Inhibitory plasticity balances excitation and inhibition in sensory processing and Hebbian assemblies

T.P. Vogels$^{1*}$, H. Sprekeler$^{1*}$, F. Zenke$^1$, C. Clopath$^{1,2}$, and W. Gerstner$^1$

$^1$School of Computer and Communication Sciences and Brain-Mind Institute, École Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL

SWITZERLAND

$^2$ CNRS, UMR 8119, Universite Paris Descartes, 45 Rue des Saints Peres

75270 Paris Cedex 06

FRANCE

$^*$These authors contributed equally to this work.

$^\dagger$To whom correspondence should be addressed; E-mail: tim.vogels@epfl.ch.

“A Hebbian form of synaptic plasticity at inhibitory synapses generates balanced input currents and sparse neuronal responses that stabilize memory traces in neuronal networks”

Cortical neurons receive balanced excitatory and inhibitory membrane currents. Here, we show that such a balance can be established and maintained in an experience-dependent manner by synaptic plasticity at inhibitory synapses. The mechanism we put forward provides an explanation for the sparse firing patterns observed in response to natural stimuli and fits well with a recently observed interaction of excitatory and inhibitory receptive field plasticity. We show that the introduction of inhibitory plasticity in suitable recurrent networks provides a homeostatic mechanism that leads to asynchronous irregular network states. Further, it can accommodate synaptic memories with activity patterns that become indiscernible from the background state, but can be re-activated by external stimuli. Our results suggest an essential role of inhibitory plasticity in the formation and maintenance of functional cortical circuitry.
The balance of excitatory and inhibitory membrane currents a neuron experiences during stimu-
lated and ongoing activity has been the topic of many recent studies (1–14). This balance, first
defined as equal average amounts of de- and hyperpolarizing membrane currents (from hereon
referred to as “global balance”) is thought to be essential for maintaining stability of cortical net-
works (1, 2). In the balanced state networks display asynchronous irregular (AI) dynamics that
mimic activity patterns observed in cortical neurons. Such asynchronous network states facili-
tate rapid responses to small changes in the input (2–4), providing an ideal substrate for cortical
signal processing (5, 15, 16). Pathologies that disrupt the balance of excitation and inhibition
have often been implicated in neurological diseases such as epilepsy or schizophrenia (17, 18).

Moreover, the input currents to a given cortical neuron are not merely globally balanced. Exci-
tatory and inhibitory inputs are coupled also in time (6–8) and co-tuned for different stimulus
features (9, 10). The tight coupling of excitation and inhibition suggests a more precise, detailed
balance, in which each excitatory input arrives at the cell together with an inhibitory counter-
part, supposedly supplied through feedforward inhibition (Fig. 1A). These observations fit well
with models of cortical processing in which balanced sensory inputs are left unattended, but can
be transiently (11), or persistently turned on by targeted disruptions of the balance (12–14).

Although it is widely thought that the excitatory-inhibitory balance plays an important role
for stability and information processing in cortical networks, it is still not understood by which
mechanisms this balance is established and maintained in the presence of ongoing sensory expe-
riences. Inspired by recent experimental results (9), we investigate the hypothesis that synaptic
plasticity at inhibitory synapses plays a central role in balancing the excitatory and inhibitory
inputs a cell receives.

To investigate the effect of inhibitory synaptic plasticity in networks with feedforward inhibition,
we simulated a single postsynaptic integrate-and-fire neuron receiving correlated excitatory and
inhibitory input signals. The cell received input through 1000 synapses (Fig. 1B), which were
divided into eight independent groups of 100 excitatory and 25 inhibitory synapses, each. All
excitatory and inhibitory synapses within each group follow the same temporally modulated
rate signal (time constant $\tau \sim 50$ ms) to mimic ongoing sensory activity (15,19,20). Spikes are
generated from independent Poisson processes, leading to 125 different spike trains per signal. This architecture allows each signal to reach the cell simultaneously through both, excitatory and inhibitory synapses (Fig. 1B). To mimic glutamatergic and GABAergic transmission, the synapses were conductance-based with reversal potentials $V^E = 0$ mV and $V^I = -80$ mV and time constants $\tau^E = 5$ ms, and $\tau^I = 10$ ms for excitation and inhibition respectively (see Methods). In our simulations, the strength of the inhibitory synapses was initially weak but could change due to a spike timing dependent plasticity rule, in which near coincident pre- and postsynaptic action potentials induce potentiation of the synapse (21–26). In addition to this Hebbian term of the learning rule, every presynaptic spike leads to synaptic depression (21,22) (Fig. 1C, see also Methods). Schematically this learning rule can be summarized as

$$\Delta w = \eta \left[ \text{pre} \times \text{post} - \rho_0 \times \text{pre} \right],$$

where $\Delta w$ denotes the change in synaptic efficacy, $\text{pre}$ and $\text{post}$ are the pre- and postsynaptic activity, $\eta$ is the learning rate, and $\rho_0$ is a constant that acts as a target rate for the postsynaptic neuron (see Supplementary Materials for a mathematical analysis).

While inhibitory synapses are plastic, the efficacies of the excitatory model synapses are fixed for each signal channel at the beginning of a simulation and left unchanged unless otherwise noted. Analogous to frequency- or orientation-tuned sensory neurons, excitatory synapses were tuned to have a preferred signal (Fig. 1E). Since all excitatory synapses are set to non-zero strengths, the postsynaptic neuron fires at high rates when the inhibitory synapses were weak at the beginning of a simulation (Fig. 1D, E (upper panels), F). The resulting high number of pairs of pre- and postsynaptic action potentials automatically leads to relatively indiscriminate strengthening of all inhibitory synapses (Fig. 1D, E (middle panels)), until excitatory and inhibitory membrane currents become approximately balanced and the postsynaptic firing rate is dramatically reduced (Fig. 1F). In this globally balanced state, only unbalanced excitatory signals lead to coincident pairs of pre- and postsynaptic spikes, consequently strengthening underpowered inhibitory synapses. Those inhibitory synapses that are stronger than their excitatory counterparts will keep the postsynaptic side unresponsive and will thus be weakened (due to sole presynaptic firing) until they allow postsynaptic spiking again. Over time, this leads to a precise, detailed balance of excitatory and inhibitory synaptic weights for each chan-
nel (Fig. 1 D, E (lower panels)). In agreement with the mathematical analysis (Supplementary Materials), the postsynaptic firing rate is determined mainly by the depression factor $\rho_0$, but not by the average input firing rate to the postsynaptic neuron (Fig. 1 G). The mechanism is robust to plausible delays of several milliseconds. However, a gradual deterioration of the detailed balance occurs when the delay between excitation and inhibition increases to values larger than the autocorrelation time of the input signals and the coincidence time of the Hebbian learning rule, but global balance across all channels persists (Supplementary Materials).

To investigate how the state of the balance affects the neuron’s response properties we presented a fixed stimulus sequence to the neuron (Fig. 2 A) and compared the spiking response over 50 trials to the input rates of each signal. In the globally balanced state (Fig. 2 B, upper panels) in which inhibitory synapses are distributed so that excitation and inhibition are balanced only on average across all channels, the peristimulus time histogram (PSTH) faithfully reproduces the firing rates of the preferred signals. The other, non-preferred input signals evoke more inhibition than excitation and have thus no impact on the cell’s firing behavior. An additional step-like input rate protocol, in which 100 ms long pulses of various step sizes (Fig. 2 C) are presented to one channel at a time, reveals that spiking responses are largely insensitive to stimulus intensity and indeed narrowly tuned to the preferred stimulus, giving rise to an all-or-none response (Fig. 2 D, E.)

In the detailed balanced state (established by inhibitory plasticity), the response of the cell is sparse (Fig. 2 B, lower panels) and reminiscent of experimental observations (20, 27–30) across many sensory systems. Action potentials are caused primarily by transients in the input signals, during which the faster dynamics of the excitatory synapses momentarily overcome inhibition. Sustained episodes of presynaptic firing, on the other hand, cause steady membrane currents that cancel each other and thus fail to evoke a postsynaptic response. Seemingly indifferent to the tuning of the excitatory synapses, each signal now contributes an equal part to the variance of the output signal, but the effect of the excitatory synaptic weights is uncovered by the step-like input protocol (Fig. 2 F). The broad, graded responses (as opposed to all-or-none) to preferred and non-preferred stimuli (Fig. 2 G, H) are in accord with experimental results (6, 9, 10, 31, 32) and confirm earlier theoretical studies arguing that sharp tuning is not a necessary feature for a sparse sensory representation (33–35). The sparsity of the response to each signal is a direct
consequence of the detailed balance of correlated excitatory and inhibitory synapses as described above, not of the specificity of the tuning curve. In short, sparse responses in time are possible despite a broad tuning in feature space.

The self-organizing dynamics of inhibitory plasticity imply that the excitatory-inhibitory balance is maintained, even in the presence of ongoing excitatory plasticity (Fig. 3). Experiments (9) in which a stimulus that altered the frequency tuning of excitatory input currents to pyramidal neurons in primary auditory cortex of adult rat point in a similar direction: The disrupted co-tuning of excitatory and inhibitory input currents (Fig. 3 A) prompted a compensatory response that subsequently changed the amplitude of the inhibitory input currents. After 180 minutes the cell had returned to a co-tuned state, albeit with a different preferred frequency (Fig. 3 B). If we disturb the co-tuning of a simulated neuron in a similar way (Fig. 3 C), inhibitory plasticity rebalances the excitatory input currents (Fig. 3 D, E) and stabilizes the output firing rates of the postsynaptic neurons with the same time course as experimentally observed. Quantitative agreement of the re-balancing dynamics (for both, synaptic depression and potentiation) is achieved by adjusting the learning rate $\eta$, depression factor $\rho_0$, and the average firing rate of the inhibitory input neurons (Methods).

The learning rule for inhibitory synapses does not rely on a feedforward structure to achieve low firing rates. It simply matches excitatory and inhibitory synapses with correlated activity as the optimal solution to achieve postsynaptic firing rates of $\rho_0$. We therefore tested whether inhibitory plasticity would be able to stabilize the dynamics of recurrent networks. In simulations of such networks (15, 36) (Methods) with plastic inhibitory synapses that are initially weak (Fig. 4 A), the resulting high firing rates and subsequent increase in inhibitory synaptic strengths caused by the plasticity rule indeed produce globally balanced input currents that lead to a self-organized AI network state (Fig. 4 B) with firing rates between 3 and 15 Hz.

We wondered if it were possible to introduce associative memories to the stabilized network by strengthening specific excitatory connections within dedicated groups of neurons. First proposed by Hebb (37), such “cell assemblies” aim to provide a physiologically plausible explanation of how groups of neurons receiving common inputs, form a memory. Groups of highly connected neurons have since been successfully embedded into large spiking networks (38) and shown to
self-sustain their activity without disrupting the global dynamics of the host network (15, 39–41) but the parameter space that guarantees stable performance is narrow and tuning is arduous. Additionally the question has been raised how useful such memory attractors can be for long term memory systems if only one of all stored memories can be active at a time, and potentially remain active for long periods of time, broadcasting the stored information into the network (39).

Intriguingly, inhibitory plasticity can solve some of these problems. After two arbitrarily chosen groups of excitatory neurons are turned into Hebbian assemblies by strengthening the excitatory connections within the groups five-fold, the assemblies temporarily fire at high rates and raise the background firing rate across the network (Fig. 4 C). The resulting increase of coincident spike pairs causes the inhibitory plasticity to increase the inhibitory synapses onto neurons in both assemblies until the global AI state is re-established (Fig. 4 D). After the excitatory and inhibitory inputs onto these neurons have been rebalanced, the firing rates of neurons in the cell assemblies become indistinguishable from the rest of the network, despite strongly imprinted memory traces in the excitatory synapses. Electrophysiological recordings of neuronal activity would thus not reveal the presence of a synaptic memory trace in this state.

Such quiescent memories would be useless unless they can be recalled. Memory retrieval can be achieved by momentarily disrupting the balance within a cell assembly, for example through additional excitatory input. It is sufficient to drive a small fraction of the cells of one assembly with a short pulse to re-activate all cells of that assembly (Supplementary Materials). Notably, the recall is asynchronous and irregular, as indicated by low correlations between neurons and large variability of the inter-spike intervals (Fig. 4 E). Despite the fact that we embedded two overlapping assemblies into the network, only one is activated. The rest of the network remains unperturbed in the AI state. Unlike traditional attractor networks, both cell assemblies can also be activated in unison by driving cells of both patterns simultaneously and their activity decays to the background state after the stimulus is turned off (Supplementary Materials).

Our results offer an explanation for how long-term memories can be stably embedded into networks as quiescent and overlapping Hebbian assemblies. Unlike previous studies, our network does not exhibit the behavior of an attractor network, in which activated cell assemblies will compete with each other, and the winning pattern often exhibits persistent elevated activity.
Instead, the network remains quiet unless the balance of one or more assemblies is modulated in favor of the excitation, and returns to the background state when balance is regained. We have shown this effect here by driving a subset of cells with an external stimulus, but there are several conceivable methods to modulate the balance of excitation and inhibition (42–44). The possibility to activate several patterns simultaneously allows for the analog combination of patterns and suggests the existence of larger composite memories, or classes of memories. The capacity of storable, and retrievable, patterns is likely to depend on complex interactions between dynamics, size, and connectivity of the assemblies and the host network (40), as well as several other parameters and is subject to future studies.

In summary, we have shown here that a simple, Hebbian plasticity rule on inhibitory synapses leads to robust and self-organized balance of excitation and inhibition that requires virtually no fine-tuning and captures a surprising number of recent experimental findings. The precision of the learned balance depends on the degree of correlation between the excitatory and the inhibitory inputs to the cell, ranging from a global balance in the absence of correlated inputs to a detailed balance for strong correlations. The phenomenon is robust to the shape of the learning rule, as long as it obeys two fundamental requirements: postsynaptic activity must potentiate activated inhibitory synapses, while in the absence of postsynaptic activity inhibitory synapses must decay. Since the balance is self-organized, inhibitory plasticity will most likely maintain balance also in the presence of excitatory plasticity, as long as excitation changes more slowly than inhibition, or when excitatory plasticity events are rare.

The mammalian brain hosts a wide variety of inhibitory cell types with different synaptic time scales, response patterns, and morphological target regions. Presumably, these cell types serve different functions, and consequently their synapses may obey several different plasticity rules (26). In a simplified model we have shown here that the dynamics of such inhibitory plasticity can powerfully contribute to the functional state of cortical architectures and may have a strong impact on cortical coding schemes.

Acknowledgments

Research was supported by Swiss National Science Foundation grants no 200020 13287 (Cod-
ing Characteristics) and CRSIK0 122697 (Sinergia). Additionally, TPV was supported by the European Community’s Seventh Framework Marie Curie International Reintegration grant no. 268436, and HS and FZ by the European Community’s Seventh Framework Program under grant agreement no. 243914 (BRAIN-I-NETS) and 237955 (FACETS-ITN), respectively. Thanks to Guillaume Hennequin and Alan Woodruff for helpful discussions.

Conflict of Interest

The authors declare no conflict of interest.

References


Figure 1: Inhibitory synaptic plasticity balances excitation and inhibition.

(A) Feedforward inhibition: Excitatory input reaches a target region through both direct excitation and indirect disynaptic inhibition. (B) Feedforward inhibition for a single postsynaptic cell: Eight groups of 100 excitatory and 25 inhibitory synapses each deliver spikes to a single postsynaptic cell. Spiking probabilities are homogeneous within the groups but vary in time, simulating eight separate (color coded) signal channels reaching the cell simultaneously through excitatory and inhibitory synapses. (C) Spike timing dependent learning rule: Near-coincident pre- and postsynaptic spikes potentiate inhibitory synapses (marked with * in A & B), while every presynaptic spike causes synaptic depression. (D) Total excitatory, inhibitory, and net membrane currents (plotted in black, grey, and green respectively) before, during, and after inhibitory synaptic plasticity. The resulting spikes are indicated as dots underneath each current plot. (E) Excitatory and inhibitory membrane currents (black and white symbols, respectively) evoked by each signal channel, averaged over 4s, before, during, and after inhibitory synaptic plasticity (upper, middle, and lower panel respectively). (F) Temporal evolution of the postsynaptic firing rate (solid line) and the average synaptic weights of the inhibitory synapses associated with 3 representative signals (dotted lines). *, #, and ⋄ indicate the times at which the upper, middle, and lower panels of D and E were recorded. (G) Average firing rate of the postsynaptic neuron after learning, plotted for different values of target firing rate $\rho_0$ (left), and different input rates (right). The dashed lines in both panels show the analytical predictions.
**Figure 2**: Inhibitory synaptic plasticity sparsifies and democratizes receptive fields.

(A) A fixed sequence of 8 stimuli of varying firing rates is fed repetitively into a postsynaptic cell. Excitatory synapses are strength-tuned by signal group (see conductance graph on the right) so that signal 5 (marked also by dashed lines) is the preferred signal. (B) Postsynaptic action potentials over 50 trials with globally or detailed balanced inhibitory synapses (upper and lower panels, respectively) as indicated also by the schematics on the left (compare Fig. 1E). The squared cross-correlation coefficient, between each input signal and the peri-stimulus histogram (PSTH) (right side of panel B). (C) Schematic of a step stimulus delivered with large and small step size (solid and dotted black lines respectively); Sample PSTHs for non-preferred (red) and preferred stimuli (blue) to both step sizes, for a globally balanced cell. (D, E) Iso-response contour lines of the postsynaptic cell in the globally balanced regime, during the onset [0 - 50 ms] (D) and sustained [50 - 100 ms] (E) parts of the response. (F) Sample responses for non-preferred (red) and preferred stimuli (blue) to both step stimuli (as in C), and (G, H) Iso-response contour lines (as in D,E) for a detailed balanced cell.
Figure 3: Temporal dynamics of inhibitory plasticity, experiment and model.  
Frequency-tuned excitatory and inhibitory membrane currents (black and white symbols, respectively) as recorded from pyramidal cells in the primary auditory cortex of adult rat (9), (A) 30 min. and (B) 180 min. after a stimulus protocol shifted the preferred frequency of the excitatory membrane currents from 16 kHz to 4 kHz. Similarly stimulus-tuned input currents in a simulation, (C) 30 min. and (D) 180 min. after (manually) changing the excitatory tuning curve. Open and solid arrowheads indicate the previous and the new preferred stimuli in all panels. (E) Summary plot of the ratios of excitatory and inhibitory current amplitudes of previously preferred stimuli and new preferred stimuli as indicated in A-D, in the experiment (open and solid symbols, respectively) and simulations (blue and red lines, respectively). (F) Firing rate of the simulated neuron over the time of the simulation in E. (Panel A,B and E, with permission, adapted from (9))
Figure 4: Inhibitory plasticity in recurrent networks.

Five consecutive snapshots of the momentary activity of a network of 10 000 integrate-and-fire cells with inhibitory plasticity. (A) Synchronous regular network dynamics with high firing rates at the beginning of the simulation with weak inhibitory-synapses. (B) Establishment of the asynchronous irregular (steady) state with low firing rates through up-regulation of inhibitory synaptic weights by the synaptic plasticity rule. (C) The introduction of two synaptic memory patterns (cell assemblies) by 5-fold increased excitatory synaptic weights between neurons outlined in red and blue in panel A leads to high firing rates. (D) Recovery of the AI state at low firing rates. (E) Memory retrieval through externally driving the lower left quarter of the red cell assembly with additional excitatory stimulus. Each snapshot (A - E) shows (from top to bottom): - The momentary (1s) average firing rate of all neurons, on a grid of $10^2$ cells and separated into excitatory and inhibitory cells (left & right of the vertical line in A, respectively). Three groups of neurons play the role of either a cell assembly (red and blue outlines) or a control group (black outline). - A raster plot of 30 randomly drawn neurons from one (red) cell assembly and the control group, indicated by a red and a black square in the plot above. - The distributions of coefficients of variation of interspike intervals (ISI CVs) recorded from the neurons in the red and black groups. - The distributions of spiking correlations between spike trains from neurons in the same designated groups. For methods and additional statistics please see Supplementary Materials.
Supplementary Online Materials

In the supplementary online materials we provide a mathematical analysis of the plasticity mechanism and present some additional figures and the methods to our simulations. We finish with a tabular summary of all parameters following Nordlie et al. and Kunkel et al. (45, 46).

Mathematical Analysis of Inhibitory Synaptic Plasticity

The goal of this analysis is to provide a mathematical background for the findings presented in the main article. To this end, we study a network of neurons with plastic feedforward inhibition (Suppl. Figure 1). The analysis is done for linear Poisson neurons, which allows an analytic treatment of most phenomena observed in the simulations. The limitations of the approach are discussed.

Network architecture

We model a network consisting of a population of inhibitory interneurons connected to a single output neuron through plastic synapses with a weight vector $W^{I\rightarrow E}$. All neurons are linear Poisson neurons. The input to this network consists of a set of time-dependent analog input signals $x(t)$, which are received by both the output neuron (with a weight vector $W^{E\rightarrow E}$) and the interneurons (weight matrix $W^{E\rightarrow I}$). To allow for the possibility that the input signals arrive at the output neuron and the interneuron with different delays, we introduce a (negative or positive) delay $\tau^{E\rightarrow E}$ in the direct excitatory projection from the input to the output neuron that summarizes the difference in the delays along the two pathways.

Neuron model

The interneurons and the output neuron are linear Poisson neurons, i.e. they emit action potentials with inhomogeneous firing rates that are given by a linear weighted sum of their input signals. We denote the spikes trains of the output neuron and the interneurons by $Y$ and $Z_j$, respectively. The spike trains are modeled as sums of $\delta$-pulses: $Z_j(t) = \sum_f \delta(t - t_{jf})$, where $t_{jf}$ denote the time of the $f$-th spike emitted by the $j$-th interneuron.
We denote the firing rate of the output neuron by \( \rho^Y \) and those of the interneurons by \( z_j \):

\[
\rho^Y(t) = \sum_i W^{E\to E}_{ji} x_i(t) - \sum_{j,f} W^{I\to E}_{j} \int \epsilon(t - t') z_j(t') dt', 
\]

\[
z_j(t) = \sum_i W^{E\to I}_{ji} x_i(t) 
\]

where \( \epsilon(t) \) is a kernel (e.g. the inhibitory postsynaptic potential) that describes the time course of the output rate change caused by spike arrival at the synapse from an interneuron. For simplicity, we assume that \( \epsilon \) is normalized such that its integral is equal to one: \( \int \epsilon(t) dt = 1 \).

For later use, let us also introduce the firing rate of the output neuron that arises after taking the ensemble average over the activity of the interneurons for a given set of input signal:

\[
y(t) := \sum_j W^{E\to E}_{ji} x_i(t - \tau_{E\to E}) - \sum_j W^{I\to E}_{j} \int \epsilon(t - t') z_j(t') dt'. 
\]

and the correlation between the output neuron and a spike at the interneuron \( j \):

\[
c(t|t_j^f) = y(t|t_j^f) - y(t) = -W^{I\to E}_{j} \epsilon(t - t_j^f). 
\]

**STDP model**

As shown in electrophysiological work in cultures and acute slice (21–26), GABAergic synapses can be modified in a Hebbian manner by near-coincident activation of the pre- and postsynaptic neuron, with a coincidence time window of about 20ms duration. Moreover, presynaptic spikes alone induce a reduction of synaptic efficacy. We model this behavior by a spike timing-dependent (STDP) learning rule, in which the weight change of an inhibitory synapse from interneuron \( j \) to an excitatory postsynaptic neuron within a given time window of duration \( T \) is determined by:

\[
\Delta W^{I\to E}_{j} = \eta \int_0^T L(t - t') Y(t') Z_j(t) dt' dt - \eta \rho_0 \int_0^T Z_j(t) dt, 
\]
where \( L(t) = [2\tau_{\text{STDP}}]^{-1}e^{-|t|/\tau_{\text{STDP}}} \) denotes a symmetric learning window with a coincidence time \( \tau_{\text{STDP}} \), \( \eta \) is a learning rate and \( \rho_0 \) is a constant that controls the relative strength of the non-Hebbian weight decrease in relation to the Hebbian weight increase.

**From STDP to rate-based learning**

To derive a rate-based learning rule from the STDP rule we consider the weight change in Eq. 5 and take the ensemble average over the activity of the output neuron \( Y \) and the interneurons \( Z \), given the input signals \( x(t) \):\(^47\):

\[
\langle \Delta W_{ij}^{E \rightarrow I} \rangle_{Y,Z|X} = \eta \int L(t-t')\langle \langle Y(t) \rangle_{Y|X,Z} Z_j(t') \rangle_{Z|X} dt \, dt' - \eta \rho_0 \int \langle Z_j(t) \rangle_{Z|X} dt
\]

\[
= \eta \int L(t-t') \langle \rho^s(t) Z_j(t') \rangle_{Z|X} dt \, dt' - \eta \rho_0 \int z_j(t) dt. \tag{7}
\]

Using that the spike trains \( Z_j \) arise from inhomogeneous Poisson processes with a correlation function \( \langle Z_j(t) Z_k(t') \rangle_{Z|X} = z_j(t)z_k(t') + \delta(t-t')\delta_{jk}z_j(t) \), Eq. 7 can be simplified to

\[
\langle \Delta W_{ij}^{I \rightarrow E} \rangle_{Z,Y|X} = \eta \int L(t-t')y(t)z_j(t') dt \, dt' - \eta(\rho_0 + W_{ij}^{I \rightarrow E} \rho_s) \int z_j(t) dt, \tag{8}
\]

with \( \rho_s := \int L(\tau)\epsilon(\tau) d\tau \).

To simplify this expression further, we assume that the correlation functions \( c_j(s) = \frac{1}{T} \int y(t)z_j(t+s) dt \) change on a time scale that is slower than the coincidence time \( \tau_{\text{STDP}} \) of the learning window. In this case, we can perform a Taylor expansion of the correlation function around \( s = 0 \) and neglect terms of order \( s \) or higher:

\[
\int L(t-t')y(t)z_j(t') dt \, dt' = T \int L(s)c_j(s) ds
\]

\[
= T \int L(s)(c_j(0) + c_j'(0)s + ...) \, ds \tag{10}
\]

\[
\approx \int L(s) ds \int y(t)z_j(t) dt
\]

\[
= \int y(t)z_j(t) dt, \tag{12}
\]
where in the last line we used that the integral over the double-exponential learning window is normalized.

Inserting this expression into Eq. 8 leads to the following simplified expression for the weight dynamics:

$$\langle \Delta W_{j}^{I \rightarrow E} \rangle_{Z,Y|X} = \eta \int (y(t)z_{j}(t) - (\rho_{0} + \rho_{s} W_{j}^{I \rightarrow E})z_{j}(t))\,dt.$$  \hspace{1cm} (13)

Theoretical Results

Energy function

The simulations suggest that the learning rule tries to balance excitation and inhibition at any given moment in time and that the neuron aims to maintain a given target firing rate. These findings can be substantiated by a mathematical analysis.

Let us for a moment neglect the delay that is introduced by the synaptic kernel at the inhibitory synapses: $\epsilon(t) = \delta(t)$. In this case, it can be shown that the learning rule Eq. 13 performs a gradient descent on the following energy function:

$$\Psi(W_{I \rightarrow E}) = \frac{1}{2}\langle (y(t) - \rho_{0})^{2} \rangle_{t} + \frac{1}{2}\rho_{s}\sum_{j} \bar{z}_{j}(t) \langle W_{j}^{I \rightarrow E} \rangle^{2}, \hspace{1cm} (14)$$

where $\langle \cdot \rangle_{t}$ denotes a temporal average and $\bar{z}_{j} = \langle z_{j}(t) \rangle_{t}$ is the mean firing rate of interneuron $j$.

For the mathematically inclined reader, we defer the proof that a gradient descent on this energy functions reproduces the learning rule Eq. 13 to the end of this Supplementary. Instead we discuss the two terms of the objective function $\Psi$. The second term, which we denote by $\Psi_{\text{spike}}$ in following, arises from spike-spike correlations and plays the role of a cost term that punishes large inhibitory weights for synapses that are active. The effect of the first term is discussed in the following.
Rate stabilization

The first term of the energy function measures the quadratic deviation of the output firing rate from $\rho_0$. Therefore, the constant $\rho_0$ acts as a target firing rate. The learning rule thus implements a form of homeostatic plasticity that stabilizes the postsynaptic firing rate. This is reflected by the simulations, which show that the postsynaptic firing rate after convergence depends linearly on the strength $\rho_0$ of the non-Hebbian synaptic depression.

For large numbers of inhibitory synapses and/or high firing rates of the inhibitory neurons, spike-spike correlations can be neglected. The output firing rate is then simply given by $y = \rho_0$. This theoretical prediction of the firing rate fits the simulations well (see Main Fig. 1G of the main text).

The fact that the firing rate is controlled by a simple ratio of two parameters is advantageous for simulation studies, particularly for recurrent networks, because it allows to automatically tune a balanced network to a desired firing rate by simply choosing the appropriate parameters for the learning rule.

Current balance

The firing rate of the output neurons is given by the difference between excitatory input $E(t) = \sum_i W^{E\rightarrow E}_{i} x_i(t - \tau^{E\rightarrow E}) = W^{E\rightarrow E} \cdot x(t - \tau^{E\rightarrow E})$ and inhibitory input $\sum_j W^{I\rightarrow E}_{j} z_j = W^{I\rightarrow E} \cdot z(t)$. Therefore, the first term of the energy function $\Psi$ measures a quadratic error in the balance between excitation and inhibition, corrected by the target firing rate $\rho_0$:

$$\Psi = \langle (E(t) - \rho_0 - W^{I\rightarrow E} \cdot z(t))^2 \rangle_t + \Psi_{\text{spike}}$$

$$= \langle (\tilde{E}(t) - W^{I\rightarrow E} \cdot z(t))^2 \rangle_t + \Psi_{\text{spike}}$$

In our simulations, the target firing rate is smaller than the excitatory drive: $\rho_0 < E(t)$ (i.e. in the absence of inhibition, the neurons fire at much higher rates than the target rate $\rho_0$). Therefore, the subtraction of the target rate can be seen as a relatively small correction of the excitatory drive: $\tilde{E}(t) \approx E(t)$. Then, the first term of the energy function measures the mean
square difference between the excitatory and the inhibitory input to the output cell. Minimizing this term corresponds to balancing excitation and inhibition for any given moment in time. Moreover, because the inhibitory input is linear in the inhibitory weights, minimizing the first part of the objective function is therefore equivalent to a linear regression of the excitatory input by a linear combination of the inhibitory rates.

**Stimulus co-tuning**

Intuitively, it is clear that a detailed balance between excitation and inhibition can only be reached if stimulus-evoked excitatory and inhibitory currents are balanced on the level of individual stimuli, i.e. that excitation and inhibition are co-tuned in terms of stimulus selectivity. To find a mathematical formalization of this intuition, let us assume that the input neurons are ordered according to their stimulus preference along an arbitrary stimulus dimension (e.g. auditory frequency, visual orientation). Then, the excitatory weights $W_{E \rightarrow E}$ determine the stimulus tuning curve for the excitatory input the neuron receives.

The inhibitory input $I(t)$ the output neuron receives is determined by the indirect propagation of the input activity $x_i(t)$ via the interneurons:

$$I(t) = W_{I \rightarrow E}^j \cdot z(t) = W_{I \rightarrow E}^j \cdot W_{E \rightarrow I}^E x(t) \quad (17)$$

$$= ((W_{E \rightarrow I})^T W_{I \rightarrow E}) \cdot x(t). \quad (18)$$

Therefore, the stimulus tuning of the inhibitory input is determined by the product $(W_{E \rightarrow I})^T W_{I \rightarrow E}$ of the weights along the indirect inhibitory pathway. Whether excitation and inhibition are co-tuned is therefore determined by the relation of the excitatory weights $W_{E \rightarrow E}$ and the effective inhibitory weight vector $(W_{E \rightarrow I})^T W_{I \rightarrow E}$.

A perfect balance can only be reached if the information that is propagated along the direct excitatory pathway is available in the activity of the interneurons. If this is not the case, the balance must remain approximate. A way to see if and how the learning rule Eq. 13 tries to approximate the perfect co-tuning is to rewrite the energy function Eq. 14 in terms of the difference of the excitatory and the inhibitory weights. To this end, let us first split the first
term of the energy into mean and variance:

\[\Psi = \langle (y - \bar{y})^2 \rangle_t + (y - \rho_0)^2 + \Psi_{\text{spike}}.\] (19)

If we neglect the delay \(\tau^{E\rightarrow E}\) on the direct excitatory pathway for a moment, the output rate is linear in the input and depends on the difference \(\Delta := W^{E\rightarrow E} - (W^{E\rightarrow I})^T W^{I\rightarrow E}\) between the excitatory and the cumulative inhibitory weights:

\[y(t) = \sum_i (W_i^{E\rightarrow E} - \sum_j W_j^{I\rightarrow E} W_j^{E\rightarrow I}) x_i(t) = \Delta \cdot \bar{x}(t).\] (20)

Inserting this into the energy function yields:

\[\Psi = \Delta^T C \Delta + (\Delta \cdot \bar{x} - \rho_0)^2 + \Psi_{\text{spike}},\] (21)

where \(\bar{x} = \langle x \rangle_t\) denotes the mean input rate and \(C := \langle (x - \bar{x})(x - \bar{x})^T \rangle_t\) is the covariance matrix of the input.

The second term of this energy function punishes deviations from the *global* balance, i.e. it aims at a balance of excitation and inhibition *on average*. The first term is a positive semi-definite quadratic form in the difference of excitation and inhibition. It therefore punishes deviations from a more precise *detailed* balance, i.e. imbalances between excitation and inhibition on the level of individual inputs. The covariance matrix \(C\) of the input introduces a weighting: imbalances along dimensions of high input variance are punished more severely than along low input variance.

In summary, the reformulation of the energy function in terms of the excitatory and inhibitory weights shows that inhibitory plasticity seeks to establish a co-tuning of excitation and inhibition. If information is lost along the indirect pathway, the system will establish an approximation that minimizes the squared deviation from the balance, weighted according to the variance in the input.
The analysis presented in the last paragraphs neglects transmission and conduction delays. The effect of delays is twofold. Firstly, they can make a balance between of excitatory and inhibitory input to the output neuron impossible, because the two signals arrive at different moments in time. Secondly, they can disturb Hebbian learning, because the correlations between pre- and postsynaptic activity are weakened.

In the presence of delays, there is no energy function for the system dynamics. To identify under which conditions inhibitory plasticity can establish a balance between excitation and inhibition in the presence of delays, we have to consider the stationary solution of the learning dynamics Eq. 13.

For clarity, we neglect the term that arises from spike-spike correlations, i.e. we assume that the number of inhibitory neurons is large:

\[
\Delta W_{I \rightarrow E} = \eta_0 \int [y(t)z(t) - \rho_0 z(t)] \, dt = 0, \tag{22}
\]

and summarize the effect of the synaptic kernel \(\epsilon\) at the inhibitory synapse into a synaptic delay:

\[
y(t) = W^{E \rightarrow E} \cdot x(t - \tau^{E \rightarrow E}) - W^{I \rightarrow E} \cdot z(t - \tau^{I \rightarrow E}). \tag{23}
\]

By inserting Eqs. 1 and 23 into the stationarity condition Eq. 22, we get

\[
W^{E \rightarrow I} C(\tau^{E \rightarrow E}) W^{E \rightarrow E} - W^{E \rightarrow I} C(\tau^{I \rightarrow E}) \left( W^{E \rightarrow I} \right)^T W^{I \rightarrow E} - \rho_0 W^{E \rightarrow I} \bar{x} = 0, \tag{24}
\]

where \(C(\tau) := \langle x(t) x^T(t - \tau) \rangle_t\) denotes the time-delayed correlation matrix of the input signals.

Because we are now interested merely in whether delays can disrupt the balance, we assume that the full input information is present in the inhibitory population, i.e. that the weight matrix \(W^{E \rightarrow I}\) has full rank. Then, Eq. 24 can be resolved for the effective inhibitory weights:

\[
\left( W^{E \rightarrow I} \right)^T W^{I \rightarrow E} = C^{-1}(\tau^{I \rightarrow E}) C(\tau^{E \rightarrow E}) W^{E \rightarrow E} - \rho_0 C^{-1}(\tau^{I \rightarrow E}) \bar{x}. \tag{25}
\]
The first observation that is consistent with our previous statements on the stimulus co-tuning is that when the delay along the indirect inhibitory pathway is the same as on the direct excitatory pathway $\tau_{I\rightarrow E} = \tau_{E\rightarrow E}$, the product $C^{-1}(\tau_{I\rightarrow E})C(\tau_{E\rightarrow E})$ is reduced to the unit matrix. The effective inhibitory weights $(W^{E\rightarrow I})^T W^{I\rightarrow E}$ are then the same as the excitatory weights $W^{E\rightarrow E}$, apart from a correction that depends on $\rho_0$ and maintains the target firing rate.

To get a clearer picture on the effect of delays, let us assume that different input signals are decorrelated and that their autocorrelation function decays exponentially, i.e. that the time-delayed correlation matrix $C(\tau)$ of the input signals has the following simple structure:

$$C(\tau) = \bar{x}\bar{x}^T + \sigma^2 e^{-|\tau|/\tau_c} \cdot E,$$

where $\sigma$ and $\tau_c$ denote the variance and autocorrelation time of the inputs and $E$ is the unit matrix.

With this assumption, the matrix $C(\tau)$ can be inverted analytically and the effective inhibitory weights can be written as a linear combination of the excitatory weights and the (untuned) mean firing rates of the input neurons:

$$(W^{E\rightarrow I})^T W^{I\rightarrow E} = AW^{E\rightarrow E} + B\bar{x},$$

where $A$ and $B$ are the following expressions that depend on the delays in the system

$$A = e^{\tau_{I\rightarrow E}/e^{\tau_{E\rightarrow E}}}$$

and

$$B = \frac{(1 - A)W^{E\rightarrow E} \cdot \bar{x}}{|x|^2 + \sigma^2 e^{-|\tau_{E\rightarrow E}|/\tau_c}} - \frac{\rho_0}{|x|^2 + \sigma^2 e^{-|\tau_{I\rightarrow E}|/\tau_c}}.$$

The dependence of the factor $A$ on the delays is worth discussing, because it determines the strength of the inhibitory tuning and because it qualitatively captures the effects observed in the simulations (see also Additional Simulation Results, below).

The key to understanding how $A$ depends on the delays lies in the Hebbian learning rule that controls the inhibitory weights. Because the activity of the output neuron is the difference between excitation and inhibitory inputs, inhibition is strengthened by correlations between the
activity of the interneurons and the excitatory drive to the output neuron, while correlations between the activity of the interneurons and the inhibitory drive to the output neuron decrease the weights. The correlation between the activity of the interneurons and the excitatory drive to the output neuron decreases with the difference in the delays along the two excitatory pathways. Therefore, inhibition is weakened with increasing $|\tau^{E\rightarrow E}|$. Conversely, the delay along the inhibitory pathway decreases the correlation between the activity of the interneurons and the inhibitory drive to the output neuron. Because this correlation limits the growth of the inhibitory weights, the inhibitory weights grow as the delay $|\tau^{I\rightarrow E}|$ increases. The sensitivity of the effective inhibitory weights on the delays depends on the autocorrelation time of the input signals, because this is the time scale that controls by how much the signals can be delayed in time, before the correlation along the respective pathways is lost.

The factor $B$ that controls the untuned contribution to the effective inhibitory weights increases as the inhibitory tuning strength $A$ decreases. This effect is qualitatively confirmed by the simulations (Suppl. Fig. 2D), although the increase in the untuned inhibition is stronger than theoretically predicted. This is due to the limitations of the rate picture that are discussed below.

Discussion

The present theoretical treatment shows that the Hebbian learning rule in inhibitory synapses can be interpreted as a gradient descent on the mean square difference between excitation and inhibition. Although the theory is based on a simple linear picture of a network with feedforward inhibition, it is able to capture the key effects that are observed in the simulations: rate homeostasis, current balance, stimulus co-tuning and the gradual loss of co-tuning with increasing delays along the excitatory pathway.

In a system with balanced excitation and inhibition, the difference between excitatory and inhibitory drive to the output neuron fluctuates around zero. Therefore, the output neuron will frequently encounter a net negative drive. The present linear treatment of the system will therefore often assume biologically implausible negative rates of the output neuron. From this perspective, it is surprising that such a simplified picture captures the key effects present in the
simulations. Because the inhibitory weights are subject to a Hebbian learning rule, negative output rates decrease the inhibitory weights, while the more realistic integrate-and-fire neuron used in the simulations simply stops firing and thereby evades Hebbian changes in the inhibitory weights. Thus, the theory systematically underestimates the inhibitory weights. This is most evident in the simulations with delays: If inhibition lags behind excitation, the output neuron receives a period of unbalanced positive input whenever the input neurons start firing (onset transient; \(6, 13, 14\)) and, conversely, a period of negative input every time the input neurons stop firing. Because of the output rectification of the integrate-and-fire neuron, the Hebbian learning rule “sees” only the onset transient and increases the inhibitory weights. This effect becomes more prominent with increasing delay (Suppl. Fig. 2 C, D).

The mathematical analysis suggests that the effects that are observed in the simulations are not sensitive to details of the STDP learning rule that is used. As long as the integral of the learning window is positive, the rate-based description of the learning dynamics remains the same. Therefore, asymmetric learning windows, as observed for most excitatory synapses, would most likely not change the results qualitatively, as long as the LTP component dominates over LTD.

Proofs

Proof that the learning rule is a gradient descent

We assume that synaptic transmission at the inhibitory synapse is fast compared to the auto-correlation time of the inhibitory rate variations, so that we can replace the synaptic kernel \(\epsilon\) by a \(\delta\)-function. Then, the output firing rate \(y(t)\) is given by

\[
y(t) = \sum_j W^E \rightarrow E x_j(t - \tau^E \rightarrow E) - \sum_j W^I \rightarrow E z_j(t) =: E(t) - \sum_j W^I \rightarrow E z_j(t) .
\]

Consequently, the derivative of the output rate \(y\) with respect to the inhibitory weights \(W^I \rightarrow E\)
is simply the negative firing rate $z_j$ of the inhibitory neuron $j$:

$$\frac{\partial}{\partial W_{j \rightarrow E}} y(t) = -z_j(t).$$

(32)

We can now calculate the partial derivative of the energy function Eq. 14 with respect to the inhibitory weights:

$$\frac{\partial}{\partial W_{j \rightarrow E}} \Psi = \langle (y(t) - \rho_0) \frac{\partial}{\partial W_{j \rightarrow E}} y(t) \rangle_t + \rho_s \langle z_j(t) \rangle_t W_{j \rightarrow E}$$

(33)

$$= -\langle y(t)z_j(t) \rangle_t - (\rho_0 + \rho_s W_{j \rightarrow E})z_j(t)$$

(34)

$$= -\frac{1}{T} \int y(t)z_j(t) - (\rho_0 + \rho_s W_{j \rightarrow E})z_j(t) dt$$

(35)

A comparison with Eq. 13 shows that the inhibitory plasticity rule is indeed a gradient descent on the energy function $\Psi$:

$$\langle \Delta W_{j \rightarrow E} \rangle_{Z,Y|X} = -\eta_0 T \frac{\partial}{\partial W_{j \rightarrow E}} \Psi.$$  

(36)

Additional Simulation Results

Uncorrelated signals

To investigate the effect of uncorrelated signals on inhibitory tuning we stimulated the single integrate-and-fire cell we used for Main Fig. 1, 2, & 3 with two additional protocols. In the absence of any temporal structure, i.e. if each synapse received a Poisson process with a constant rate, the plasticity rule rapidly established a global balance with identical inhibitory weights for all channels and firing rates of nearly $\rho_0$ (Suppl. Fig. 2 A). When we stimulated the cell with temporally structured input as in the main part of the paper but removed the correlation between excitatory and inhibitory signals, the learning rule could still establish a global balance of all input currents (albeit with higher inhibitory synaptic weights) (Suppl. Fig. 2 B), but failed to bring the firing rates to the target rate $\rho_0$, because in this scenario some excitatory spikes cannot be balanced since they lack inhibitory partner spikes.
Delayed signals

Similarly, but much less pronounced, a delay between the excitatory and the inhibitory component of the signals also caused a deviation of the postsynaptic firing rate from the target rate $\rho_0$ (Suppl. Fig. 2 C). The detailed balance of each input channel on the other hand was maintained for delays smaller than or equal to the autocorrelation time of the input signals (50 ms in our simulations): Although the learning rule compensated for the delay between excitation and inhibition through up-regulating the weights of all inhibitory synapses equally, the tuning shape of the excitatory synapse population was maintained (Suppl. Fig. 2 D). For delays much larger than the autocorrelation time of the input signals, the correlation between excitation and inhibition is lost, leading to the above case of structured, but uncorrelated input signals (Suppl. Fig. 2 B).

Additional network statistics

To complete the statistics for the network simulations (Main Fig. 4), we additionally supply the distributions of inhibitory synaptic weights onto all neurons in the red cell assembly and the black control assembly, as well as their distributions of firing rates, averaged over 1 s, for each snapshot (Suppl. Fig. 3).

Additional recall protocols

Recall of an imprinted pattern in recurrent networks as shown in Main Figure 4 was not limited to one assembly, but could also be achieved in the other assembly (Suppl. Fig. 4 F), or even in both patterns simultaneously, if cells in both patterns were driven in equal numbers (Suppl. Fig. 4 G). Elevated firing rates during recall are a direct consequence of the strengthened excitatory synapses within each assembly (Suppl. Fig. 5). When cells that are part of the un-strengthened control assembly are stimulated, the background firing rate of (the other) cells in the control assembly is raised slightly. The firing rate of the two cell assemblies remains low, since they receive greater amounts of inhibitory currents through strengthened synapses. When cells that are part of a cell assembly are stimulated, the average firing rate of the (other) cells in the stimulated assembly increases as a function of the number of stimulated cells. The firing rate in the other pattern tends to decrease because patterns inhibit each other slightly through inhibition that originates from inhibitory cells servicing both patterns (Suppl. Fig. 5 B, C) when
activated exclusively. When equal numbers of cells in two separate patterns are driven simultaneously, both assemblies can be activated (Suppl. Fig. 5 D).

Simulation Methods

Neuron Model

The model used in all our simulations is a leaky integrate-and-fire neuron, characterized by a time constant, $\tau = 20 \text{ ms}$, and a resting membrane potential, $V_{\text{rest}} = -60 \text{ mV}$. Whenever the membrane potential crosses a spiking threshold of $-50 \text{ mV}$, an action potential is generated and the membrane potential is reset to the resting potential, where it remains clamped for a 5 ms refractory period. To set the scale for currents and conductances in the model, we use a membrane resistance of 100 MΩ.

We model synapses onto each neuron as conductances, so the sub-threshold membrane voltage obeys

$$\tau \frac{dV_i}{dt} = (V_{\text{rest}} - V_i) + g^E_i (V^E - V_i) + g^I_i (V^I - V_i) + I_b .$$

Reversal potentials are $V^E = 0 \text{ mV}$ and $V^I = -80 \text{ mV}$. The synaptic conductances $g^E_i$ and $g^I_i$ are expressed in units of the resting membrane conductance. For the single cell simulations for Main Fig. 1, 2, and 3, $I_b$ was set to 0 pA; in the network simulations for Main Fig. 4 a constant current $I_b = 200 \text{ pA}$ was used to maintain a minimum amount of activity.

When the neuron $i$ receives a presynaptic action potential from neuron $j$, the appropriate post-synaptic conductance is increased, $g^E_i \rightarrow g^E_i + \Delta g^E_{ij}$ for an excitatory spike and $g^I_i \rightarrow g^I_i + \Delta g^I_{ij}$ for an inhibitory spike. Otherwise, these parameters obey the equations

$$\tau^E \frac{dg^E_i}{dt} = -g^E_i \quad \text{and} \quad \tau^I \frac{dg^I_i}{dt} = -g^I_i ,$$

with synaptic time constants $\tau^E = 5 \text{ ms}$ and $\tau^I = 10 \text{ ms}$. The conductance of each synapse is constructed such that $\Delta g_{ij} = \bar{g} W_{ij}$ where $\bar{g}$ is a constant. $W_{ij}$ can be plastic or fixed, depending on the identity of the synapse (see below). The integration time step for our simulations was 0.1 ms. All Simulations were programmed in C.
Inhibitory synaptic plasticity

Recent experimental results (21–26) show that inhibitory synapses can be modified by coincident pre- and postsynaptic activity with a coincidence time window $\Delta t$. Additionally, sole presynaptic spikes lead to a reduction of synaptic efficacy. For the sake of simplicity and in accordance with the discussed theoretical results, we model this behavior by a symmetric spike timing-dependent learning rule, in which potentiation occurs as a function of $\Delta t = |t^{jf}_i - t^{jf}_f|$ (in which $t^{jf}_i$ and $t^{jf}_f$ denote the time of a pre- and postsynaptic spike respectively) and depression is a constant $\alpha = 2 \times \rho_0 \times \tau_{STDP}$ that occurs for each presynaptic spike. This STDP rule was implemented for inhibitory synapses projecting onto excitatory cells. In order to calculate the changes to each $W_{ij}$, a synaptic trace $x_i$ is assigned for each neuron (or spike train, in case of the single cell simulations for Fig. 1, 2, & 3). $x_i$ increases with each spike $x_i \rightarrow x_i + 1$ and decays otherwise, following

$$\tau_{STDP} \frac{dx}{dt} = -x,$$

with the fitted time constants $\tau_{STDP} = 20$ ms. The synaptic weight $W^{ij}$ is updated for every pre- or postsynaptic event such that

$$W_{ij} = W_{ij} + \eta (x_i - \alpha) \quad \text{for pre} \text{synaptic spikes at time } t^{jf}_i$$

and

$$W_{ij} = W_{ij} + \eta x_j \quad \text{for postsynaptic spikes at time } t^{jf}_f$$

where $\eta$ is the learning rate and $\alpha = 0.2$ ($\rho_0 = 5$ Hz) is the depression factor as discussed above.

Single Cell Simulations

In the first part of the paper, we model the arrival of multiple signals at a single integrate-and-fire cell. 800 excitatory and 200 inhibitory spike trains, divided into 8 groups of 100 excitatory and 25 inhibitory inputs, and drawn from Poisson processes with changing spike probabilities $s_k(t)$, where $k = \{1..8\}$ denotes signal identity. Spike trains within each of the groups share the same time-dependent spike probability $s_k$, given by the input signals (scaled by a coupling strength), added to a background of 5 Hz, i.e. they are different realizations of the same inhomogeneous Poisson process. The resulting input spike trains are delivered to the neuron through synapses with varying strengths $\Delta g_{ij} = \bar{g} W_{ij}$. Excitatory synapses are set to values between 50 pS and
200 pS, resulting in typical PSP amplitudes on the order of \( \sim 0.1 \text{mV} \) at \( V^\text{rest} \). Inhibitory synapses are tuned so that with \( W_{ij} = 1, \Delta g^I_{ij} = \bar{g}W_{ij} = 50 \text{pS} \), resulting in \(-0.2 \text{mV}\) PSPs at \( V^\text{rest} \). At the beginning of a simulation, we set \( W^I_{ij} = 0.1 \).

To mimic sensory input during learning, we used 8 independent traces of Gaussian-distributed white noise, low-pass filtered at 50 ms and half-wave rectified \((19)\) as input signals. Unless otherwise noted, the coupling strength between the input signal traces and the resulting input spike probability was set to achieve average firing rates of 13 Hz with peak firing rates at \( \sim 150 \text{Hz} \) for each spike train. In places, we used the same realization of the input signal for multiple trials \(\text{(for averaging, Fig. 2 A, B)}\). We used step-like stimuli with step sizes between 5 Hz and 50 Hz to map the receptive field properties \(\text{(Fig. 2 C-H)}\).

The variables that determine the dynamics of inhibitory synaptic plasticity \( (\eta = 10^{-5}, \text{the learning rate, and } \alpha = 0.2 (\rho_0 = 5 \text{Hz}), \text{the depression factor}) \) are kept fixed unless otherwise noted. To reproduce the temporal dynamics of the discussed experiment \( (9)\) in Fig. 3, we set \( \eta = 10^{-6}, \alpha = 0.35 (\rho_0 = 8.75 \text{Hz}), \) and increased the average firing rate of the inhibitory spike trains 4 fold in comparison to their excitatory counterparts from 13 Hz to approximately 50 Hz to adjust the temporal dynamics of depression and potentiation relative to each other.

We used a correlation measure \((13, 15, 19)\) to determine the impact of each input signal on the output. To do this, we calculated firing rate histograms \( r(t) \) \(\text{(bin size = 5 ms)}\) of the output signals over 100 trials with identical input signals and determined its time-averaged firing rate \( \bar{r} \). The correlation is then

\[
C_k = \frac{\langle (s_k(t) - \bar{s}_k) (r(t) - \bar{r}) \rangle_t}{\sigma_{s_k} \sigma_r},
\]

where the brackets denote an average over time, \( s_k(t) \) and \( \bar{s}_k \) are the firing rate and its average for a given input signal, and \( \sigma_r, \sigma_{s_k} \) are the standard deviations of the corresponding firing rates.

We define the impact of each signal as \( C_k^2 \), the total amount of variance of the output firing pattern that can be explained by each input signal.

**Network Simulations**

For the last part of the paper we study the effect of inhibitory synaptic plasticity in a large recur-
rent network. To this end, we simulated a network of 8,000 excitatory and 2,000 inhibitory leaky integrate-and-fire neurons, laid out on a 100 × 100 grid. All neurons have random connectivity of 2% to the rest of the network. The connectivity remains fixed during a simulation, although the weights of the inhibitory synapses onto excitatory cells can change according to inhibitory plasticity (see below). We chose the specific network because of its publication history (15, 36) and because it is small enough to study within reasonable computation times but large enough to avoid boundary effects (48). Network parameters were chosen in keeping with both general properties of cortical circuits and previous work (15, 19, 48, 49) and can be found in tabular summaries following the standard form (45, 46) below.

Memory Pattern
In addition to the general architecture, we introduce specific patterns into the connectivity matrix by defining two groups of 28 × 28 = 784 excitatory neurons as Hebbian assemblies. We strengthen all existing connections between the neurons within each group by a factor of five. We allow the patterns to overlap by selecting 8 × 8 = 64 neurons to be part of both groups. Furthermore, we define a third, control assembly of neurons which don’t take part in either pattern. The strength of intra-group synapses of this third pattern remained un-strengthened.

Synaptic Plasticity
As discussed above, we can distinguish between excitatory to excitatory, excitatory to inhibitory, inhibitory to inhibitory and inhibitory to excitatory connections in our network. Only the latter group is plastic. For simplicity, we assume that the structure of the connectivity matrix remains fixed after the network has been initialized. Particularly this means we will restrict inhibitory plasticity operations to already existing connections. Note however that the weight of an existing connection $w^{ij}$ can decay to zero. For the simulations shown here, we set $\eta = 10^{-4}$ and $\alpha = 0.3$ ($\rho_0 = 7.5$ Hz).

Network Dynamics
The irregular asynchronous network activity that is thought to mimic cortical dynamics has a
roughly constant population firing rate with low spiking correlation values and ISI CVs near 1. To characterize the global state of the network (Fig. S2) we monitored individual membrane potentials, the population firing rate (the average of firing rates across the network), average membrane potentials, and interspike intervals (ISIs). The ISI CV for a neuron is the ratio of the standard deviation of the ISI distribution to its mean. ISI CV values close to zero indicate regular spiking patterns, values near 1 indicate irregular spiking, and values larger than 1 indicate, in our simulations, burstiness in the firing pattern (1, 50).

Additionally, we calculated distributions of the spiking correlations (5) and the ISI CVs of neurons in two groups: We collected data from 392 neurons within one memory pattern and an equal number of cells from the control group. ISI CV histograms we calculated as above. Following Renart et al. (5) we computed the spiking correlation coefficient $X_{ij}$ between spike trains $S_i(t)$ and $S_j(t)$. We then constructed filtered spike trains $F_i$ defined as

$$F_i(t) = S_i(t) \ast K(t),$$

in which the spiketrain $S_i = \sum \delta(t-t_i)$ is convoluted with a symmetric bi-exponential kernel $K(t)$ defined as

$$K(t) = \frac{1}{\tau_1} \exp \left( -\frac{|t|}{\tau_1} \right) - \frac{1}{\tau_2} \exp \left( -\frac{|t|}{\tau_2} \right)$$

with $\tau_1 = 50$ ms and $\tau_2 = 4 \times \tau_1$. The unnormalized covariance $V_{ij} = \sum_t F_i(t)F_j(t)$ over all discrete times $t$ then leads to

$$X_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}}. $$

To gain insight into ensemble statistics we calculate all possible $(N^2 - N)/2$ correlation coefficients between the filtered spiketrains of a given group and plotted them in histograms (Fig. 4). All values were computed in discrete time with a resolution of $dt = 1$ ms. The combination of ISI CV and spiking correlations provides a good measure of the (ir)regularity and the (a)synchrony of the local network activity as we change the connectivity matrix of the network as described below.

To recall a stored pattern we externally stimulated a subset of neurons within the pattern. Stimulated neurons were randomly connected (5% connectivity) to a group of 1000 independent
Poisson processes with a mean firing rate of 100 Hz. For all recall experiments we only stimulated a set of neurons that was disjunct to the neurons used for computing the spike train statistics.

Annotated Protocol for Figure 4

The simulation protocol for Fig. 4 was structured as follows:

0: $t = -1$ min: The AI network dynamics of the original network (15) without inhibitory plasticity. This phase serves as a reference and is not shown.

A: $t = 0$ min: Inhibitory plasticity is turned on. Inhibitory to excitatory synapses are turned to 0 efficacy. The network is forced out of the asynchronous irregular (AI) regime and begins to fire at high rates.

B: $t = 60$ min: Inhibitory plasticity has restored AI activity.

C: $t = 60$ min, 5 s: The excitatory non-zero weights of the 2 designated memory patterns are increased ad-hoc by a factor of $\chi = 5$. The neurons of the subset begin to exhibit elevated and more synchronized activity.

D: $t = 120$ min: Inhibitory plasticity has successfully suppressed any elevated activity from the pattern and restored the global background state.

E: $t = 120$ min, 5 s: By delivering an additional stimulus as described above to 25% of the cells within one memory pattern, the whole pattern is activated. Activity inside the pattern stays asynchronous and irregular, and the rest of the network, including the other pattern, remains unaffected.

F: $t = 120$ min, 11 s: An additional stimulus was delivered to 25% of the cells within the other (blue) memory pattern, the other pattern is activated similar to E.

G: $t = 120$ min, 17 s: An additional stimulus was delivered to to 25% of the cells of both memory patterns, including the cells shared between both patterns. This results in elevated asynchronous irregular activity in both patterns.
Supplementary Figure 1: Overview of the variables used in the theoretical analysis.
An excitatory cell population (in black) with firing rate $x$ delivers postsynaptic currents to an inhibitory (grey) and an excitatory (green) population of neurons. The currents are proportional to the strength of the excitatory synapses $W_{E \rightarrow I}$ and $W_{E \rightarrow E}$, respectively. Currents to the excitatory (green) population are delivered with a delay $\tau_{E \rightarrow E}$. Additionally the excitatory target population receives inhibitory input currents with delay $\tau_{I \rightarrow E}$. These currents are the product of $z$, the firing rate of the inhibitory neuron population, and $W_{I \rightarrow E}$ the synaptic strength of the inhibitory synapses, here subject to plasticity. Consequently, the firing rate $y$ of the excitatory population is determined by the difference of excitatory and inhibitory currents.
Supplementary Figure 2: The effect of uncorrelated or delayed signals on inhibitory tuning. Tuning solutions for (A) unstructured input noise (on all synapses) and (B) structured but uncorrelated input signals (i.e. different signals for excitatory and inhibitory synapses): Upper panel: Excitatory and inhibitory membrane currents (black and white symbols, respectively) evoked by each signal channel, averaged over 8s after inhibitory synaptic plasticity dynamics reached steady state. Lower panel: Temporal evolution of the postsynaptic firing rate (solid line) and the average synaptic weights of the inhibitory synapses associated with 3 representative signals (dotted lines). The * symbols indicate the time at which the respective upper panels were recorded. (C) Output firing rate as a function of the delay between the excitatory and inhibitory signal stream. All other parameters remained as in Main Fig. 1. The dashed lines shows the firing rates for unstructured noisy inputs (red) and structured but uncorrelated inputs (blue, see also panels A & B). Open symbols mark the delays investigated in (D) ((D)). Shape of the inhibitory synaptic weight tuning for all stimuli as a function of different delays.
Supplementary Figure 3: Additional statistics for Main Figure 4.
Five consecutive snapshots of network states as in Main Figure 4. Each snapshot shows: The
distributions of inhibitory synaptic weights $W^{I\rightarrow E}$ onto all cells of the red cell assembly and
the unstrengthened (black) control assembly (upper panels) and the distribution of average
firing rates recorded for 1 s (lower panels). Temporal development identical to Main Fig.
4: (A) Synchronous regular network dynamics at the beginning of the simulation with weak
inhibitory-synapses. (B) Establishment of the asynchronous irregular (steady) state through
up-regulated inhibitory synaptic weights. (C) Introduction of an excitatory synaptic memory
as strongly enhanced excitatory synaptic strengths between neurons of the red designated
neuron group in Main Fig. 4A leads to high regular firing of these neurons. (D) Recovery of
the AI state. (E) Memory recall as pattern completion by means of externally driving one
quarter of the memory pattern with additional excitatory stimulus.
Supplementary Figure 4: Two additional instances of recall similar to Main Fig. 4E. Each snapshot, as in Main Fig. 4, contains (from top to bottom): - The momentary (1s) average firing rate of each neuron - A raster plot of 30 randomly drawn neurons from one (red) cell assembly and the control group - The distributions of coefficients of variation of interspike intervals (ISI CVs) recorded from all neurons in the red and black groups. - The distributions of spiking correlations between spike trains from neurons in the same designated groups - The distributions of inhibitory synaptic weights $W_{I\rightarrow E}$ in the network. - The distribution of average firing rates recorded for 1 s. Temporal development continuing from to Main Fig. 4E: **(F)** Memory recall of the blue cell assembly as pattern completion by means of externally driving cells in the second (blue) memory pattern with additional excitatory current. **(G)** Memory recall of both patterns as pattern completion by means of externally driving cells in both patterns symmetrically with additional excitatory stimulus.
Supplementary Figure 5: Recall firing rates as a function of the number of externally driven neurons. Each subfigure shows a schematic of the network, indicating which cells in which pattern are driven (upper panel), and a plot of the resulting firing rates in each group as a function of the number of cells stimulated, following the same color code as above. (A) Different numbers of cells of an arbitrarily chosen control group are driven with an external stimulus. (B, C) Different numbers of cells belonging exclusively to either (red) pattern 1, or (blue) pattern 2 are driven with an external stimulus. (D) Cells belonging to both patterns are driven with an additional stimulus. The symbol indicates the maximum number of shared cells that can be driven. Beyond this point the cell population is driven asymmetrically, i.e. additionally cells belong to (blue) pattern 2 but not to (red) pattern 1.
## A Model Summary

<table>
<thead>
<tr>
<th>Model Summary</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Populations</strong></td>
<td>Two: excitatory, inhibitory</td>
</tr>
<tr>
<td><strong>Topology</strong></td>
<td>Random all-to-all connections</td>
</tr>
<tr>
<td><strong>Connectivity</strong></td>
<td>Random all-to-all connections</td>
</tr>
<tr>
<td><strong>Neuron model</strong></td>
<td>Leaky integrate-and-fire, fixed voltage threshold, fixed absolute refractory time</td>
</tr>
<tr>
<td><strong>Synapse model</strong></td>
<td>Conductance based inputs (exponentially decaying PSC)</td>
</tr>
<tr>
<td><strong>Plasticity</strong></td>
<td>Inhibitory plasticity</td>
</tr>
<tr>
<td><strong>Input</strong></td>
<td>Fixed input current to all units</td>
</tr>
<tr>
<td><strong>Measurements</strong></td>
<td>Spike activity</td>
</tr>
</tbody>
</table>

## B Populations

<table>
<thead>
<tr>
<th>Name</th>
<th>Elements</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Iaf neuron</td>
<td>$N_E = 4N_I$</td>
</tr>
<tr>
<td>I</td>
<td>Iaf neuron</td>
<td>$N_I$</td>
</tr>
</tbody>
</table>

## C Connectivity

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Target</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>E</td>
<td>E</td>
<td>Random with sparseness $\epsilon$, weight $\tilde{g}^E$ ( $\chi \tilde{g}^E$ for connections in a pattern)</td>
</tr>
<tr>
<td>IE</td>
<td>E</td>
<td>I</td>
<td>Random with sparseness $\epsilon$, weight $g^E$</td>
</tr>
<tr>
<td>EI</td>
<td>I</td>
<td>E</td>
<td>Random with sparseness $\epsilon$, weight plastic</td>
</tr>
<tr>
<td>II</td>
<td>I</td>
<td>I</td>
<td>Random with sparseness $\epsilon$, weight $\tilde{g}^I$</td>
</tr>
</tbody>
</table>

## D1 Neuron and Synapse Model

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td>Leaky integrate-and-fire, exponential conductance based input</td>
</tr>
</tbody>
</table>

**Subthreshold dynamics**

$$C \frac{dV}{dt} = (V_{rest} - V) + g^E_i (V^E - V) + g^I_i (V^I - V) + I_b R$$

$$V(t) = V_{rest} \quad \text{otherwise}$$

**Synaptic dynamics**

$$\tau_E \frac{dg^E(t)}{dt} = -g^E(t) + \delta(t - t^*)$$

$$\tau_I \frac{g^I(t)}{dt} = -g^I(t) + \delta(t - t^*)$$

**Spiking**

If $V(t^-) < \theta \land V(t^+) \geq \theta$

1. set $t^* = t$
2. emit spike with time-stamp $t^*$

## D2 Plasticity Model

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td>Inhibitory Spike Timing Dependend Plasticity (iSTDP)</td>
</tr>
</tbody>
</table>

**Acts on**

IE

**Synaptic traces**

$$\tau_{\text{STDP}} \frac{dx_i}{dt} = -x_i + \delta(t - t^*_i)$$

**Online rule**

$$W_{ij} = W_{ij} + \eta (x_i - \alpha) \quad \text{for presynaptic spikes neuron at time } t^*_f$$

and

$$W_{ij} = W_{ij} + \eta x_j \quad \text{for postsynaptic spikes at time } t^*_f$$

## E Input

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current input</strong></td>
<td>Fixed current $I$ to all neurons</td>
</tr>
</tbody>
</table>

Table 1: Tabular description of network model (Figure 4).
### Populations

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_E$</td>
<td>8000</td>
<td>Size of excitatory population E</td>
</tr>
<tr>
<td>$N_I$</td>
<td>2000</td>
<td>Size of inhibitory population I</td>
</tr>
</tbody>
</table>

### Connectivity

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$</td>
<td>0.02</td>
<td>Probability of any connection (EE, EI, IE, II)</td>
</tr>
<tr>
<td>$\bar{g}$</td>
<td>3 nS</td>
<td>Basic weight unit</td>
</tr>
<tr>
<td>$\bar{g}^E$</td>
<td>$\bar{g}$</td>
<td>Weight of all excitatory synapses</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>10</td>
<td>Scaling factor for inhibitory weights</td>
</tr>
<tr>
<td>$\bar{g}^H$</td>
<td>$\gamma \bar{g}$</td>
<td>Weight of inhibitory to inhibitory synapses</td>
</tr>
<tr>
<td>$\bar{g}^{HE}$</td>
<td>$W_{ij} \bar{g}'$</td>
<td>Weight of inhibitory to excitatory synapses</td>
</tr>
<tr>
<td>$\chi$</td>
<td>5</td>
<td>Potentiation factor of excitatory weights belonging to one pattern</td>
</tr>
</tbody>
</table>

### Neuron Model

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{mem}$</td>
<td>10 nS</td>
<td>Membrane capacitance</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>−50 mV</td>
<td>Spiking threshold</td>
</tr>
<tr>
<td>$V_{\text{rest}}$</td>
<td>−60 mV</td>
<td>Resting potential</td>
</tr>
<tr>
<td>$V_E$</td>
<td>0 mV</td>
<td>Excitatory reversal potential</td>
</tr>
<tr>
<td>$V_I$</td>
<td>−80 mV</td>
<td>Inhibitory reversal potential</td>
</tr>
<tr>
<td>$g_{\text{leak}}$</td>
<td>10 nS</td>
<td>Leak conductance</td>
</tr>
<tr>
<td>$I_b$</td>
<td>200 pA</td>
<td>Background current to each cell (unless stated otherwise)</td>
</tr>
<tr>
<td>$R$</td>
<td>1 Ω</td>
<td>Resistance to background current</td>
</tr>
<tr>
<td>$\tau_{\text{ref}}$</td>
<td>5 ms</td>
<td>Absolute refractory period</td>
</tr>
</tbody>
</table>

### Synapse Model

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{\text{ampa}}$</td>
<td>5 ms</td>
<td>Decay constant of AMPA-type conductance</td>
</tr>
<tr>
<td>$\tau_{\text{gaba}}$</td>
<td>10 ms</td>
<td>Decay constant of GABA-type synapse</td>
</tr>
</tbody>
</table>

### Plasticity Model

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{\text{stdp}}$</td>
<td>20 ms</td>
<td>Decay constant of (pre and post) synaptic trace</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$5 \times 10^{-3}$</td>
<td>Learning rate</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.14</td>
<td>Presynaptic offset</td>
</tr>
<tr>
<td>$W_{\text{min}}$</td>
<td>0</td>
<td>Minimum inhibitory synaptic weight</td>
</tr>
<tr>
<td>$W_{\text{max}}$</td>
<td>$3 \bar{g}'$</td>
<td>Maximum inhibitory synaptic weight</td>
</tr>
</tbody>
</table>

Table 2: Simulation parameters for network model (Figure 4).