Locally balanced inhibition allows for robust learning of input-output associations in feedforward networks with Hebbian plasticity

Gloria Cecchini[1,2] and Alex Roxin[1]

[1]Centre de Recerca Matematica, Campus de Bellaterra, Edifici C, Bellaterra, 08193, Barcelona, Spain.

[2]Department of Information and Communications Technologies, Universitat Pompeu Fabra, Carrer Roc Boronat 138, Barcelona, 08018, Spain.

Abstract

In neural networks within the brain, the activity of a post-synaptic neuron is determined by the combined influence of many pre-synaptic neurons. This distributed processing enables mechanisms like Hebbian plasticity to associate sensory inputs with specific internal states, as seen in feedforward structures such as the CA1 region of the hippocampus. By modifying synaptic weights through Hebbian rules, sensory inputs can subsequently elicit outputs that consistently reflect their corresponding internal states. When input and output patterns are uncorrelated, this approach allows for the encoding of a large number of distinct associations, enabling efficient memory storage.

Our study demonstrates a critical limitation when output patterns become weakly correlated with input patterns through the intrinsic feedforward network's connectivity. In these cases, the Hebbian rule preferentially strengthens synaptic weights shared across patterns, leading to a "freezing" of the network's structure. This results in highly correlated output patterns over time, effectively reducing the network's capacity to store diverse associations and limiting its flexibility in learning.

To address this challenge, we find that including a mechanism of locally balanced inhibition, which has been shown to be a key feature of cortical circuits in-vivo, counteracts the undesired correlations between inputs and outputs. By dynamically regulating inhibitory input, locally balanced inhibition prevents the over-strengthening of shared weights, restoring the network's ability to maintain robust and flexible learning. This finding underscores the importance of inhibitory mechanisms in enabling efficient and adaptive information processing in neural circuits, offering insights into how biological networks maintain their remarkable capacity for associative learning.

Keywords: Plasticity, Inhibition, Feedforward Network, Associative Learning

1 Introduction

Associative learning in cortex is often modeled with Hebbian plasticity, by which synapses strengthen between co-active pre- and post-synaptic neurons, and weaken otherwise (Hebb (1949); Trappenberg (2002); Shatz (1992)). In such frameworks, feedforward networks are used as a tool to store mappings between input cues and output responses by progressively sculpting synaptic connectivity. However, continual learning

presents a fundamental challenge: as new associations are acquired, they interfere with and overwrite older ones, leading to degraded recall.

Classical Hebbian models typically assume that patterns of activity at the output layer are imposed a priori, effectively decoupling postsynaptic activity from the synaptic structure being learned. Biologically, however, neuronal circuits receive inputs from a number of sources which might include sensory input, top-down inputs and neuromodulatory effects. Consequently, any hetero-associative mapping from a given pathway to the network's output must be robust to input–output correlation. When the output reflects the concurrent activity of multiple pathways, this coupling may be small, but it is unavoidable, and models should be able to operate reliably in its presence. In this manuscript, we adapt classical Hebbian hetero-associative models to capture this mixture of drives. Doing so necessarily introduces correlations between synaptic input and postsynaptic output.

Here, we revisit Hebbian associative learning in a minimal feedforward setting where the output patterns are explicitly influenced by the input pathway targeted by the hetero-associative plasticity mechanism. We consider binary units and a stochastic Hebbian rule, and we compute the postsynaptic output as the sum of the current synaptic drive from the target input and additional inputs, which can be approximated as a Gaussian noise term of adjustable amplitude. This design interpolates between two regimes: (i) strong input-output correlations when outputs depend only on the targeted input pathway, and (ii) vanishing input-output correlations when the influence of the targeted pathway on outputs is negligible. This interpolation allows us to isolate how input-output correlation shapes plasticity, network dynamics, and memory.

We show that when input-output coupling is strong (i.e., the postsynaptic output is dominated by the current synaptic drive), the network undergoes a *freezing transition*: successive readouts for uncorrelated inputs become highly similar. Mechanistically, binarized outputs preferentially activate high in-degree; Hebbian updates then potentiate those rows and depress low in-degree rows, producing a progressive polarization of the connectivity (row-sum separation). Increasing the contribution of additional inputs weakens this correlation and recovers classical behavior, only when the influence of the targeted pathway on outputs is almost negligible. This raises the question: what mechanism preserves flexibility when output–input coupling cannot be neglected? Motivated by circuit evidence for inhibitory balance (Vogels et al. (2011); Renart et al. (2010); Chistiakova et al. (2015); Bannon et al. (2020)), as well as in in-vivo recordings (Haider et al. (2006); Shu et al. (2003); Dehghani et al. (2016); Okun and Lampl (2008); Murphy and Miller (2009)), we introduce a locally balanced inhibition mechanism: each neuron receives an inhibitory term proportional to its own excitatory in-degree. When the inhibitory gain is tuned to the network's activity level, the local balancing term effectively recenters postsynaptic drive, so output variability is determined by the current input rather than by pre-existing differences in synaptic in-degree. This breaks the bias toward repeatedly activating the same highly connected neurons, and consequently it decorrelates successive readouts. Moreover, in this regime we observe a clear improvement in memory strength.

All together, by introducing a mechanism of locally balanced inhibition, we demonstrate a simple yet powerful solution that restores flexibility and preserves memory capacity. These findings emphasize the importance of inhibitory regulation in maintaining the balance between stability and adaptability in neural learning systems.

2 Results

2.1 Dynamics of associative learning and interference in Hebbian networks

In this section, we introduce the main results and features of Hebbian plasticity as a mechanism for learning in feedforward neural networks (Hebb (1949); Trappenberg (2002); Shatz (1992)). First, Hebbian plasticity enables the formation of associations between input and output patterns. For instance, the presentation of a specific sensory cue, such as the sight of rain, can become associated with a behavioral response, like reaching for an umbrella (Fig. 1a). In a Hebbian framework, such associations are stored by modifying synaptic weights in a way that strengthens the connection between co-active pre- and post-synaptic neurons,

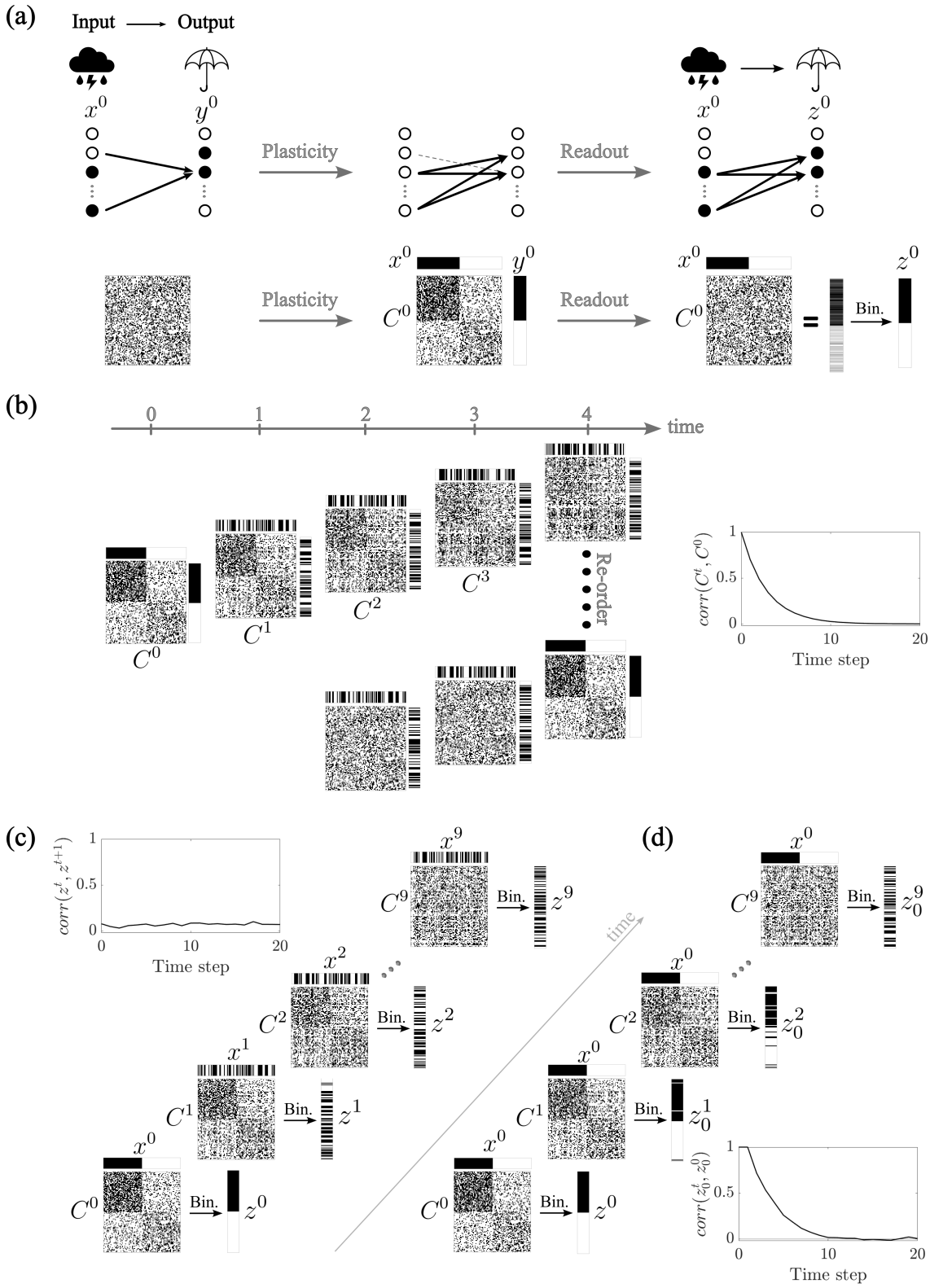

Fig. 1: Caption continues on next page

Fig. 1: continued from previous page

Dynamics of Associative Learning and Interference in Hebbian Networks. **(a)** Hebbian plasticity allows the association of input patterns with corresponding output patterns in a feedforward network. **(b)** As new associations are learned, the synaptic structure formed by previous associations is progressively overwritten, as shown by the decay in the correlation between the connectivity matrix C at time 0 and time t. The connectivity matrix is always correlated with the most recently stored patterns, as revealed by reordering the cells. **(c)** After learning, driven readouts remain largely uncorrelated, e.g. distinct sensory inputs lead to distinct outputs. The correlation between consecutive readouts z^t of random inputs x^t remains low. **(d)** Over time, the autocorrelation of an elicited readout z_0^t decays over time due to the overwriting from ongoing learning, leading to a progressive loss in recall fidelity. As a result, the correlation $corr(z_0^t, z_0^0)$ decreases and eventually approaches zero after a sufficient number of learning steps. – Parameters: $f = 0.5$, $p = 0.4$, $N = 100$ ($N = 1000$ for the correlation plots).

and weakens it if one is inactive. Specifically, in this work, we consider the connectivity matrix C^t at time t to be binary. If $C_{ij}^t = 1$, then the pre-synaptic neuron x_j^t is connected to the post-synaptic neuron y_i^t. Moreover, we also consider the activity of the neurons to be binary, i.e., we focus only on the active/inactive status of each neuron, therefore $x_j^t, y_i^t \in \{0, 1\}$. Finally, Hebbian plasticity follows a stochastic rule

- if $C_{ij}^t = 0$ and $x_i^t = y_j^t = 1$, then $C_{ij}^{t+1} = 1$ with probability p;
- if $C_{ij}^t = 1$ and $x_i^t \neq y_j^t$, then $C_{ij}^{t+1} = 0$ with probability p;
- if $x_i^t = y_j^t = 0$, then $C_{ij}^{t+1} = C_{ij}^t$.

As the network continues to learn new associations, previously formed synaptic structures are progressively altered (Fig. 1b). This process of interference is evident when examining the correlation $corr(C^t, C^0)$ between the connectivity matrix C at the initial learning time ($t = 0$) and its state at a later time t. As new input-output pairs are encoded, the correlation between these matrices decays, reflecting the ongoing overwriting of old associations by new ones. Despite this turnover, the synaptic matrix remains consistently aligned with the most recently stored associations. This is demonstrated by reordering the neurons according to their current activity patterns: a clear structural organization emerges that reflects the latest learning episodes (Fig. 1b).

As shown in Fig. 1c, when the network is probed with novel learned inputs, the resulting readouts are distinct from one another. This ensures that different inputs elicit distinct outputs, an essential feature of reliable memory

storage. However, the persistence of these associations is limited. Over time, as additional patterns are encoded, the output corresponding to an initially learned input pattern progressively changes (Fig. 1d). This is captured by the decay of the autocorrelation $corr(z_0^t, z_0^0)$, where z_0^0 is the initial output elicited by an input x^0, and z_0^t is the readout of the same input after t learning steps. The correlation decreases exponentially with each new association, eventually approaching zero, indicating a complete loss of recall fidelity. This phenomenon illustrates the inherent limitations of associative memory in ongoing learning in Hebbian networks, highlighting the trade-off between learning new information and retaining previously acquired associations.

2.2 An Input-Driven Framework for Hebbian Learning

As described in the previous section, classical formulations of Hebbian learning typically rely on the assumption that both the input and the output patterns are imposed a-priori, in other words they are predefined rather than emerging dynamically within the network. This assumption can be justified in many scenarios, especially when modeling simplified systems, since a neuron's output is generally shaped by multiple converging inputs. Take, for example, the hippocampal CA1 region. Its activity is driven not only by input from CA3, but also by significant projections from other brain areas, such as the entorhinal cortex layer III (ECIII). Classical Hebbian models often implicitly assume that these additional influences are dominant, effectively treating the contribution from CA3 as negligible.

To address this limitation, we introduce a more biologically grounded modeling approach in which the output activity is not imposed a-priori but is instead computed as the result of two additive components: the driven response evoked by a specific input pattern (e.g., from CA3), and additional inputs representing the influence of other brain areas (e.g., ECIII), see Fig. 2a. Mathematically, we model the output y^t as the result of the sum of the target input $C^{t-1}x^t$ and additional inputs $v_1^t, \cdots, v_n^t$, see Fig. 2a. Assuming the pathways are uncorrelated, the sum of these additional inputs can be modeled as Gaussian noise $\sigma\xi \sim N(0, \sigma)$. The parametrization of the standard deviation σ allows us to modulate its relative influence on the output. Therefore, the output is calculated as $\bar{y}^t = C^{t-1}x^t + \sigma\xi$ (Fig. 2b). In order to classify neurons as active or inactive, we set a threshold and binarize $\bar{y}^t$. The threshold value is chosen such that the fraction of active cell f remains constant across time. Once the binarized output y^t is calculated, the network undergoes plasticity and the connectivity matrix is updated as described in the previous section, resulting in C^t. Finally, the readout z^t is driven by the target input x^t and the corresponding connectivity matrix C^t (Fig. 2c-d). Specifically, the readout z^t is the binarization of $\bar{z}^t = C^t x^t$, again keeping constant the fraction of active cell f.

To investigate how the partial correlation between output and synaptic input strength influences network plasticity, we varied the strength of the non-target input pathways relative to the target input pathway, i.e., σ. This manipulation simulates a spectrum of scenarios, ranging from fully input-driven outputs (low σ) to outputs dominated by the other inputs (high σ), see Fig. 2e. Our simulations reveal that even when the correlation between output and synaptic input is weak (e.g., $\sigma = 0.1$), the network exhibits a pathological dynamic: successive readouts z^t remain highly similar over time. In the extreme case where the output is entirely determined by the structured input ($\sigma = 0$), all readouts z^t become identical. This leads to a complete loss of learning flexibility, as the network effectively freezes into a fixed response. In contrast, for large σ values, the output is no longer systematically biased by the synaptic input structure, thereby recovering the conditions assumed in classical Hebbian paradigms. In this regime, the network behaves as previously described in Fig. 1, where uncorrelated inputs give rise to distinct outputs, enabling robust associative learning.

These results underscore the critical role of output-input correlations in shaping the long-term plasticity of feedforward networks and highlight the potential stabilizing role of additional inputs in maintaining learning flexibility. In the next section, we perform a more systematic analysis on the impact of an output that is correlated to the synaptic input strength, on network plasticity.

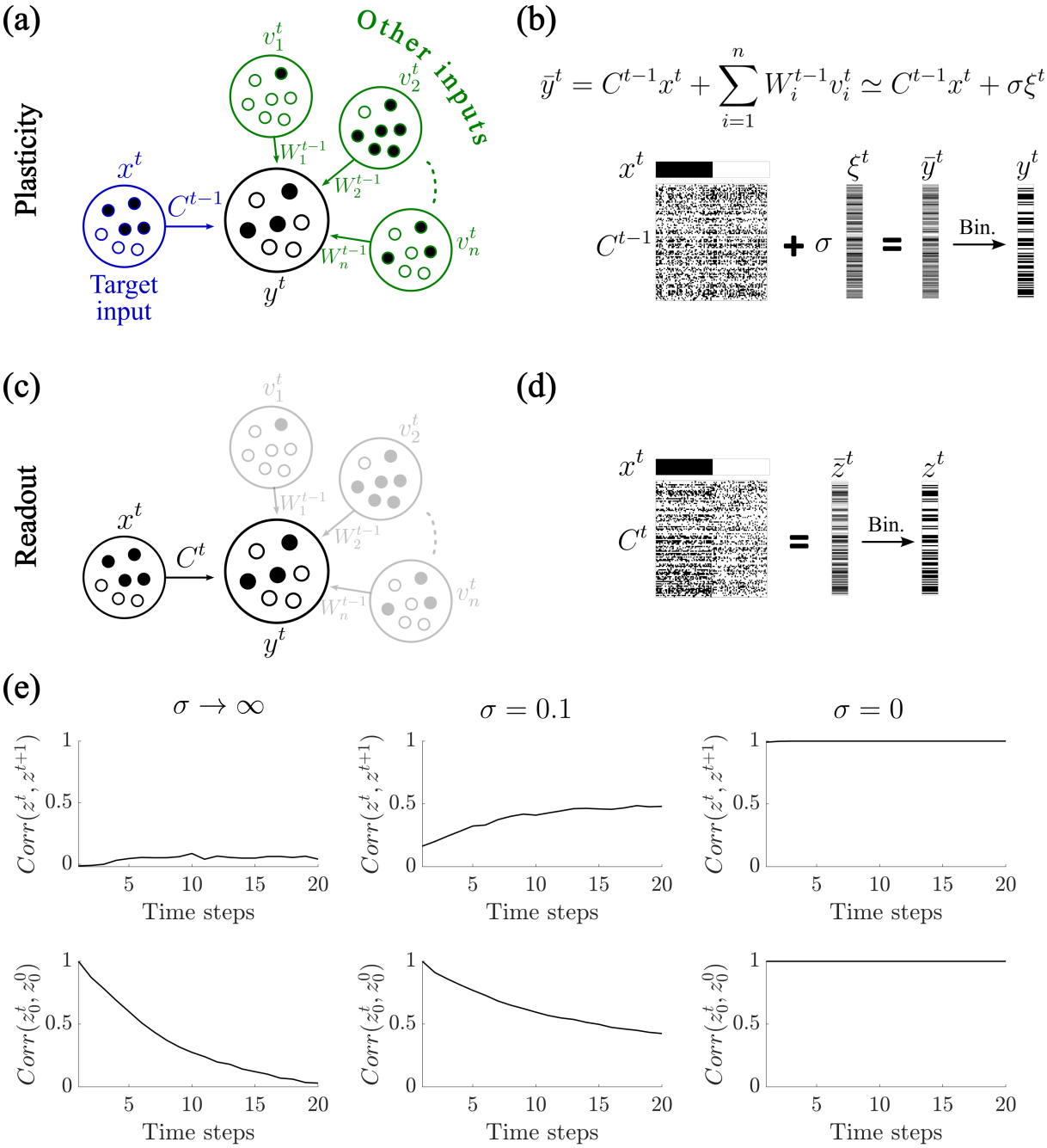

Fig. 2: An Input-Driven Framework for Hebbian Learning. **(a-b)** The output y^t reflects the sum of a target input pattern and additional inputs, resulting in partial input-output correlations. These additional inputs can be modeled as Gaussian noise. **(c-d)** The readout is driven by the target input pattern and the corresponding connectivity matrix, after plasticity. **(e)** For $\sigma \to \infty$, we have the same scenario as in Fig. 1. For $\sigma = 0.1$, i.e., in the presence of weak correlation between output and synaptic input strength, the network produces readouts z^t that remain strongly similar across time. Finally, for $\sigma = 0$, the readouts z^t become identical. – Parameters: (b,d) $N = 100$, $f = 0.5$, $p = 0.4$, $\sigma = 0.1$ (e) $N = 1000$, $f = 0.5$, $p = 0.2$.

2.3 Correlation between output and synaptic input strength impairs associative learning

In the previous section, we presented some illustrative examples demonstrating how correlation between output activity and synaptic input strength can profoundly affect the dynamics of Hebbian plasticity. In this section, we extend our analysis with a more systematic and quantitative investigation, aimed at characterizing the network's plastic behavior across a continuum of input-output correlation strengths. To this end, we varied the amplitude σ of the noise term, which modulates the extent to which the output is driven by the target input versus other inputs. Small values of σ correspond to outputs tightly determined by the synaptic input, whereas large values of σ simulate conditions in which outputs are dominated by additional, non-specific inputs.

To evaluate the effects of σ on the network's learning dynamics, we focus on the correlation $corr(\bar{z}^t, \bar{z}^{t+1})$ between consecutive readouts before binarization $\bar{z}$, as a metric to quantify the degree of freezing in the network's output. Since we consider uncorrelated inputs, high correlations between successive readouts indicate reduced flexibility and a tendency toward rigid, self-reinforcing representations.

We first analyze the case of $\sigma = 0$, where the output is entirely determined by the target input. As shown in Fig. 3a, in this regime the correlation between consecutive readouts progressively increases and eventually converges to 1, regardless of other network parameters such as the sparsity of activation f or the synaptic plasticity probability p. This convergence implies that, over time, the network settles into a fixed readout pattern that is insensitive to the input identity, indicating a complete loss of representational flexibility. Interestingly, this convergence follows a power-law trajectory, as illustrated in Fig. 3b. We extracted the power-law exponent across different values of f and p and found that both parameters modulate the rate at which this rigidity emerges: higher values of f or p accelerate the convergence (Fig. 3c).

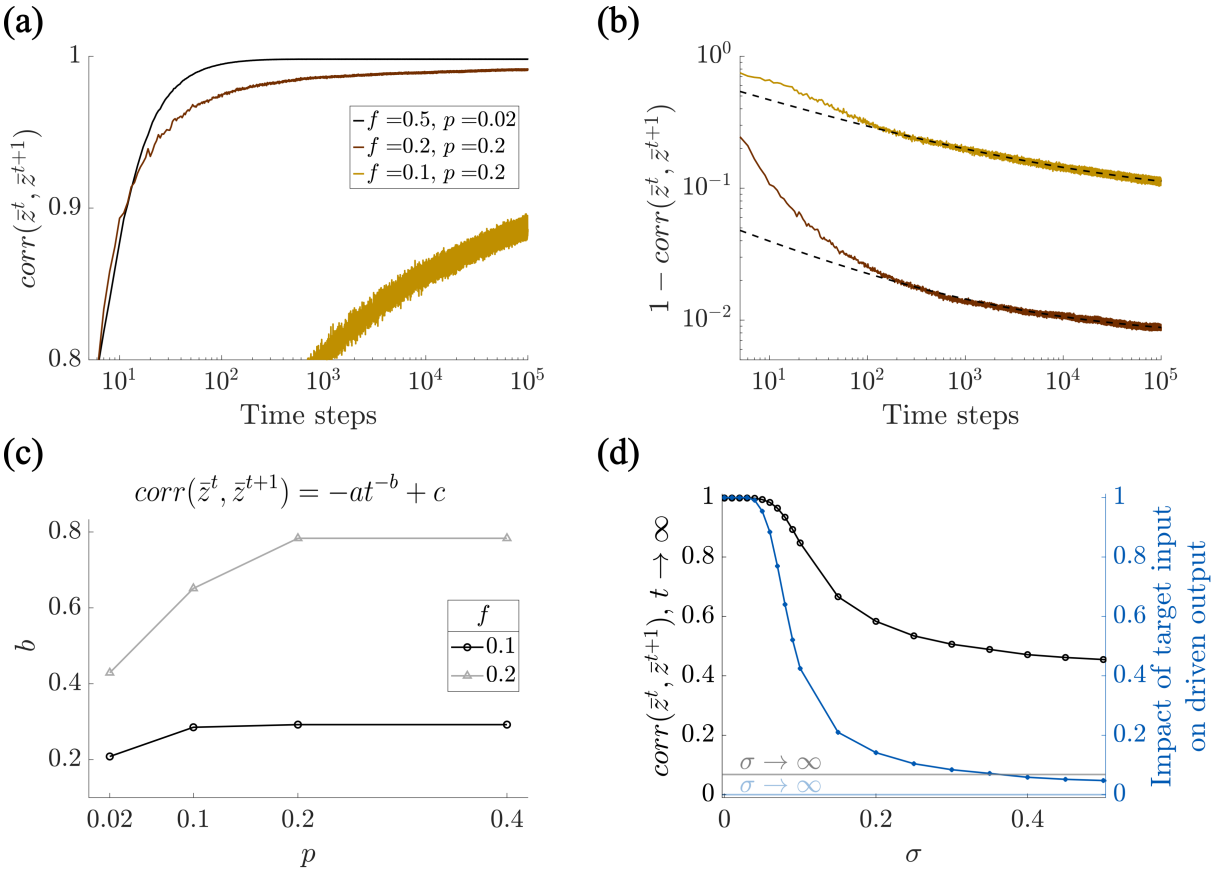

Fig. 3: Correlation between output and synaptic input strength impairs associative learning. **(a)** Uncorrelated inputs yield identical outputs after learning when the effect of other input pathways is zero, for any fraction of active cells f and probability of potentiation/depression p. **(b)** The correlation $corr(\bar{z}^t, \bar{z}^{t+1})$ converges to one following a power-law. **(c)** The extracted power-law parameters for the power reveal the dynamics of readout convergence. **(d)** (Black) The influence of other inputs prevents convergence of the correlation $corr(\bar{z}^t, \bar{z}^{t+1})$, which decreases with increasing σ. (Blue) Impact of target input on the driven output, calculated as $corr(C^{t-1}x^t + \sigma\xi^t \ , \ C^{t-1}x^t)$. – Each curve shows the mean over 50 simulations. Parameters: $N = 1000$, (a-c): $\sigma = 0$, (d): $f = 0.5$, $p = 0.2$.

Next, we explored how increasing the amplitude of the noise term σ, representing an increasing contribution of other inputs, affects the network's tendency toward freezing. We systematically varied σ and measured the asymptotic value of $corr(\bar{z}^t, \bar{z}^{t+1})$, both numerically and via power-law fitting. The results, summarized in Fig. 3d (black), show a clear trend: as σ increases, the correlation between successive readouts decreases monotonically. These findings imply that additional inputs can prevent the network from entering a fully frozen state, thereby preserving its ability to encode diverse patterns.

Finally, we quantified the impact of the additional inputs on the driven output, specifically, how much the purely driven output $C^{t-1}x^t$ differs from the total output $C^{t-1}x^t + \sigma\xi^t$ as a function of the noise amplitude σ. To this end, we computed the correlation between these two signals. i.e., $corr(C^{t-1}x^t + \sigma\xi^t \ , \ C^{t-1}x^t)$, which serves as a measure of the impact of the target input on the driven output. As shown in Fig. 3d (blue), increasing the noise leads to a progressive decorrelation between the two, as expected. However, a comparison with the correlation $corr(\bar{z}^t, \bar{z}^{t+1})$ (in black) reveals a more nuanced picture: to fully decorrelate the output from the intrinsic bias of

the network, the noise must be strong enough to almost completely override the contribution of the target input. In other words, mitigating network freezing through noise alone requires driving the output toward a nearly random pattern, essentially reverting to the classical Hebbian learning scenario in which outputs are externally imposed. Since the aim of this study is to develop a model in which outputs are determined by a target input, we must seek an alternative solution to this problem.

In the next section, we investigate in greater detail the mechanisms underlying network freezing, particularly in scenarios where the output remains strongly correlated with the synaptic input strength.

2.4 Mechanism of network freezing due to high correlation between output and synaptic input strength

To understand the emergence of rigid dynamics in a feedforward network under Hebbian plasticity, we examined the consequences of strong correlations between output activity and synaptic input strength. Intuitively, once a neuron loses enough synapses, its input drops below threshold and it falls silent. Because the plasticity rule potentiates only co-active pairs, a silent postsynaptic neuron cannot regain synapses and instead tends to undergo further depression, effectively freezing it in a no-input state. Conversely, a neuron with sufficiently many synapses keeps firing, and co-activity sustains its potentiated inputs.

In Fig. 4a, we begin with a randomly initialized synaptic matrix. For visualization purposes, we reorder the rows of the matrix such that the resulting output vector $\bar{y}^0$ exhibits a descending gradient of values. This reordering allows us to visualize, in later steps, how Hebbian learning affects different subsets of the network. The distribution of output entries in $\bar{y}^0$ is shown using a histogram, where the first part of the vector is depicted in black and the second part in gray. In order to classify neurons as active or inactive, we set a threshold and binarize $\bar{y}^0$. The threshold value is chosen such that the fraction of active cell f remains constant across time. By construction, the binarization of this graded output yields a binary pattern y^0, with the first part consisting entirely of ones and the second part of zeros. This binarized output pattern y^0 is then used to do Hebbian plasticity together with the input x^0, as explained in Sec. 2.1.

In subsequent learning steps, the asymmetry of the $\bar{y}^t$ entries becomes amplified (Fig. 4b). As the network continues to process new inputs and undergo plasticity, already at time $t = 1$, the output vector $\bar{y}^1$ exhibits an even stronger separation between the first and second parts, as shown by the histograms. Consequently, its binarized version y^1 remains almost identical to y^0. This pattern of activity leads to further strengthening and weakening of the same subsets of synapses, reinforcing the structure already imprinted in the matrix during previous steps. Over time, the network's synaptic matrix becomes increasingly polarized, with a clear division between potentiated and depressed rows (see C^{100} in Fig. 4b).

This progressive separation of output activity and its corresponding synaptic changes establishes a positive feedback loop: once a pattern of activity becomes dominant, it increasingly biases future outputs and plasticity events toward reinforcing the same structure. As a result, the network enters a frozen state in which the output becomes insensitive to new inputs. After a few time steps, regardless of the input presented, the network produces identical readouts. This collapse of output diversity reflects a pathological rigidity in the network's dynamics, fundamentally impairing its capacity for associative learning and memory storage.

In the next section, we build upon the insights gained from the analysis of network freezing dynamics. Specifically, we aim to develop and investigate a biologically inspired mechanism capable of counteracting the onset of rigidity in network dynamics. By introducing a targeted intervention, our goal is to preserve the network's capacity for flexible associative learning while mitigating the detrimental effects of excessive input-driven reinforcement.

2.5 Locally balanced inhibition counteracts input-driven rigidity

As shown in the previous sections, in feedforward networks governed by Hebbian plasticity, output

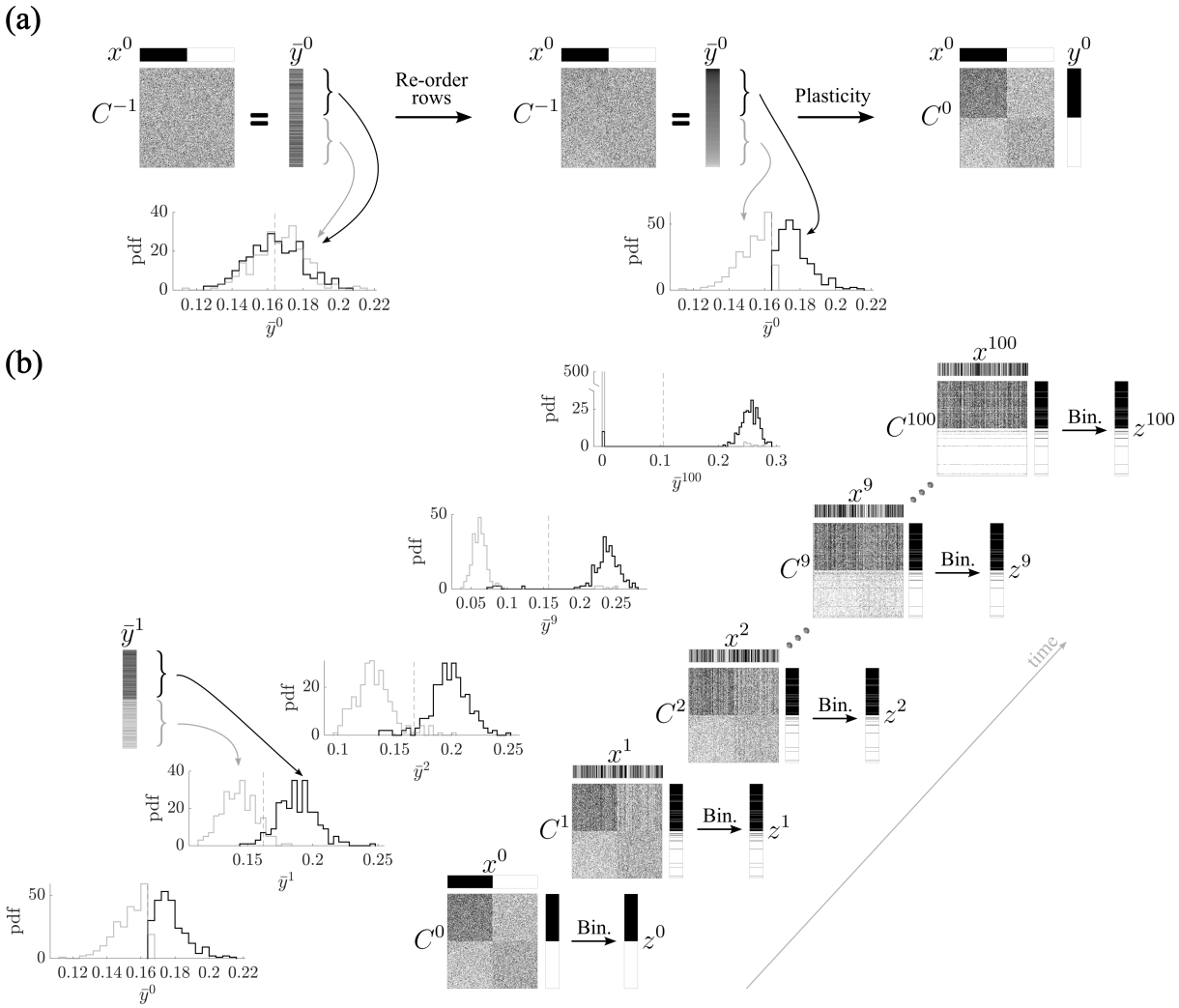

Fig. 4: Mechanism of network freezing due to high correlation between output and synaptic input strength. **(a)** Starting from a random synaptic matrix, the rows are reordered to produce an output vector $\bar{y}^0$ with values arranged in decreasing order. The histogram of the first part of its entries is shown in black, and the second one in gray. Binarizing $\bar{y}^0$ yields the vector y^0, where the first part consists of 1s and the second one of 0s. Following Hebbian plasticity, the corresponding synaptic matrix C^0 is updated, by potentiation and depression with probability p. **(b)** In the subsequent steps, the two parts of the output $\bar{y}^t$ become increasingly separated. As a result, the binarized vector y^t consists mainly of 1s in the first part and 0s in the second one, further reinforcing the synaptic structure established in the previous step and amplifying the network's rigidity. As the output entries become increasingly separated, the synaptic matrix develops a more pronounced structure, with potentiated and depressed rows clearly diverging. This growing separation causes the network to produce identical readouts over time. – Parameters: $N = 500$, $f = 0.5$, $p = 0.2$, $\sigma = 0$.

activity that is exclusively determined by excitatory input can lead to highly stereotyped and inflexible activation patterns. In such scenarios, neurons with a large excitatory synaptic input are consistently activated, while those with fewer inputs are persistently suppressed. This biased pattern of activity reinforces the same subset of

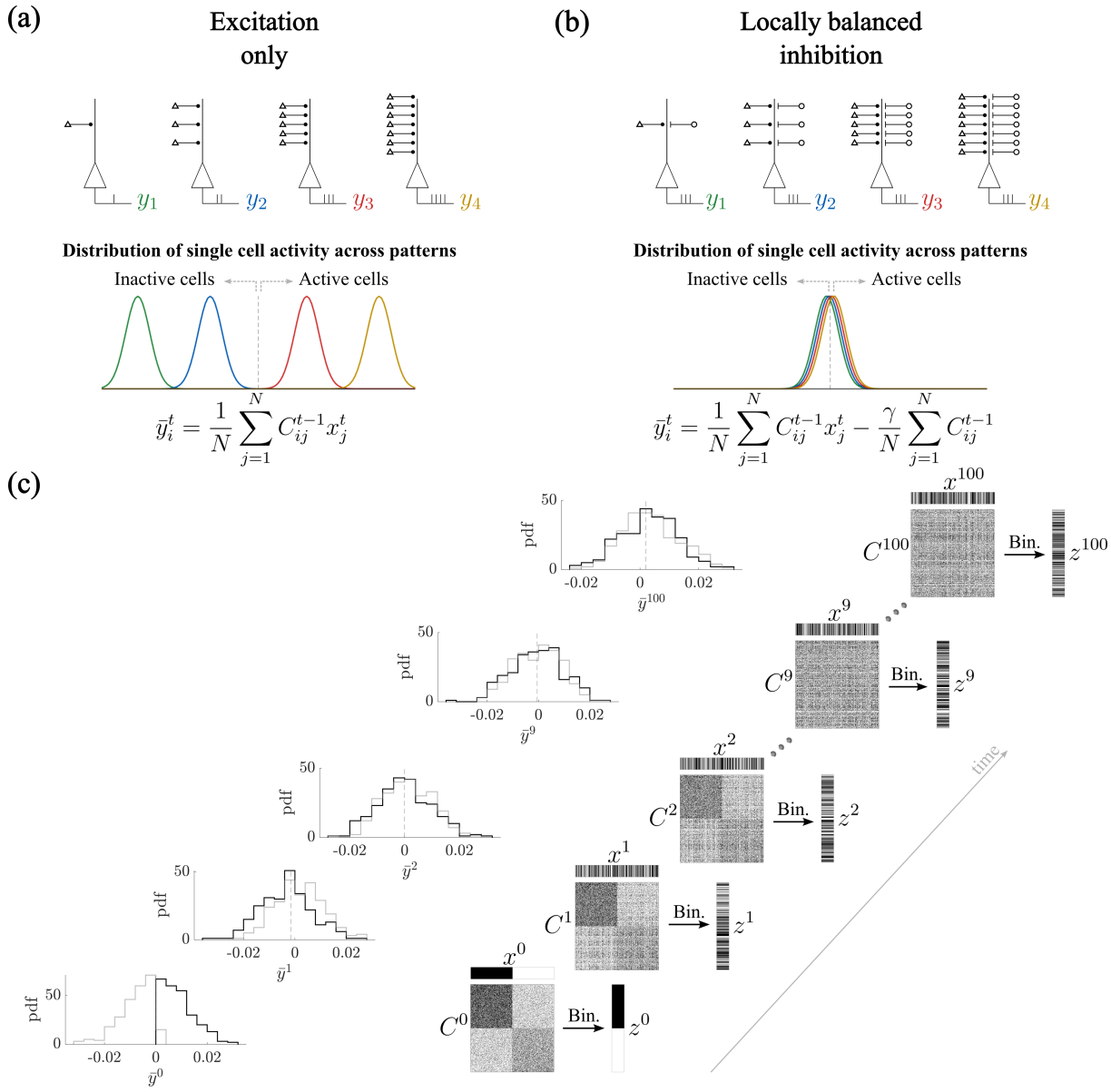

Fig. 5: Locally balanced inhibition counteracts input-driven rigidity. **(a)** When output activity is solely driven by excitatory input, neurons with high in-degree are consistently selected as active, while those with low in-degree remain inactive, leading to biased activation patterns. **(b)** Each output neuron receives inhibitory input scaled to its total excitatory input (in-degree), providing a neuron-specific balancing mechanism. **(c)** This local inhibitory balancing prevents progressive separation of output values, thereby avoiding matrix freezing and preserving the network's capacity for flexible learning. – Parameters: $N = 500$, $f = 0.5$, $p = 0.4$, $\gamma = 0.5$.

neurons across learning episodes, ultimately leading to network rigidity and reduced capacity for storing diverse associations.

To address this issue, we introduce a mechanism of locally balanced inhibition, which has been shown to be a key feature of cortical circuits both theoretically (Vogels et al. (2011); Renart et al.

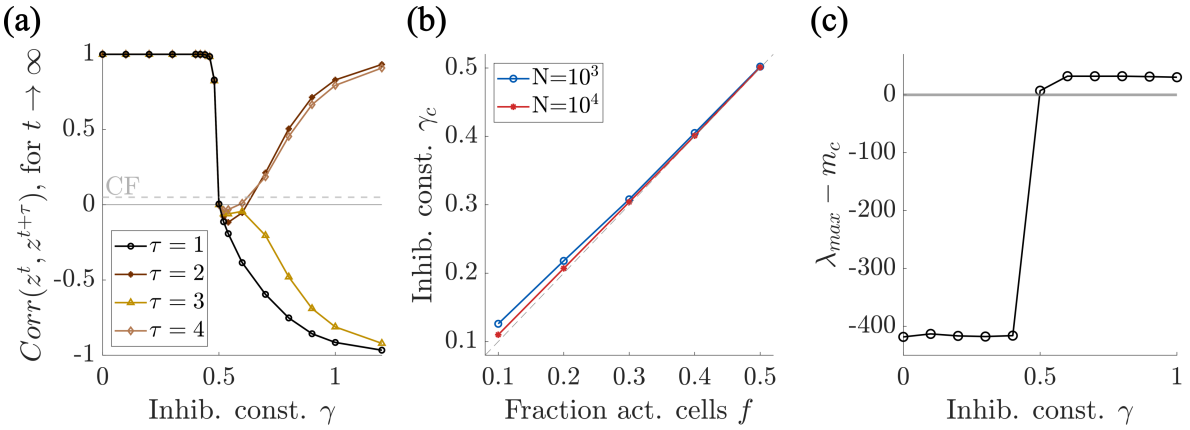

Fig. 6: Locally balanced inhibition enables robust learning of input-output associations. **(a)** Appropriate levels of inhibition restore the network's ability to form reliable input-output associations, as shown by the correlation of the first ($\tau = 1$) and subsequent ($\tau = 2, 3, 4$) nearest neighbor. **(b)** The inhibitory constant γ that minimize, in absolute value, the correlation between neighboring readouts is proportional to the fraction of active cells f. **(c)** The rescaled largest eigenvalue vanishes only when $\gamma = f$, showing that full network flexibility is restored when inhibition matches the fraction of active cells. – Each curve shows the mean over 10 simulations. Parameters: $p = 0.2$, $f = 0.5$, $N = 5000$.

(2010); Vreeswijk and Sompolinsky (1996); Murphy and Miller (2009); Rubin et al. (2015); Koch et al. (2010)), and in-vivo (Haider et al. (2006); Shu et al. (2003); Dehghani et al. (2016); Okun and Lampl (2008); Murphy and Miller (2009); Ozeki et al. (2009)). In (Chistiakova et al. (2015); Bannon et al. (2020)), the authors show that Hebbian plasticity at excitatory inputs to inhibitory neurons risks runaway potentiation, but weight-dependent heterosynaptic plasticity counteracts this by inversely adjusting other synapses, normalizing weights and preventing instability. Along these lines, in our manuscript, each output neuron receives inhibitory input that is proportionally scaled to its own excitatory synaptic input (Fig. 5a-b). Specifically, the output at time t is the binarization of

$$\bar{y}_i^t = \frac{1}{N}\sum_{j=1}^{N} C_{ij}^{t-1} x_j^t - \frac{\gamma}{N}\sum_{j=1}^{N} C_{ij}^{t-1}$$

where N is the number of neurons, and γ is the inhibitory constant. This neuron-specific inhibitory feedback effectively normalizes the total synaptic drive each cell experiences, thereby dampening the competitive advantage conferred by high in-degree. By counteracting the dominance of highly connected neurons, this inhibitory balancing mechanism prevents the progressive separation of output values that would otherwise emerge during repeated learning (compare Fig. 5c with Fig.4b).

The inhibitory mechanism just described enables robust learning of input-output associations only when appropriately tuned. Fig. 6a shows how the correlation $corr(z^t, z^{t+\tau})$ between neighboring readouts evolves by changing the inhibitory constant γ; for this figure we set the fraction of active cells to $f = 0.5$. For γ not large enough ($\gamma < 0.5$), the inhibitory feedback is not able to counterbalance the synaptic drive, and therefore, eventually, the system converges into a rigid state, i.e., $corr(z^t, z^{t+1}) = 1$. On the other hand, if the inhibition is too strong ($\gamma > 0.5$), the correlation becomes negative, i.e., cells with low in-degree are first boosted and then strongly depressed at the following time step, causing $corr(z^t, z^{t+1}) \to -1$ and $corr(z^t, z^{t+2}) \to 1$ for $\gamma >> 0.5$. When appropriately tuned ($\gamma = 0.5$), the network retains the ability to generate distinct and reliable output patterns in response to different input stimuli, i.e., $corr(z^t, z^{t+\tau}) = 0$ for every τ.

We investigated if changing the parameters for the fraction of active cells f alters the trend of the correlation $corr(z^t, z^{t+1})$ as a function of

γ. We found that the curve profile is qualitatively the same, but the value of γ that returns $corr(z^t, z^{t+1}) = 0$, called γ_c, equals f, see Fig. 6b. In order to have an intuition of why setting $\gamma = f$ leads to $corr(z^t, z^{t+1}) = 0$, we calculate the mean of the output $\bar{y}_i^t$

$$\langle \bar{y}_i^t \rangle = \langle \frac{1}{N} \sum_{j=1}^{N} C_{ij}^{t-1} x_j^t - \frac{\gamma}{N} \sum_{j=1}^{N} C_{ij}^{t-1} \rangle = \frac{d_i}{N}(f - \gamma)$$

where d_i is the in-degree of each neuron i. We noticed that for $\gamma = f$, then $\langle \bar{y}_i^t \rangle = 0$, which means that the variability due to the heterogeneity in mean in-degrees is eliminated, and therefore strong/weak synapses will not necessarily be strengthened/weakened even more.

As an additional demonstration that $\gamma = f$ restores full flexibility in the network, we calculated the eigenvalues of the connectivity matrix (C) after the system reached a steady state ($t = 10^3$). For every value of γ, the eigenvalues are distributed within a disk centered at zero, except for one outlier lying on the real axis at varying distances depending on γ (see studies on non-Hermitian (Ginibre) matrices, e.g., O'Rourke and Renfrew (2013); Girko (1985)). In particular, this outlier, i.e., the largest eigenvalue λ_{max}, scales as $N\langle C \rangle$. To assess the role of this outlier, we calculated the theoretical mean m_c expected if the network had evolved according to the classical framework (Sec. 2.1). This value was derived in Devalle et al. (2025), see supplementary materials. We then plotted the distance of the outlier eigenvalue λ_{max} from the expected scaled mean Nm_c as a function of γ, shown in Fig. 6c for $f = 0.5$. We observed that $\lambda_{\max} - m_c = 0$ only when $\gamma = f = 0.5$. The disappearance of the rescaled outlier at $\gamma = f$ indicates that the network statistics are consistent with the classical Hebbian process described in Sec. 2.1, demonstrating once again that the system fully recovers flexibility only when inhibition is locally balanced with the fraction of active cells.

2.6 Locally balanced inhibition enhances memory strength

As a step further, we wondered if the inclusion of locally balanced inhibition recovers the network's effective memory capacity. To this aim, after a period of initialization, the system was probed with the same input x^0, giving a different readout $z_0^t = \frac{1}{N} C^t x^0 - \frac{\gamma}{N} \sum_{j=1}^{N} C_{ij}^t$ at each time step. As shown in Fig. 7a-b, the correlation $corr(z_0^t, z_0^0)$ decays in time, for both the classical framework (CF) and the locally balanced inhibition (BI) case, i.e., with $\gamma = f$. Surprisingly, the memory performance for BI improved when compared to CF. This suggests that locally balanced inhibition not only preserves the network's representational diversity but also extends its operational regime for stable and flexible learning.

To assess the extent to which memory performance improves, we computed the memory strength of the system under these two conditions (CF vs. BI), varying the system size, i.e., the number of neurons N. For this purpose, we used the signal-to-noise ratio (SNR), which is defined as the ratio between the contribution of the stored memory pattern being retrieved and the cumulative interference from all other stored patterns (Fusi and Abbott (2007); Benna and Fusi (2016); Amit and Fusi (1992, 1994)). Specifically, we defined the signal as the expected correlation with the target pattern we are trying to retrieve, and the noise as the standard deviation of correlations. Mathematically, we computed the correlation between the readout z_0^t at time t elicited by input x^0 with the original readout z_0^0 at time $t = 0$. Accordingly, the signal is the correlation $\text{Mean}(corr(z_0^t, z_0^0))$, and the noise is $\text{Std}(corr(z_0^t, z_0^0))$. Based on this definition, we set the SNR threshold to 1, meaning successful retrieval requires the signal to exceed the noise. As shown in Fig. 7b, if we had defined the memory strength as the first time step that gives SNR= 1, the two scenarios would have yielded very similar results. However, by visual inspection, the BI case would be preferable. To take this into account, we defined the *Memory strength* as a cumulative measure of the SNR (Benna and Fusi (2016); Fusi et al. (2021)). Mathematically, let m^t and s^t be the mean (signal) and standard deviation (noise), respectively, at time step t. Then, call T the first time step such that SNR= 1, or equivalently $m^t = s^t$. We defined the memory strength as the following cumulative measure: $\sum_{t=1}^{T}(m^t - s^t)$. This metric indeed accounts for the difference in shape between the two curves (Fig. 7b), and keeps track of the SNR. As a result,

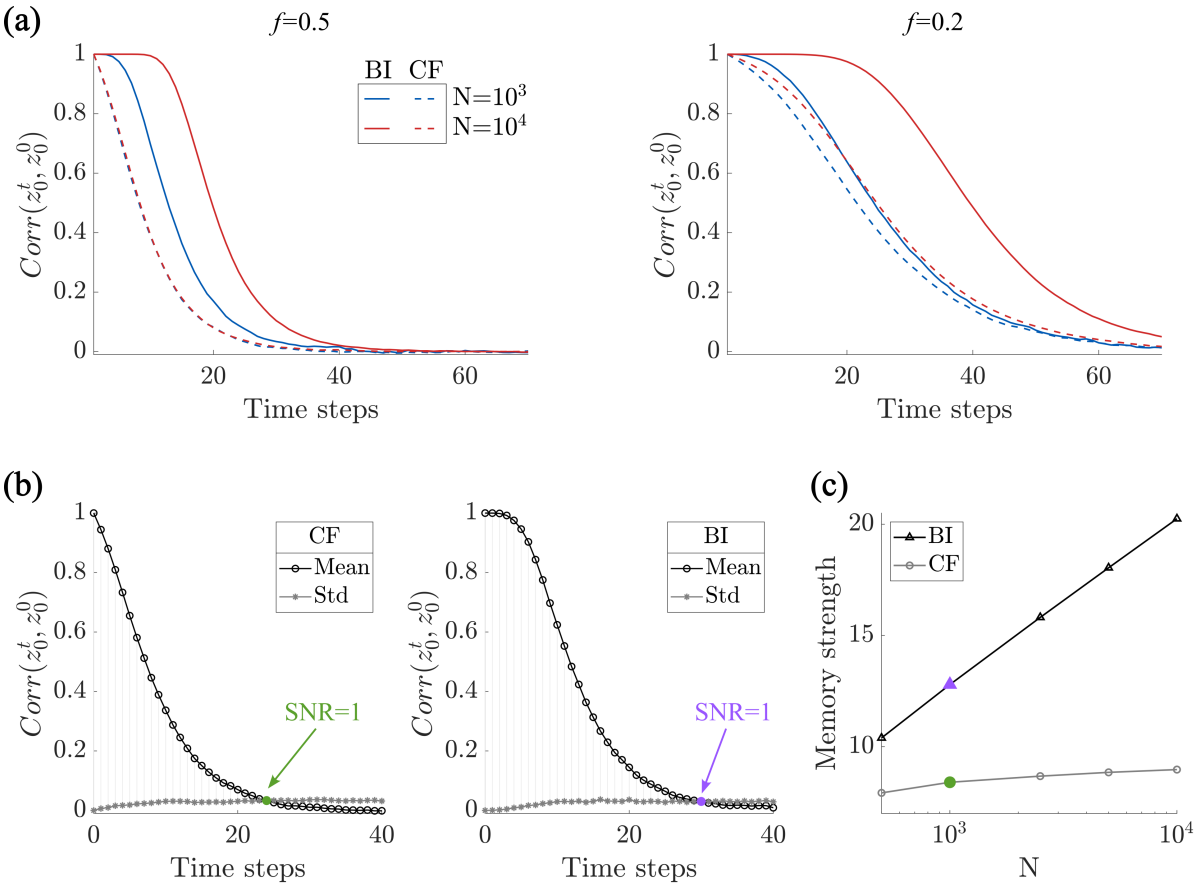

Fig. 7: Locally balanced inhibition enhances memory strength. **(a)** The autocorrelation of an elicited output decays over time due to the overwriting from ongoing learning. Such decay is slower for larger N and for the locally balanced inhibition case (BI), compared to classical framework (CF). **(b)** Mean and standard deviation of $corr(z_0^t, z_0^0)$ for CF and BI. The color dot identifies the first value of t such that the signal-to-noise ratio (SNR) equals 1. **(c)** Memory strength increases logarithmically with the system size, and it does it faster for BI, compared to CF. – Each curve shows the mean over 100 simulations. Parameters: $\gamma = f$, $f = 0.5$, $p = 0.2$.

we found that the memory strength scaled logarithmically with N, see Fig. 7c. Moreover, BI increases the memory strength compared to CF.

Taken together, these results show that locally balanced inhibition not only preserves representational diversity but also improves retrieval. This yields to higher memory-strength metric, thereby extending the effective capacity of Hebbian feedforward networks under continual learning.

In summary, locally balanced inhibition serves as an effective mechanism to prevent network rigidity induced by Hebbian plasticity in feedforward architectures. By normalizing synaptic drive at the level of individual neurons, it preserves the diversity of output patterns, and enables the network to store a greater number of input-output associations, while maintaining learning flexibility across time. This highlights the critical role of balanced excitation and inhibition in sustaining functional plasticity within learning systems.

3 Discussion and conclusion

In this manuscript, we studied Hebbian heteroassociative networks with stochastic synapses. We first showed how continual learning provokes interference and turnover of synaptic structure, which leads to a decrease in the correlation of

driven output patterns over time. We then demonstrated that when the postsynaptic output is driven mainly by a targeted input pathway, i.e., the contribution of additional pathways is negligible, the network enters a pathological freezing regime in which successive readouts become highly correlated. However, even when the contribution of additional inputs is increased, thereby weakening correlations between successive readouts, the targeted pathway still induces output–input correlations, which in turn degrades network performance. To eliminate input-output correlations, we proposed a locally balanced inhibition mechanism, whereby each neuron receives inhibitory input proportional to its own excitatory in-degree. When the inhibitory gain is matched to the network's activity level, this local balancing recenters the postsynaptic drive. This removes the bias toward repeatedly recruiting the same highly connected neurons and, in turn, decorrelates successive readouts, restoring flexibility in the system. Moreover, in this balanced regime, we observe a clear increase in memory strength.

Classical Hebbian formulations (Hebb (1949); Shatz (1992); Trappenberg (2002)) often assume that the outputs are imposed a priori. By explicitly mixing targeted input drive with additional inputs and varying their relative strength, we bridge imposed-output regimes and biologically motivated input-driven regimes in the same model. Within this continuum, locally balanced inhibition emerges as the minimal circuit-level correction that preserves input-driven readouts while avoiding rigidity, aligning with theoretical (Vreeswijk and Sompolinsky (1996); Murphy and Miller (2009); Renart et al. (2010); Koch et al. (2010); Vogels et al. (2011); Rubin et al. (2015)), and in-vivo (Shu et al. (2003); Haider et al. (2006); Okun and Lampl (2008); Murphy and Miller (2009); Ozeki et al. (2009); Dehghani et al. (2016)) evidence for tight excitation/inhibition balance and heterosynaptic normalization as stabilizers of plasticity.

As mentioned, in this manuscript, we introduced a biologically motivated input-driven regimes. This necessarily introduces correlations between synaptic inputs and postsynaptic outputs. To the best of our knowledge, prior work on the impact of correlations on Hebbian learning has focused on correlations among input patterns and/or among output patterns, not on correlations between inputs and outputs. In (Clopath et al. (2012)), the authors assumed imposed input–output patterns with hand-tuned correlations. In (Földiák (1990); Oja (1982)), to eliminate the correlation between the outputs, it has been proposed to introduce an anti-Hebbian rule between output units. Other works focused on counteracting Hebbian runaway dynamics by distributing changes and maintaining overall balance (Chistiakova et al. (2015); Bannon et al. (2020)). Our work here goes further: in our model, output correlations emerge naturally from uncorrelated inputs, as outputs are driven rather than specified a-priori.

Our results connect directly to representational drift, i.e., the gradual reconfiguration of neural codes despite stable behavior (Ziv et al. (2013); Rubin et al. (2015); Khatib et al. (2023); Geva et al. (2023); Devalle et al. (2025); Schoonover et al. (2021)). In the classical Hebbian regime (with weak output–input coupling), continual learning produces turnover of synaptic structure and a slow decay of recall, echoing drift-like dynamics. By contrast, strong coupling collapses drift into freezing: the code no longer reconfigures but instead converges to a stereotyped pattern, impairing memory. Locally balanced inhibition prevents freezing while preserving ongoing variability driven by current inputs, supporting flexible, but not runaway, reorganization of representations. This suggests a circuit-level account of how cortex can exhibit drift without loss of function: balanced E/I suppresses correlation-driven positive feedback, maintaining diversity of readouts and extending the regime where behavior remains stable as codes evolve.

Although our model is a strictly feedforward network with binary states, we expect that using real-valued activities would yield qualitatively similar results. On the other hand, future work should examine how alternative plasticity rules shape the dynamics. In this regard, the recent work of Confavreux et al. (2025), which introduces an automated search algorithm to generate thousands of diverse plasticity-rule quadruplets, offers a useful starting point.

All together, by introducing a mechanism of locally balanced inhibition, we demonstrate a simple yet powerful solution that restores flexibility and preserves memory capacity. These findings

emphasize the importance of inhibitory regulation in maintaining the balance between stability and adaptability in neural learning systems.

4 Acknowledgement

This project has received funding from Proyectos De Generación De Conocimiento 2021 (PID2021-124702OB-I00) and the Proyecto de Colaboración Internacional PCI2023-145967-2 from the Spanish Ministry of Science and Innovation. This work is supported by the Spanish State Research Agency, through the Severo Ochoa and Maria de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M). We thank CERCA Programme/Generalitat de Catalunya for institutional support.

References

Amit, D., Fusi, S.: Constraints on learning in dynamic synapses. Network: Computation in Neural Systems **3**, 443 (1992) https://doi.org/10.1088/0954-898X/3/4/008

Amit, D.J., Fusi, S.: Learning in neural networks with material synapses. Neural Computation **6**, 957–982 (1994) https://doi.org/10.1162/NECO.1994.6.5.957

Bannon, N.M., Chistiakova, M., Volgushev, M.: Synaptic plasticity in cortical inhibitory neurons: What mechanisms may help to balance synaptic weight changes? Frontiers in Cellular Neuroscience **14**, 543415 (2020) https://doi.org/10.3389/FNCEL.2020.00204/XML

Benna, M.K., Fusi, S.: Computational principles of synaptic memory consolidation. Nature Neuroscience **19**, 1697–1706 (2016) https://doi.org/10.1038/NN.4401;SUBJMETA=1595,2167,2638,378,631;KWRD=CONSOLIDATION,LONG-TERM+MEMORY

Chistiakova, M., Bannon, N.M., Chen, J.Y., Bazhenov, M., Volgushev, M.: Homeostatic role of heterosynaptic plasticity: Models and experiments. Frontiers in Computational Neuroscience **9**, 134920 (2015) https://doi.org/10.3389/FNCOM.2015.00089/XML

Confavreux, B., Harrington, Z.P.M., Kania, M., Ramesh, P., Krouglova, A.N., Bozelos, P.A., Macke, J.H., Saxe, A.M., Gonçalves, P.J., Vogels, T.P.: Memory by a thousand rules: Automated discovery of multi-type plasticity rules reveals variety & degeneracy at the heart of learning. bioRxiv (2025) https://doi.org/10.1101/2025.05.28.656584 https://www.biorxiv.org/content/early/2025/06/04/2025.05.28

Clopath, C., Nadal, J.P., Brunel, N.: Storage of correlated patterns in standard and bistable purkinje cell models. PLOS Computational Biology **8**, 1002448 (2012) https://doi.org/10.1371/JOURNAL.PCBI.1002448

Dehghani, N., Peyrache, A., Telenczuk, B., Quyen, M.L.V., Halgren, E., Cash, S.S., Hatsopoulos, N.G., Destexhe, A.: Dynamic balance of excitation and inhibition in human and monkey neocortex. Scientific Reports 2016 6:1 **6**, 1–12 (2016) https://doi.org/10.1038/srep23176

Devalle, F., Zou, L., Cecchini, G., Roxin, A.: Representational drift as the consequence of ongoing memory storage. Scientific Reports 2025 15:1 **15**, 1–20 (2025) https://doi.org/10.1038/s41598-025-11102-x

Fusi, S., Abbott, L.F.: Limits on the memory storage capacity of bounded synapses. Nature Neuroscience **10**, 485–493 (2007) https://doi.org/10.1038/NN1859;KWRD=BIOMEDICINE

Fusi, S., Zuckerman, M.B., Kahana, M., Wagner, A.: Memory capacity of neural network models. The Oxford Handbook of Human Memory: Foundations and Applications **1**, 740–764 (2021) https://doi.org/10.1093/oxfordhb/9780190917982.013.26

Földiák, P.: Forming sparse representations by local anti-hebbian learning. Biological Cybernetics 1990 64:2 **64**, 165–170 (1990) https://doi.org/10.1007/BF02331346

Geva, N., Deitch, D., Rubin, A., Ziv, Y.: Time and experience differentially affect distinct aspects of hippocampal representational drift. Neuron **111**, 2357–23665 (2023) https://doi.org/10.1016/J.NEURON.2023.05.005/ASSET/

DE552EC1-2E48-4B87-BCE5-94E6C7F91FF0/MAIN.ASSETS/GR4.JPG

Girko, V.L.: Circular law. Theory of Probability & Its Applications **29**, 694–706 (1985) https://doi.org/10.1137/1129095

Haider, B., Duque, A., Hasenstaub, A.R., McCormick, D.A.: Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. Journal of Neuroscience **26**(17), 4535–4545 (2006) https://doi.org/10.1523/JNEUROSCI.5297-05.2006 https://www.jneurosci.org/content/26/17/4535.full.pdf

Hebb, D.O.: The Organization of Behavior. A Neuropsychological Theory, (1949). https://www.scirp.org/(S(ny23rubfvg45z345vbrepxrl))/reference/referencespapers?referenceid=1293339

Koch, I., Gade, M., Schuch, S., Philipp, A.M.: The role of inhibition in task switching: A review. Psychonomic Bulletin & Review 2010 17:1 **17**, 1–14 (2010) https://doi.org/10.3758/PBR.17.1.1

Khatib, D., Ratzon, A., Sellevoll, M., Barak, O., Morris, G., Derdikman, D.: Active experience, not time, determines within-day representational drift in dorsal ca1. Neuron **111**, 2348–23565 (2023) https://doi.org/10.1016/j.neuron.2023.05.014

Murphy, B.K., Miller, K.D.: Balanced amplification: A new mechanism of selective amplification of neural activity patterns. Neuron **61**, 635–648 (2009) https://doi.org/10.1016/J.NEURON.2009.02.005/ATTACHMENT/61EAFF01-14EF-48D6-86A8-83CD8E4B631A/MMC1.PDF

Ozeki, H., Finn, I.M., Schaffer, E.S., Miller, K.D., Ferster, D.: Inhibitory stabilization of the cortical network underlies visual surround suppression. Neuron **62**, 578–592 (2009) https://doi.org/10.1016/j.neuron.2009.03.028

Oja, E.: Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology **15**, 267–273 (1982) https://doi.org/10.1007/BF00275687/METRICS

Okun, M., Lampl, I.: Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nature Neuroscience **11**, 535–537 (2008) https://doi.org/10.1038/NN.2105;KWRD=BIOMEDICINE

O'Rourke, S., Renfrew, D.: Low rank perturbations of large elliptic random matrices. Electronic Journal of Probability **19** (2013) https://doi.org/10.1214/EJP.v19-3057

Rubin, A., Geva, N., Sheintuch, L., Ziv, Y.: Hippocampal ensemble dynamics timestamp events in long-term memory. eLife **4** (2015) https://doi.org/10.7554/ELIFE.12247

Renart, A., Rocha, J.D.L., Bartho, P., Hollender, L., Parga, N., Reyes, A., Harris, K.D.: The asynchronous state in cortical circuits. Science **327**, 587–590 (2010) https://doi.org/10.1126/SCIENCE.1179850

Rubin, D.B., VanHooser, S.D., Miller, K.D.: The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. Neuron **85**, 402–417 (2015) https://doi.org/10.1016/j.neuroN.2014.12.026

Shatz, C.J.: The developing brain. Scientific American **267**, 61–67 (1992) https://doi.org/10.1038/SCIENTIFICAMERICAN0992-60,

Shu, Y., Hasenstaub, A., McCormick, D.A.: Turning on and off recurrent balanced cortical activity. Nature **423**, 288–293 (2003) https://doi.org/10.1038/NATURE01616;KWRD=SCIENCE

Schoonover, C.E., Ohashi, S.N., Axel, R., Fink, A.J.P.: Representational drift in primary olfactory cortex. Nature 2021 594:7864 **594**, 541–546 (2021) https://doi.org/10.1038/s41586-021-03628-7

Trappenberg, T.P.: Fundamentals of Computational Neuroscience (2002). https://philpapers.org/rec/PTRFOC

Vreeswijk, C.V., Sompolinsky, H.: Chaos in neuronal networks with balanced excitatory and

inhibitory activity. Science **274**, 1724–1726 (1996) https://doi.org/10.1126/SCIENCE.274.5293.1724;ISSUE:ISSUE:DOI

Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C., Gerstner, W.: Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science **334**, 1569–1573 (2011) https://doi.org/10.1126/SCIENCE.1211095

Ziv, Y., Burns, L.D., Cocker, E.D., Hamel, E.O., Ghosh, K.K., Kitch, L.J., Gamal, A.E., Schnitzer, M.J.: Long-term dynamics of ca1 hippocampal place codes. Nature Neuroscience **16**, 264–266 (2013) https://doi.org/10.1038/NN.3329;SUBJMETA=1554,1595,1697,378,631;KWRD=HIPPOCAMPUS,NEURONAL+PHYSIOLOGY