\ (\epsilon_I-\epsilon_E)\boldsymbol{a}^* & \epsilon_I - \epsilon_E\boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}} & \\ \boldsymbol{0} & & \ddots \end{pmatrix}\begin{pmatrix} 0 & \boldsymbol{0} & \\ \boldsymbol{0} & \mathbb{1}-\boldsymbol{M}^* & \cdots \\ & \boldsymbol{0} & \mathbb{1} \end{pmatrix}\left[\bar{\boldsymbol{\Lambda}} - \bar{\lambda}^*\mathbb{1}\right], \tag{91}$$

where we used Eq. 80 and Eq. 89. Therefore, to find the eigenvalues, we consider each diagonal block separately. We make the simplifying assumption that there is exactly one inhibitory neuron tuned to each feedforward eigenvector. Then, $\boldsymbol{a}^* \to a^*$ becomes a scalar, $\boldsymbol{N}$ and $\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}$ become diagonal, and $\boldsymbol{M}^* \to 1$. The transformed Jacobian remains triangular and becomes

$$\Rightarrow \quad \bar{\boldsymbol{V}}^{-1}\,\bar{\boldsymbol{J}}\Big|_*\,\bar{\boldsymbol{V}} = \boldsymbol{N}\begin{pmatrix} \epsilon_E - \epsilon_I a^{*2} & (\epsilon_E-\epsilon_I)a^* & \boldsymbol{0} \\ (\epsilon_I-\epsilon_E)a^* & \epsilon_I - \epsilon_E a^{*2} & \\ \boldsymbol{0} & & \ddots \end{pmatrix}\begin{pmatrix} 0 & 0 & \\ 0 & 0 & \cdots \\ \boldsymbol{0} & & \mathbb{1} \end{pmatrix}\left[\bar{\boldsymbol{\Lambda}} - \bar{\lambda}^*\mathbb{1}\right], \tag{92}$$

with 2×2 blocks on the diagonal of which we only show the first, that corresponds to perturbations in the direction of the fixed point eigenvector or its null eigenvector[1]. From the matrix product above, we see that their corresponding eigenvalues must be zero since the first two columns of the second to last matrix are zero. For perturbations in the direction of a non-fixed point eigenvector $\bar{\boldsymbol{v}}^\dagger$ or its null eigenvector we have to consider the block matrix

$$\boldsymbol{J}_*^\dagger \equiv \bar{\boldsymbol{V}}^{\dagger-1}\,\bar{\boldsymbol{J}}\Big|_*\,\bar{\boldsymbol{V}}^\dagger = \frac{1}{1-a^{\dagger2}}\begin{pmatrix} \epsilon_E - \epsilon_I a^{\dagger2} & \epsilon_E - \epsilon_I \\ (-\epsilon_E+\epsilon_I)a^{\dagger2} & -\epsilon_E a^{\dagger2} + \epsilon_I \end{pmatrix}\begin{pmatrix} \bar{\lambda}^\dagger - \bar{\lambda}^* & 0 \\ 0 & -\bar{\lambda}^* \end{pmatrix}, \tag{93}$$

where $\bar{\boldsymbol{V}}^\dagger$ is a two-column matrix that holds $\bar{\boldsymbol{v}}^\dagger$ 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. For trace and

[1] Since we assumed that there is exactly one inhibitory neuron per feedforward eigenvector, there is also exactly one null eigenvector per feedforward eigenvector (cf. Eq. 52)

determinant, we find

$$\det(\bar{\boldsymbol{J}}_*^\dagger) = \frac{1}{\left(1-a^{\dagger 2}\right)^2}\left(1-2a^{\dagger 2}+a^{\dagger 4}\right)\epsilon_E\epsilon_I\left(\bar{\lambda}^* - \bar{\lambda}^\dagger\right)\bar{\lambda}^*, \tag{94}$$

$$\mathrm{tr}(\boldsymbol{J}_*^\dagger) = \frac{1}{\left(1-a^{\dagger 2}\right)}\left(\epsilon_E\left[\bar{\lambda}^\dagger - \left(1-a^{\dagger 2}\right)\bar{\lambda}^*\right] + \epsilon_I\left[-a^{\dagger 2}\bar{\lambda}^\dagger - \left(1-a^{\dagger 2}\right)\bar{\lambda}^*\right]\right) \tag{95}$$

Finally, the two stability conditions read

$$\det(\boldsymbol{J}_*^\dagger) \overset{!}{>} 0 \quad \Rightarrow \quad -\left(\bar{\lambda}^\dagger - \bar{\lambda}^*\right)\bar{\lambda}^*\epsilon_E\epsilon_I > 0, \tag{96}$$

$$\mathrm{tr}(\boldsymbol{J}_*^\dagger) \overset{!}{<} 0 \quad \Rightarrow \quad \epsilon_E(\lambda^\dagger - \bar{\lambda}^*) - \epsilon_I\left(a^{\dagger 2}\lambda^\dagger + \bar{\lambda}^*\right) < 0, \tag{97}$$

where, for the trace term, we made use of the equality $\bar{\lambda}^\dagger = \lambda^\dagger\left(1-a^{\dagger 2}\right)$ (cf. Eq. 53) to replace $\bar{\lambda}^\dagger$.

2.2.1 Principal component analysis in inhibition modified input space

The first stability condition above states that only the fixed point $\bar{\boldsymbol{v}}^*$ with the largest eigenvalue, $\bar{\lambda}^* > \bar{\lambda}^\dagger, \forall\bar{\lambda}^\dagger$, can be stable, and then only if it is not repulsive, i.e., provided that its corresponding eigenvalue is larger than zero. An eigenvector can become repulsive if inhibition is sufficiently strong, i.e., if $\bar{\lambda}^* = \lambda^*(1-a^{*2}) < 0 \;\Rightarrow\; a^{*2} > 1$. This implies that post-synaptic neurons tuned to repulsive eigenvectors receive more inhibition than excitation, which results in negative firing rates $r^* = \boldsymbol{y}_E{}^\mathsf{T}\boldsymbol{v}^* - \boldsymbol{y}_E{}^\mathsf{T}\boldsymbol{v}^*a^{*2} < 0$, for $a^{*2} > 1$ (cf. Eq. 55). However, in biology, neurons with larger inhibitory than excitatory input are hyperpolarized and remain silent, which is why we assume $\bar{\lambda}^* > 0$. In the following, we call the combination of the excitatory feedforward attraction of an eigenvector λ^* (cf. Sec. 1.1) plus any contribution of laterally projecting neurons, in this case, minus the lateral inhibitory repulsion $a^{*2}\lambda^*$, the effective attraction, $\bar{\lambda}^*$, of a feedforward input mode.

For $\epsilon_I = \epsilon_E$, the second condition reduces to[1] $\bar{\lambda}^\dagger - 2\bar{\lambda}^* < 0$, which holds if the first condition is met. Therefore, the post-synaptic neuron becomes tuned to the eigenvector of the modified covariance matrix with the largest eigenvalue, i.e., it performs principal component analysis on a modified feedforward input space, where the attraction of feedforward eigenvectors is modified by laterally projecting inhibitory neurons (cf. Eq. 53). We will further discuss the notion of a modified input space in Section 3.

2.2.2 Fast inhibition increases stability

In our networks, stationary states can still emerge when inhibitory plasticity is slower than excitatory plasticity. In the extreme case of static inhibition, $\epsilon_I = 0$, the second stability condition is still satisfied if the fixed point attraction $\bar{\lambda}^*$ is larger than the feedforward attraction $\lambda^\dagger$ of any other eigenvector, $\lambda^\dagger - \bar{\lambda}^* < 0$. When inhibitory weights are static, they remain tuned to the fixed point and the repulsive component of competing eigenvectors $a^{\dagger 2}\lambda^\dagger$ do not matter for stability. This explains why we have to consider only the attractive part $\lambda^\dagger$ of the effective attraction $\bar{\lambda}^\dagger$ in the first term of the second stability condition. However, for growing $\epsilon_I > 0$, the influence of the inhibitory part of competing eigenvectors increases, corresponding to an increasingly negative second term in the second stability condition[2]. Then, for sufficiently fast inhibitory plasticity $\epsilon_I > \epsilon_E$, the second condition always holds. Therefore, we consider slightly faster inhibitory than excitatory plasticity in our numerical simulations (cf. Table 1).

2.2.3 Stability of non-eigenvector fixed points

Before, we considered the stability of fixed points $\bar{\boldsymbol{w}}^*$ that are eigenvectors $\bar{\boldsymbol{v}}$ of the modified covariance matrix $\bar{\boldsymbol{C}}$. Weight vectors of that shape put a strong constraint on the choice of the weight norms, as the ratio between the excitatory and the inhibitory weight norms is given by the norms of the excitatory and the inhibitory parts of the eigenvector (cf. Eq. 52). The issue was solved in that we found that arbitrary combinations of multiples of the excitatory and inhibitory components of regular eigenvectors are also fixed points (cf. Sec. 2.1.3). We will now consider the stability of such non-eigenvector fixed points.

[1] Here, we make use of the equality $\bar{\lambda}^\dagger = \lambda^\dagger\left(1-a^{\dagger 2}\right)$.

[2] Note that we consider non-repulsive fixed points $\bar{\lambda}^* > 0$ and inhibitory neurons with positive firing rates, i.e., $a > 0, \forall a$, such that the second term in the second stability condition is always negative.

Let the shape of a fixed point $\overline{\boldsymbol{w}}^*$ be (cf. Eq. 56)

$$\overline{\boldsymbol{w}}^* = \begin{pmatrix} k_E \boldsymbol{v}_E^* \\ k_I \boldsymbol{v}_I^* \end{pmatrix} = \begin{pmatrix} \mathbb{1} k_E & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{1} k_I \end{pmatrix} \overline{\boldsymbol{v}}^* \equiv \boldsymbol{K} \overline{\boldsymbol{v}}^*, \tag{98}$$

where k_E and k_I are scalar constants. We recapitulate the general weight dynamics as given in Eq. 45:

$$\overline{\tau} \dot{\overline{\boldsymbol{w}}} = \left[\mathbb{1} - \frac{\overline{\boldsymbol{w}}_E \overline{\boldsymbol{c}}_E{}^\mathsf{T}}{\overline{\boldsymbol{c}}_E{}^\mathsf{T} \overline{\boldsymbol{w}}_E} - \frac{\overline{\boldsymbol{w}}_I \overline{\boldsymbol{c}}_I{}^\mathsf{T}}{\overline{\boldsymbol{c}}_I{}^\mathsf{T} \overline{\boldsymbol{w}}_I} \right] \overline{\boldsymbol{C}} \overline{\boldsymbol{w}}. \tag{99}$$

Instead of evaluating the eigenvalues of the Jacobian, we now switch to a new coordinate system in which the Jacobian will have a familiar shape. This is possible since fixed points and their stability do not depend on the choice of coordinates. We define:

$$\overline{\boldsymbol{w}}' \equiv \boldsymbol{K}^{-1/2} \overline{\boldsymbol{w}}, \quad \Rightarrow \overline{\boldsymbol{w}} = \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}', \tag{100}$$

from which the weight dynamics can be written as

$$\dot{\overline{\boldsymbol{w}}}' = \boldsymbol{K}^{-1/2} \dot{\overline{\boldsymbol{w}}} = \boldsymbol{K}^{-1/2} \overline{\tau}^{-1} \left[\mathbb{1} - \frac{\boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_E \overline{\boldsymbol{c}}_E{}^\mathsf{T}}{\overline{\boldsymbol{c}}_E{}^\mathsf{T} \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_E} - \frac{\boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_I \overline{\boldsymbol{c}}_I{}^\mathsf{T}}{\overline{\boldsymbol{c}}_I{}^\mathsf{T} \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_I} \right] \boldsymbol{K}^{-1/2} \boldsymbol{K}^{1/2} \overline{\boldsymbol{C}} \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}', \tag{101}$$

where we inserted $\boldsymbol{K}^{-1/2} \boldsymbol{K}^{1/2} = \mathbb{1}$. We now make use of the following identities:

$$\overline{\boldsymbol{c}}_A{}^\mathsf{T} \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_A = k_A^{1/2} \overline{\boldsymbol{c}}_A{}^\mathsf{T} \overline{\boldsymbol{w}}'_A, \quad \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'_A \overline{\boldsymbol{c}}_A{}^\mathsf{T} = k_A^{1/2} \overline{\boldsymbol{w}}'_A \overline{\boldsymbol{c}}_A{}^\mathsf{T}, \quad \frac{\overline{\boldsymbol{w}}'_A \overline{\boldsymbol{c}}_A{}^\mathsf{T}}{\overline{\boldsymbol{c}}_A{}^\mathsf{T} \overline{\boldsymbol{w}}'_A} \boldsymbol{K}^{-1/2} = \boldsymbol{K}^{-1/2} \frac{\overline{\boldsymbol{w}}'_A \overline{\boldsymbol{c}}_A{}^\mathsf{T}}{\overline{\boldsymbol{c}}_A{}^\mathsf{T} \overline{\boldsymbol{w}}'_A}, \quad \forall A \in \{E, I\}. \tag{102}$$

We find that the $\boldsymbol{K}^{1/2}$ matrices inside the bracket cancel, and we can pull $\boldsymbol{K}^{-1/2}$ from the right side to the left side of the bracket:

$$\dot{\overline{\boldsymbol{w}}} = \boldsymbol{K}^{-1/2} \dot{\overline{\boldsymbol{w}}} = \boldsymbol{K}^{-1/2} \overline{\tau}^{-1} \boldsymbol{K}^{-1/2} \left[\mathbb{1} - \frac{\overline{\boldsymbol{w}}'_E \overline{\boldsymbol{c}}_E{}^\mathsf{T}}{\overline{\boldsymbol{c}}_E{}^\mathsf{T} \overline{\boldsymbol{w}}'_E} - \frac{\overline{\boldsymbol{w}}'_I \overline{\boldsymbol{c}}_I{}^\mathsf{T}}{\overline{\boldsymbol{c}}_I{}^\mathsf{T} \overline{\boldsymbol{w}}'_I} \right] \boldsymbol{K}^{1/2} \overline{\boldsymbol{C}} \boldsymbol{K}^{1/2} \overline{\boldsymbol{w}}'. \tag{103}$$

We introduce the following definitions

$$\overline{\tau}' = \overline{\tau} \boldsymbol{K}, \quad \overline{\boldsymbol{C}}' = \boldsymbol{K}^{1/2} \overline{\boldsymbol{C}} \boldsymbol{K}^{1/2} = \left\langle \begin{pmatrix} k_E^{1/2} \mathbf{y}_E \\ k_I^{1/2} \mathbf{y}_I \end{pmatrix} \left(k_E^{1/2} \mathbf{y}_E{}^\mathsf{T}, -k_I^{1/2} \mathbf{y}_I{}^\mathsf{T} \right)^\mathsf{T} \right\rangle. \tag{104}$$

Note that $\overline{\boldsymbol{C}}'$ is *not* the modified covariance matrix expressed in the new coordinate system but a new modified covariance matrix that corresponds to an altered input space where excitatory and inhibitory input firing rates $\boldsymbol{y}_E, \boldsymbol{y}_I$ are scale by $k_E^{1/2}, k_I^{1/2}$, respectively. In summary, we can write the plasticity of the weight vector in the new coordinate system as[1]

$$\boxed{\overline{\tau}' \dot{\overline{\boldsymbol{w}}}' = \left[\mathbb{1} - \frac{\overline{\boldsymbol{w}}'_E \overline{\boldsymbol{c}}_E{}^\mathsf{T}}{\overline{\boldsymbol{c}}_E{}^\mathsf{T} \overline{\boldsymbol{w}}'_E} - \frac{\overline{\boldsymbol{w}}'_I \overline{\boldsymbol{c}}_I{}^\mathsf{T}}{\overline{\boldsymbol{c}}_I{}^\mathsf{T} \overline{\boldsymbol{w}}'_I} \right] \overline{\boldsymbol{C}}' \overline{\boldsymbol{w}'}}. \tag{105}$$

We are interested in the stability of the fixed points given in Eq. 98. In the new coordinate system, they become

$$\overline{\boldsymbol{w}}'^* = \boldsymbol{K}^{-1/2} \overline{\boldsymbol{w}}^* = \boldsymbol{K}^{-1/2} \boldsymbol{K} \overline{\boldsymbol{v}}^* = \boldsymbol{K}^{1/2} \overline{\boldsymbol{v}}^*. \tag{106}$$

It is straightforward to proof that $\overline{\boldsymbol{w}}'^*$ is an eigenvector of the new modified covariance matrix $\overline{\boldsymbol{C}}'$ with eigenvalue

[1] Here, $\overline{\tau}^{-1}$ and $\boldsymbol{K}^{-1/2}$ are both diagonal matrices and commute.

$\bar{\lambda}'^* = (k_E - k_I a^{*2})\lambda^*$: With $\bar{\boldsymbol{C}}$ defined in Eq. 51 we get

$$\bar{\boldsymbol{C}}'\bar{\boldsymbol{w}}'^* = \boldsymbol{K}^{1/2}\bar{\boldsymbol{C}}\boldsymbol{K}^{1/2}\boldsymbol{K}^{1/2}\bar{\boldsymbol{v}}^* \quad [107]$$

$$= \boldsymbol{K}^{1/2}\begin{pmatrix} \boldsymbol{C} & -\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{A} \\ \boldsymbol{A}^\top\boldsymbol{\Lambda}\boldsymbol{V}^\top & -\boldsymbol{A}^\top\boldsymbol{\Lambda}\boldsymbol{A} \end{pmatrix}\begin{pmatrix} \boldsymbol{V}\boldsymbol{e}^* k_E \\ \boldsymbol{A}^\top\boldsymbol{e}^* k_I \end{pmatrix} \quad [108]$$

$$= \boldsymbol{K}^{1/2}\begin{pmatrix} \boldsymbol{V}\boldsymbol{e}^*\lambda^* k_E - \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{e}^* k_I \\ \boldsymbol{A}^\top\boldsymbol{e}^*\lambda^* k_E - \boldsymbol{A}^\top\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{e}^* k_I \end{pmatrix} \quad [109]$$

$$= \boldsymbol{K}^{1/2}\begin{pmatrix} \boldsymbol{V}\boldsymbol{e}^* \\ \boldsymbol{A}^\top\boldsymbol{e}^* \end{pmatrix}(k_E - k_I a^{*2})\lambda^* = \bar{\boldsymbol{w}}'^*\bar{\lambda}'^*, \quad [110]$$

where we defined a^{*2} as the entry of the diagonal matrix $\boldsymbol{A}\boldsymbol{A}^\top$ that corresponds to the eigenvector $\boldsymbol{v}^*$[1]. are a regular eigenvectors of a new modified covariance matrix $\bar{\boldsymbol{C}}'$. Note that this is independent of the change of variables, however, only in the new coordinate system one can identify the new modified covariance matrix with an actual input space[2], where pre-synaptic firing rates are scaled by k_E, k_I (Eq. 104). In theory, we can now proceed in finding the eigenvalues of the Jacobian[3], as explained in Section 2.2. As before, one finds that stability is largely determined by the eigenvalues of the modified covariance matrix, which now are $\bar{\lambda}'$.

Apart from providing a principled way to determine if a non-eigenvector fixed point is stable, our formulation provides additional insight: Let's assume the total synaptic inhibitory weight of a neuron is very small, much smaller than any eigenvector of $\bar{\boldsymbol{C}}$ 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 repulsion of the inhibitory neurons (cf. Eq. 104 for $k_I \ll 1$), and its stability would be primarily determined by the excitatory attraction of the different eigenvector modes, i.e., $\bar{\lambda}' = (k_E - k_I a^{*2})\lambda \approx \lambda$. 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.

While the effective plasticity timescale $\bar{\boldsymbol{\tau}}' = \bar{\boldsymbol{\tau}}\boldsymbol{K}$ in Eq. 105 depends on the magnitude of the excitatory and the inhibitory part of the specific fixed point under consideration, this does *not* mean that the speed of synaptic plasticity is different from the original formulation in Eq. 45. For example, when we consider a fixed point with a decreased inhibitory weight norm $k_I < 1$, the effective inhibitory plasticity appears to increase, since $\bar{\tau}'_I = \tau_I k_I$. However, this effect is balanced by the decrease in pre-synaptic inhibitory firing rates, which decreases with decreasing k_I[4]. Similarly, the coordinate system in which we describe the weight dynamics also does not affect the speed of plasticity[5].

3 Lateral input stretches and compresses the feedforward input space

Before we consider how synapse-type-specific Hebbian plasticity affects learning in fully plastic recurrent networks, we first build additional intuition for how static lateral input affects the weight dynamics. From the previous section we know that in this case the eigenvalues of the modified covariance matrix are the key factors that determine fixed point stability, and from Sections 1.1 & 2.1.2 we know that these eigenvalues describe the Hebbian growth towards the corresponding eigenvector that can be attractive or repulsive, corresponding to a positive or negative eigenvalue. When a neuron receives only feedforward excitatory input (Fig. S2*A*), the weight dynamics is described by a true covariance matrix with eigenvalues equal to the variances along the principal components of the feedforward input space (cf. Sec. 1). Then the weight vector in the fixed point aligns with the direction of maximal variance in the input space (Fig.1*G*). In the following, we introduce a similar perspective and show that additional lateral input can be interpreted to stretch and compress the original feedforward input space, while the feedforward component of the weight vector performs PCA on this modified input space.

[1] $a^{*2} = \boldsymbol{a}^{*\top}\boldsymbol{a}^*$, cf. Eq. 80.

[2] The new modified covariance matrix in the original coordinates is $\boldsymbol{K}^{1/2}\bar{\boldsymbol{C}}'\boldsymbol{K}^{-1/2} = \boldsymbol{K}\bar{\boldsymbol{C}}$, with eigenvectors $\boldsymbol{K}\bar{\boldsymbol{v}}^*$.

[3] We would have to employ the eigenvector basis of the new modified covariance matrix $\bar{\boldsymbol{V}}' = \boldsymbol{K}^{1/2}\bar{\boldsymbol{V}}$ for triangularization.

[4] From Eq. 103 we see that we can freely shift scaling matrices $\boldsymbol{K}$ between the modified covariance matrix and the plasticity timescale by pulling diagonal matrices of the same shape as $\boldsymbol{K}$ through the bracket (cf. Eq. 102).

[5] A change in the overall weight norms, however, can affect the magnitude of postsynaptic activities and synaptic changes.

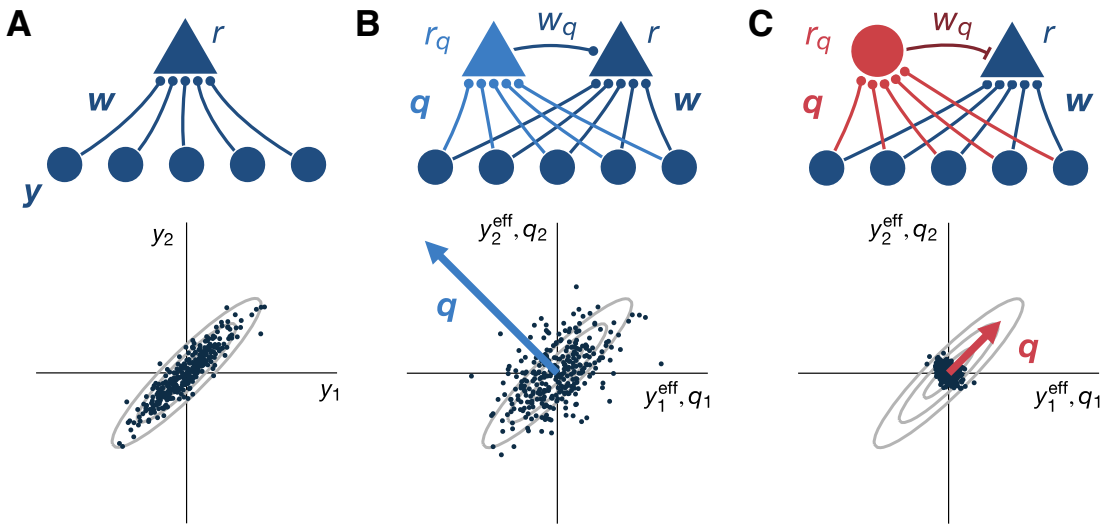

Figure S2: Input space modification due to lateral input. (*A*) Top: a single neuron with firing rate r receives synaptic inputs $\boldsymbol{w}$ from a population of excitatory neurons $\boldsymbol{y}$. Bottom: input distribution projected onto the first two input dimensions. Each dot represents the 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 (cf. Sec. 1). (*B*) Top: Same as in *A* for a neuron that receives additional input w_q from a laterally projecting excitatory neuron r_q which is tuned to an eigenvector $\boldsymbol{q}$ of the original input covariance matrix. Bottom: the effective input space $\boldsymbol{y}^{\text{eff}}$ of the target neuron (dark blue triangle) is warped such that the variance along the eigenvector $\boldsymbol{q}$ (blue arrow) is stretched in proportion to the absolute value of the weight vector $\boldsymbol{q}$. 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.

We consider a circuit of two neurons that both receive feedforward input from a population of input neurons $\boldsymbol{y}$ (Fig. S2*B*, top). Let the first neuron have a fixed, non-plastic set of feedforward weights $\boldsymbol{q}$ and firing rate

$$r_q = \boldsymbol{q}^\mathsf{T}\boldsymbol{y} \tag{111}$$

We let the first neuron project laterally onto the second neuron via a synaptic weight w_q, without receiving any lateral input itself. Then the equilibrium firing rate of the second neuron is

$$r = w_q r_q + \boldsymbol{w}^\mathsf{T}\boldsymbol{y} \tag{112}$$

where we assume that both $\boldsymbol{w}$ and w_q are plastic according to a stabilized Hebbian rule.

From the perspective of the second neuron, the input space is increased by one dimension due to the additional lateral input, i.e., we can write Eq. 112 as

$$r = \bar{\boldsymbol{w}}^\mathsf{T}\bar{\mathbf{y}}, \qquad \bar{\mathbf{y}} = \left(\mathbf{y}^\mathsf{T}, \mathbf{q}^\mathsf{T}\mathbf{y}\right)^\mathsf{T}, \quad \bar{\boldsymbol{w}} = \left(\boldsymbol{w}^\mathsf{T}, w_q\right)^\mathsf{T}, \tag{113}$$

where, we defined the new input vector $\bar{\boldsymbol{y}}$ and the combined input weights $\bar{\boldsymbol{w}}$. Effectively this is still a feedforward network without feedback, and the static covariance matrix $\bar{\mathbf{C}}$ of the new inputs $\bar{\mathbf{y}}$ fully determines the average synaptic weight dynamics[1]:

$$\bar{\mathbf{C}} = \left\langle \bar{\mathbf{y}}\bar{\mathbf{y}}^\mathsf{T} \right\rangle = \left\langle \begin{pmatrix} \mathbf{y}\mathbf{y}^\mathsf{T} & \mathbf{y}\mathbf{y}^\mathsf{T}\mathbf{q} \\ \mathbf{q}^\mathsf{T}\mathbf{y}\mathbf{y}^\mathsf{T} & \mathbf{q}^\mathsf{T}\mathbf{y}\mathbf{y}^\mathsf{T}\mathbf{q} \end{pmatrix} \right\rangle = \begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{q} \\ \mathbf{q}^\mathsf{T}\mathbf{C} & \mathbf{q}^\mathsf{T}\mathbf{C}\mathbf{q} \end{pmatrix}, \tag{114}$$

where $\boldsymbol{C}$ is the covariance matrix of the original input $\boldsymbol{y}$. We are interested in the eigenvectors $\bar{\boldsymbol{v}}$ and eigenvalues $\bar{\lambda}$ of this matrix for two reasons. First, because they describe the attraction towards different input modes due to the Hebbian term in our competitive plasticity rule. Second, because eigenvectors are fixed points with their stability mostly determined by the eigenvalues. Eigenvectors and eigenvalues must satisfy

$$\bar{\mathbf{C}}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}, \tag{115}$$

$$\begin{pmatrix} \mathbf{C} & \mathbf{C}\mathbf{q} \\ \mathbf{q}^\mathsf{T}\mathbf{C} & \mathbf{q}^\mathsf{T}\mathbf{C}\mathbf{q} \end{pmatrix} \begin{pmatrix} \mathbf{v}_F \\ v_q \end{pmatrix} = \bar{\lambda} \begin{pmatrix} \mathbf{v}_F \\ v_q \end{pmatrix}. \tag{116}$$

[1] In this case, $\bar{\boldsymbol{C}}$ is a true covariance matrix, since the lateral projecting neuron is excitatory.

where v_q and $\boldsymbol{v}_F$ are the lateral and the feedforward components of the eigenvector, respectively. In the following, we focus on the feedforward component for different $\boldsymbol{q}$. It follows

$$\mathbf{C}\mathbf{v}_F + \mathbf{C}\mathbf{q}v_q = \bar{\lambda}\mathbf{v}_F, \tag{117}$$

$$\mathbf{q}^\mathsf{T}\mathbf{C}\mathbf{v}_F + \mathbf{q}^\mathsf{T}\mathbf{C}\mathbf{q}v_q = \bar{\lambda}v_q. \tag{118}$$

$$\mathbf{C}\left(\mathbf{v}_F + \mathbf{q}v_q\right) = \bar{\lambda}\mathbf{v}_F, \tag{119}$$

$$\mathbf{q}^\mathsf{T}\mathbf{C}\left(\mathbf{v}_F + \mathbf{q}v_q\right) = \bar{\lambda}v_q. \tag{120}$$

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

$$\boxed{\mathbf{C}\left(\mathbb{1} + \mathbf{q}\mathbf{q}^\mathsf{T}\right)\mathbf{v}_F = \bar{\lambda}\mathbf{v}_F.} \tag{121}$$

This is again an eigenvector equation, where the feedforward components of the original eigenvector $\boldsymbol{v}_F$, are themselves eigenvectors of the matrix $\mathbf{C}(\mathbb{1} + \mathbf{q}\mathbf{q}^\mathsf{T})$. Note that for $\boldsymbol{q} = \mathbf{0}$ we recover the case without lateral projections and feedforward components are multiples of eigenvectors of $\boldsymbol{C}$ with attractions $\bar{\lambda} = \lambda$. For general $\boldsymbol{q}$, the solution is not straightforward: We consider the equation in the input eigenspace, where Eq. 121 becomes

$$\Lambda\left(\mathbb{1} + \boldsymbol{q}_v\boldsymbol{q}_v{}^\mathsf{T}\right)\boldsymbol{v}_{Fv} = \bar{\lambda}\boldsymbol{v}_{Fv}, \tag{122}$$

with Λ being the diagonal matrix of feedforward eigenvalues λ and the subscript $(\cdot)_v$ indicates a vector in the eigenbasis of $\boldsymbol{C}$. In this basis, eigenvectors of $\boldsymbol{C}$ are unit vectors, i.e., $\boldsymbol{v}_v = \boldsymbol{e}$, where $\boldsymbol{e}$ is a vector of zeros with a single entry equal to one, corresponding to the respective eigenvector. When $\boldsymbol{q}$ contains components of more than one eigenvector, the matrix $\boldsymbol{q}_v\boldsymbol{q}_v{}^\mathsf{T}$ is not diagonal and eigenvectors of $\boldsymbol{C}$, do not solve the equation. Here we consider a simplified case: When the first neuron had plastic feedforward input, we know from Section 1 that it would converge to a multiple of an eigenvector of the feedforward covariance matrix[1], $\boldsymbol{q} \propto \boldsymbol{v}^\dagger$, with $\boldsymbol{C}\boldsymbol{v}^\dagger = \lambda^\dagger\boldsymbol{v}^\dagger$. Then, $\boldsymbol{q}_v\boldsymbol{q}_v{}^\mathsf{T} = e^\dagger e^{\dagger\mathsf{T}}$ is diagonal with a single non-zero entry, and it is obvious that feedforward eigenvector components of $\bar{\boldsymbol{C}}$ are eigenvectors of the feedforward covariance matrix $\boldsymbol{C}$ that solve Eq. 121.

To find the eigenvalues $\bar{\lambda}$, we first consider feedforward eigenvector components $\boldsymbol{v}_F$ that are orthogonal to $\boldsymbol{q}$:

$$\boldsymbol{v}_F \propto \boldsymbol{v}^\ddagger \perp \boldsymbol{q} \propto \boldsymbol{v}^\dagger \quad \Rightarrow \quad \boldsymbol{q}^\mathsf{T}\boldsymbol{v}_F \propto \boldsymbol{v}^{\dagger\mathsf{T}}\boldsymbol{v}^\ddagger = 0, \quad \Rightarrow \quad \bar{\lambda}^\ddagger = \lambda^\ddagger = \sigma^{\ddagger 2}. \tag{123}$$

Therefore, input modes that are orthogonal to the tuning of the laterally projecting neuron maintain their attractions, equal to the respective eigenvalues of $\boldsymbol{C}$ that is equal to the variances $\sigma^{\ddagger 2}$ of the input distribution along the respective orthogonal eigenvectors. Therefore, the laterally projecting neuron does not affect the Hebbian growth dynamics in the input subspace orthogonal to $\boldsymbol{q}$[2]. The remaining feedforward eigenvector component is proportional to $\boldsymbol{q}$:

$$\boldsymbol{v}_F \propto \boldsymbol{v}^\dagger \propto \boldsymbol{q} \quad \Rightarrow \quad \boldsymbol{q}^\mathsf{T}\boldsymbol{v}_F \propto \boldsymbol{v}^{\dagger\mathsf{T}}\boldsymbol{v}^\dagger = 1, \quad \Rightarrow \quad \bar{\lambda}^\dagger = \lambda^\dagger + \lambda^\dagger a_q^2 = \sigma^{\dagger 2} + \sigma_q^2, \tag{124}$$

where a_q is equal to $\|\boldsymbol{q}\|$, the L2-norm of $\boldsymbol{q}$, and $\sigma_q^2 = \lambda^\dagger a_q^2$ is the firing rate variance of the laterally projecting neuron. Therefore, the second neuron adjusts its feedforward weights $\boldsymbol{w}$ as if the variance along the eigenvector $\boldsymbol{q}$ was increased by σ_q^2 (Fig. S2*B*, bottom). In that sense, the second neuron 'perceives' its feedforward input space as stretched and we speak of a modified input space (cf. Sec. 2.2.1) that is described by a modified covariance matrix $\bar{\boldsymbol{C}}$. We note that it is possible to choose $\boldsymbol{q}$ such that an arbitrary direction of the input space becomes stable. For $\boldsymbol{q} = \boldsymbol{C}^{-1}\boldsymbol{h}$ Eq. 121 is[3]

$$\left(\mathbf{C} + \mathbf{h}\mathbf{h}^\mathsf{T}\boldsymbol{C}^{-1}\right)\mathbf{v}_F = \bar{\lambda}\mathbf{v}_F. \tag{125}$$

For increasing $\|\boldsymbol{h}\|$, the principal eigenvector transitions from $\boldsymbol{v}_F \propto \boldsymbol{v}$, for $\|\boldsymbol{h}\| = 0$, to $\boldsymbol{v}_F^\infty \propto \boldsymbol{h}$ for $\|\boldsymbol{h}\| \to \infty$. In the following, we only consider the case when $\boldsymbol{q}$ is parallel to one of the eigenvectors of $\boldsymbol{C}$. Then, for sufficiently large

[1] More precisely, $\boldsymbol{q}$ would converge to a multiple of the principal eigenvector. Here, we consider the more general case where $\boldsymbol{q}$ is proportional to an arbitrary eigenvector. We will see that with suitable lateral input, any feedforward eigenvector can be stable.

[2] However, the constraint term in the weight dynamics introduces interactions between the subspaces orthogonal and parallel to $\boldsymbol{q}$.

[3] Note that $\boldsymbol{C}^{-1} = (\boldsymbol{C}^{-1})^\mathsf{T}$ since $\boldsymbol{C}$ is a true covariance matrix, i.e., $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ are symmetric.

a_q and $\boldsymbol{q} \propto \boldsymbol{v}^\dagger$, an arbitrary non-principle eigenvector $\boldsymbol{v}^\dagger$ with attraction $\bar{\lambda}^\dagger = \lambda^\dagger(1 + a_q^2)$ can become stable. In that case, the corresponding fixed point is of the following shape[1]:

$$\Rightarrow \quad \bar{\boldsymbol{w}}^* = \begin{pmatrix} \boldsymbol{w}^* \\ w_q^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{w}^* \\ \boldsymbol{q}^\mathsf{T} \boldsymbol{w}^* \end{pmatrix} \propto \begin{pmatrix} \boldsymbol{v}^\dagger \\ a_q \end{pmatrix}, \tag{126}$$

When the laterally projecting neuron is inhibitory (Fig. S2*C*, top), the modified covariance matrix becomes (cf. Eq. 51)

$$\bar{\boldsymbol{C}} = \begin{pmatrix} \mathbf{C} & -\mathbf{Cq} \\ \mathbf{q}^\mathsf{T}\mathbf{C} & -\mathbf{q}^\mathsf{T}\mathbf{Cq} \end{pmatrix}, \tag{127}$$

and it follows that the input space is compressed along $\boldsymbol{q} \propto \boldsymbol{v}^\dagger$ (Fig. S2*C*, bottom):

$$\bar{\lambda}^\dagger = \lambda^\dagger - \lambda^\dagger a_q^2 = \sigma^{\dagger 2} - \sigma_q^2. \tag{128}$$

In the case of lateral inhibition and sufficiently large vector norms a_q, an eigenvector can become repulsive, i.e., its eigenvalue becomes negative. Geometrically, this corresponds to a reflection of the input space along $\boldsymbol{q}$ through the origin, which can no longer be visualized as intuitively as in Fig. S2.

We can generalize the overall approach to multiple excitatory and inhibitory neurons such that the effective attraction towards a feedforward eigenvector becomes

$$\bar{\lambda} = \lambda \left(1 + \|\boldsymbol{a}_E\|^2 - \|\boldsymbol{a}_I\|^2\right), \tag{129}$$

$$\Rightarrow \quad \boxed{\bar{\lambda} = \sigma^2 + \|\boldsymbol{\sigma}_E\|^2 - \|\boldsymbol{\sigma}_I\|^2}, \tag{130}$$

where $\lambda = \sigma^2$, the vectors $\boldsymbol{a}_E$, $\boldsymbol{a}_I$ hold the feedforward vector norms of the laterally projecting neurons that are tuned to the respective feedforward eigenvector, and $\boldsymbol{\sigma}_A = \lambda \boldsymbol{a}_A$, $A \in \{E, I\}$, hold the standard deviations of their firing rates. This allows writing the regular fixed points as[2]

$$\boxed{\bar{\boldsymbol{w}}^* = \begin{pmatrix} \boldsymbol{w}^* \\ \boldsymbol{w}_E^* \\ \boldsymbol{w}_I^* \end{pmatrix} \propto \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{a}_E \\ \boldsymbol{a}_I \end{pmatrix} = \sigma^{-1} \begin{pmatrix} \sigma_{\boldsymbol{v}} \\ \boldsymbol{\sigma}_E \\ \boldsymbol{\sigma}_I \end{pmatrix}}, \tag{131}$$

This implies that for regular fixed points, the total synaptic weight is distributed among synapses in proportion to the standard deviation of their pre-synaptic activities. Note that the different weight norms of the excitatory and inhibitory part of non-eigenvector fixed points can distort this relation (cf. Sec. 2.1.3).

In summary, we demonstrated how static lateral input can be interpreted to reshape the feedforward attraction landscape of afferent neurons. Note that these results are independent of what causes the laterally projecting neurons' tuning. The second, afferent neuron does not 'see' what inputs to the laterally projecting neurons cause their tuning For example, in addition to feedforward input, laterally projecting neurons might be integrated into a recurrent circuit of neurons that are all tuned to the same eigenvector[3]. Then $\boldsymbol{\sigma}_E^2, \boldsymbol{\sigma}_I^2$ result from recurrent interaction in addition to the norm of the feedforward weight vectors. However, for the dynamics of the second neuron, it would not make any difference as long as the firing rate statistics of its pre-synaptic inputs were the same. In the following sections, we will consider circuits where the firing rate statistics emerge from recurrent interactions.

4 Eigencircuits

In the previous section we considered neurons that receive feedforward input from an excitatory population and lateral input from neurons with fixed feedforward tuning (Fig. S2). We found that the attraction of different feedforward input modes is determined by the eigenvalues of a modified covariance matrix, composed of a feedforward

[1] If a_q is too small so that $\bar{\lambda}^\dagger < \bar{\lambda}^\ddagger = \lambda^\ddagger$, the principal feedforward eigenvector $\boldsymbol{v}^\ddagger$ of $\boldsymbol{C}$ with eigenvalue $\lambda^\ddagger$ is stable and $\bar{w}^* = (\boldsymbol{v}^{\ddagger\mathsf{T}}, 0)^\mathsf{T}$.

[2] If none of the laterally projecting neurons is tuned to a specific feedforward eigenvector $\boldsymbol{v}^\ddagger$, i.e., $\boldsymbol{v}^\ddagger \perp \boldsymbol{q}_i \ \forall i$, the corresponding fixed point becomes $\bar{v}^\ddagger = (\boldsymbol{v}^{\ddagger\mathsf{T}}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T})^\mathsf{T}$.

[3] 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.

contribution and a contribution due to the laterally projecting neurons that is proportional to the variance of their firing rates (Eq. 130). In this section, we consider networks of recurrently connected, laterally projecting neurons and explore the variances of their firing rates.

First, we consider a network of excitatory and inhibitory neurons $\boldsymbol{y}_E, \boldsymbol{y}_I$ that are laterally connected to themselves and each other and receive feedforward input from the same excitatory population $\boldsymbol{y}$. We assume that the activity in the network is dominated by feedforward input such that neurons become selective for different eigenvectors of the feedforward covariance matrix $\boldsymbol{C} = \langle \boldsymbol{y}\boldsymbol{y}^\mathsf{T} \rangle$, e.g., the steady state firing rate y_a of a neuron that is tuned to an eigenvector $\boldsymbol{v}_a$ is proportional to $\boldsymbol{v}_a{}^\mathsf{T}\boldsymbol{y}$ (Fig. 4*A*), where the proportionality factor depends on the number and firing rates of other neurons that are tuned to the same eigenvector (see Sec. 4.1). Then the average Hebbian growth of a synapse that connects two neurons that are tuned to different eigenvectors is[1]:

$$\langle \dot{w}_{ab} \rangle \propto \langle y_a y_b \rangle \propto \langle \boldsymbol{v}_a{}^\mathsf{T}\boldsymbol{y}\boldsymbol{y}^\mathsf{T}\boldsymbol{v}_b \rangle = \boldsymbol{v}_a{}^\mathsf{T}\boldsymbol{C}\boldsymbol{v}_b = \lambda_b \boldsymbol{v}_a{}^\mathsf{T}\boldsymbol{v}_b = 0. \qquad [132]$$

Due to the competition for synaptic resources, the synapse loses out to the non-zero growth of synapses that connect neurons that are tuned to the same eigenvector, and decays to zero over time (Fig. 4*B*). Eventually, the circuit is separated into sub-circuits that are tuned to different eigenvectors with recurrent connections within, but not between sub-circuits. Since there is one sub-circuit per eigenvector of the feedforward covariance matrix, we call these decoupled sub-circuits 'eigencircuits' (cf. Fig. 4).

4.1 Variance propagation

In Section 3 we have seen that the attraction and the stability of a feedforward eigenvector is closely related to the firing rate variances of laterally projecting neurons, independent from how these variances arise. In the effective feedforward circuits that we considered, it was straightforward to compute variances based on feedforward weight norms (Eq.130 f.). We now show how variances can be determined in recurrent eigencircuits, which allows to quantify the effective attraction of an input mode.

We consider a generic eigencircuit and investigate how variances propagate through the network, i.e., our goal is to express the standard deviation of a neuron's firing rate as a function of the standard deviations of its pre-synaptic input firing rates. For a neuron in an eigencircuit, all pre-synaptic inputs with non-zero synaptic weight are tuned to the same feedforward eigenvector $\boldsymbol{v}$. We only consider these non-zero entries and assume that the steady state firing rate of an arbitrary neuron can be written as (Fig. S3*A*)

$$r = \boldsymbol{w}^\mathsf{T}\boldsymbol{y} + \boldsymbol{w}_E{}^\mathsf{T}\boldsymbol{y}_E - \boldsymbol{w}_I{}^\mathsf{T}\boldsymbol{y}_I, \qquad [133]$$

$$\boldsymbol{y}_E = \boldsymbol{a}_E(\boldsymbol{v}^\mathsf{T}\boldsymbol{y}), \quad \boldsymbol{y}_I = \boldsymbol{a}_I(\boldsymbol{v}^\mathsf{T}\boldsymbol{y}), \qquad [134]$$

Note that before, $\boldsymbol{a}_E$ and $\boldsymbol{a}_I$ referred to feedforward weight norms (cf. Sec. 3). Now these vectors more generally express how firing rate variances relate to the input variance along the eigencircuit's feedforward eigenvector, without making any assumptions about how this tuning arises. We will show in Section 5 that this assumption is correct and specify how the entries of $\boldsymbol{a}_E$, $\boldsymbol{a}_I$ relate to the recurrent excitatory and inhibitory weights (cf. Eqs. 160 & 161). For the weight vectors, we require that the excitatory and inhibitory parts are normalized to maintain the total amount of inhibitory and excitatory synaptic resources:

$$\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w}_E \end{pmatrix} = W_E \frac{\boldsymbol{v}_E}{\|\boldsymbol{v}_E\|_p}, \quad \boldsymbol{w}_I = W_I \frac{\boldsymbol{v}_I}{\|\boldsymbol{v}_I\|_p}, \qquad [135]$$

$$\boldsymbol{v}_E = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{a}_E \end{pmatrix}, \quad \boldsymbol{v}_I = \boldsymbol{a}_I, \qquad [136]$$

where W_E, W_I are scalar weight norms, and $\boldsymbol{v}_E$, $\boldsymbol{v}_I$ are the excitatory and inhibitory parts of the fixed point eigenvector (cf. Sec. 5), with entries that are proportional to the pre-synaptic standard deviations (cf. Eq. 131). Then, the p-norm, $\|\cdot\|_p$, is maintained due to competition for synaptic resources[2]. For the post-synaptic firing rate, it follows

$$r = \left(\frac{1 + \|\boldsymbol{a}_E\|^2}{\|\boldsymbol{v}_E\|_p} W_E - \frac{\|\boldsymbol{a}_I\|^2}{\|\boldsymbol{v}_I\|_p} W_I \right) (\boldsymbol{v}^\mathsf{T}\boldsymbol{y}). \qquad [137]$$

[1] Since we assume Hebbian plasticity between all types of neurons, excitatory and inhibitory, we do not specify the neuron type. y_a and y_b are the firing rates of two arbitrary vectors that are part of two different eigencircuits.

[2] The vectors $(\boldsymbol{w}^\mathsf{T}, \boldsymbol{w}_E{}^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{w}_I$ are normalized such that $\left\|(\boldsymbol{w}^\mathsf{T}, \boldsymbol{w}_E{}^\mathsf{T})^\mathsf{T}\right\|_p = W_E$, and $\|\boldsymbol{w}_I\|_p = W_I$. This is achieved by scaling the excitatory and inhibitory part of the regular eigenvector, i.e., scaling $\boldsymbol{v}_E$ by $k_E = W_E/\|\boldsymbol{v}_E\|_p$, and $\boldsymbol{v}_I$ by $k_I = W_I/\|\boldsymbol{v}_I\|_p$ (cf. Sec. 2.1.3)

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

$$\sigma_r = \left(\frac{1 + \|\boldsymbol{a}_E\|^2}{\|\boldsymbol{v}_E\|_p} W_E - \frac{\|\boldsymbol{a}_I\|^2}{\|\boldsymbol{v}_I\|_p} W_I \right) \sigma = \frac{\sigma^2 + \|\boldsymbol{a}_E\|^2 \sigma^2}{\|\boldsymbol{v}_E\|_p \sigma} W_E - \frac{\|\boldsymbol{a}_I\|^2 \sigma^2}{\|\boldsymbol{v}_I\|_p \sigma} W_I, \quad [138]$$

$$\Rightarrow \quad \boxed{\sigma_r = \frac{\|\boldsymbol{\sigma}^E\|^2}{\|\boldsymbol{\sigma}^E\|_p} W_E - \frac{\|\boldsymbol{\sigma}^I\|^2}{\|\boldsymbol{\sigma}^I\|_p} W_I}, \quad [139]$$

$$\boldsymbol{\sigma}^E = \left(\sigma, \boldsymbol{\sigma}_E{}^{\mathsf{T}} \right)^{\mathsf{T}}, \quad \boldsymbol{\sigma}^I = \boldsymbol{\sigma}_I, \quad [140]$$

For a network in the steady state, i.e., when synaptic weights converged, this equation puts the standard deviation of neural firing rates in relation to each other, i.e., it provides the standard deviation of a post-synaptic neuron's activity as a function of the standard deviations of its pre-synaptic input neurons' activities[1]. It describes how standard deviations and variances 'propagate' through the network. In the next section, we will use this variance propagation equation (Eq. 139) to express the standard deviations in terms of only the weight norms and the feedforward standard deviation σ.

4.2 Consistency conditions provide eigencircuit firing rate variances

We now consider a single eigencircuit where n_E excitatory and n_I inhibitory neurons are recurrently connected, and are tuned to the same feedforward eigenvector with standard deviation σ (Fig. S3*B*). In their steady state, all neurons have to fulfil the variance propagation equation (Eq. 139). In the fully connected eigencircuit, the firing rate variance of each neuron depends on the firing rate variances of all other neurons, and all neurons have the same set of non-zero pre-synaptic inputs. This provides $N = n_E + n_I$ consistency conditions for the N unknown standard deviations. For example, the condition for a single excitatory neuron i reads

$$\sigma_E^i = W_{EE}^i \left(\frac{\sigma^2 + \|\boldsymbol{\sigma}_E\|^2}{\sigma + \|\boldsymbol{\sigma}_E\|_1} \right) - W_{EI}^i \left(\frac{\|\boldsymbol{\sigma}_I\|^2}{\|\boldsymbol{\sigma}_I\|_1} \right), \quad [141]$$

where we chose the L1-norm, $p = 1$, for normalization (but see Sec. 4.3), and $W_{AB},\ A, B \in \{E, I\}$ are the total synaptic weight that a neuron of type A receives from neurons of type B. We make the simplifying assumption that all neurons have similar weight norms, i.e., $W_{AB}^i \approx W_{AB},\ \forall i,\ A, B \in \{E, I\}$. Then, also the standard deviations of their activities are similar, and we approximate $\sigma_A^i \approx \sigma_A,\ \forall i,\ A \in \{E, I\}$:

$$\|\boldsymbol{\sigma}_A\|^2 = \sum_i \sigma_A^{i2} \approx n_A \sigma_A^2, \quad \Rightarrow \quad \frac{\sigma^2 + \|\boldsymbol{\sigma}_A\|^2}{\sigma + \|\boldsymbol{\sigma}_A\|_1} \approx \frac{\sigma^2 + n_A \sigma_A^2}{\sigma + n_A \sigma_A}. \quad [142]$$

The standard deviations of excitatory and inhibitory neural firing rates become

$$\sigma_E = W_{EE} \left(\frac{\sigma^2 + n_E \sigma_E^2}{\sigma + n_E \sigma_E} \right) - W_{EI} \left(\frac{n_I \sigma_I^2}{n_I \sigma_I} \right), \qquad \sigma_I = W_{IE} \left(\frac{\sigma^2 + n_E \sigma_E^2}{\sigma + n_E \sigma_E} \right) - W_{II} \left(\frac{n_I \sigma_I^2}{n_I \sigma_I} \right). \quad [143]$$

After some algebra, this yields the standard deviations as

$$\boxed{\sigma_I = \frac{W_{IE}}{1 + W_{II}} \frac{1}{\Phi} \sigma_E}, \quad \Phi \equiv \left[W_{EE} - \frac{W_{EI} W_{IE}}{1 + W_{II}} \right], \quad [144]$$

$$\Rightarrow \quad \boxed{\sigma_E = \frac{1}{2\,(1 - \Phi)\, n_E} \left(-1 \pm \sqrt{1 + 4\Phi\,(1 - \Phi)\, n_E} \right) \sigma}. \quad [145]$$

This provides standard deviations as a function of the number of neurons in the eigencircuit[2], n_E, n_I, their weight norms, W_{AB}, and the standard deviation of the feedforward input activity along the corresponding eigenvector, σ.

[1] Note that we allow self-excitation and self-inhibition, i.e., in a fully connected recurrent network, σ_r also appears on the right sides of the equation, as an entry of $\boldsymbol{\sigma}^E$ or $\boldsymbol{\sigma}^I$.

[2] Note that for $0 > \Phi < 1$ there exists a real solution for σ_E, independent of n_E.

Via Eq. 130, we can determine how the eigencircuit modifies the attraction of its feedforward eigenvector, i.e., the effective attraction $\bar{\lambda}$ is

$$\bar{\lambda} = \sigma^2 + n_E\sigma_E^2 - n_I\sigma_I^2 \equiv \lambda + \lambda_{\text{eig}}, \tag{146}$$

where we defined the attraction of the eigencircuit, λ_{eig}, and λ is the attraction of the respective feedforward eigenvector. In the following, we refer to $\bar{\lambda}$ interchangeably as the effective attraction of the eigencircuit or the effective attraction of the feedforward input mode.

In summary, we assumed that neurons are tuned to feedforward eigenvectors (Eq.134) and showed how the network decomposes into recurrent eigencircuits. We demonstrated how variances propagate through such eigencircuits, and quantified how eigencircuits modify the attraction of their feedforward eigenvector (cf. Sec. 3) by laterally projecting onto other neurons (cf. Fig. 4*C*). In the following (Sec. 5), we will show that eigencircuits are indeed stable fixed points of fully plastic recurrent networks and investigate their stability.

4.3 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. 139) this corresponds to $p = 2$ which becomes

$$\sigma_r = \left\|\boldsymbol{\sigma}^E\right\| W_E - \left\|\boldsymbol{\sigma}^I\right\| W_I. \tag{147}$$

Following a similar logic as in Section 4.2, the eigencircuit consistency condition for a single inhibitory neuron becomes (cf. Eq. 141):

$$\sigma_I = \frac{W_{IE}}{1 + W_{II}} \left(\sigma^2 + \|\boldsymbol{\sigma}_E\|^2\right)^{\frac{1}{2}}, \tag{148}$$

where we once more assumed that all neurons have similar weight norms, $W^i_{AB} \approx W_{AB}, \forall i$. The variance of an excitatory neuron becomes

$$\sigma_E^2 = \Phi^2 \left(\sigma^2 + \|\boldsymbol{\sigma}_E\|^2\right) = \Phi^2 \left(\sigma^2 + n_E\sigma_E^2\right), \tag{149}$$

$$\Rightarrow \quad \boxed{\sigma_E^2 = \frac{\Phi^2}{1 - \Phi^2 n_E}\sigma^2}. \tag{150}$$

For an increasing number of excitatory neurons n_E, the firing rate 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 fulfil 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.

5 E-I networks with fully plastic recurrent connectivity

We now consider fully connected networks of excitatory and inhibitory neurons where all connections, feedforward and recurrent, are plastic according to the competitive Hebbian learning rule we introduced in Section 2. We will first show that eigencircuits are fixed points and then consider their stability with respect to a weight perturbation. Specifically, we would like to know when a neuron from one eigencircuit becomes attracted to a different feedforward eigenvector. We start with some simplifying assumptions.

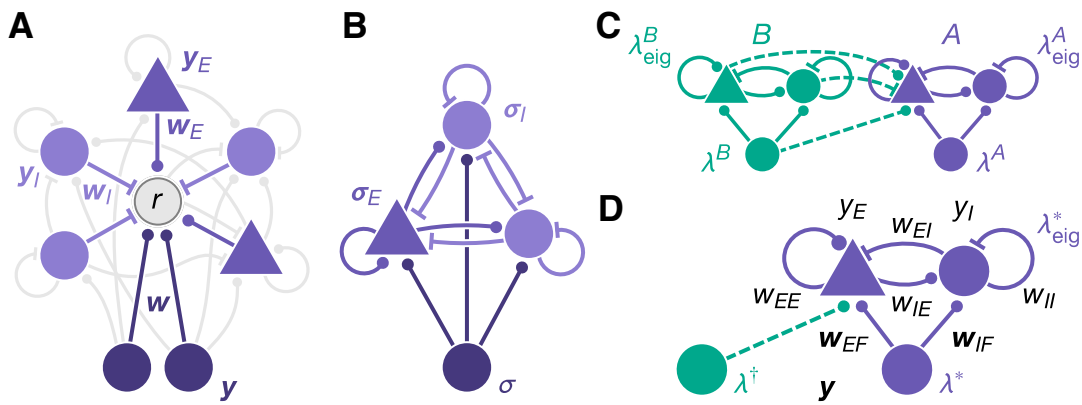

Figure S3: (*A*) A neuron with firing rate r (gray, center) receives synaptic inputs as part of a recurrent eigencircuit. The neuron receives excitatory synapses $\boldsymbol{w}$ from a population of input neurons $\boldsymbol{y}$ (dark purple, bottom). Excitatory (purple, triangles) and inhibitory neurons (light purple, circles) with firing rates, $\boldsymbol{y_E}, \boldsymbol{y_I}$, that are part of the same eigencircuit, project laterally onto neuron r via excitatory $\boldsymbol{w}_E$ and inhibitory $\boldsymbol{w}_I$ synapses. Recurrent synaptic connections that are not inputs of neuron r are shown in light gray – Not all synaptic connections are shown, for clarity. (*B*) Recurrently connected eigencircuit of $n_E = 1$ excitatory neuron (purple triangle) and $n_I = 2$ inhibitory neurons (light purple circles) that are tuned to the same feedforward eigenvector (dark purple circle, bottom). The standard deviation σ of input firing rates along the input eigenvector propagates through the network and results in firing rate standard deviations of $\boldsymbol{\sigma_E}$ and $\boldsymbol{\sigma_I}$ (cf. Eq. 144 f.). (*C*) Two excitatory neurons (triangles, top) and two inhibitory neurons (circles, top) in a recurrent circuit receive feedforward excitation from two input neurons (purple and green circles, bottom) that correspond to two different eigenvectors with eigenvalue λ^A, λ^B. Neurons are configured in a fixed point with two eigencircuits, *A* and *B*, with eigencircuit attractions λ^A_{eig}, λ^B_{eig} (cf. Eq. 146). Neurons that are part of the same eigencircuit are recurrently connected to each other. Synaptic weights between neurons that are tuned to different eigencircuits are zero. The excitatory neuron in eigencircuit *A* is perturbed in the direction of eigencircuit *B* (dashed lines). (*D*) Equivalent circuit with one excitatory and one inhibitory neuron. We consider a fixed point, where both neurons are tuned to the same feedforward eigenvector with eigenvalue λ^*. The neurons form an eigencircuit with attraction λ^*_{eig}. The excitatory neuron is perturbed in the direction of another feedforward eigenvector with attraction $\lambda^\dagger$ (dashed line). Firing rates, $y_E, y_I, \boldsymbol{y}$, and recurrent and feedforward synaptic weights, $w_{EE}, w_{EI}, w_{IE}, w_{II}, \boldsymbol{w}_{EF}, \boldsymbol{w}_{IF}$, are shown in black (cf. Eqs.152 & 152). See text for details.

Since each neuron can be bidirectionally connected to all other neurons, the dimensionality of the weight dynamics grows quadratically with the number of neurons. We are only interested in the general principles and consider a simplified circuit of two excitatory and two inhibitory neurons. One possible fixed point configuration is shown in Figure S3*C* (without dashed lines), where neurons are configured in two eigencircuits, *A*,*B*, with one excitatory and one inhibitory neuron per eigencircuit[1]. In this fixed point, all neurons receive feedforward input from a population of input neurons but synapses that connect neurons of different eigencircuits are zero (cf. Sec. 4). When a neuron in eigencircuit *A* is perturbed towards the other eigencircuit *B* (Fig. S3*C*, dashed lines), the tuning and the firing rates of all neurons in eigencircuit *A* change. However, neurons in eigencircuit *B* are unaffected because there are no connections projecting from eigencircuit *A* to eigencircuit *B*. Therefore, we only consider the recurrence within eigencircuit *A*, and think of input from other eigencircuits as effectively feedforward and static: That is, we construct an equivalent circuit where we perturb an excitatory neuron that is part of an eigencircuit, '$*$', in the direction of another eigencircuit, '$\dagger$', that does not contain any neurons and has feedforward attraction equal to the effective attraction of eigencircuit *B*, that is[2] (Fig. S3*D*)

$$\lambda^\dagger = \bar{\lambda}^B = \lambda^B + \lambda^B_{\text{eig}}, \quad \lambda^\dagger_{\text{eig}} = 0. \tag{151}$$

The configuration and attraction of the perturbed eigencircuit '$*$' is equal to eigencircuit *A*, i.e., $\lambda^* = \lambda^A$, $\lambda^*_{\text{eig}} = \lambda^A_{\text{eig}}$. In Section 5.2.3 we will explain in more detail why these two circuits (Fig. S3*C* & *D*) are highly similar with regards to their stability.

In the equivalent circuit (Fig. S3*D*), we now consider the generic equilibrium firing rates of the $n_E = 1$ excitatory and $n_I = 1$ inhibitory neuron without taking any tuning into account (Fig. S3*D*)

$$y_E = \boldsymbol{w}_{EF}{}^\mathsf{T}\boldsymbol{y} + w_{EE}y_E - w_{EI}y_I, \tag{152}$$

$$y_I = \boldsymbol{w}_{IF}{}^\mathsf{T}\boldsymbol{y} + w_{IE}y_E - w_{II}y_I, \tag{153}$$

where $\boldsymbol{y}$ holds the firing rates of a population of N_F input neurons and we did not assume any specific tuning of the

[1] We will show in Section 5.1 that eigencircuits are in fact fixed points of the weight dynamics.
[2] See Eq. 146 for the definition of the eigencircuit attraction λ_{eig}.

feedforward weights $\boldsymbol{w}_{EF}$, $\boldsymbol{w}_{IF}$. Since the network is linear, we can solve for the firing rates:

$$y_E = \frac{1}{1 - w_{EE} + \frac{w_{EI}w_{IE}}{1+w_{II}}} \left(\boldsymbol{w}_{EF}{}^{\mathsf{T}} - \frac{w_{EI}\boldsymbol{w}_{IF}{}^{\mathsf{T}}}{1+w_{II}} \right) \boldsymbol{y} \equiv \boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{y}, \tag{154}$$

$$y_I = \frac{1}{1 + w_{II} + \frac{w_{IE}w_{EI}}{1-w_{EE}}} \left(\boldsymbol{w}_{IF}{}^{\mathsf{T}} + \frac{w_{IE}\boldsymbol{w}_{EF}{}^{\mathsf{T}}}{1-w_{EE}} \right) \boldsymbol{y} \equiv \boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{y}, \tag{155}$$

where we defined the effective feedforward vectors $\boldsymbol{a}_E$, $\boldsymbol{a}_I$. The weight dynamics is

$$\boldsymbol{\tau}\dot{\overline{\boldsymbol{w}}} = \begin{pmatrix} \dot{\boldsymbol{w}}_{EF} \\ \dot{w}_{EE} \\ \dot{w}_{EI} \\ \vdots \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} & \boldsymbol{y}y_E & -\boldsymbol{y}y_I & \\ y_E\boldsymbol{y}^{\mathsf{T}} & y_Ey_E & -y_Ey_I & \boldsymbol{0} \\ y_I\boldsymbol{y}^{\mathsf{T}} & y_Iy_E & -y_Iy_I & \\ & \boldsymbol{0} & & \ddots \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_{EF} \\ w_{EE} \\ w_{EI} \\ \vdots \end{pmatrix} - \begin{pmatrix} \gamma_E & 0 & 0 & \\ 0 & \gamma_E & 0 & \boldsymbol{0} \\ 0 & 0 & \rho_E & \\ & \boldsymbol{0} & & \ddots \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_{EF} \\ w_{EE} \\ w_{EI} \\ \vdots \end{pmatrix}, \tag{156}$$

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

$$\overline{\boldsymbol{C}} = \begin{pmatrix} \langle \boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} \rangle & \langle \boldsymbol{y}y_E \rangle & -\langle \boldsymbol{y}y_I \rangle & \\ \langle y_E\boldsymbol{y}^{\mathsf{T}} \rangle & \langle y_Ey_E \rangle & -\langle y_Ey_I \rangle & \boldsymbol{0} \\ \langle y_I\boldsymbol{y}^{\mathsf{T}} \rangle & \langle y_Iy_E \rangle & -\langle y_Iy_I \rangle & \\ & \boldsymbol{0} & & \ddots \end{pmatrix} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{C}\boldsymbol{a}_I & \\ \boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C} & \boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_I & \boldsymbol{0} \\ \boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C} & \boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_I & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}, \tag{157}$$

and write the average synaptic change as[1] (cf. Eq. 40)

$$\Rightarrow \quad \overline{\boldsymbol{\tau}}\dot{\overline{\boldsymbol{w}}} \equiv \overline{\boldsymbol{C}}\overline{\boldsymbol{w}} - \boldsymbol{\Gamma}\overline{\boldsymbol{w}}, \tag{158}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix that holds the scalar constraint factors, and $\overline{\boldsymbol{\tau}}$ holds the timescales for excitatory synapses, $\boldsymbol{\tau}_E = \mathbb{1}\tau_E$, and inhibitory synapses, $\boldsymbol{\tau}_I = \mathbb{1}\tau_I$, on the diagonal. We make the simplifying assumption that the plasticity of excitatory and inhibitory synapses is equally fast, $\tau_E = \tau_I = \tau$. Then $\overline{\boldsymbol{\tau}} = \tau\mathbb{1}$, which does not affect the fixed points or the stability of the system[2], and we set $\overline{\boldsymbol{\tau}} = \mathbb{1}$.

Note that this is a highly non-linear dynamical system since the modified covariance matrix not only depends on the feedforward inputs $\boldsymbol{y}$ but also on the plastic synaptic weights $\overline{\boldsymbol{w}}$, in addition to the weight dependence of the normalization factors $\boldsymbol{\Gamma}$. Next, we show that the eigencircuit configuration we discussed in the introduction to this section is in fact a fixed point of the weight dynamics.

5.1 Fixed points

In general, fixed points $\overline{\boldsymbol{w}}^*$ must fulfil the following condition

$$\overline{\boldsymbol{C}}^*\overline{\boldsymbol{w}}^* - \boldsymbol{\Gamma}^*\overline{\boldsymbol{w}}^* \stackrel{!}{=} 0. \tag{159}$$

where $\overline{\boldsymbol{C}}^*$ is the modified covariance matrix evaluated in the fixed point. We consider the special case when the two neurons form a single eigencircuit, tuned to the feedforward eigenvector $\boldsymbol{v}^*$. Then we can write the excitatory and inhibitory firing rates as[3]

$$y_E^* = \boldsymbol{a}_E^{*\mathsf{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{a}_E^*, \quad \boldsymbol{a}_E^* = a_E^*\boldsymbol{v}^*, \tag{160}$$

$$y_I^* = \boldsymbol{a}_I^{*\mathsf{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{a}_I^*, \quad \boldsymbol{a}_I^* = a_I^*\boldsymbol{v}^*, \tag{161}$$

where a_E and a_I depend on the excitatory and inhibitory weights and can be determined via Eq. 154 & 155. This demonstrates that when neurons are tuned to the same feedforward eigenvector $\boldsymbol{v}^*$, their firing rate is proportional to

[1] We omitted the angle notation $\langle\cdot\rangle$ to improve readability.

[2] 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 (cf. Sec. 2.2).

[3] Note that here the superscript '*' indicates a variable that is evaluated in the fixed point of the weight dynamics and not a fixed point of the firing rate activity. Different input patterns $\boldsymbol{y}$ result in different neural activities y_E^*, y_I^*.

the projection of the activity vector $\boldsymbol{y}$ onto the eigenvector $\boldsymbol{v}^*$, and justifies our assumption in Eq. 134. The modified covariance matrix in the fixed point becomes

$$\bar{\boldsymbol{C}}^* = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{C}\boldsymbol{v}^* a_E^* & -\boldsymbol{C}\boldsymbol{v}^* a_I^* & \\ a_E^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C} & a_E^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C}\boldsymbol{v}^* a_E^* & -a_E^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C}\boldsymbol{v}^* a_I^* & \mathbf{0} \\ a_I^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C} & a_I^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C}\boldsymbol{v}^* a_E^* & -a_I^* \boldsymbol{v}^{*\mathsf{T}} \boldsymbol{C}\boldsymbol{v}^* a_I^* & \\ & \mathbf{0} & & \ddots \end{pmatrix} = \begin{pmatrix} \boldsymbol{C} & \lambda^* a_E^* \boldsymbol{v}^* & -\lambda^* a_I^* \boldsymbol{v}^* & \\ \lambda^* a_E^* \boldsymbol{v}^{*\mathsf{T}} & \lambda^* a_E^{*2} & -\lambda^* a_E^* a_I^* & \mathbf{0} \\ \lambda^* a_I^* \boldsymbol{v}^{*\mathsf{T}} & \lambda^* a_I^* a_E^* & -\lambda^* a_I^{*2} & \\ & \mathbf{0} & & \ddots \end{pmatrix}. \quad [162]$$

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

$$\bar{\boldsymbol{V}}^* = \begin{pmatrix} \boldsymbol{V}_{\setminus *} & \boldsymbol{v}^* & \boldsymbol{v}^* a_E^* & \boldsymbol{v}^* a_I^* & \\ \mathbf{0} & a_E^* & -1 & 0 & \mathbf{0} \\ \mathbf{0} & a_I^* & 0 & 1 & \\ & \mathbf{0} & & & \ddots \end{pmatrix}, \qquad \bar{\boldsymbol{V}}^{*-1} = \mathcal{N}^{-1} \begin{pmatrix} \mathcal{N}\, \boldsymbol{V}_{\setminus *}{}^{\mathsf{T}} & \mathbf{0} & \mathbf{0} & \\ \boldsymbol{v}^{*\mathsf{T}} & a_E^* & -a_I^* & \mathbf{0} \\ a_E \boldsymbol{v}^{*\mathsf{T}} & -(1 - a_I^{*2}) & -a_E^* a_I^* & \\ -a_I \boldsymbol{v}^{*\mathsf{T}} & -a_I^* a_E^* & 1 + a_E^{*2} & \\ & \mathbf{0} & & \ddots \end{pmatrix}, \quad [163]$$

$$\mathcal{N} \equiv 1 + a_E^{*2} - a_I^{*2}, \quad [164]$$

where the subscript $(\cdot)_{\setminus *}$ indicates that a matrix does not contain an entry that corresponds to the input mode $\boldsymbol{v}^*$.

In general, $\bar{\boldsymbol{C}}^*$ is of dimension $2D \times 2D$ where $D = N_F + N_E + N_I$. Therefore, to diagonalize $\bar{\boldsymbol{C}}^*$, we require $2D$ eigenvectors. In this specific circuit, we have $N_E = N_I = 1$. Because $\bar{\boldsymbol{C}}^*$ has a block diagonal structure (Eq. 162), with the first $D \times D$ block driving development of weights onto the excitatory neuron and the second $D \times D$ block driving development of weights onto the inhibitory neuron, the eigenvector matrix $\bar{\boldsymbol{V}}^*$ and its inverse have the same block diagonal structure. Since each block has the same sub-structure, we only show the first block in Eq. 163. Then for each eigencircuit there are $1 + n_E + n_I$ eigenvectors of $\bar{\boldsymbol{C}}^*$, where n_E and n_I are the number of excitatory and inhibitory neurons in the respective eigencircuit: One regular eigenvector, and $n_E + n_I$ null-eigenvectors (cf. Sec. 2.1.2). The first column of $\bar{\boldsymbol{V}}^*$ in Eq. 163 corresponds to the $N_F - 1$ eigencircuits with $n_E = n_I = 0$, i.e., with one regular eigenvector per feedforward eigenvector $\boldsymbol{v} \neq \boldsymbol{v}^*$, but without null-eigenvectors. The corresponding eigenvalues are $\boldsymbol{\Lambda}_{\setminus *}$, which are also eigenvalues of the feedforward covariance matrix $\boldsymbol{C}$. For the eigencircuit corresponding to $\boldsymbol{v}^*$ there are $1 + n_E + n_I = 3$ eigenvectors of $\bar{\boldsymbol{C}}^*$. The first is a regular eigenvector and the last two are null eigenvectors, where the excitatory feedforward component is cancelled by either a negative lateral excitatory component[1], or a positive lateral inhibitory component. The null eigenvectors have eigenvalues equal to zero, and the eigenvalue of the regular eigenvector is $1 + a_E^{*2} - a_I^{*2}$.

Similar to the feedforward case, arbitrary multiples of the separately normalized parts of eigenvectors of $\bar{\boldsymbol{C}}^*$ are fixed points. The only exception is the rightmost null eigenvector (cf. Sec. 2.1.3). There, the inhibitory and the excitatory weights are aligned such that the post-synaptic activity is zero, which does not allow for arbitrary scaling of the weight norms. Inserting these fixed points into Equations 154 & 155, provides conditions to determine a_E^* and a_I^*.

5.2 Stability analysis

We are interested in the stability of the circuit described in the introduction of Section 5 and consider the stability of a regular eigenvector $\bar{\boldsymbol{v}}^*$

$$\bar{\boldsymbol{w}}^* = \bar{\boldsymbol{v}}^* = \begin{pmatrix} \boldsymbol{v}^* \\ a_E^* \\ a_I^* \\ \boldsymbol{v}^* \\ a_E^* \\ a_I^* \end{pmatrix}, \quad \bar{\lambda}^* = \lambda^* \left(1 + a_E^{*2} - a_I^{*2}\right), \quad [165]$$

[1] In our simulations, we constrain synaptic weights to be positive. Then null eigenvectors with negative weights are only relevant in combination with regular eigenvectors: When a null eigenvector is added to a regular eigenvector, the net synaptic input does not change. 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.

This means we do *not* consider arbitrary scalings of the excitatory and inhibitory parts of eigenvectors of $\bar{\boldsymbol{C}}^*$, but assume that weight norms are fine tuned to match the norms of the excitatory and inhibitory parts of the regular eigenvector[1] $\bar{\boldsymbol{v}}^*$.

When are such eigenvectors stable, and when are they attracted to a different input mode? To answer this question, we consider small fixed point perturbations $\Delta\bar{\boldsymbol{w}}(t_0)$, where the excitatory neuron shifts its tuning in the direction of a different feedforward input eigenvector $\boldsymbol{v}^\dagger$:

$$\Delta\bar{\boldsymbol{w}}(t_0) \propto \begin{pmatrix} \boldsymbol{v}^\dagger \\ 0 \\ 0 \\ \boldsymbol{0} \\ 0 \\ 0 \end{pmatrix} = \bar{\boldsymbol{V}}^* \boldsymbol{e}^\dagger. \tag{166}$$

where $\boldsymbol{e}^\dagger$ is a vector of zeros with a single non-zero entry that corresponds to the feedforward eigenvector $\boldsymbol{v}^\dagger$ (cf. Eq. 163). The system is stable with respect to a perturbation if the perturbation decays to zero over time. To check this, we consider the following differential equation that holds for small perturbations (cf. Sec. 1.2.2)

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\bar{\boldsymbol{w}}(t) = \bar{\boldsymbol{J}}^* \Delta\bar{\boldsymbol{w}}(t), \tag{167}$$

where $\bar{\boldsymbol{J}}^*$ is the Jacobian evaluated in the fixed point. We will consider the dynamics in the non-orthogonal eigenbasis $\bar{\boldsymbol{V}}^*$ of the modified covariance matrix $\bar{\boldsymbol{C}}^*$ evaluated in the fixed point $\bar{\boldsymbol{w}}^* = \bar{\boldsymbol{v}}^*$. Note that $\bar{\boldsymbol{V}}^*$ is not time-dependent, because it is evaluated in the fixed point. In this static basis, we can express perturbations as

$$\Delta\bar{\boldsymbol{w}}_v(t) = \bar{\boldsymbol{V}}^{*-1}\Delta\bar{\boldsymbol{w}}(t), \tag{168}$$

$$\Rightarrow \Delta\bar{\boldsymbol{w}}_v(t_0) = \bar{\boldsymbol{V}}^{*-1}\Delta\bar{\boldsymbol{w}}(t_0) \propto \boldsymbol{e}^\dagger, \tag{169}$$

where the subscript $(\cdot)_v$ indicates a vector or matrix expressed in this basis. The perturbation dynamics becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\bar{\boldsymbol{w}}_v(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\bar{\boldsymbol{V}}^{*-1}\Delta\bar{\boldsymbol{w}}(t)\right) = \bar{\boldsymbol{V}}^{*-1}\frac{\mathrm{d}}{\mathrm{d}t}\Delta\bar{\boldsymbol{w}}(t) = \bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\Delta\bar{\boldsymbol{w}}(t) = \bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\bar{\boldsymbol{V}}^*\Delta\bar{\boldsymbol{w}}_v(t) = \bar{\boldsymbol{J}}_v^*\Delta\bar{\boldsymbol{w}}_v(t), \tag{170}$$

where we defined the transformed Jacobian, $\bar{\boldsymbol{J}}_v^* = \bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\bar{\boldsymbol{V}}^*$. Without loss of generality, we assume that eigenvectors in $\bar{\boldsymbol{V}}^*$ are sorted such that the first entry of $\boldsymbol{e}^\dagger$ is non-zero, i.e., the first column of $\bar{\boldsymbol{V}}^*$ is proportional to the initial perturbation $\Delta\bar{\boldsymbol{w}}(t_0)$ (cf. Eq. 166). Next, we will derive the transformed Jacobian.

5.2.1 The transformed Jacobian

First, we consider the regular Jacobian $\bar{\boldsymbol{J}}^*$. We rewrite the dynamics in Eq. 158 as[2]

$$\dot{\bar{\boldsymbol{w}}} = \left[\mathbb{1} - \frac{(\bar{\boldsymbol{w}}_{EF} + \bar{\boldsymbol{w}}_{EE})\bar{\boldsymbol{c}}_{EE}^\mathsf{T}}{\bar{\boldsymbol{c}}_{EE}^\mathsf{T}(\bar{\boldsymbol{w}}_{EF} + \bar{\boldsymbol{w}}_{EE})} - \ldots\right]\bar{\boldsymbol{C}}\bar{\boldsymbol{w}}, \qquad \bar{\boldsymbol{w}}_{EF} = \begin{pmatrix} \boldsymbol{w}_{EF} \\ 0 \\ 0 \\ \boldsymbol{0} \\ 0 \\ 0 \end{pmatrix}, \quad \bar{\boldsymbol{w}}_{EE} = \begin{pmatrix} \boldsymbol{0} \\ w_{EE} \\ 0 \\ \boldsymbol{0} \\ 0 \\ 0 \end{pmatrix}, \quad \bar{\boldsymbol{c}}_{EE} = \begin{pmatrix} \boldsymbol{c} \\ 1 \\ 0 \\ \boldsymbol{0} \\ 0 \\ 0 \end{pmatrix}, \tag{171}$$

where the second term in the bracket corresponds to the normalization of all excitatory synapses onto the excitatory neuron, additional normalization terms are indicated by ellipsis[3] (cf. Eq. 45), and $\boldsymbol{c}$ is a vector of ones. Then the Jacobian has the following shape (cf. Eq. 29)

$$\bar{\boldsymbol{J}}^* = \left.\frac{\mathrm{d}\dot{\bar{\boldsymbol{w}}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_* = \left[\mathbb{1} - \frac{(\bar{\boldsymbol{v}}_{EF}^* + \bar{\boldsymbol{v}}_{EE}^*)\bar{\boldsymbol{c}}_{EE}^\mathsf{T}}{\bar{\boldsymbol{c}}_{EE}^\mathsf{T}(\bar{\boldsymbol{v}}_{EF}^* + \bar{\boldsymbol{v}}_{EE}^*)} - \ldots\right]\left(\bar{\boldsymbol{C}}^* - \bar{\lambda}^*\mathbb{1} + \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_* \bar{\boldsymbol{w}}^*\right), \tag{172}$$

[1] We presume that when considering the stability of non-eigenvector fixed points, it is possible to make a similar argument as in Section 2.2.3 and consider regular eigenvectors of a different modified covariance matrix $\bar{\boldsymbol{C}}'$ with adjusted plasticity timescales, $k_E\tau_E$, $k_I\tau_I$. Here we consider the case of $\tau_E = \tau_I$ and regular eigenvectors of $\bar{\boldsymbol{C}}^*$, i.e., $k_E = k_I = 1$.

[2] Remember that we set $\bar{\boldsymbol{\tau}} = \mathbb{1}$.

[3] In general, there are $2 \times (n_E + n_I)$ normalization terms.

where $\bar{\boldsymbol{v}}^*_{EF}$, $\bar{\boldsymbol{v}}^*_{EE}$ have the same shape as $\bar{\boldsymbol{w}}_{EF}$, $\bar{\boldsymbol{w}}_{EE}$ in Eq. 171 with entries corresponding to the respective entries of the regular eigenvector $\bar{\boldsymbol{v}}^*$ (cf. Eq. 165). Note that we accounted for the weight dependence of the modified covariance matrix $\bar{\boldsymbol{C}}$ which results in the tensor $\mathrm{d}\bar{\boldsymbol{C}}/\mathrm{d}\bar{\boldsymbol{w}}$. To find the transformed Jacobian $\bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\bar{\boldsymbol{V}}^*$, we consider the first bracket:

$$\bar{\boldsymbol{V}}^{*-1}\left[\mathbb{1} - \frac{(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}}{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})} - \dots\right]\bar{\boldsymbol{V}}^* \tag{173}$$

The first entry remains equal to the identity matrix, as the eigenvector matrix and its inverse cancel. We consider the columns $\bar{\boldsymbol{v}}^*_b$ of $\bar{\boldsymbol{V}}^*$ separately. Then, we can write

$$\left[-\frac{(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}}{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})} - \dots\right]\bar{\boldsymbol{v}}^*_b = -(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})h^b_{EE} - \bar{\boldsymbol{v}}^*_{EI}h^b_{EI} - (\bar{\boldsymbol{v}}^*_{IF} + \bar{\boldsymbol{v}}^*_{IE})h^b_{IE} - \bar{\boldsymbol{v}}^*_{II}h^b_{II} = -\boldsymbol{H}_b\bar{\boldsymbol{v}}^*, \tag{174}$$

$$\boldsymbol{H}_b \equiv \begin{pmatrix} \mathbb{1}h^b_{EE} & & & & & \\ & h^b_{EE} & & & & \\ & & h^b_{EI} & & & \\ & & & \mathbb{1}h^b_{IE} & & \\ & & & & h^b_{IE} & \\ & & & & & h^b_{II} \end{pmatrix}, \quad h^b_{EE} \equiv \frac{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}\bar{\boldsymbol{v}}^*_b}{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})}, \quad h^b_{EI} \equiv \frac{\bar{\boldsymbol{c}}_{EI}{}^{\mathsf{T}}\bar{\boldsymbol{v}}^*_b}{\bar{\boldsymbol{c}}_{EI}{}^{\mathsf{T}}\bar{\boldsymbol{v}}^*_{EI}}, \tag{175}$$

where $\boldsymbol{H}_b$ is a diagonal matrix with entries corresponding to the respective normalization constraint, of which we give h^b_{EE} and h^b_{EI} as examples. Then each column $\bar{\boldsymbol{v}}^*_b$ of $\bar{\boldsymbol{V}}^*$ is transformed into a multiple of the separately normalized parts of the fixed point eigenvector $\bar{\boldsymbol{v}}^*$ (Eq. 165). After transformation, the bth column becomes

$$\bar{\boldsymbol{V}}^{*-1}\boldsymbol{H}_b\bar{\boldsymbol{v}}^* = \mathcal{N}^{-1}\begin{pmatrix} \mathcal{N}\boldsymbol{V}_{\setminus *}{}^{\mathsf{T}} & \boldsymbol{0} & \boldsymbol{0} & \\ \boldsymbol{v}^{*\mathsf{T}} & a^*_E & -a^*_I & \boldsymbol{0} \\ a_E\boldsymbol{v}^{*\mathsf{T}} & -(1-a^{*2}_I) & -a^*_E a^*_I & \\ -a_I\boldsymbol{v}^{*\mathsf{T}} & -a^*_I a^*_E & 1+a^{*2}_E & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}\begin{pmatrix} h^b_{EE}\boldsymbol{v}^* \\ h^b_{EE}a^*_E \\ h^b_{EI}a^*_I \\ \vdots \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \vdots \end{pmatrix}, \tag{176}$$

where, as before, ellipsis indicate potentially non-zero entries. Importantly, after the transformation, the first $N_F - 1$ entries are zero, independent of the column index, b, because $\boldsymbol{v}^*$ is orthogonal to the columns of $\boldsymbol{V}_{\setminus *}$. Overall, we can write

$$\Rightarrow \quad \bar{\boldsymbol{V}}^{*-1}\left[\mathbb{1} - \frac{(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}}{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}(\bar{\boldsymbol{v}}^*_{EF} + \bar{\boldsymbol{v}}^*_{EE})} - \dots\right]\bar{\boldsymbol{V}}^* = \begin{pmatrix} \mathbb{1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \vdots & \vdots & \vdots & \vdots & \boldsymbol{0} \\ & \boldsymbol{0} & & & \ddots \end{pmatrix}, \tag{177}$$

where the block structure arises from the block structure of $\bar{\boldsymbol{V}}^{*-1}$ (cf. Eq. 176).

After transformation, the second bracket of Eq. 172 becomes

$$\bar{\boldsymbol{V}}^{*-1}\left(\bar{\boldsymbol{C}}^* - \bar{\lambda}^*\mathbb{1} + \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_* \bar{\boldsymbol{w}}^*\right)\bar{\boldsymbol{V}}^* = \left(\bar{\Lambda}^* - \bar{\lambda}^*\mathbb{1} + \bar{\boldsymbol{V}}^{*-1}\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_* \bar{\boldsymbol{w}}^*\bar{\boldsymbol{V}}^*\right). \tag{178}$$

We next consider the first columns of $\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_* \bar{\boldsymbol{w}}^*$, for which we compute the matrix $\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}w^b_{EF}}\right|_*$, where w^b_{EF} is the bth feed-

forward weight onto the excitatory neuron.

$$\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}w_{EF}^{b}}\right|_{*} = \left.\frac{\mathrm{d}}{\mathrm{d}w_{EF}^{b}}\begin{pmatrix} \boldsymbol{C} & \boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{C}\boldsymbol{a}_I & \\ \boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C} & \boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{a}_E{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_I & \boldsymbol{0} \\ \boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C} & \boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_E & -\boldsymbol{a}_I{}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{a}_I & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}\right|_{*} \tag{179}$$

$$= \begin{pmatrix} \boldsymbol{0} & \boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_E}{\mathrm{d}w_{EF}^{b}}\right|_{*} & -\boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_I}{\mathrm{d}w_{EF}^{b}}\right|_{*} & \\ \left.\frac{\mathrm{d}\boldsymbol{a}_E{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C} & \left(\left.\frac{\mathrm{d}\boldsymbol{a}_E{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C}\boldsymbol{a}_E^{*} + \boldsymbol{a}_E^{*}{}^{\mathsf{T}}\boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_E}{\mathrm{d}w_{EF}^{b}}\right|_{*}\right) & -\left(\left.\frac{\mathrm{d}\boldsymbol{a}_E{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C}\boldsymbol{a}_I^{*} + \boldsymbol{a}_E^{*}{}^{\mathsf{T}}\boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_I}{\mathrm{d}w_{EF}^{b}}\right|_{*}\right) & \boldsymbol{0} \\ \left.\frac{\mathrm{d}\boldsymbol{a}_I{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C} & \left(\left.\frac{\mathrm{d}\boldsymbol{a}_I{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C}\boldsymbol{a}_E^{*} + \boldsymbol{a}_I^{*}{}^{\mathsf{T}}\boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_E}{\mathrm{d}w_{EF}^{b}}\right|_{*}\right) & -\left(\left.\frac{\mathrm{d}\boldsymbol{a}_I{}^{\mathsf{T}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\boldsymbol{C}\boldsymbol{a}_I^{*} + \boldsymbol{a}_I^{*}{}^{\mathsf{T}}\boldsymbol{C}\left.\frac{\mathrm{d}\boldsymbol{a}_I}{\mathrm{d}w_{EF}^{b}}\right|_{*}\right) & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}, \tag{180}$$

where we used the definition of $\bar{\boldsymbol{C}}$ from Eq. 157. The vectors $\boldsymbol{a}_E$ and $\boldsymbol{a}_I$ are defined in Eq. 154 & 155. It follows:

$$\left.\frac{\mathrm{d}\boldsymbol{a}_E}{\mathrm{d}w_{EF}^{b}}\right|_{*} = \frac{1}{1 - w_{EE}^{*} + \dfrac{w_{EI}^{*}w_{IE}^{*}}{1 + w_{II}^{*}}}\boldsymbol{e}_b \equiv \mu_E\boldsymbol{e}_b, \tag{181}$$

$$\left.\frac{\mathrm{d}\boldsymbol{a}_I}{\mathrm{d}w_{EF}^{b}}\right|_{*} = \frac{1}{1 + w_{II}^{*} + \dfrac{w_{IE}^{*}w_{EI}^{*}}{1 - w_{EE}^{*}}}\frac{w_{IE}^{*}}{1 - w_{EE}^{*}}\boldsymbol{e}_b \equiv \mu_I\boldsymbol{e}_b. \tag{182}$$

Additionally, we have (cf. Eqs.160 & 161)

$$\boldsymbol{C}\boldsymbol{a}_E^{*} = \lambda^{*}a_E^{*}\boldsymbol{v}^{*}, \quad \boldsymbol{C}\boldsymbol{a}_I^{*} = \lambda^{*}a_I^{*}\boldsymbol{v}^{*}, \tag{183}$$

which results in

$$\Rightarrow \quad \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}w_{EF}^{b}}\right|_{*} = \begin{pmatrix} \boldsymbol{0} & \mu_E\boldsymbol{C}\boldsymbol{e}_b & -\mu_I\boldsymbol{C}\boldsymbol{e}_b & \\ \mu_E\boldsymbol{e}_b{}^{\mathsf{T}}\boldsymbol{C} & 2\lambda^{*}a_E^{*}\mu_E\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b & -\lambda^{*}(\mu_E a_E^{*} + \mu_I a_E^{*})\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b & \boldsymbol{0} \\ \mu_I\boldsymbol{e}_b{}^{\mathsf{T}}\boldsymbol{C} & \lambda^{*}(\mu_I a_E^{*} + \mu_E a_I^{*})\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b & 2\lambda^{*}a_I^{*}\mu_I\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}, \tag{184}$$

where $\boldsymbol{e}_b$ is a vector of dimension N_F with entries equal to zero, except for the bth entry equal to one.

$$\Rightarrow \quad \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}w_{EF}^{b}}\right|_{*}\bar{\boldsymbol{w}}^{*} = \begin{pmatrix} \beta_E\boldsymbol{C}\boldsymbol{e}_b \\ g_1\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b \\ g_2\boldsymbol{v}^{*\mathsf{T}}\boldsymbol{e}_b \\ \boldsymbol{0} \end{pmatrix}, \quad \bar{\boldsymbol{w}}^{*} = \bar{\boldsymbol{v}}^{*} = \begin{pmatrix} \boldsymbol{v}^{*} \\ w_{EE}^{*} \\ w_{EI}^{*} \\ \vdots \end{pmatrix}, \quad \beta_E = \mu_E w_{EE}^{*} - \mu_I w_{EI}^{*}, \tag{185}$$

where $g_{(\cdot)}$ are scalars.

$$\Rightarrow \quad \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\boldsymbol{w}_{EF}}\right|_{*}\boldsymbol{w}^{*} = \begin{pmatrix} \beta_E\boldsymbol{C} \\ g_1\boldsymbol{v}^{*\mathsf{T}} \\ g_2\boldsymbol{v}^{*\mathsf{T}} \\ \boldsymbol{0} \end{pmatrix}. \tag{186}$$

We find other columns in a similar fashion and write

$$\Rightarrow \quad \left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_{*}\boldsymbol{w}^{*} = \begin{pmatrix} \beta_E\boldsymbol{C} & g_3\boldsymbol{v}^{*} & g_6\boldsymbol{v}^{*} & \\ g_1\boldsymbol{v}^{*\mathsf{T}} & g_4 & g_7 & \boldsymbol{0} \\ g_2\boldsymbol{v}^{*\mathsf{T}} & g_5 & g_8 & \\ & \boldsymbol{0} & & \ddots \end{pmatrix}, \tag{187}$$

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

$$\bar{\boldsymbol{V}}^{*-1}\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_{*}\bar{\boldsymbol{w}}^{*}\bar{\boldsymbol{V}}^{*}=\bar{\boldsymbol{V}}^{*-1}\begin{pmatrix}\beta_E\boldsymbol{C} & g_3\boldsymbol{v}^* & g_6\boldsymbol{v}^* & & \\ g_1\boldsymbol{v}^{*\mathsf{T}} & g_4 & g_7 & \boldsymbol{0} & \\ g_2\boldsymbol{v}^{*\mathsf{T}} & g_5 & g_8 & & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_{\setminus *} & \boldsymbol{v}^* & \boldsymbol{v}^*a_E^* & \boldsymbol{v}^*a_I^* & \\ \boldsymbol{0} & a_E^* & -1 & 0 & \boldsymbol{0} \\ \boldsymbol{0} & a_I^* & 0 & 1 & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix} \tag{188}$$

$$=\bar{\boldsymbol{V}}^{*-1}\begin{pmatrix}\beta_E\boldsymbol{C}\boldsymbol{V}_{\setminus *} & g_9\boldsymbol{v}^* & g_{12}\boldsymbol{v}^* & g_{15}\boldsymbol{v}^* & \\ \boldsymbol{0} & g_{10} & g_{13} & g_{16} & \boldsymbol{0} \\ \boldsymbol{0} & g_{11} & g_{14} & g_{17} & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix} \tag{189}$$

$$=\mathcal{N}^{-1}\begin{pmatrix}\mathcal{N}\boldsymbol{V}_{\setminus *}{}^{\mathsf{T}} & \boldsymbol{0} & \boldsymbol{0} & & \\ \boldsymbol{v}^{*\mathsf{T}} & a_E^* & -a_I^* & \boldsymbol{0} & \\ a_E\boldsymbol{v}^{*\mathsf{T}} & -(1-a_I^{*2}) & -a_E^*a_I^* & & \\ -a_I\boldsymbol{v}^{*\mathsf{T}} & -a_I^*a_E^* & 1+a_E^{*2} & & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}\begin{pmatrix}\beta_E\boldsymbol{V}_{\setminus *}\boldsymbol{\Lambda}_{\setminus *} & g_9\boldsymbol{v}^* & g_{12}\boldsymbol{v}^* & g_{15}\boldsymbol{v}^* & \\ \boldsymbol{0} & g_{10} & g_{13} & g_{16} & \boldsymbol{0} \\ \boldsymbol{0} & g_{11} & g_{14} & g_{17} & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix} \tag{190}$$

$$=\begin{pmatrix}\beta_E\boldsymbol{\Lambda}_{\setminus *} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \boldsymbol{0} & g_{18} & g_{21} & g_{24} & \boldsymbol{0} \\ \boldsymbol{0} & g_{19} & g_{22} & g_{25} & \\ \boldsymbol{0} & g_{20} & g_{23} & g_{26} & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}. \tag{191}$$

The fully transformed Jacobian is (cf. Eq. 178)

$$\bar{\boldsymbol{J}}_v^*=\bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\bar{\boldsymbol{V}}^*=\bar{\boldsymbol{V}}^{*-1}\left[\mathbb{1}-\frac{(\bar{\boldsymbol{v}}_{EF}^*+\bar{\boldsymbol{v}}_{EE}^*)\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}}{\bar{\boldsymbol{c}}_{EE}{}^{\mathsf{T}}(\bar{\boldsymbol{v}}_{EF}^*+\bar{\boldsymbol{v}}_{EE}^*)}-\ldots\right]\bar{\boldsymbol{V}}^*\left(\bar{\boldsymbol{\Lambda}}^*-\bar{\lambda}^*\mathbb{1}+\bar{\boldsymbol{V}}^{*-1}\left.\frac{\mathrm{d}\bar{\boldsymbol{C}}}{\mathrm{d}\bar{\boldsymbol{w}}}\right|_{*}\bar{\boldsymbol{w}}^*\bar{\boldsymbol{V}}^*\right) \tag{192}$$

Finally, by inserting Eq. 177 & 191 we find

$$\bar{\boldsymbol{V}}^{*-1}\bar{\boldsymbol{J}}^*\bar{\boldsymbol{V}}^*=\begin{pmatrix}\mathbb{1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \vdots & \vdots & \vdots & \vdots & \boldsymbol{0} \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}\left(\bar{\boldsymbol{\Lambda}}^*-\bar{\lambda}^*\mathbb{1}+\begin{pmatrix}\beta_E\bar{\boldsymbol{\Lambda}}_{\setminus *} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \boldsymbol{0} & g_{18} & g_{21} & g_{24} & \boldsymbol{0} \\ \boldsymbol{0} & g_{19} & g_{22} & g_{25} & \\ \boldsymbol{0} & g_{20} & g_{23} & g_{26} & \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}\right). \tag{193}$$

$$\Rightarrow\quad \bar{\boldsymbol{J}}_v^*=\begin{pmatrix}\left(\bar{\boldsymbol{\Lambda}}_{\setminus *}-\mathbb{1}\bar{\lambda}^*+\beta_E\boldsymbol{\Lambda}_{\setminus *}\right) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \vdots & \vdots & \vdots & \vdots & \boldsymbol{0} \\ & \boldsymbol{0} & & & \ddots\end{pmatrix}, \tag{194}$$

where $\bar{\boldsymbol{\Lambda}}_{\setminus *}^*$ contains eigenvalues of $\bar{\boldsymbol{C}}^*$ that correspond to regular, non-fixed point eigenvectors.[1]

[1] Note that in our specific network the top left block of $\bar{\boldsymbol{\Lambda}}^*$ is equal to $\boldsymbol{\Lambda}_{\setminus *}$, i.e., $\bar{\boldsymbol{\Lambda}}_{\setminus *}=\boldsymbol{\Lambda}_{\setminus *}$, because there are no neurons tuned to the respective feedforward eigenvectors. In particular, $\bar{\lambda}^\dagger=\lambda^\dagger$ (cf. Eq. 151)

5.2.2 Stability conditions

The dynamics of a general fixed point perturbation $\Delta \boldsymbol{w}_v$ in the eigenbasis of $\bar{\boldsymbol{C}}^*$ is (cf. Eq.170)

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\bar{\boldsymbol{w}}_v = \bar{\boldsymbol{J}}_v^* \Delta\bar{\boldsymbol{w}}_v = \begin{pmatrix} \left(\bar{\boldsymbol{\Lambda}}_{\setminus *} - \mathbb{1}\bar{\lambda}^* + \beta_E \boldsymbol{\Lambda}_{\setminus *}\right) & \mathbf{0} & \mathbf{0} & \\ \vdots & \vdots & \vdots & \mathbf{0} \\ \mathbf{0} & & & \ddots \end{pmatrix} \Delta\bar{\boldsymbol{w}}_v. \tag{195}$$

Note that the transformed Jacobian (Eq. 194) has a triangular block structure, and each row of $\bar{\boldsymbol{J}}_v^*$ corresponds to the growth of a perturbation in the direction of a different eigenvector of $\bar{\boldsymbol{C}}^*$. We are only interested in perturbations that grow in the direction of a non-fixed point feedforward eigenvector, $\boldsymbol{V}_{\setminus *}$. Therefore, we focus on the first rows of $\bar{\boldsymbol{J}}_v^*$, which correspond to growth in these directions. Except for the first diagonal block, these rows are zero. It follows that perturbations $\Delta\bar{\boldsymbol{w}}_v(t_0)$ that do not already contain components in the direction of non-fixed point eigenvectors, also do not develop such components in their later dynamics. In contrast, perturbations in the direction of a non-fixed point feedforward input mode, e.g., $\Delta\bar{\boldsymbol{w}}_v \propto \boldsymbol{e}^\dagger$, can induce perturbations within the original eigencircuit that corresponds to the feedforward eigenvector $\boldsymbol{v}^*$ [1]. For example, a decrease in feedforward and recurrent excitatory synaptic weights within the eigencircuit balances the increase of feedforward excitatory synaptic weights due to the perturbation towards a different eigencircuit, to maintain the weight norm. However, as explained above, these 'second-order' perturbations, without components in the direction of a non-fixed point feedforward eigenvectors, $\boldsymbol{V}_{\setminus *}$, are contained within the eigencircuit, i.e., they can not induce subsequent perturbations in the direction of non-fixed point feedforward input modes, $\boldsymbol{V}_{\setminus *}$. Therefore, to answer the question of when an eigencircuit is stable, we only consider the dynamics along the direction of the original perturbation (cf. Eq.166) by projecting the dynamics onto the perturbation vector at time zero, $\Delta\bar{\boldsymbol{w}}_v(t_0) \propto \boldsymbol{e}^\dagger$ (cf. Eq.169):

$$\boldsymbol{e}^{\dagger\mathsf{T}} \frac{\mathrm{d}}{\mathrm{d}t}\Delta\bar{\boldsymbol{w}}_v = \left(\bar{\lambda}^\dagger - \bar{\lambda}^* + \beta_E \lambda^\dagger\right) \boldsymbol{e}^{\dagger\mathsf{T}} \Delta\bar{\boldsymbol{w}}_v, \tag{196}$$

which provides the eigencircuit stability condition for the excitatory neuron

$$\boxed{\bar{\lambda}^\dagger - \bar{\lambda}^* + \beta_E \lambda^\dagger < 0}. \tag{197}$$

If Eq. 197 holds, perturbations in the direction of non-fixed point eigenvectors decay to zero, and the eigencircuit is stable. For β_E we have (cf. Eqs. 185 & 181 f.)

$$\beta_E = \frac{1}{1 - w_{EE}^* + \dfrac{w_{EI}^* w_{IE}^*}{1 + w_{II}^*}} w_{EE}^* - \frac{1}{1 + w_{II}^* + \dfrac{w_{IE}^* w_{EI}^*}{1 - w_{EE}^*}} \left(\frac{w_{EI}^* w_{IE}^*}{1 - w_{EE}^*}\right). \tag{198}$$

From Eq. 154 & 155 we find

$$\left.\frac{\mathrm{d}y_E}{\mathrm{d}(\boldsymbol{w}_{EF}{}^\mathsf{T}\boldsymbol{y})}\right|_* = \frac{1}{1 - w_{EE}^* + \dfrac{w_{EI}^* w_{IE}^*}{1 + w_{II}^*}}, \tag{199}$$

$$\left.\frac{\mathrm{d}y_I}{\mathrm{d}(\boldsymbol{w}_{EF}{}^\mathsf{T}\boldsymbol{y})}\right|_* = \frac{1}{1 + w_{II}^* + \dfrac{w_{IE}^* w_{EI}^*}{1 - w_{EE}^*}} \frac{w_{IE}^*}{1 - w_{EE}^*}, \tag{200}$$

and we get

$$\Rightarrow \quad \boxed{\beta_E = \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\boldsymbol{w}_{EF}{}^\mathsf{T}\boldsymbol{y})}\right|_* w_{EE}^* - \left.\frac{\mathrm{d}y_I}{\mathrm{d}(\boldsymbol{w}_{EF}{}^\mathsf{T}\boldsymbol{y})}\right|_* w_{EI}^*}. \tag{201}$$

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

$$\boxed{\bar{\lambda}^\dagger - \bar{\lambda}^* + \beta_I \lambda^\dagger < 0}, \tag{202}$$

[1] This is due to the potentially non-zero elements in the block below the top left diagonal block of $\bar{\boldsymbol{J}}_v^*$ in Equation 195.

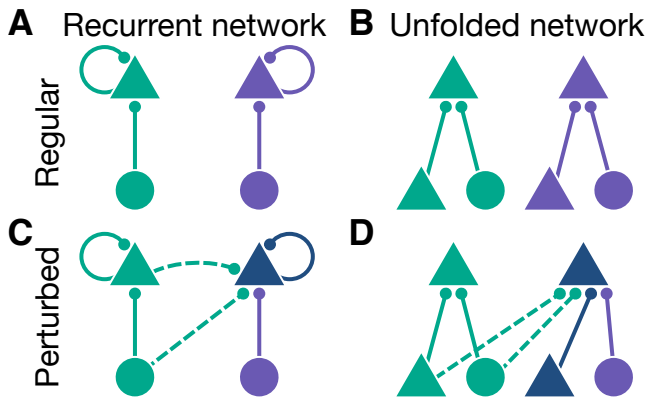

Figure S4: (*A*) Two excitatory neurons (triangles) are tuned to two different, but equally attractive input modes (circles, green and purple). (*B*) The same circuit as in *A*, unfolded to highlight pre-synaptic partners. Both input modes are balanced in their attraction. (*C*) Perturbing the purple excitatory neuron towards the green input mode (dashed lines) shifts its tuning (dark blue) such that it now responds to both the green and the purple input modes. (*D*) The unfolded circuit from *C*. Due to the perturbation, the green input mode is now more attractive, and the previously purple excitatory neuron shifts its tuning. See text for details.

$$\Rightarrow \quad \boxed{\beta_I = \left.\frac{\mathrm{d}y_E}{\mathrm{d}(\boldsymbol{w}_{IF}{}^{\mathsf{T}}\boldsymbol{y})}\right|_* w^*_{IE} - \left.\frac{\mathrm{d}y_I}{\mathrm{d}(\boldsymbol{w}_{IF}{}^{\mathsf{T}}\boldsymbol{y})}\right|_* w^*_{II}} \,. \qquad [203]$$

We will now interpret these results.

5.2.3 Eigencircuit stability depends on recurrent connectivity

We first consider the case when the effective attraction of all eigencircuits is the same, i.e., $\bar{\lambda}^* = \bar{\lambda}^\dagger$ (cf. Eqs. 197 & 202). Then the stability is fully determined by β_E, and β_I. In feedforward circuits we have not found any β-terms, because in that case, the modified covariance matrix does not depend on any plastic synaptic weights (cf. Eq. 51 & Sec. 2.2). This is not the case in recurrent circuits where the perturbation induces a change in the tuning of laterally projecting neurons.

To build some intuition, we consider a simple example: Think of a recurrent network of two excitatory neurons with identical weight norms that receive feedforward input from the same population of excitatory neurons. In the fixed point, the neurons are tuned to two different feedforward input eigenvectors of equal attraction and are recurrently connected to themselves but not each other (Fig. S4*A*). Then the effective attraction of the two eigencircuits is the same. In general, neurons receive synaptic inputs, but have no information about the overall network structure, e.g., which synaptic inputs are feedforward or recurrent. Taking this perspective, we unfold the recurrent network and observe that the effective mode attraction is a combination of the feedforward input and the recurrent self-excitation (Fig. S4*B*). When we perturb one neuron towards the opposing input mode (Fig. S4*C*, dashed lines), the tuning of the perturbed neuron changes slightly in the direction of that mode (Fig. S4*C*, dark blue). From the perspective of the perturbed neuron, this tuning change leads to an attraction increase of the opposing eigencircuit, which is now more attractive than the original eigencircuit of the perturbed neuron (Fig. S4*D*), and the perturbation grows in the direction of the more attractive mode – the fixed point is unstable. Similarly, if the neurons were inhibitory instead, the perturbation would decrease the attraction towards the opposite input mode which would stabilize the network.

In our mathematical analysis of the circuit shown in Figure S3*D*, the attraction increase or decrease due to the tuning change of recurrently projecting neurons is reflected in the β-terms in Equations 197 & 202, which emerge from the weight dependence of the modified covariance matrix $\bar{\boldsymbol{C}}$ (cf. Eq. 172). For example, when perturbing the excitatory neuron, the increase in attraction from the perspective of the perturbed neuron is (cf. Eq. 197)

$$\beta_E \lambda^\dagger = \left(\left.\frac{\mathrm{d}y_E}{\mathrm{d}(\boldsymbol{w}_{EF}{}^{\mathsf{T}}\boldsymbol{y})}\right|_* \lambda^\dagger \right) w^*_{EE} - \left(\left.\frac{\mathrm{d}y_I}{\mathrm{d}(\boldsymbol{w}_{EF}{}^{\mathsf{T}}\boldsymbol{y})}\right|_* \lambda^\dagger \right) w^*_{EI}, \qquad [204]$$

where the brackets reflect the tuning shifts of the excitatory and the inhibitory neuron[1] in response to the perturbation of $\boldsymbol{w}_{EF}$ in the direction of $\boldsymbol{v}^\dagger$, which is then weighted by the respective synaptic connection onto the excitatory neuron, w^*_{EE}, w^*_{EI}. When the inhibitory neuron is perturbed instead, the terms for β_I follow the same logic (cf. Eq. 203).

[1] Note that also the tuning of the inhibitory neuron changes, although it is not directly perturbed.

Without going through the lengthy mathematical derivation, we now give some intuition about β-terms of the network perturbation in Figure S3*C*. In the fixed point, the perturbed excitatory neuron receives recurrent input from all neurons in its eigencircuit, including itself. In the following, superscripts indicate the corresponding eigencircuit, *A* or *B*. Then, as for the equivalent circuit (cf. Fig. S3*D*), β_E^A comprises two terms, one due to the tuning shift of y_E^A, and a second due to the tuning shift of y_I^A. Assuming the same weight norms, this is exactly equal to the β_E for the equivalent circuit (Eq. 197). Different from β_E, β_E^A is weighted with the effective attraction $\bar{\lambda}^B = \lambda^B + \lambda_{\text{eig}}^B$, instead of only the feedforward attraction (cf. $\lambda^\dagger$ in Eq. 204), because the perturbation comprises not only the feedforward eigenvector component but the whole eigencircuit (cf. dashed lines in Figs. S3*C* & *D*). This is why, for the equivalent circuit, we chose the feedforward attraction $\lambda^\dagger = \bar{\lambda}^B = \lambda^B + \lambda_{\text{eig}}^B$ (Eq. 151). Then, the diagonal entries corresponding to the respective perturbations in the upper left blocks of the transformed Jacobians are the same[1] (cf. Eq. 194), i.e.,

$$\bar{\lambda}^\dagger - \bar{\lambda}^* + \beta_E \lambda^\dagger = \bar{\lambda}^B - \bar{\lambda}^A + \beta_E^B \bar{\lambda}^B. \qquad [205]$$

We find that perturbations in both circuits initially follow the same dynamics, while the later dynamics diverges: At time t_0, there are no lateral projections from eigencircuit *B* towards eigencircuit *A* (cf. Fig. S3*C*), since in the fixed point there are no recurrent connections between eigencircuits (cf. Sec. 4), and the perturbation at time t_0 only introduces connections from eigencircuit *B* onto eigencircuit *A*. However, as we just discussed, the perturbation introduces a tuning shift in neurons of eigencircuit *A* in the direction of eigencircuit *B*. This shift leads to non-zero correlations between neurons of both eigencircuits, and synaptic weights from eigencircuit *A* onto eigencircuit *B* start to grow. These growing synapses shift the attraction of neurons in eigencircuit *B* and thus impact the dynamics of perturbation components in the direction of eigencircuit *B*. Therefore, the transformed Jacobian of the original circuit (Fig. S3*C*) has a more complex structure than the Jacobian for the equivalent circuit[2]. However, since we consider an initial perturbation that is aligned with a regular eigenvector, i.e., $\Delta\bar{\boldsymbol{w}}_v(t_0) \propto \boldsymbol{e}^B$ is one-hot (cf. Eq. 169), the top left diagonal block of the Jacobian still determines the initial dynamics[3].

In summary, recurrent synapses can stabilize or destabilize a circuit with respect to small perturbations away from a fixed point. These stabilizing and destabilizing effects are described by β-terms that depend on the specific weight configuration in the fixed point (cf. Eq. 198), which again depends on the weight norms that constrain the total synaptic weights. In the following, we consider the case when synaptic weights are tuned such that β-terms are small. In the equivalent circuit (Fig. S3*D*) this is the case when the influence of the tuning shifts of the excitatory and the inhibitory neuron balance each other[4] (cf. the first and second terms in Eqs. 201 & 203).

5.2.4 Decorrelation condition

We now consider how neurons self-organize to represent all parts of their input space instead of clustering all their tuning curves around a dominant input mode. We consider the fixed point stability of different eigencircuit configurations. In particular, we consider the case when recurrent connectivity motifs do not influence the stability of an eigencircuit, i.e., β_E and β_I are approximately zero. This can be achieved by a suitable choice of weight norms[5]. Then, all eigencircuits are marginally stable when they are equally attractive, i.e., (cf. Eqs. 197 & 202 for $\beta_{E/I} = 0$)

$$\bar{\lambda}^a = \lambda^a + \lambda_{\text{eig}}^a \overset{!}{=} \bar{\lambda}^b, \quad \forall a, b. \qquad [206]$$

For homogeneous input spaces, where the feedforward attraction of all input modes is the same, i.e., $\lambda^a = \lambda^b = \lambda, \forall a, b$, the only alternative stable configuration is when all neurons are tuned to the same feedforward input mode and form a single eigencircuit. Such a configuration does not reflect the tunings of biological neural populations, where all parts of the input space are represented. To prevent such a global clustering of neural tunings, we require that the corresponding eigencircuit is unstable. When all β-terms are approximately zero, this is the case when the effective attraction of the only occupied eigencircuit, $\bar{\lambda}^*$, is smaller than the attraction of one of the $N_F - 1$ unoccupied[6]

[1] Recall that $\lambda_{\text{eig}}^\dagger = 0$ and, therefore, $\bar{\lambda}^\dagger = \lambda^\dagger$ (Eq. 151).

[2] The Jacobian of the original circuit (Fig. S3*C*) has additional entries to the right of the top left diagonal block in Equation 195 that are non-zero. These non-zero entries result in the growth of synapses of y_E^A in the direction of eigencircuit *B* due to 'second-order' perturbations of recurrent synapses from eigencircuit *A* to eigencircuit *B*.

[3] Non-zero entries of the Jacobian to the right of the top left diagonal block are cancelled by the zero entries in the initial perturbation vector $\Delta\bar{\boldsymbol{w}}_v(t_0)$ (cf. Eq. 195).

[4] Note that the term in Equation 203 that corresponds to the tuning shift of the inhibitory neuron can be positive since $\mathrm{d}y_I/\mathrm{d}(\boldsymbol{w}_{IF}{}^\top \boldsymbol{y})$ is negative when the circuit is inhibition stabilized (*6*).

[5] See Section 5.2.3 for a discussion of the case $\beta_{E/I} \neq 0$.

[6] An unoccupied input mode corresponds to $\lambda_{\text{eig}}^\dagger = 0$ (cf. Eq. 146).

input modes, $\bar{\lambda}^{\dagger} = \lambda^{\dagger}$ (cf. Eq. 197 & 202):

$$\bar{\lambda}^{*} < \lambda^{\dagger}, \quad [207]$$

$$\Rightarrow \quad \sum_{i} \sigma_{E,i}^{2} - \sum_{l} \sigma_{I,j}^{2} + \lambda < \lambda, \quad [208]$$

$$\Rightarrow \quad N_E \sigma_E^2 - N_I \sigma_I^2 < 0, \quad [209]$$

$$\Rightarrow \quad \boxed{N_E \sigma_E^2 < N_I \sigma_I^2}, \quad [210]$$

where σ_E^2, σ_I^2 are the average variance, and N_E, N_I the total number of inhibitory and excitatory neurons. When this condition is satisfied, the only stable solution is when the effective attraction of all eigencircuits is identical. The simplest configuration where this is the case is when each eigencircuit contains the same number of excitatory and inhibitory neurons.

6 Movie captions

Movie M1: Decorrelation of feedforward tuning curves of excitatory neurons in plastic recurrent networks. Development of feedforward tuning curves of $N_E = 10$ excitatory neurons (cf. Figs. 3*A* & *B*). Synaptic weights were initialized randomly. Different color shades indicate weights of different post-synaptic neurons.

Movie M2: Decorrelation of feedforward tuning curves of inhibitory neurons in plastic recurrent networks. Development of feedforward tuning curves of $N_I = 10$ inhibitory neurons (cf. Figs. 3*A* & *B*). Synaptic weights were initialized randomly. Different color shades indicate weights of different post-synaptic neurons.

References

1. E. Oja, Simplified neuron model as a principal component analyzer. *Journal of mathematical biology* (1982).
2. K. D. Miller, D. J. MacKay, The role of constraints in Hebbian learning. *Neural computation* (1994).
3. S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering* (CRC press, 2018).
4. T. P. Vogels *et al.*, Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* (2011).
5. C. Clopath *et al.*, Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. *BioRxiv* (2016).
6. M. V. Tsodyks *et al.*, Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience* (1997).