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

Samuel Eckmann and Julijana Gjorgjieva

1Max Planck Institute for Brain Research, Frankfurt am Main, Germany
2TUM School of Life Sciences, Technical University Munich, Germany

Corresponding author. Email: ec.sam@outlook.com

Cortical networks exhibit complex stimulus-response patterns. Previous work has identified the balance between excitatory and inhibitory currents as a central component of cortical computations. However, how these requirements are self-organized into biologically plausible plasticity rules remains an open question. Here we present a Hebbian learning framework that (i) enables the development of stimulus selectivity and inhibitory synapses, (ii) demonstrates how E-I balance and stimulus selectivity can simultaneously develop in a single neuron, and (iii) provides a foundational understanding of plasticity in recurrent networks.

Equation 1:

$$\Delta w = y_A r, \quad A \in \{E, I\},$$

where $r$ is the post-synaptic firing rate, and $y_A$ is a vector that holds the pre-synaptic firing rates of excitatory ($A = E$) and inhibitory ($A = I$) neurons. Experimental results have shown that after the induction of long-term plasticity neither the total excitatory nor the total inhibitory synaptic area change. This suggests that a synapse can only grow at the expense of another synapse – a competitive mechanism potentially mediated by the limited supply of synaptic proteins (Fig. 1D).

Motivated by these results, we adopted a competitive normalization rule for both excitatory and inhibitory synapses:

$$w_A \leftarrow W_A \frac{w_A + \Delta w_A}{|w_A + \Delta w_A|}$$

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

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

$$\langle W_E \rangle \propto C W_E - y W_E.$$

where $C = \langle y_E y_I^\top \rangle$ is the input covariance matrix, with $\langle \cdot \rangle$ being the temporal average, and $y$ is a scalar normalization factor that regulates Hebbian growth. Then, fixed points, for which $\langle W_E \rangle = 0$, are eigenvectors of the covariance matrix, i.e., the neuron becomes selective to the first principal component of its input data. For a non-linear activation function $f(u)$, neurons become selective for higher-order correlations in their inputs. In the following, we call the fixed points of such pure feedforward circuits ‘input modes’.

Figure 1: Synapse-type-specific competitive Hebbian learning enables the development of stimulus selectivity and inhibitory balance. (A) Feedforward input to a pyramidal neuron consists of direct excitation and disinhibition (represented as linear and dashed lines, respectively). (B) A single post-synaptic pyramidal neuron, targeted by excitatory (E) and inhibitory (I) synapses. (C) Excitatory and inhibitory inputs are equally tuned to the orientation of a stimulus grating (bottom), tuning curve of neurons tuned to 60° highlighted in dark gray) and exhibit a Gaussian-shaped population response when a single grating of 30° is presented (orange plate, dashed line). (D) Excitatory synapses are plastic, whereas inhibitory synaptic weights are fixed. (E) Synaptic weights change their strength, after training, as a function of the tuning peak of the pre-synaptic neurons. (F & G) Same as (E) and (F), but for classic inhibitory plasticity.
that excitatory and inhibitory synaptic weights eventually have identical shape, $w_E \propto w_I$, equal to a multiple of an eigenvector of $C$. Neurons become selective for activity along one particular input direction, while excitatory and inhibitory neural inputs are balanced, which explains the joint development of stimulus selectivity and E-I balance in feedforward circuits, in agreement with our numerical simulations (Fig. 1).

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

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

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

\[ w_i \propto y^T(r - r_0). \]

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

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

\[ \langle w \rangle \propto Cw - \gamma - \rho w. \]

where $\gamma$ and $\rho$ are scalars that ensure normalization, and we defined the modified covariance matrix $\tilde{C}$. Now the fixed points of the weight dynamics are multiples of the excitatory and the inhibitory part of the eigenvectors of the modified covariance matrix $\tilde{C}$. When excitatory and inhibitory inputs are equally stimulus selective, such that one can approximate $y_E \propto y_I$, the modified covariance matrix $\tilde{C}$ is composed of multiples of the original covariance matrix $C$. This implies

\[ \begin{pmatrix} \bar{y}_E \bar{y}_I^T \bar{y}_E^T \bar{y}_I^T \end{pmatrix} \frac{\Delta \delta}{\rho}. \]

where $\Delta \delta$ is the largest synaptic weight between population $A$ and $B$, $\rho = \rho (\bar{E} \bar{I})$. Normalized (norm.) recurrent weight strengths as a function of the difference between preferred orientations of pre- and post-synaptic neurons, $\Delta \delta \propto \Delta \delta - \bar{E} \bar{I}$ averaged over all neuron pairs. (F) Average firing rate response of inhibitory and excitatory neurons for a stimulus orientation $\theta$ relative to their preferred orientation, $\Delta \delta = \bar{E} - \bar{I}$, averaged over all neurons. (G) Same as in (F) for excitatory and inhibitory inputs to excitatory neurons.

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

Figure 3: Tuning curve decorrelation in plastic recurrent networks. (A) Top: A population of recurrently connected excitatory and inhibitory neurons receives input from a set of input neurons that are tuned to different stimulus orientations (compare Fig. 1B, bottom). Every neuron has a different orientation is presented to the network (vertical gray lines). At the same time, all synapses exhibit plasticity according to a synapse-type-specific Hebbian rule (see Appendix 1 for details). Bottom: Exemplary firing rate activity of one excitatory and one inhibitory neuron before and after training. (B) Feedforward tuning curves of $N_E = 10$ excitatory neurons before (top), during (center), and after (bottom) training. (C) Connectivity matrices for $N_E = 80$ excitatory (blue) and $N_I = 20$ inhibitory (red) neurons after training. Neurons are sorted according to their preferred orientation $\bar{E}$. $w_{AB}^\theta$ is the largest synaptic weight between population $A$ and $B$. $\theta \in \{ \bar{E}, \bar{I} \}$. (D) Normalized (norm.) recurrent weight strengths as a function of the difference between preferred orientations of pre- and post-synaptic neurons, $\Delta \delta \propto \bar{E} - \bar{I}$, averaged over all neuron pairs. (E) Average firing rate response of inhibitory and excitatory neurons to a stimulus orientation $\theta$ relative to their preferred orientation, $\Delta \delta = \bar{E} - \bar{I}$, averaged over all neurons. (F) Same as in (E) for excitatory and inhibitory inputs to excitatory neurons. (G) Inhibitory input to an excitatory neuron with preferred orientation close to $90^\circ$. Each curve corresponds to the input from one pre-synaptic inhibitory neuron.
functions, input modes are eigenvectors of the input covariance matrix (compare Eq.3). Since these eigenvectors are orthogonal by definition, the activities of neurons that are tuned to different eigenvectors are uncorrelated, and their reciprocal connections decay to zero under Hebbian plasticity (Fig. 4A). Then, neurons that are tuned to the same eigenvector form recurrent ‘eigencircuits’ that are otherwise separated from the rest of the network (SM Sec. 5). Crucially, this decomposition of the network into eigencircuits allows us to write the effective attraction $\lambda$ of an input mode as the sum of a feedforward component $\lambda_{eq}$ and the variances of the neurons that reside in the eigencircuit (Fig. 4B):

$$\lambda = \lambda_{eq} + \lambda = \sum_{i} \sigma_{E,i}^2 - \sum_{i} \sigma_{I,i}^2,$$

(7)

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

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

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

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

In the brain, total synaptic strengths are dynamic and homeostatically regulated on a timescale of hours to days. 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. However, theoretical models suggest that homeostatic scaling is too slow to balance rapid synaptic plasticity. In our networks, Hebbian growth is instead thought to be stabilized by the rapid competition for a limited pool of synapse-type specific resources, while total synaptic strengths remain fixed. In line with...
these results, we suggest that scaling of synaptic strengths may not be required for immediate network stability but instead controls the operating regime of the network.

Our results suggest that synapse-type-specific competitive Hebbian learning is the key mechanism that enables the formation of functional recurrent networks. Rather than hand-tuning connectivity to selectively explain experimental data, our networks emerge from unsupervised, biologically plausible learning rules. In a single framework, we readily explain multiple experimental observations, including the development of stimulus selectivity, excitation-inhibition balance, decorrelated neural activity, assembly structures, response normalization, and orientation-specific surround suppression. Our results demonstrate how the connectivity of balanced networks is shaped by their input statistics and explain the experience-dependent formation of extra-classical receptive fields.

In our model, circuit formation depends only on the statistical regularities between input streams and is agnostic to their origin. Therefore, we expect our approach to extend beyond sensory cortices and to provide a fundamental framework for plasticity in recurrent neural networks.
Supplementary Material for

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

Samuel Eckmann*1 and Julijana Gjorgjieva1,2
1Max Planck Institute for Brain Research, Frankfurt am Main, Germany
2TUM School of Life Sciences, Technical University Munich, Germany
*Corresponding author. Email: ec.sam@outlook.com

Contents

1 Methods & simulation parameters .................................................. 2
   Plasticity and Normalization .................................................. 2
   Input model ........................................................................ 2

2 Linear competitive Hebbian learning finds principal components .... 3
   2.1 Hebbian plasticity without normalization is unstable .............. 3
   2.2 Weight constraints stabilize unlimited Hebbian growth ........... 4
   Classic Inhibitory plasticity prevents stimulus selectivity .......... 6

3 Subtype-specific normalization balances E-I receptive fields ..... 7
   3.1 Fixed points are multiples of eigenvector components .......... 8
   Eigenvectors and eigenvalues of the modified covariance matrix .... 8
   Non-eigenvector fixed points .............................................. 9
   3.2 Stability analysis .............................................................. 9
   Principal component analysis in inhibitory modified input spaces .. 11
   Fast inhibition increases stability ........................................ 11
   Stability of non-eigenvector fixed points .............................. 11
   Effective timescales and attraction landscapes .................... 12

4 Lateral input warps attraction landscape .................................... 12

5 Eigencircuits ............................................................................ 14
   5.1 Variance propagation ......................................................... 15
   5.2 Consistency conditions provide eigencircuit firing rate variances 16
   5.3 A note on the choice of weight norm .................................. 16

6 Fully plastic recurrent E-I networks ........................................... 17
   6.1 Fixed points .................................................................... 18
   6.2 Stability analysis .............................................................. 19
   The transformed Jacobian .................................................... 19
   Stability conditions ............................................................ 23
   6.3 Decorrelation condition .................................................... 24
   6.4 Eigencircuits are stabilized by intra-eigencircuit inhibition and destabilized by intra-eigencircuit excitation 25
# Methods & simulation parameters

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 $u$ evolve according to

$$
\tau_A u_A = -u_A + \sum_B W_{AB} r_B - W_{AI} r_I, \quad A \in \{E,I\}, \quad B \in \{E,I,F\}
$$

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

$$
r_A = f(u_A), \quad f(u) = a [u - b]^n, \quad A \in \{E,I\}
$$

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

## Plasticity and Normalization

Plastic weights evolve according to a Hebbian plasticity rule

$$
W_{AB} = \epsilon_{AB} r_A \odot r_B, \quad A \in \{E,I\}, \quad B \in \{E,I,F\}
$$

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

$$
W_{\text{total}}^{(i)} \leftarrow \frac{W_{\text{total}}^{(i)}}{\sum_j W_{AE}^{(i)} + \sum_k W_{AF}^{(i)}}, \quad W_{\text{total}}^{(i)} \leftarrow \frac{W_{\text{total}}^{(i)}}{\sum_j W_{AI}^{(i)}}, \quad A \in \{E,I\}, \quad B \in \{E,F\},
$$

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

In Fig. 1, we set the activity of the inhibitory input neurons equal to the activity of the excitatory input neurons, i.e., $r_I = r_E$. For panels E & F of Fig. 1, inhibitory weights evolved according to the classic inhibitory plasticity rule\(^1\) without normalization:

$$
W_{EI} = \epsilon_{EI}(r_E - r_0) r_I,
$$

where $r_0$ is a target firing rate.

## Input model

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

$$
r_F = c A_F \exp \left( \frac{|\theta - \theta_F|^2}{2 \sigma_F^2} \right),
$$

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

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

$$
W_{AF} = \exp \left( \frac{|\hat{\theta} - \theta_F|^2}{2 \sigma_W^2} \right).
$$

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

See Table S1 for an overview of used simulation parameters. Python code will be made available after journal publication.
Before considering inhibitory plasticity, we recapitulate how linear Hebbian learning finds the principal eigenvector of a neuron’s inputs. Although first described by Oja, we will mostly follow the derivation by Miller and MacKay that we will later extend to inhibitory neurons.

### Table S1: Simulation parameters

<table>
<thead>
<tr>
<th>Figure</th>
<th>E,F</th>
<th>G,H</th>
<th>A,B</th>
<th>C,D,E,F,G</th>
<th>A,B,C</th>
<th>D,E,F,G</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_E$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$N_I$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>20 x 2</td>
</tr>
<tr>
<td>$N_F$</td>
<td>10</td>
<td>10</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>80 x 2</td>
</tr>
<tr>
<td>$a$</td>
<td>1</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>$b$</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$n$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$W_{EE}$</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>0.6</td>
<td>3.51</td>
<td>3.51</td>
</tr>
<tr>
<td>$W_{EI}$</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0.85</td>
<td>3.35</td>
<td>3.35</td>
</tr>
<tr>
<td>$W_{II}$</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>0.3</td>
<td>1.44</td>
<td>1.44</td>
</tr>
<tr>
<td>$W_{EF}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$W_{IF}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$W_{IF}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$c$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_F$</td>
<td>1</td>
<td>1</td>
<td>35</td>
<td>140</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>$\sigma_F$</td>
<td>20°</td>
<td>20°</td>
<td>12°</td>
<td>12°</td>
<td>30°/√2</td>
<td>30°/√2</td>
</tr>
<tr>
<td>$\sigma_A$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30°/√2</td>
<td>30°/√2</td>
</tr>
<tr>
<td>$\Delta t$</td>
<td>200ms</td>
<td>200ms</td>
<td>10ms</td>
<td>10ms</td>
<td>10ms</td>
<td>10ms</td>
</tr>
<tr>
<td>$\tau_E$</td>
<td>200ms</td>
<td>200ms</td>
<td>20ms</td>
<td>20ms</td>
<td>25ms</td>
<td>25ms</td>
</tr>
<tr>
<td>$\tau_I$</td>
<td>-</td>
<td>-</td>
<td>17ms</td>
<td>17ms</td>
<td>12.5ms</td>
<td>12.5ms</td>
</tr>
<tr>
<td>$e_{EE}$</td>
<td>-</td>
<td>-</td>
<td>$2 \times 10^{-9}ms^{-1}$</td>
<td>$1.0 \times 10^{-10}ms^{-1}$</td>
<td>$1.0 \times 10^{-9}ms^{-1}$</td>
<td>$1.0 \times 10^{-9}ms^{-1}$</td>
</tr>
<tr>
<td>$e_{EI}$</td>
<td>-</td>
<td>-</td>
<td>$3 \times 10^{-9}ms^{-1}$</td>
<td>$1.5 \times 10^{-10}ms^{-1}$</td>
<td>$1.5 \times 10^{-9}ms^{-1}$</td>
<td>$1.5 \times 10^{-9}ms^{-1}$</td>
</tr>
<tr>
<td>$e_{II}$</td>
<td>-</td>
<td>-</td>
<td>$4 \times 10^{-4}ms^{-1}$</td>
<td>$4 \times 10^{-4}ms^{-1}$</td>
<td>$2.0 \times 10^{-4}ms^{-1}$</td>
<td>$2.0 \times 10^{-4}ms^{-1}$</td>
</tr>
<tr>
<td>$e_{EF}$</td>
<td>-</td>
<td>-</td>
<td>$5 \times 10^{-4}ms^{-1}$</td>
<td>$2.5 \times 10^{-4}ms^{-1}$</td>
<td>$2.5 \times 10^{-4}ms^{-1}$</td>
<td>$2.5 \times 10^{-4}ms^{-1}$</td>
</tr>
<tr>
<td>$e_{IF}$</td>
<td>-</td>
<td>-</td>
<td>$2 \times 10^{-4}ms^{-1}$</td>
<td>$2 \times 10^{-4}ms^{-1}$</td>
<td>$\epsilon_{EE}$</td>
<td>$\epsilon_{EE}$</td>
</tr>
<tr>
<td>$r_0$</td>
<td>0.25</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 2 Linear competitive Hebbian learning finds principal components

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

#### 2.1 Hebbian plasticity without normalization is unstable

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

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

where $\mathbf{w}$ is a vector that holds the synaptic weights, and $\tau$ defines the timescale at which the activity changes. The transpose is denoted by $^T$. In the following, lowercase letters in bold indicate vectors and uppercase letters in bold
matrices or tensors. Following Hebb’s principle, synaptic weight changes depend on the pre- and postsynaptic firing rates. In vector notation,

\[ \Delta \mathbf{w} = \mathbf{r} \mathbf{y} \]

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

\[ \langle \Delta \mathbf{w} \rangle = \langle \mathbf{r} \rangle = \langle \mathbf{y} \mathbf{y}^\top \mathbf{w} \rangle = \mathbf{C} \mathbf{w}, \quad \mathbf{C} \equiv \langle \mathbf{y} \mathbf{y}^\top \rangle, \tag{10} \]

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

\[ \mathbf{w}_r \equiv \mathbf{V}^\top \mathbf{w} = \mathbf{V}^\top \mathbf{C} \mathbf{V}^\top \mathbf{w} = \Lambda \mathbf{w}_r, \quad \mathbf{w}_r = \mathbf{w}_r(t_0) \exp(\Lambda \Delta t). \tag{11} \]

\[ \Rightarrow \mathbf{w}_r = \mathbf{w}_r(t_0) \exp(\Lambda \Delta t). \tag{12} \]

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

\[ \mathbf{w} = \mathbf{w}(t) + \mathbf{C} \Delta \mathbf{w} = \frac{\mathbf{W}}{\mathbf{c}^\top \mathbf{w}} \mathbf{w}, \quad \mathbf{W} \equiv \mathbf{c}^\top \mathbf{w}(t). \tag{13} \]

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

\[ \mathbf{w}(t+\Delta t) = \beta \left[ \mathbf{w}(t) + \mathbf{C} \mathbf{w}(t) \Delta t \right], \quad \beta(\mathbf{w}(t), \Delta t) = \frac{\mathbf{c}^\top \mathbf{w}(t)}{\mathbf{c}^\top \left[ \mathbf{C} \mathbf{w}(t) \Delta t + \mathbf{w}(t) \right]}. \tag{14} \]
where \( \beta \) describes the multiplicative normalization that depends on the size of the timestep \( \Delta t \) and the previous weight change \( \Delta t \). Then, the effective weight change \( \Delta t \) is given as

\[
\Delta t = \lim_{\Delta t \to 0} \frac{1 - \beta}{\beta \Delta t} \frac{\mathbf{C} \mathbf{w}(t) - \mathbf{1} \mathbf{w}}{\mathbf{c}^\top \mathbf{w}}.
\]

After taking the limit, one gets

\[
\lim_{\Delta t \to 0} \frac{1 - \beta}{\beta \Delta t} = \frac{\mathbf{c}^\top \mathbf{C} \mathbf{w}}{\mathbf{c}^\top \mathbf{w}}.
\]

And finally (compare Fig. S1B)

\[
\Rightarrow \Delta t = \mathbf{C} \mathbf{w} - \mathbf{c} \Delta t = \mathbf{C} \mathbf{w} - \mathbf{c} \mathbf{w} = 0.
\]

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

\[
\frac{d}{dt}(\mathbf{c} \mathbf{w}) = \mathbf{c} \mathbf{w}' = \mathbf{c} \mathbf{C} \mathbf{w} - \gamma \mathbf{c} \mathbf{w} = 0.
\]

Note that for \( \mathbf{c} \) being a constant vector of ones, the L1-norm of the weight vector is maintained. However, \( \mathbf{c} = \mathbf{w} \) 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 \( \mathbf{w} = \mathbf{C} \mathbf{w} - \mathbf{\zeta} \mathbf{k} \), where \( \mathbf{k} \) is a vector of ones.

**Fixed points**

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

\[
\mathbf{w}^* = a \mathbf{C} \mathbf{v} - \frac{\mathbf{c} \mathbf{v}}{\mathbf{c}^\top \mathbf{v}} \mathbf{a} \mathbf{v} = a \mathbf{v} \mathbf{\lambda} - \frac{\mathbf{c}^\top \mathbf{v}}{\mathbf{c}^\top \mathbf{v}} a \mathbf{v} = 0.
\]

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

**Stability analysis**

Multiplicative normalization constrains the norm of the weight vector and therefore prevents the otherwise unlimited growth of Hebbian plasticity. However, in theory, it is still possible that the system is unstable and never settles into a fixed point. Following Miller and MacKay,\(^4\) we will now 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.\(^5\) The weight dynamics around a fixed point \( \mathbf{w}^* \) can be approximated with its Taylor expansion

\[
\mathbf{w} \approx \mathbf{w}^* + \sum_i \frac{\partial \mathbf{w}}{\partial \mathbf{w}_i} (\mathbf{w}_i - \mathbf{w}^*_i),
\]

\[
\frac{d \mathbf{w}}{d \mathbf{w}_i} = \frac{\delta \mathbf{w}}{\delta \mathbf{w}_i} = \mathbf{J}^*_i (\mathbf{w} - \mathbf{w}^*).
\]

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

\[
\mathbf{J}^* = \left( \begin{array}{cccc}
\frac{\partial \mathbf{w}_1}{\partial \mathbf{w}_1} & \cdots & \frac{\partial \mathbf{w}_1}{\partial \mathbf{w}_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial \mathbf{w}_n}{\partial \mathbf{w}_1} & \cdots & \frac{\partial \mathbf{w}_n}{\partial \mathbf{w}_n}
\end{array} \right) \equiv \frac{\partial \mathbf{w}}{\partial \mathbf{w}_i}.
\]
A fixed point is stable if small perturbations away from the fixed point, $\Delta w = w - w^*$, decay to zero, i.e.,

$$\frac{\Delta w}{\Delta t} = \Delta w(0) \exp(J^* t),$$

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

$$\Delta w(t) = \Delta w(0) \exp(J^* t).$$

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

$$w = Cw - w y,$$

$$= Cw - \frac{w c^T C w}{c^T w},$$

$$= \left[ 1 - \frac{w c^T}{c^T w} \right] Cw.$$  

It follows:

$$\frac{d w}{d w^*} = \left[ 1 - \frac{v^* c^T}{c^T v^*} \right] C + \left[ -1 \frac{v^* c^T}{c^T v^*} + \frac{v^* c^T}{c^T v^*} \right] C v^*,$$

$$= \left[ 1 - \frac{v^* c^T}{c^T v^*} \right] [C - \lambda^* 1].$$

where $w|_0 = w^* = a v^*$ is the fixed point with $v^*$ being an eigenvector of $C$. The scalar $a$ is the length of the fixed point weight vector $w^*$ (which cancels) and $\lambda^*$ is the eigenvalue to $v^*$. To find the eigenvalues of the Jacobian, $\lambda_J$, we diagonalize $J$ by switching to the eigenbasis of $C$. When $V$ is the matrix that holds the eigenvectors of $C$ as columns one gets

$$V^T \frac{d w}{d w^*} V = \left[ 1 - \frac{v^* c^T}{c^T v^*} \right] V^T C V - \lambda^* 1, $$

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

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

$$\lambda_J = \lambda - \lambda^*.$$  

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

**Classic inhibitory plasticity prevents stimulus selectivity**

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

$$w_i \propto y_i (r - r_0).$$

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. Classic inhibitory plasticity is

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

2. To make sense of the vector notation, it helps to first consider the $b'$th column of $\Delta w$, which is equal to $\Delta w_{b'}$, where $w_0$ is the $b'$th vector component of $w$.

3. Because the eigenvalues of a product of two triangular matrices is equal to the product of their eigenvalues.
required to act faster than excitatory plasticity to enable stable weight dynamics. For much faster inhibitory plasticity, the dynamics of excitatory and inhibitory weights decouples and fixed points of the inhibitory weights 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 $r^*$. 

$$w^*_E = 0 \implies r^* = r_0 \implies w_E \propto y_{E0} - \text{normalization} \quad (36)$$

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

$$\langle w_E \rangle \propto cy_{00} - \text{normalization}, \quad (37)$$

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

3 Subtype-specific normalization balances E-I receptive fields

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

$$r = y_E^Tw_E - y_I^Tw_I \equiv y^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} w,$$

$$w = \begin{pmatrix} w_E \\ w_I \end{pmatrix}, \quad y = \begin{pmatrix} y_E \\ y_I \end{pmatrix}, \quad (38)$$

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

$$\tau \langle w \rangle = \langle yr \rangle = \begin{pmatrix} yE \\ yI \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} w_E \\ w_I \end{pmatrix} \equiv Cw, \quad (39)$$

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

$$\tau w = \hat{C}w - yw^s_E - \rho w^s_I \quad (40)$$

where $0$ are vectors of zeros of appropriate dimension. The constraint factors $y$ and $\rho$ follow from the requirement that the weight vector does not grow along the direction of the constraint vectors $c_E^*$ and $c_I^*$. 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:

$$\begin{align*}
    c_E^T w &= 0, \\
    c_I^T w &= 0, \\
    c_E^T \equiv (1, \ldots, 1, 0, \ldots, 0), \quad c_I^T \equiv (0, \ldots, 0, 1, \ldots, 1).
\end{align*} \quad (41)$$

$$\Rightarrow y = c_E^T \hat{C}w, \quad \rho = c_I^T \hat{C}w / c_I^T w_I, \quad (42)$$

\[1\] The choice of the L1-norm is motivated by the synaptic competition for a fixed amount of resources, where, in the simplest case, each unit of resource linearly increases synaptic strengths. Higher-order L-norms do not affect results in the feedforward learning. However in recurrent networks they can lead to instabilities (see Section 5.3).
where each column of $C_w$ is equal to the number of excitatory and inhibitory neurons, respectively. Finally, we can write the weight dynamics as

$$
\tau \bar{w} = \begin{bmatrix}
-w_E^c c_E^c & -w_E^i c_I^i \\
c_E^c w_E^c & c_I^i w_I^i
\end{bmatrix} \bar{C} \bar{w}.
$$

(47)

### 3.1 Fixed points are multiples of eigenvector components

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

$$
w^* = \bar{C}w^* - \gamma w_E^E - \rho w_I^E
$$

(48)

$$
= \bar{C}w^* - \frac{c_E^c}{c_E^c} \bar{C}w^* - \frac{c_I^i}{c_I^i} \bar{C}w^* + \lambda_E w_E^E + \lambda_I w_I^E = 0.
$$

(49)

This equation holds if $w^*$ solves

$$
\bar{C}w^* = \lambda_E w_E^E + \lambda_I w_I^E.
$$

(50)

where $\lambda_E$ and $\lambda_I$ are any scalars. It is straightforward to check that multiples of eigenvectors $\bar{v}$ of the modified covariance matrix $\bar{C}$ are fixed points.

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

(51)

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

$$
y_I = \mathbf{Q}^T y_E = A^T \bar{V} y_E.
$$

(52)

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

### Eigenvectors and eigenvalues of the modified covariance matrix

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

$$
\bar{C} = \begin{pmatrix}
y_E y_E^T & -y_E y_I^T \\
y_I y_E^T & -y_I y_I^T
\end{pmatrix} = \begin{pmatrix}
C & -CVA \\
A^T V^T C & -A^T V^T CVA
\end{pmatrix} = \begin{pmatrix}
C & -VAA \\
A^T V^T C & -A^T V^T CVA
\end{pmatrix}.
$$

(53)

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

$$
\bar{V} = \begin{pmatrix}
V & V \\
A^T & A^T
\end{pmatrix}, \quad \bar{V}^{-1} = \begin{pmatrix} (I - AA^T)^{-1} & 0 \\
0 & (1 - AA^T)^{-1} \end{pmatrix}.
$$

(54)

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

$$
r = y^T \begin{pmatrix} 1 & 0 \\
0 & -I \end{pmatrix} \bar{w}.
$$

(55)

$$
y = \begin{pmatrix} y_E \\ y_I \end{pmatrix}, \quad y_I = A^T V^T y_E.
$$

(56)
Since the postsynaptic firing rate is zero for these eigenvectors, we call them “null eigenvectors” or “null fixed points”. For usual choices for the excitatory and inhibitory weight norms where excitation is larger than inhibition, the set of fixed point solutions is irrelevant. Finally, the eigenvector and its null-eigenvector, i.e., all eigenvectors with the same excitatory component. To better see this, we write explicitly
\[ \mathbf{\tilde{C}}\mathbf{v} \equiv \mathbf{v}\lambda, \quad \lambda = \begin{pmatrix} \Lambda(1 - \mathbf{A}\mathbf{A}^\top) & 0 \\ 0 & 0 \end{pmatrix}. \] (57)

Note that null-eigenvectors define the null-space of the \( \mathbf{\tilde{C}} \) matrix.

**Non-eigenvector fixed points**

We first consider eigenvectors that result in non-zero postsynaptic activity and show that they meet the condition. We make the ansatz that the matrix of fixed point vectors \( \mathbf{W}^* \) has the shape
\[ \mathbf{W}^* = \begin{pmatrix} \mathbf{V}\mathbf{K}_E \\ \mathbf{A}^\top \mathbf{K}_I \end{pmatrix}, \quad \mathbf{\tilde{C}}\mathbf{W}^* = \begin{pmatrix} \mathbf{W}_E^*\mathbf{\Lambda}_E \\ \mathbf{W}_I^*\mathbf{\Lambda}_I \end{pmatrix}, \] (58)
where \( \mathbf{K}_E \) and \( \mathbf{K}_I \) are diagonal scaling matrices of arbitrary constants and the second equation follows from the fixed point condition in Eq. 50. We now show that for any \( \mathbf{K}_E \), \( \mathbf{K}_I \) we can find diagonal matrices \( \mathbf{\Lambda}_E \), \( \mathbf{\Lambda}_I \) that fulfill this condition. Note that for \( \mathbf{K}_E = \mathbb{I} \) and \( \mathbf{K}_I = (\mathbf{AA}^\top)^{-1} \) columns of \( \mathbf{W}^* \) are null-eigenvectors. Therefore, if true, the theorem holds for regular as well as null-eigenvectors. We write explicitly
\[ \mathbf{\tilde{C}}\mathbf{W}^* = \begin{pmatrix} \mathbf{C} & -\mathbf{V}\mathbf{\Lambda}\mathbf{A} \\ \mathbf{A}^\top \mathbf{\Lambda}\mathbf{\tilde{V}} & -\mathbf{A}^\top \mathbf{\Lambda}\mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{V}\mathbf{K}_E \\ \mathbf{A}^\top \mathbf{K}_I \end{pmatrix} = \begin{pmatrix} \mathbf{V}\mathbf{K}_E \mathbf{\Lambda}_E \\ \mathbf{A}^\top \mathbf{K}_I \mathbf{\Lambda}_I \end{pmatrix}. \] (59)

\[ \mathbf{C}\mathbf{V}\mathbf{K}_E - \mathbf{V}\mathbf{\Lambda}\mathbf{A}\mathbf{K}_I = \mathbf{V}\mathbf{K}_E \mathbf{\Lambda}_E, \] (60)
\[ \mathbf{A}^\top \mathbf{\Lambda}\mathbf{\tilde{V}}\mathbf{V}\mathbf{K}_E - \mathbf{A}^\top \mathbf{\Lambda}\mathbf{A}\mathbf{K}_I = \mathbf{A}^\top \mathbf{K}_I \mathbf{\Lambda}_I, \] (61)
\[ \mathbf{V}\mathbf{K}_E \mathbf{\Lambda}_E - \mathbf{V}\mathbf{K}_E \mathbf{\Lambda}_I = \mathbf{V}\mathbf{K}_E \mathbf{\Lambda}_E, \] (62)
\[ \mathbf{A}^\top \mathbf{K}_I \mathbf{\Lambda}_E - \mathbf{A}^\top \mathbf{K}_I \mathbf{\Lambda}_I = \mathbf{A}^\top \mathbf{K}_I \mathbf{\Lambda}_I, \] (63)
where we made use of the fact that independent of their subscript, the \( \mathbf{K} \), \( \lambda \), and \( \mathbf{AA}^\top \) matrices are diagonal and commute. By comparing the left and right sides of the equations, we find
\[ \mathbf{\Lambda}_E = \Lambda \left( \mathbf{I} - \mathbf{K}_E^{-1}\mathbf{K}\mathbf{AA}^\top \right), \] (64)
\[ \mathbf{\Lambda}_I = \Lambda \left( \mathbf{K}_I^{-1}\mathbf{K}_E - \mathbf{AA}^\top \right), \] (65)
which are diagonal matrices, as required.

### 3.2 Stability analysis

We first consider the stability of fixed point eigenvectors of the modified covariance matrix and discuss the general case afterwards. With Eq. 47, for the Jacobian \( \mathbf{J} \) it follows (compare Eq. 31)
\[ \tau \mathbf{J}_J = \tau \frac{\mathbf{d}\mathbf{w}_J}{\mathbf{d}t} = \begin{bmatrix} \mathbf{v}_E^\top \mathbf{c}_E^\top - \mathbf{v}_I^\top \mathbf{c}_I^\top & \mathbf{v}_E^\top \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{v}_I^\top \mathbf{c}_I^\top \mathbf{v}_I^\top \\ \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{c}_I^\top \mathbf{v}_I^\top & \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{c}_I^\top \mathbf{v}_I^\top \end{bmatrix} \mathbf{C} - \lambda^* \mathbb{I}. \] (66)
where \( \mathbf{v}_E \) and \( \mathbf{v}_I \) are the excitatory and the inhibitory part of the eigenvector fixed point \( \mathbf{v} \) with eigenvalue \( \lambda^* \) and the superscript \( \mathbb{I}_{\text{h.o.}} \) indicates an additional set of zeros to reach the correct dimensionality of the vector (compare Eq. 43). To find the eigenvalues \( \mathbf{\Lambda}_J \) of the Jacobian, we switch to the eigendecomposition of the modified covariance matrix\(^1\):
\[ \Rightarrow \mathbf{v}^{-1} \mathbf{J}_J \mathbf{v} = \mathbf{v}^{-1} \tau^{-1} \mathbf{v}^{-1} + \begin{bmatrix} \mathbf{v}_E^\top \mathbf{c}_E^\top - \mathbf{v}_I^\top \mathbf{c}_I^\top & \mathbf{v}_E^\top \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{v}_I^\top \mathbf{c}_I^\top \mathbf{v}_I^\top \\ \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{c}_I^\top \mathbf{v}_I^\top & \mathbf{c}_E^\top \mathbf{v}_E^\top - \mathbf{c}_I^\top \mathbf{v}_I^\top \end{bmatrix} \mathbf{C}\mathbf{v} \mathbf{v}^{-1} \] (67)
where \( \mathbf{v}^{-1} \mathbf{v}^{-1} = \mathbb{I} \). The result is a block diagonal matrix where each block corresponds to one regular eigenvector and its null-eigenvector, i.e., all eigenvectors with the same excitatory component. To better see this,
\(^1\)Note that we must make use of the inverse instead of the transpose since, in general, the eigenvector matrix \( \mathbf{v} \) is not orthonormal.
Where we define $\epsilon \equiv \tau^{-1}$ and write

\[
\mathbf{V}^{-1}\mathbf{\tau}^{-1}\mathbf{V} = \begin{pmatrix}
\mathbf{0} & (1 - \mathbf{AA}^T)^{-1} & -\mathbf{A}^T\mathbf{V}^T & \mathbf{A}^T \mathbf{V}^T & \epsilon_i \mathbf{A}^T \\
\mathbf{0} & (1 - \mathbf{AA}^T)^{-1} & -\mathbf{A}^T & \mathbf{A}^T & \epsilon_i \\
\mathbf{0} & (1 - \mathbf{AA}^T)^{-1} & \mathbf{AA}^T & \mathbf{A}^T & \epsilon_i \\
\mathbf{0} & (1 - \mathbf{AA}^T)^{-1} & \mathbf{A}^T & \mathbf{A}^T & \epsilon_i \\
\mathbf{0} & (1 - \mathbf{AA}^T)^{-1} & \mathbf{A}^T & \mathbf{A}^T & \epsilon_i \\
\end{pmatrix}
\]  

(68)

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

\[
\mathbf{V}^{-1}\mathbf{\tau}^{-1}\mathbf{V} = \begin{pmatrix}
(1 - a^{-2})^{-1} & 0 & 0 & 0 & \epsilon_E - \epsilon_i a^{-2} \\
0 & (1 - a^{-2})^{-1} & 0 & 0 & \epsilon_E - \epsilon_i (-\epsilon_E + \epsilon_i) a^{-2} + \epsilon_i \\
0 & 0 & (1 - a^{-2})^{-1} & 0 & \epsilon_E - \epsilon_i - \epsilon_E a^{-2} + \epsilon_i \\
0 & 0 & 0 & (1 - a^{-2})^{-1} & \epsilon_E - \epsilon_i \\
0 & 0 & 0 & 0 & \epsilon_E - \epsilon_i \\
\end{pmatrix}
\]  

(70)

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

\[
\mathbf{V}^{-1}\begin{pmatrix}
\mathbf{v}_{\mathbf{e}}^T \mathbf{c}^T \\
\mathbf{c}^T \mathbf{v}_{\mathbf{e}}^T \\
\mathbf{v}_{\mathbf{d}}^T \mathbf{c}^T \\
\mathbf{c}^T \mathbf{v}_{\mathbf{d}}^T \\
\mathbf{v}_{\mathbf{e}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{e}}^T \\
\mathbf{v}_{\mathbf{d}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{d}}^T \\
\end{pmatrix} = \mathbf{N} \begin{pmatrix}
\mathbf{V}^T & -\mathbf{A} \\
-\mathbf{A}^T & \mathbf{V}^T \\
\mathbf{v}_{\mathbf{e}}^T \mathbf{d}^T & \mathbf{v}_{\mathbf{d}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{e}}^T & \mathbf{d}^T \mathbf{v}_{\mathbf{d}}^T \\
\end{pmatrix}
\]  

\[
\mathbf{v}_{\mathbf{e}}^T = \mathbf{V} \mathbf{e}^*, \quad \mathbf{v}_{\mathbf{d}}^T = \mathbf{A}^T \mathbf{e}^*.
\]  

(71)

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

\[
\mathbf{V}^{-1}\begin{pmatrix}
\mathbf{v}_{\mathbf{e}}^T \mathbf{d}^T \\
\mathbf{v}_{\mathbf{d}}^T \mathbf{d}^T \\
\end{pmatrix} = \mathbf{N} \begin{pmatrix}
\mathbf{V}^T & -\mathbf{A} \\
-\mathbf{A}^T & \mathbf{V}^T \\
\mathbf{v}_{\mathbf{e}}^T \mathbf{d}^T & \mathbf{v}_{\mathbf{d}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{e}}^T & \mathbf{d}^T \mathbf{v}_{\mathbf{d}}^T \\
\end{pmatrix}
\]  

\[
\mathbf{v}_{\mathbf{e}}^T = \mathbf{V} \mathbf{e}^*, \quad \mathbf{v}_{\mathbf{d}}^T = \mathbf{A}^T \mathbf{e}^*.
\]  

(73)

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

\[
\mathbf{V}^{-1}\begin{pmatrix}
\mathbf{v}_{\mathbf{e}}^T \mathbf{c}^T \\
\mathbf{c}^T \mathbf{v}_{\mathbf{e}}^T \\
\mathbf{v}_{\mathbf{d}}^T \mathbf{c}^T \\
\mathbf{c}^T \mathbf{v}_{\mathbf{d}}^T \\
\mathbf{v}_{\mathbf{e}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{e}}^T \\
\mathbf{v}_{\mathbf{d}}^T \mathbf{d}^T \\
\mathbf{d}^T \mathbf{v}_{\mathbf{d}}^T \\
\end{pmatrix} = \mathbf{N} \begin{pmatrix}
1 - a^{-2} d_i^* & 1 - a^{-2} d_i^* & \ldots \\
-a^{-2} + a^{-2} d_i^* & -a^{-2} + a^{-2} d_i^* & \ldots \\
0 & 0 & \ldots \\
\end{pmatrix}
\]  

(76)

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

\[
\mathbf{d}_i^T = \frac{c_i^T \mathbf{V}}{c_i^T \mathbf{v}_{\mathbf{e}}^T} = \frac{1}{a^{2}} \left( c_i^T \mathbf{A}^T, c_i^T \mathbf{A}^T \right).
\]  

(77)
Remembering that $A^{-1} = A^T (AA^T)^{-1}$ we rearrange and get
\[
\begin{align*}
\frac{d}{dt} \begin{pmatrix}
\mathbf{w} \\
\mathbf{v}
\end{pmatrix}
= 
\begin{pmatrix}
A^T & \mathbf{a}^T \\
\mathbf{a} & \mathbf{v}^T
\end{pmatrix}
\begin{pmatrix}
\mathbf{w} \\
\mathbf{v}
\end{pmatrix}
+ \begin{pmatrix}
\mathbf{0} \\
\mathbf{0}
\end{pmatrix}
\end{align*}
\]
from which the equality follows.

In summary, we find that after rearrangement Eq. 67 is a block triangular matrix.
\[
\begin{align*}
\mathbf{V}^{-1} \mathbf{J} \mathbf{V} &= \mathbf{V}^{-1} \tau^{-1} \begin{pmatrix} 0 & \cdots & 0 \\ 0 & 0 & 0 \\ 1 & 1 & \ddots \end{pmatrix} \begin{pmatrix} \lambda & \cdots & \lambda \\ \cdots & 0 & \cdots \\ 0 & \cdots & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},
\end{align*}
\]
Therefore, to find the eigenvalues, we consider each diagonal block separately. The first block corresponds to disturbances in the direction of the fixed point eigenvector or its null-eigenvector. From the matrix product above, we see immediately that their corresponding eigenvalues must be zero. For disturbances in the direction of a non-fixed point eigenvector $\mathbf{v}^*$ or its null-eigenvector we consider an exemplary block matrix:
\[
\mathbf{J}' \equiv \mathbf{V}^{-1} \mathbf{J} \mathbf{V} = \frac{1}{1-a^2} \begin{pmatrix} \epsilon_E - \epsilon_1 a^2 I & \epsilon_E - \epsilon_2 I \\ -\epsilon_E a^2 \lambda & -\epsilon_E a^2 + \epsilon_2 I \end{pmatrix} \begin{pmatrix} \lambda - \lambda^* & 0 \\ 0 & -\lambda^* \end{pmatrix},
\]
where $\mathbf{V}^*$ is a two-column matrix that holds $\mathbf{v}^*$ and its null-eigenvector. The eigenvalues of this matrix are negative under two conditions. First, its determinant must be positive, and second, its trace must be negative. After some algebra, these two conditions read
\[
\begin{align*}
\det(\mathbf{J}') > 0 & \Rightarrow - (\lambda^* - \lambda^*), \\
\tau(\mathbf{J}') < 0 & \Rightarrow (\lambda^* - \lambda^*) - \epsilon_2 \left( a^2 \lambda^* + \lambda^* \right) < 0.
\end{align*}
\]

### Principal component analysis in inhibitory modified input spaces

First, we assume that excitatory and inhibitory plasticity are equally fast, i.e., $\epsilon_1 = \epsilon_E$. Then the first stability condition above states that the fixed point $\mathbf{v}^*$ with the largest eigenvalue, $\lambda^* > \lambda^*$, is stable, provided that it is not repulsive, i.e., provided that its eigenvalue is larger than zero: $\lambda^* > 0$. For $\epsilon_2 = \epsilon_E$, the second condition reduces to $\lambda^* - 2\lambda^* < 0$ which holds if the first condition is met. Therefore, the neuron finds the principal component of the modified input space while it takes the tuning of the inhibitory input population into account.

### Fast inhibition increases stability

Unlike many other inhibitory plasticity rules, we do not require that inhibitory plasticity is faster than excitatory plasticity. In the extreme case of static inhibition, $\epsilon_E = 0$, the second condition is still satisfied if the fixed point attraction $\lambda^*$ is larger than the excitatory attraction $\lambda^*$ of any other eigenvector alone. For growing $\epsilon_2 > 0$, the repulsive influence of the inhibitory part of competing eigenvectors increases until they become overamplified, and the second condition is always fulfilled. In practice, fast inhibition helps to stabilize the system and otherwise does not crucially affect the dynamics. Therefore, we consider slightly faster inhibition than excitatory plasticity for numerical simulations.

### Stability of non-eigenvector fixed points

In principle, we can choose the total synaptic excitatory and inhibitory weights maintained during plasticity. Until now, we assumed that fixed points $\mathbf{w}^*$ are multiples of eigenvectors $\mathbf{v}$ of the modified covariance matrix $\mathbf{C}$, which puts a strong constraint on our choice for the weight norms. As shown before, the general shape of a fixed point $\mathbf{w}^*$ is
\[
\mathbf{w}^* = \begin{pmatrix} kE \mathbf{v}_E \\ kI \mathbf{v}_I \end{pmatrix} = \begin{pmatrix} kE & 0 \\ 0 & kI \end{pmatrix} \mathbf{v} = \begin{pmatrix} kE & 0 \\ 0 & kI \end{pmatrix} \mathbf{v} = K \mathbf{v},
\]
where $kE$ and $kI$ are scalar constants. We recapitulate the general weight dynamics as given in Eq. 42:
\[
\tau \mathbf{w} = \mathbf{C} \mathbf{w} - \gamma \mathbf{w}_E^p - \rho \mathbf{w}_I^p.
\]
Instead of evaluating the Jacobian in a general fixed point, we switch to a new coordinate system in which the fixed point is an eigenvector of the modified covariance matrix. We define:

\[ \hat{\mathbf{w}} = \mathbf{K}^{12} \mathbf{w} \]

from which the weight dynamics becomes

\[ \dot{\mathbf{w}} = \hat{\mathbf{C}} \mathbf{w} - \hat{\mathbf{y}} E - \rho \hat{\mathbf{w}}. \]

\[ \dot{\mathbf{y}} = \mathbf{K}^{12} \hat{\mathbf{C}} \mathbf{w} = \frac{c_E^T \hat{\mathbf{C}} \mathbf{w}}{E} = \frac{c_I^T \hat{\mathbf{C}} \mathbf{w}}{I}, \quad \hat{\mathbf{y}} = \frac{\mathbf{c}_E^T \hat{\mathbf{C}} \mathbf{w}}{E} \]

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

\[ \mathbf{w}^* = \mathbf{K}^{-12} \mathbf{w}^* = \mathbf{K}^{-12} \mathbf{K} \mathbf{y} = \mathbf{K}^{12} \mathbf{y} \]

It is straightforward to prove that \( \mathbf{w}^* \) is an eigenvector of \( \hat{\mathbf{C}} \) with eigenvalue \( \lambda^* = (k_E - k_I \sigma^2) \lambda \). In principle, we can now proceed in finding the eigenvalues of the Jacobian, as explained before. However, we would have to employ the eigenvector basis in the new coordinates \( \mathbf{V} = \mathbf{K}^{-12} \mathbf{V} \) for triangularization. As before, one finds that stability is largely determined by the eigenvalues \( \lambda \).

**Effective timescales and attraction landscapes**

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

**4 Lateral input warps attraction landscape**

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

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

\[ \hat{\mathbf{y}} = (y^T, q^T y)^T \]

and the input weights onto the second neuron as

\[ \mathbf{w} = (w^T, w_q)^T \]

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

\[ \check{\mathbf{C}} = \langle \hat{\mathbf{y}} \hat{\mathbf{y}}^T \rangle = \begin{pmatrix} yy^T & yy^T q \\ q' y y^T & q' y y^T q' \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{C} q \\ q^T \mathbf{C} & q^T q^T \mathbf{C} \end{pmatrix} \]
**Figure S2: Input space warping due to lateral connectivity.** (A) Top: a single neuron with firing rate $r$ receives synaptic inputs $w$ from a population of excitatory neurons $y$. Bottom: input distribution projected onto the first two input dimensions. Each dot represents the firing rates of the first two neurons during one input pattern. (Contour lines in light gray). Under a linear Hebbian learning rule, the neuron becomes selective for the direction of maximum variance, the first principal component (see Section 2). (B) Top: Same as (A) for a neuron that receives input from a laterally projecting excitatory neuron which is tuned to an eigenvector $q$ of the original input covariance matrix. Bottom: the effective input space $y_{eff}$ of the target neuron (dark blue triangle) is warped such that the variance along the eigenvector $q$ (blue arrow) is stretched in proportion to the absolute value of the weight vector $q$. The contour lines of the original input distribution from (A) are shown in light gray for reference. (C) Top: Same as (B) for a laterally projecting inhibitory neuron. Bottom: Now, the effective input space is compressed. See text for details.

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

$$\begin{align*}
\tilde{w}^* &= \tilde{v} \equiv \begin{pmatrix} w^T \cdot w_q^* \end{pmatrix}^T, \\
\tilde{C}v &= \lambda v.
\end{align*}$$

(93)

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

$$\begin{align*}
\tilde{C}v &= \begin{pmatrix} C & Cq \\
qu^TC & q^TCq
\end{pmatrix} \begin{pmatrix} w \\
w_q
\end{pmatrix} = \lambda \begin{pmatrix} w \\
w_q
\end{pmatrix}.
\end{align*}$$

(94)

$$Cw + Cqw_q = \lambda w,$$

(95)

$$q^TCw + q^TCqw_q = \lambda wq,$$

(96)

$$C(w + qw_q) = \lambda w,$$

(97)

$$q^TC(w + qw_q) = \lambda wq.$$

(98)

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

$$\begin{align*}
C \left( I + qq^T \right) w &= \lambda w.
\end{align*}$$

(99)

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

$$\Lambda \left( I + q_vq_v^T \right) w_v = \lambda w_v,$$

(100)

with $\Lambda$ being the diagonal matrix of eigenvalues and the subscript $(\cdot)_v$ indicates a vector in the eigenbasis of $C$. In this basis, eigenvectors of $C$ are unit vectors, i.e., $v_v = e_v$, where $e_v$ is a vector of zeros with one entry equal to one that corresponds to the respective eigenvector. When $q$ contains components of more than one eigenvector, the matrix $q_vq_v^T$ is no longer diagonal and eigenvectors of $C$, $w_v = e_v$, do not solve the equation. However, when we assume that the first neuron has plastic feedforward input, we know that it will converge to a multiple of an eigenvector of the feedforward input $q \propto v_i$, where $Cv_i = \lambda_i v_i$. Then multiples of eigenvectors of $C$ solve Eq. 99. To find the eigenvalues...
\[ w \propto v \neq v \propto q \quad \Rightarrow \quad q^T w \propto v, v = 0, \quad \Rightarrow \quad \lambda_j = \lambda = \sigma_q^2. \] (101)

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

\[ w \propto v \propto q \quad \Rightarrow \quad \lambda = \lambda + i = \sigma^2 + \sigma_q^2, \] (102)

where \( a_q \) is equal to \( \|q\|_2 \), the L2-norm of \( q \). Input modes aligned with the tuning of the laterally projecting neuron increase their attraction by the variance of that neuron. Together this can be interpreted as a stretching of the original input space in the direction of the laterally projecting neuron’s feedforward weight vector \( q \) (Fig. S2B). Then fixed points are of the following general shape:

\[ w' = v, \quad q = a_q v, \] (103)

\[ \Rightarrow \quad \bar{w} = \begin{pmatrix} w' \\ a_q \end{pmatrix} = k \begin{pmatrix} v \\ a_q \end{pmatrix}, \] (104)

where \( k \) is a scalar constant. Similarly, when the laterally projecting neuron is inhibitory, the modified covariance matrix becomes \( \bar{C} = \begin{pmatrix} \bar{C} & -Cq \\ Cq^T & -q^T \bar{C} \end{pmatrix} \) (compare Eq. 53) and it follows that the effective input space is compressed\(^1\)(Fig. S2C):

\[ \lambda = \lambda - \lambda a_q^2 = \sigma^2 - \sigma_q^2, \] (105)

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

\[ \bar{\lambda} = \lambda \left( 1 + \|a_E\|^2 - |a_i|^2 \right), \] (106)

\[ \Rightarrow \quad \bar{\lambda} = \sigma^2 + \|\sigma_E\|^2 - |\sigma_i|^2, \] (107)

where \( a_E, a_i \) hold the vector norms of the laterally projecting neurons and \( \sigma_E, \sigma_i \) hold their standard deviations.

In this general case, one can write the fixed points as

\[ \bar{w} = K \begin{pmatrix} v \\ a_E \\ a_i \end{pmatrix} = K' \begin{pmatrix} \sigma v \\ \sigma_E \\ \sigma_i \end{pmatrix}, \] (108)

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

Note that these results are independent of what causes the laterally projecting neurons’ tuning. For example, in addition to feedforward input, a neuron can be integrated into a recurrent circuit of neurons that are all tuned to the same eigenvector. Then \( \sigma_q^2 \) results from recurrent interaction in addition to the norm of the feedforward weight vector \( \|q\|_2 \). We will consider such circuits in the following sections.\(^2\)

5 Eigencircuits

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

---

\(^1\) For sufficiently large vector norms \( \|q\|_2 \) the eigenvector mode becomes repulsive. Then the transformation of the input space can no longer be visualized as intuitively as in Fig. S2.

\(^2\) Another example is neurons that project from outside the local circuit, e.g., from another brain area that is higher up in the processing hierarchy.
eigenvectors is:
\[ \begin{align*}
\langle v^T y y^T v_j \rangle &= v_i^T C v_j, \\
\langle v_i^T \lambda_i v_j \rangle &= \lambda_i v_i^T v_j, \\
\langle v_i^T \lambda_i v_j \rangle &= 0.
\end{align*} \]

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

### 5.1 Variance propagation

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

\[ r = w^T y + w_E^T y_E - w_i^T y_i. \] (114)

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

\[ y_E = a_E(v^T y), \quad y_i = a_i(v^T y), \] (115)

where the vectors \( a_E \) and \( a_i \) set the vectors’ response magnitudes which are proportional to their standard deviations. For the weight vector, we require that the excitatory and inhibitory parts are normalized to maintain the total amount of inhibitory and excitatory synaptic resources.

\[ \begin{pmatrix} w \\ w_E \end{pmatrix} = W_E \frac{\bar{v}_E}{\|\bar{v}_E\|_p}, \quad w_i = W_i \frac{\bar{v}_i}{\|\bar{v}_i\|_p}. \] (116)

\[ \bar{v}_E = \begin{pmatrix} v \\ a_E \end{pmatrix}, \quad \bar{v}_i = a_i. \] (117)

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

\[ r = \left( 1 + \frac{\|a_E\|^2}{\|\bar{v}_E\|_p^2} W_E - \frac{\|a_i\|^2}{\|\bar{v}_i\|_p^2} W_i \right) (v^T y). \] (118)

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

\[ \sigma_r = \left( 1 + \frac{\|a_E\|^2}{\|\bar{v}_E\|_p^2} W_E - \frac{\|a_i\|^2}{\|\bar{v}_i\|_p^2} W_i \right) \sigma = \left( 1 + \frac{\|a_E\|^2}{\|\bar{v}_E\|_p^2} \sigma^2 W_E - \frac{\|a_i\|^2}{\|\bar{v}_i\|_p^2} \sigma W_i \right) \] (119)

\[ \Rightarrow \sigma_r = \frac{\|\sigma^E\|^2}{\|\sigma^E\|_p^2} W_E - \frac{\|\sigma^I\|^2}{\|\sigma^I\|_p^2} W_i, \] (120)

\[ \sigma^E = (\sigma, \sigma^E)^T, \quad \sigma^I = \sigma_i. \] (121)

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

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

\[ \sigma_i = W_{EE} \left( \frac{\sigma^2 + n_i \sigma_E^2}{\sigma + n_i \sigma_E} \right) - W_{EI} \left( \frac{n_i \sigma_i^2}{\sigma_i} \right), \]

which becomes

\[ \sigma_i \approx \sigma_i^* \approx \sigma_i \forall i. \]  

The standard deviations of excitatory and inhibitory neurons become

\[ \sigma_E = W_{EE} \left( \frac{\sigma^2 + n_E \sigma_E^2}{\sigma + n_E \sigma_E} \right) - W_{EI} \left( \frac{n_E \sigma_i^2}{\sigma_i} \right), \]

\[ \sigma_I = W_{IE} \left( \frac{\sigma^2 + n_E \sigma_E^2}{\sigma + n_E \sigma_E} \right) - W_{II} \left( \frac{n_I \sigma_i^2}{\sigma_i} \right). \]

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

\[ \sigma_i = \frac{W_{IE}}{1 + W_{II} \Phi} \sigma_E. \]

\[ \Phi = \left| \frac{W_{EE} - W_{EI} W_{II}}{1 + W_{II}} \right|. \]

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

Note that for \( \Phi < 1 \) there exists a real solution for \( \sigma_E \), independent of \( n_E \).

5.3 A note on the choice of weight norm

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

\[ \sigma_i = \frac{W_{IE}}{1 + W_{II}} \sigma. \]

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

\[ \sigma_i = \frac{W_{IE}}{1 + W_{II}} \left( \sigma^2 + \left( \frac{\sigma_E^2}{\sigma + \sigma_E} \right)^2 \right)^{\frac{1}{2}}, \]

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

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

6 Fully plastic recurrent E-I networks

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

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

\[
\tau \frac{d w}{d w} = \left[ I - \ldots \right] \left( \hat{C}^* - \hat{\lambda}^* I + \frac{d \hat{C}}{d w} w^* \right),
\]

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

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

\[
y_E = w_{EF}^T y + w_{EE} y_E - w_{EI} y_I,
\]

\[
y_I = w_{IF}^T y + w_{IE} y_E - w_{II} y_I,
\]

\[
y_E = \frac{1}{1 - w_{EE} + w_{EF} w_{IE}} \left( w_{EF}^T - \frac{w_{EE}}{w_{II}} w_{IF}^T \right) y \equiv a_{E}^T y,
\]

\[
y_I = \frac{1}{1 + w_{II} + w_{IE} w_{EI}} \left( w_{IF} + w_{IE} w_{EF}^T \right) y \equiv a_{I}^T y.
\]

The weight dynamics is

\[
\tau \dot{w} = \left( \begin{array}{c} w_{EF} \\ w_{EE} \\ w_{EI} \\ y_E y_I \end{array} \right) = \left( \begin{array}{cccc} y y^T & y y_E & -y y_I & 0 \\ y_E y_I & y y_E & -y y_I & 0 \\ y_I y_E & -y_I y_E & y_I y_I & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \left( \begin{array}{c} y_E \\ y_I \\ y_E \\ y_I \end{array} \right) - \frac{\Phi}{\rho_e} \left( \begin{array}{c} 0 \\ y_E \\ 0 \\ 0 \end{array} \right) \left( \begin{array}{c} 0 \\ 0 \\ \rho_e \\ 0 \end{array} \right).
\]
where ellipsis indicate similar terms for afferent weights of the inhibitory neuron. We define the modified covariance matrix
\[
\hat{C} = \begin{pmatrix}
\langle yy' \rangle & \langle yE \rangle & \langle y_i y_i \rangle \\
\langle yE y' \rangle & \langle yE^2 \rangle & \langle y_i y_i \rangle \\
\langle y_i y_i \rangle & \langle y_i y_i \rangle & 0 \\
0 & 0 & \ldots
\end{pmatrix}
= \begin{pmatrix}
C & CA_E & -Ca_i \\
Ca^T C & a_i^T Ca_E & -a_i^T Ca_i \\
0 & a_i^T Ca_E & -a_i^T Ca_i \\
\end{pmatrix}
\]
(139)
and write the average synaptic change as
\[
\tau(w) \equiv \hat{C}w - \Gamma w.
\]
(140)
where \(\Gamma\) is a diagonal matrix that holds the scalar constraint factors (compare Eq. 42). Note that this is a non-linear dynamical system since the modified covariance matrix depends on plastic synaptic weights. We make the simplifying assumption that the plasticity of excitatory and inhibitory synapses is equally fast, \(\tau_E = \tau_I\). Then \(\tau = \tau I\), which does not affect the fixed points or the stability of the system\(^1\). Therefore, we set \(\tau = 1\).

### 6.1 Fixed points

Fixed points \(w^*\) must fulfill the following condition
\[
\hat{C}^* w^* - \Gamma^* w^* \perp 0.
\]
(141)
where \(\hat{C}^*\) is the modified covariance matrix evaluated in the fixed point. We consider a fixed point where all neurons form a single eigencircuit and are tuned to the same input mode \(v^*\). Then we can write the excitatory and inhibitory firing rates as
\[
y_E' = a_E^T y = \bar{y}^T a_E^*, \quad a_E^* = a_E^* v^*,
\]
(142)
y_i' = a_i^T y = y_i^T a_i^*, \quad a_i^* = a_i^* v^*.
(143)
Note that the superscript \(^*\) indicates a value in the fixed point of the weight dynamics and not a fixed point of the firing rate activity. Different input patterns \(y\) still result in different neural activities \(y_E'^*\). The modified covariance matrix in the fixed point becomes
\[
\hat{C}^* = \begin{pmatrix}
C & Cv^* a_E^* & -Cv^* a_i^* \\
\lambda a_E^* v^* C & a_E^* v^* C v^* a_E^* & -a_E^* v^* C v^* a_i^* \\
\lambda a_i^* v^* C & a_i^* v^* C v^* a_E^* & -a_i^* v^* C v^* a_i^* \\
\end{pmatrix}
\]
(144)
\[
= \begin{pmatrix}
C & \lambda a_E^* v^* & -\lambda a_i^* v^* \\
\lambda a_i^* v^* & \lambda a_E^* v^* & -\lambda a_i^* v^*
\end{pmatrix}
\]
(145)
which can be diagonalized by the eigenvector matrix \(V^*\) and its inverse:
\[
V^* = \begin{pmatrix}
1 & v^* a_E^* & v^* a_i^* \\
0 & a_E^* -1 & 0 \\
0 & a_i^* & 1 \\
0 & 0 & \ldots
\end{pmatrix}, \quad V^{-1} = N^{-1} \begin{pmatrix}
N v^* \top & 0 & 0 \\
v^* \top & a_E^* & -a_i^* \\
y_i^* \top & -(1 - a_i^2) & -a_i^* a_i^* \\
0 & -a_i^* v^* \top & -a_i^* a_i^* + 1 + a_i^2
\end{pmatrix},
\]
(146)
\[
N \equiv 1 + a_E^2 - a_i^2.
\]
(147)
where the subscript \((\cdot)_v\) indicates that a matrix does not contain an entry that corresponds to the input mode \(v^*\). The first set of eigenvectors has eigenvalues \(\lambda_{1v}\), which are eigenvalues of the feedforward covariance matrix \(C\), since

\(^1\)It does not affect the sign of the eigenvalues of the Jacobian, since \(\tau\) is always positive. In principle, however, different timescales for excitatory and inhibitory weights can affect stability (compare Section 3.2).
there are no neurons in the recurrent circuit\(^1\) that are tuned to these input modes. The last two eigenvectors in the first block are null eigenvectors with eigenvalue zero. Their excitatory feedback and recurrent components are balanced by either an inhibitory component or a negative recurrent excitatory component. The remaining eigenvector has eigenvalue \(1 + a_E^2 - a_I^2\). Similar to the feedforward case (compare section 3.1), arbitrary multiples of the separately normalized parts of eigenvectors of \(\mathbf{\hat{C}}\) are fixed points\(^3\). Inserting these fixed points into Eq. 136 and 137, provides conditions to determine \(a_E^*\) and \(a_I^*\). We consider the simplest case of an eigenvector of the following shape:

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

### 6.2 Stability analysis

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

\[
\Delta \mathbf{w} \propto \begin{pmatrix} \mathbf{v}^* \\ 0 \end{pmatrix}. \quad (149)
\]

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

\[
\frac{d}{dt}(\Delta \mathbf{w}) = \mathbf{J}^*(\Delta \mathbf{w}). \quad (150)
\]

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

\[
\Delta \mathbf{w} = \mathbf{\hat{V}} \Delta \mathbf{w}_r. \quad (151)
\]

and its dynamics becomes

\[
\frac{d}{dt}(\Delta \mathbf{w}_r) = \frac{d}{dt} \left(\mathbf{\hat{V}}^{-1} \Delta \mathbf{w}\right) = \mathbf{\hat{V}}^{-1} \mathbf{J}^* \mathbf{\hat{V}}^* \mathbf{\hat{V}}^{-1} (\Delta \mathbf{w}) = \mathbf{\hat{V}}^{-1} \mathbf{J}^* \mathbf{\hat{V}}^* \mathbf{e}^i, \quad (152)
\]

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

#### The transformed Jacobian

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

\[
\mathbf{w} = \begin{pmatrix} \mathbf{1} - (\mathbf{w}_{EF} + \mathbf{w}_{EE}) \mathbf{C}_{EE}^T \\ \mathbf{C}_{EE}^T (\mathbf{w}_{EF} + \mathbf{w}_{EE}) \end{pmatrix} ... \mathbf{\hat{C}} \mathbf{w} \quad (153)
\]

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

---

\(^1\)No neurons that are not input neurons.

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

\(^3\)The only exception is the rightmost null-eigenvector. There, the inhibitory and the excitatory weights are aligned such that the postsynaptic activity is zero, which does not allow for arbitrary scaling of the weight norms.
Jacobian has the following shape (compare Eq. 31)

\[
J^* = \left[ \frac{d\mathbf{w}^*}{d\mathbf{w}} \right] = \left[ \begin{array}{c} 1 \end{array} \right] - \left( \begin{array}{c} c_{EE}^T (\mathbf{w}^* + \mathbf{w}_0) \end{array} \right) - \ldots \left[ \begin{array}{c} \mathbf{C} - \hat{\lambda}^* \mathbb{I} + \hat{\mathbf{C}} \end{array} \right] \frac{d\mathbf{C}}{d\mathbf{w}}, \tag{154}
\]

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

\[
\hat{\mathbf{V}}^{-1} \left[ 1 - \left( \frac{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0}{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0} \right) c_{EE}^T - \ldots \right] \hat{\mathbf{V}}^* \tag{155}
\]

\[
= \hat{\mathbf{V}}^{-1} \left[ 1 - \left( \frac{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0}{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0} \right) c_{EE}^T \right] \ldots \left[ \begin{array}{c} \mathbf{V}_i \v^* a_E^* \v^* a_i^* \end{array} \right] \tag{156}
\]

\[
= \left[ 1 - \hat{\mathbf{V}}^{-1} \mathbf{H} \hat{\mathbf{v}}^* \right] \mathbf{H}_l \equiv \left( \begin{array}{c} h_{EE}^{(i)} \hbar_{EE}^{(i)} \hbar_{EE}^{(i)} \hbar_{EE}^{(i)} \end{array} \right), \tag{157}
\]

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

\[
h_{EE}^{(i)} \equiv \frac{c_{EE}^T}{c_{EE}^T (\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0)} \hat{\mathbf{V}}^*_i, \tag{158}
\]

where \( \mathbf{V}^*_i \) is the \( i \)th column of \( \mathbf{V}^* \). Then

\[
\hat{\mathbf{V}}^{-1} \mathbf{H} \hat{\mathbf{v}}^* = N^{-1} \begin{pmatrix} N \mathbf{v}_E^T & 0 & 0 \\ -a_E^T v_E^T & -a_i^T & -a_i^T \\ -a_i^T v_E^T & a_E^T v_i & 1 + a_E^2 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{H} \hat{\mathbf{v}}^* = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{159}
\]

\[
\Rightarrow \hat{\mathbf{V}}^{-1} \left[ 1 - \left( \frac{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0}{\mathbf{V}_{EE}^* + \mathbf{V}_{EE}^0} \right) c_{EE}^T \right] \ldots \hat{\mathbf{V}}^* = \left( \begin{array}{c} 1 \ 0 \ 0 \ 0 \\ 0 \ 0 \ 0 \ 0 \\ \vdots \ \vdots \ \vdots \ \vdots \\ 0 \ 0 \ 0 \ 0 \end{array} \right) \tag{160}
\]

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

\[
\hat{\mathbf{V}}^{-1} \left( \hat{\mathbf{C}} - \hat{\lambda}^* \mathbb{I} + \hat{\mathbf{C}} \right) \frac{d\hat{\mathbf{C}}}{d\mathbf{w}} \hat{\mathbf{V}}^* = \left( \tilde{\lambda}^* - \hat{\lambda}^* \mathbb{I} + \hat{\mathbf{V}}^{-1} \frac{d\hat{\mathbf{C}}}{d\mathbf{w}} \hat{\mathbf{V}}^* \right). \tag{161}
\]
We next consider the first columns of \( \frac{d\hat{C}}{d w_{EF}^B} \) \( \mathbf{w}^* \), for which we compute the matrix \( \frac{d\hat{C}}{d w_{EF}^B} \), where \( w_{EF}^B \) is the \( b \)th feedforward weight onto the excitatory neuron.

\[
\frac{d\hat{C}}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \frac{d}{d w_{EF}^B} \begin{pmatrix} C & a_E^T C & -a_E^T C a_i & 0 \\ a_i^T C & a_i^T C & -a_i^T C a_i & 0 \\ 0 & 0 & \ddots & \ddots \end{pmatrix} = \begin{pmatrix} 0 & C & \frac{d a_E}{d w_{EF}^B} & \frac{d a_i}{d w_{EF}^B} \\ C & a_E^T C & \frac{d a_E}{d w_{EF}^B} & \frac{d a_i}{d w_{EF}^B} \\ \frac{d a_i}{d w_{EF}^B} & \frac{d a_i}{d w_{EF}^B} & \ddots & \ddots \end{pmatrix} \]

(162)

where we used the definition of \( \hat{C} \) from Eq. 139. The vectors \( a_E \) and \( a_i \) are defined in Eq. 136 & 137. It follows:

\[
\frac{d a_E}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \frac{1}{1 - w_{EE}^* + \frac{w_{EE} E^*}{1 + \lambda E^*}} \mathbf{e}_b \equiv \mu E \mathbf{e}_b, \quad (164)
\]

\[
\frac{d a_i}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \frac{1}{1 + \lambda E^* - \frac{w_{EE} E^*}{1 - w_{EE}^*}} \mathbf{e}_b \equiv \mu I \mathbf{e}_b, \quad (165)
\]

\[
\mathbf{C a}_E^* = \lambda a_E^* \mathbf{v}^*, \quad \mathbf{C a}_i^* = \lambda a_i^* \mathbf{v}^*. \quad (166)
\]

\[
\Rightarrow \frac{d\hat{C}}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \begin{pmatrix} 0 & \mu E C \mathbf{e}_b & -\mu C \mathbf{e}_b \\ \mu E \mathbf{e}_b^T C & 2 \lambda a_E^* \mu E \mathbf{v}^T \mathbf{e}_b & -\lambda (\mu E a_E^* + \mu a_E^* \mathbf{v}^T) \mathbf{e}_b \\ \mu a_i^* \mathbf{v}^T C & \lambda E (\mu a_i^* + \mu a_i^* \mathbf{v}^T \mathbf{e}_b) & 2 \lambda a_i^* \mu E \mathbf{v}^T \mathbf{e}_b \end{pmatrix} \]

(167)

\[
\Rightarrow \frac{d\hat{C}}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \begin{pmatrix} \beta_E \mathbf{C e}_b \\ g_1 \mathbf{v}^T \mathbf{e}_b \\ g_2 \mathbf{v}^T \mathbf{e}_b \\ 0 \end{pmatrix}, \quad \mathbf{w}^* = \begin{pmatrix} \mathbf{v}^* \\ \mathbf{w}_{EE}^* \\ \mathbf{w}_{EI}^* \\ \vdots \end{pmatrix}, \quad \beta_E = \mu E \mathbf{w}_{EE}^* - \mu I \mathbf{w}_{EI}^*. \quad (168)
\]

\[
\Rightarrow \frac{d\hat{C}}{d w_{EF}^B} \bigg|_{\mathbf{w}^*} = \begin{pmatrix} \beta_E C \\ g_1 \mathbf{v}^T \\ g_2 \mathbf{v}^T \\ 0 \end{pmatrix} \quad (169)
\]

We find other columns in a similar fashion and write

\[
\Rightarrow \frac{d\hat{C}}{d \mathbf{w}^*} = \begin{pmatrix} \beta_E C & g_3 \mathbf{v}^* \\ g_1 \mathbf{v}^T & g_4 \\ g_2 \mathbf{v}^T & g_5 \\ 0 \end{pmatrix}. \quad (170)
\]
where \(g_{i,j}\) are scalars. After applying the transformation, we get

\[
\mathbf{V}^{-1} \frac{d\mathbf{C}}{d\mathbf{w}^*} \mathbf{w}^* \mathbf{v}^* \tag{171}
\]

\[
\mathbf{V}^{-1} = \begin{pmatrix}
\beta_E C & g_3 v^* & g_6 v^* \\
g_1 v^T & g_4 & g_7 \\
g_2 v^T & g_5 & g_8 \\
0 & \ldots & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
V_1^* & v^* a_E^* & v^* a_j^* \\
0 & a_E^* & 0 \\
0 & a_j^* & 1 \\
0 & \ldots & 0
\end{pmatrix}
\tag{172}
\]

\[
\mathbf{V}^{-1} = \begin{pmatrix}
0 & g_9 v^* & g_{12} v^* & g_{15} v^* \\
g_10 & g_{13} & g_{16} & 0 \\
g_{11} & g_{14} & g_{17} & 0 \\
0 & \ldots & 0
\end{pmatrix}
\tag{173}
\]

\[
\mathbf{N}^{-1} = \begin{pmatrix}
(\mathbf{N} V_1^T) & 0 & 0 & 0 \\
V^T & a_E^* & 0 & 0 \\
a_E v^T & a_j^* & g_9 v^* & g_{12} v^* & g_{15} v^* \\
0 & \ldots & 0 & g_{10} & g_{13} & g_{16} & 0 \\
0 & g_{11} & g_{14} & g_{17} & 0
\end{pmatrix}
\tag{174}
\]

\[
\begin{pmatrix}
\beta_E \Lambda^* & 0 & 0 & 0 \\
0 & g_{18} & g_{21} & g_{24} & 0 \\
0 & g_{19} & g_{22} & g_{25} & 0 \\
0 & g_{20} & g_{23} & g_{26} & 0
\end{pmatrix}
\tag{175}
\]

The fully transformed Jacobian becomes (compare Eq. 161)

\[
\mathbf{V}^{-1} \mathbf{J} \mathbf{V}^* = \mathbf{V}^{-1} \left[ \mathbf{I} - \left( \frac{\mathbf{v}_E^{\mathbf{C}}}{\mathbf{v}_E^{\mathbf{C}}} \mathbf{C}_E^T \mathbf{C}_E \right) - \ldots \right] \mathbf{V}^{-1} \left( \hat{\lambda}^* - \hat{\lambda}^* \mathbf{I} + \hat{\mathbf{V}}^{-1} \frac{d\mathbf{C}}{d\mathbf{w}^*} \mathbf{w}^* \mathbf{v}^* \right) \tag{176}
\]

Finally, by inserting Eq. 160 and 175 we get

\[
\mathbf{V}^{-1} \mathbf{J} \mathbf{V}^* = \begin{pmatrix}
\mathbf{I} & 0 & 0 & 0 \\
0 & \ldots & 0 & 0 \\
0 & \ldots & \hat{\lambda}^* & 0 \\
0 & \ldots & 0 & \hat{\lambda}^* \mathbf{I} + \hat{\mathbf{V}}^{-1} \frac{d\mathbf{C}}{d\mathbf{w}^*} \mathbf{w}^* \mathbf{v}^* \end{pmatrix}
\tag{177}
\]

\[
\Rightarrow \mathbf{V}^{-1} \mathbf{J} \mathbf{V}^* = \begin{pmatrix}
\mathbf{I} & 0 & 0 & 0 \\
0 & \ldots & 0 & 0 \\
0 & \ldots & \hat{\lambda}^* - \hat{\lambda}^* \mathbf{I} + \beta_E \Lambda^* & 0 \\
0 & \ldots & 0 & \hat{\mathbf{V}}^{-1} \frac{d\mathbf{C}}{d\mathbf{w}^*} \mathbf{w}^* \mathbf{v}^* \end{pmatrix}
\tag{178}
\]
Stability conditions

The dynamics of a fixed point perturbation

\[
\frac{d}{dt}(\Delta \mathbf{w}_t) = \left( \begin{array}{c}
\Lambda_{\text{E}} - 1 - \hat{\lambda}^\dagger + \beta_E \Lambda_{\text{E}} \\
\vdots \\
0
\end{array} \right) 0 0 \Delta \mathbf{w}_t.
\]

(179)

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

\[
\frac{d}{dt}(\mathbf{e}_0^\dagger \mathbf{e}) = \left( \hat{\lambda}^\dagger - \hat{\lambda}^\dagger + \beta_E \hat{\lambda}^\dagger \right) (\mathbf{e}_0^\dagger \mathbf{e}).
\]

(180)

which provides the general stability condition for the excitatory neuron

\[
\left( \hat{\lambda}^\dagger - \hat{\lambda}^\dagger + \beta_E \hat{\lambda}^\dagger \right) < 0
\]

(181)

where \(\hat{\lambda}^\dagger = \lambda^\dagger + \lambda_{\text{eig}}^\dagger\). In the circuit considered here, we integrated the eigencircuit attraction into the feedforward input, i.e., \(\lambda_{\text{eig}}^\dagger\) is zero. For \(\beta_E\) we find

\[
\beta_E = \frac{1}{1 - w_{\text{EE}}^* + \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}} - \frac{1}{1 + w_{\text{ IL}}} \left( \frac{w_{\text{IE}} w_{\text{IE}}}{1 - w_{\text{EE}}} \right)},
\]

(182)

\[
\frac{dy_E}{d(\mathbf{w}_{\text{EF}} \mathbf{y})} = \left. \frac{w_{\text{EE}}^* - \frac{dy_I}{d(\mathbf{w}_{\text{IF}} \mathbf{y})}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(183)

\[
\frac{dy_E}{d(\mathbf{w}_{\text{EF}} \mathbf{y})} = \left. \frac{w_{\text{EE}} - \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(184)

\[
\Rightarrow \beta_E = \frac{\frac{dy_E}{d(\mathbf{w}_{\text{EF}} \mathbf{y})}}{1 - \frac{1}{1 - w_{\text{EE}}}},
\]

(185)

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

\[
\frac{dy_E}{d(\mathbf{w}_{\text{EF}} \mathbf{y})} = \left. \frac{w_{\text{IE}} - \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(186)

\[
\frac{dy_I}{d(\mathbf{w}_{\text{IF}} \mathbf{y})} = \left. \frac{w_{\text{IE}} - \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(187)

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

\[
\left( \hat{\lambda}^\dagger - \hat{\lambda}^\dagger + \beta_I \hat{\lambda}^\dagger \right) < 0
\]

(188)

\[
\beta_I = \frac{1}{1 - w_{\text{EE}}^* + \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}} - \frac{1}{1 + w_{\text{ IL}}} \left( \frac{w_{\text{IE}} w_{\text{IE}}}{1 - w_{\text{EE}}} \right)},
\]

(189)

\[
\frac{dy_E}{d(\mathbf{w}_{\text{IF}} \mathbf{y})} = \left. \frac{w_{\text{IE}} - \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(190)

\[
\frac{dy_I}{d(\mathbf{w}_{\text{IF}} \mathbf{y})} = \left. \frac{w_{\text{IE}} - \frac{w_{\text{IE}} w_{\text{IE}}}{1 + w_{\text{ IL}}}}{w_{\text{EE}}^*} \right|_{\mathbf{w}_{\text{IF}}},
\]

(191)

\(^1\text{Compare Eq. 158: } \lambda_{\text{EE}}^\dagger \neq 0.\)
6.3 Decorrelation condition

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

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

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

$$\lambda^a \cdot 1 - \lambda^b < 1,$$  \tag{194}

$$N_E \sigma_{E}^2 N_I \sigma_{I}^2 < 0,$$  \tag{195}

$$N_E \sigma_{E}^2 < N_I \sigma_{I}^2.$$  \tag{196}

where $\sigma_{E}^2$, $\sigma_{I}^2$ are the average variances of the population and $N_E$, $N_I$ are the total number of inhibitory and excitatory neurons.

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

$$\Delta \lambda = \lambda(n_E + 1, n_I) - \lambda(n_E, n_I).$$  \tag{198}

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

$$\lambda(n_E, n_I) = \lambda + n_E \sigma_{E}(n_E, n_I) + n_I \sigma_{I}(n_E, n_I).$$  \tag{199}

$$\eta = \frac{W_{E} - 1}{W_{I} + 1},$$  \tag{200}

The variance $\sigma_{E}$ does not depend on $n_I$ and strictly decreases with $n_E$ (compare Eq. 128)$^3$. Then, an upper bound for the change in total attraction is$^4$

$$\Delta \lambda < \left( 1 - \eta^2 \frac{N_I}{N_E} \right) \sigma_{E}(n_E + 1, n_I) \frac{N_I}{N_E} < 0.$$  \tag{201}

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

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

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

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

24
Figure S4: Input space warping due to lateral connectivity. (A) Two excitatory neurons (triangles) are tuned to two different but equally attractive input modes (circles, green and purple). Both neurons recurrently connect to themselves but not each other. (B) The same circuit as in A, unfolded to highlight presynaptic partners. Both input modes are balanced in their attraction. (C) Perturbing the purple excitatory neuron towards the green input mode (dashed lines) shifts its tuning (dark blue) such that it now responds to both the green and the purple input modes. (D) Unfolded circuit shown in C. Due to the perturbation, the green input mode is now more attractive, and the previously purple excitatory neuron shifts its tuning. (E) Stabilizing (red) and destabilizing (blue) effect of different circuit patterns in case of a perturbation of a neuron in the direction of an equally attractive input mode (dashed line). Recurrent excitation destabilizes (top, left) while monosynaptic recurrent inhibition stabilizes the circuit (top right). Similarly, disynaptic recurrent inhibition has a stabilizing effect (center). In a fully recurrent E-I network (bottom), the net effect results from recurrent excitatory and inhibitory connectivity motifs. The self-inhibition of a recurrently connected inhibitory neuron decreases the stabilizing effect of the disynaptic recurrent inhibition pattern.

\[ \eta^2 > \frac{N_E}{N_I} \iff N_E \sigma_E^2 < N_I \sigma_I^2 \]  

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

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

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

References


[3] Erkki Oja. “Simplified neuron model as a principal component analyzer”. In: Journal of mathematical biology 15.3 (1982), pp. 267–273. doi: https://doi.org/10.1101/2022.03.11.483899. This version posted March 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

