# Cooperative coding of continuous variables in networks with sparsity constraint

Paul Züge[†], Raoul-Martin Memmesheimer[*]

Institute for Genetics, University of Bonn, Germany

* rm.memmesheimer@uni-bonn.de, † pzuege@uni-bonn.de

## Abstract

A hallmark of biological and artificial neural networks is that neurons tile the range of continuous sensory inputs and intrinsic variables with overlapping responses. It is characteristic for the underlying recurrent connectivity in the cortex that neurons with similar tuning predominantly excite each other. The reason for such an architecture is not clear. Using an analytically tractable model, we show that it can naturally arise from a cooperative coding scheme. In this scheme neurons with similar responses specifically support each other by sharing their computations to obtain the desired population code. This sharing allows each neuron to effectively respond to a broad variety of inputs, while only receiving few feedforward and recurrent connections. Few strong, specific recurrent connections then replace many feedforward and less specific recurrent connections, such that the resulting connectivity optimizes the number of required synapses. This suggests that the number of required synapses may be a crucial constraining factor in biological neural networks. Synaptic savings increase with the dimensionality of the encoded variables. We find a trade-off between saving synapses and response speed. The response speed improves by orders of magnitude when utilizing the window of opportunity between excitatory and delayed inhibitory currents that arises if, as found in experiments, spike frequency adaptation is present or strong recurrent excitation is balanced by strong, shortly-lagged inhibition.

## Author summary

Neurons represent continuous sensory or intrinsic variables in their joint activity, with rather broad and overlapping individual response profiles. In particular there are often many neurons with highly similar tuning. In the cortex, these neurons predominantly excite each other. We provide a new explanation for this type of recurrent excitation, showing that it can arise in a novel cooperative coding scheme that minimizes the number of required synapses. This suggests the number of required synapses as a crucial constraining factor in biological neural networks. In our cooperative coding scheme, neurons use few strong and specific excitatory connections to share their computations with those neurons that also need it. This way, neurons can generate a large part of their response by leveraging inputs from neurons with similar responses. This allows to replace many feedforward and less specific recurrent connections by few specific recurrent connections. We find a trade-off between saving synapses and response speed. Theoretical estimates and numerical simulations show that specific features of biological single neurons and neural networks can drastically increase the response speed, improving the trade-off.

# Introduction

The brain encodes continuous sensory or intrinsic variables in the coordinated activity of populations of neurons. The tuning curves (response profiles) of individual neurons in such populations are rather broad, leading to large overlaps between them [1,2]. Further, there are often many neurons with highly similar tuning. Neuron populations with such features include simple cells in the primary visual cortex (V1) [3,4], head direction cells in the anterior thalamic nucleus [5], tactile neurons in primary somatosensory cortex [6], place cells in the hippocampus [7] and grid cells in the medial entorhinal cortex [8,9]. In machine learning, convolutional networks have overlapping receptive fields (RFs) that tile the input space [10]. RFs similar to those in visual cortex emerge by learning a sparse code for natural images [11], and RFs similar to grid cells emerge through training on navigation tasks [12,13].

Neurobiological data show that neurons with strongly overlapping receptive fields are predominantly excitatorily coupled: Synaptic connections between similarly-tuned excitatory principal neurons are more likely [14], stronger and more often bidirectional [15,16]. In line with this, the strongest incoming synapses provide excitation that matches a neuron's RF [16,17]. Furthermore, highly similarly tuned principal neurons have overall, i.e. including indirect, polysynaptic connections, a net excitatory effect on each other [18,19]. In contrast, if the tuning is barely similar or dissimilar, the net effect is inhibitory.

Such recurrent excitatory connectivity may seem unintuitive from a normative standpoint, as it amplifies noise [20] and can increase response times [21,22]. Previous studies suggested that it may support persistent activity and thus working memory [23,24] and that it may implement complicated priors [25].

Neural networks, however, evolved subject to physiological and physical constraints [26–29], including metabolic cost and available space. Optimizing for specific features can largely determine the neural network and lead to solutions that are in other aspects sub-optimal. A prominent example for this is a recent version of the efficient coding hypothesis [30–34]. It posits that neural networks greedily minimize the number of used spikes or the rate activity, which contribute to metabolic cost. The network connectivity obtained from the optimization is, however, very dense, which is not found in experiments. Further, the coding scheme is "competitive", in the sense that similarly tuned neurons compete for the opportunity to generate spikes. In other words, such neurons take away spikes and activity from each other. This predicts inhibitory couplings between very similarly tuned neurons, contrary to the experimentally observed physiological and effective excitatory interconnectivity between them.

Here, we explore the implications of "cooperative coding" in a neural network. In this newly proposed scheme, neurons avoid replicating computations through feedforward weights whose results are already accessible from the activity of other feature neurons. Instead, each feature neuron performs only a non-redundant feedforward computation. It then achieves the required response by additionally incorporating the results already obtained by similarly tuned feature neurons through recurrent connections. In other words, feature neurons do not independently replicate shared parts of the computations through feedforward weights, but they transmit them through recurrent connections to each other. The resulting connectivity is like-to-like, i.e. strong and effectively excitatory between similarly tuned principle neurons, as observed in experiments. Interestingly the scheme optimizes the number of synapses in a network, while maintaining the required neural network dynamics. Such an optimization differs from the common focus on saving spikes and may be imposed by space restrictions or cost of maintaining synapses [28,35].

# Results

To demonstrate the concept of cooperative coding, we consider a layer of feature neurons (output neurons), which receive feedforward input from an input layer as well as recurrent input. The task of the feature neurons is to generate a weighted sum of the inputs with weight strengths that decay exponentially with the distance of an input from the preferred input. We assume that the functionally relevant network response, representing the desired features (outputs), is the steady state activity. The desired outputs are linear functions of the inputs. Neural responses can hence be characterized by linear RFs and implemented by feedforward connectivity alone. Importantly, they can also be implemented using mixtures of feedforward and recurrent input.

We will compare the different network implementations in terms of the space requirement, approximated by the number of required synapses, and in terms of the metabolic cost to keep up the stationary state. Finally we will compare the response times and demonstrate how they can be substantially decreased in networks with spike frequency adaptation (SFA) or balancing inhibition.

## Models

In the following, we analyze three concrete examples of cooperative coding: (i) encoding a one-dimensional stimulus, (ii) simultaneously encoding two one-dimensional stimuli with linear mixed selectivity (MS) and (iii) encoding a two-dimensional stimulus. For ease of description, we focus on translationally invariant RFs. (Approximate) translational invariance, meaning that offset RFs have similar shapes, is a common characteristic of experimentally encountered RFs [2, 4, 5, 8, 9]. Further, it is a common characteristic of RFs that emerge in machine learning [10, 11]. Although the RFs that we consider do not have the precise shape of measured RFs, for example those of simple cells in V1 [36], they share the key properties of localized, overlapping and broadening RFs that tile the represented space. Indeed, RFs of neurons in hierarchically higher layers are often broader and constructed from those in lower layers [37, 38].

## Encoding a 1D stimulus

As a concrete, analytically tractable model that illustrates how cooperative coding works and can save synapses, we consider RFs that tile the one-dimensional parameter space of a stimulus (see Fig. 1a)). An input neuron $j$, $j = 1, ..., N$ signals the presence and strength of a stimulus with a specific parameter $j$ by nonzero activity $r_j > 0$. The task of the feature layer is to generate a response that is maximal at the preferred stimulus parameter and then decays exponentially the more different the stimulus becomes from the preferred one. This behavior is qualitatively similar to commonly observed tuning curves such as orientation tuning curves or place fields. We further assume that if multiple stimuli are present, the feature layer responses to their different parameters superpose linearly.

As an example, the inputs may be interpreted as a simple model for the representation of the orientation of a bar in the early visual system. $r_j > 0$ then means that the orientation is within the $j$th bin of the total orientation range $[0, 180°]$. The transformation from input to features in our model describes the combination of responses from hierarchically lower visual areas to hierarchically higher ones [39]. As another example, the neurons may model the activity of place cells on a periodic, closed track. The transformation then models the transformation from input neurons with smaller place fields to neurons with larger place fields. Such a transformation may take place from the hippocampal dentate gyrus to the downstream area CA3 [40]. In our model, the input generates a

simple encoding of the current location, where input neuron $j$ is active if the animal is in the $j$th location.

The desired stationary feature layer activity can be expressed as

$$x_i^{\text{resp}} = \sum_{j=1}^{N} \text{RF}_{ij} r_j, \qquad \text{RF}_{ij} = e^{-\frac{|i-j|}{d}} = \gamma^{|i-j|}. \tag{1}$$ {{eq:exp_RF_target}}

Here $r_j$ is the activity of the $j$th input neuron, $\gamma = \exp(-1/d)$ and $d$ defines the width

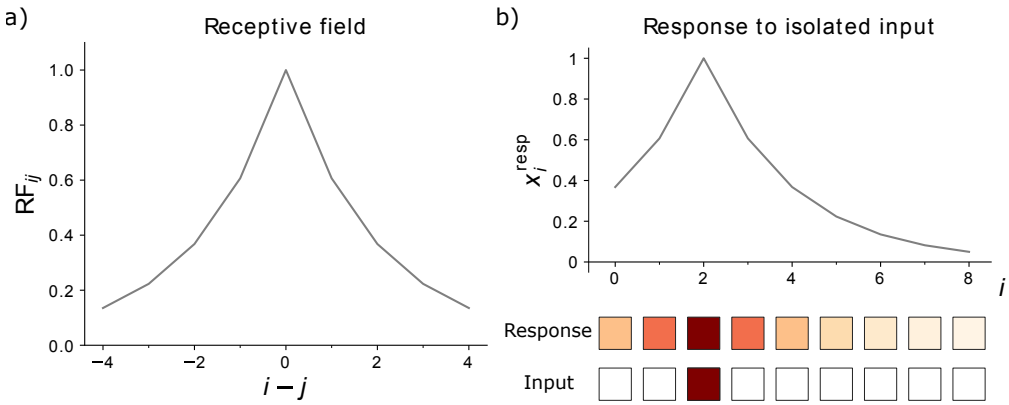


Fig 1: **Receptive field and response of 1D network. a)** The RF of neuron $i$ is peaked at $j = i$ and decays exponentially with $|i - j|$. The receptive field width parameter is $d = 2$. **b)** Top: The network response to an isolated unit input, here located at $j_0 = 2$, has the same shape and amplitude as a neuron's RF. It peaks at $i = j_0$. Bottom: Input and feature neurons are shown with color-coded activities $r_j$ and $x_i$, respectively. Increasingly dark red color represents higher activity; white squares indicate inactive neurons. {fig:1_RF_and_response_1D}

of the RF. We use periodic boundary conditions. For computations with neuron indices, this means that $|i - j|$ means $\min_{n\in\{-1,0,1\}} |i - j + nN|$. There are as many input as feature neurons. We note that, because of the symmetry $\text{RF}_{ij} = \text{RF}_{ji}$, the vector $\text{RF}_{k\cdot}$, describing the RF of feature neuron $k$, is the same as $\text{RF}_{\cdot k}$, the network response when only input neuron $k$ is active, compare Fig. 1a) and b). Summarized in a formula, we have $x_k^{\text{resp}}|_{r_j=\delta_{jl}} = \text{RF}_{kl} = \text{RF}_{lk} = x_l^{\text{resp}}|_{r_j=\delta_{jk}}$, where $k$ is fixed and $l$ variable.

### Feedforward implementation

To model the temporal dynamics of the neurons, we choose a standard simple linear rate network model [41, 42]. The purely feedforward network that generates the response eq. (1) as stationary state is then given by

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j(t), \tag{2}$$ {{eq:eom_1D_FF}}

with $W_{ij}^{\text{ff}} = \text{RF}_{ij}$ and a time constant $\tau$. In the stationary state, we have $\dot{x}_i = 0$ for all $i$ and thus

$$x_i = \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j = \sum_{j=1}^{N} \text{RF}_{ij} r_j = x_i^{\text{resp}}. \tag{3}$$ {{eq:stat response}}

This state is asymptotically stable and globally attracting; the flow is a contraction to it. These properties follow immediately from the fact that the system is linear and has a

unique fixed point, which is asymptotically stable because all eigenvalues of the matrix specifying the homogeneous differential equation are negative, equal to $-1/\tau$ [43, 44]. To approximate the network with a characteristic number of feedforward weights that is smaller than $N$, we require synapses $W^{\mathrm{ff}}_{ij} = \gamma^{|i-j|}$ only where $|i-j| \leq d$. This defines the RF size $n_{\mathrm{RF}} = 2d+1$ as the number of feedforward synapses per neuron needed to implement the RF within a distance $d$ around its center.

### Cooperative implementation

The same stationary neuronal responses can be obtained as the steady state of a recurrent network that uses cooperative coding. It requires only three synapses per feature neuron, two recurrent and one feedforward one. This network's dynamics are given by

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W^{\mathrm{rec}}_{ij} x_j(t) + \sum_{j=1}^{N} W^{\mathrm{ff}}_{ij} r_j(t) \tag{4}$$ {{eq:eom_1D_general}}

$$= -x_i(t) + w^{\mathrm{rec}}(x_{i+1}(t) + x_{i-1}(t)) + w^{\mathrm{ff}} r_i(t), \tag{5}$$ {{eq:eom_1D}}

with weights $w^{\mathrm{rec}} = \frac{1}{\gamma+\gamma^{-1}} = \frac{\gamma}{1+\gamma^2}$ and $w^{\mathrm{ff}} = 1 - 2\gamma w^{\mathrm{rec}} = \frac{1-\gamma^2}{1+\gamma^2}$. If the receptive fields are not narrow ($d$ is not small against 1), the two recurrent connections are strong, in the sense that $w^{\mathrm{rec}}$ is not small against 1. Thus the network features strong like-to-like excitation and is driven by feedforward input. One can straightforwardly verify that $x_i = x_i^{\mathrm{resp}}$ is indeed a stationary state of the network, by inserting eq. (1) into eq. (5), see appendix S1. The reason for this is ultimately that the desired response of a neuron $i$ can be largely generated by summing the responses of the two neurons $i \pm 1$ neighboring $i$, see eq. (9) and Fig. 2b). This is achieved by the recurrent connections. The missing part is contributed by the feedforward input. This state is asymptotically stable as all real parts of the eigenvalues of the matrix defining the homogeneous system are negative, see appendix S1. For broad receptive fields (where $\gamma \lesssim 1$), the recurrent connections are nearly as strong as possible: their sum $2w^{\mathrm{rec}}$ is close to 1, the value beyond which the network becomes unstable. The stationary state is also the only stationary state. Since the system is linear, the state is therefore a global attractor as for the feedforward network [43, 44]. Thus, for constant input the network forms this stable response pattern.

### Cooperative coding

Cooperative coding can be understood as sharing of the information that an individual neuron obtains from external input specifically with those neurons that also need it. This allows to generate most of the neuronal responses from sparse recurrent connectivity. Especially very similarly tuned neurons will project strongly excitatorily onto each other; oppositely tuned neurons would inhibit each other.

As a concrete example, we introduced the networks eq. (5), where it suffices that each neuron receives input from only one input and two feature neurons. Still, each neuron effectively responds to $\mathcal{O}(d)$ input neurons. This is possible because the feature neurons recurrently share their activity, and hence their access to feedforward input, with their neighbors. These in turn share it with their neighbors, thus propagating it through the network. The network response then forms dynamically through the interplay of feedforward input and recurrent interactions.

The coding relies on the fact that despite very few feedforward and recurrent synapses, poly-synaptic connectivity can still be far-reaching [45, 46]: For further clarification of

this point consider the approximate, discretized dynamics [47, 48]

$$x_i((n+1)\tau) \approx \sum_{j=1}^{N} W_{ij}^{\mathrm{rec}} x_j(n\tau) + \sum_{j=1}^{N} W_{ij}^{\mathrm{ff}} r_j \tag{6}$$ {{eq:eom discretized}}

that lead to the same steady state as the time-continuous dynamics eq. (4). The response to a constant input $r$ after $n$ time constants is

$$x(n\tau) = \Big(\mathbb{1} + W^{\mathrm{rec}} + \cdots + (W^{\mathrm{rec}})^{n-1}\Big) W^{\mathrm{ff}} r. \tag{7}$$

It is determined by $W^{\mathrm{rec}}$ and its higher powers, which reflect the redistribution of feedforward input via poly-synaptic recurrent pathways.

The coding scheme can also be understood as feedforward inputs providing a correction to the response that is mainly constructed from the sparse recurrent input. To clarify this we focus on elementary stationary responses, namely those that are driven by a single unit input from neuron $j$; the input activity is $r_k = \delta_{kj}$. Responses to more complicated input patterns are weighted linear sums of such elementary responses. Consider feature neuron $i$ and assume that all other neurons already respond correctly. The stationary activity of neuron $i$ in response to a single unit input from neuron $j$ is then $\mathrm{RF}_{ij}$, while those of the other network neurons $k$ is $\mathrm{RF}_{kj}$. Equation (4) with $\dot{x}_i = 0$ implies that $\mathrm{RF}_{ij}$ is the sum of the RFs of its presynaptic feature neurons and its feedforward connectivity,

$$\mathrm{RF}_{ij} = \sum_k W_{ik}^{\mathrm{rec}} \mathrm{RF}_{kj} + W_{ij}^{\mathrm{ff}}. \tag{8}$$ {{eq:add_rec_fields_general}}

For the specific network eq. (5) we have

$$\mathrm{RF}_{ij} = w^{\mathrm{rec}}(\mathrm{RF}_{i-1,j} + \mathrm{RF}_{i+1,j}) + w^{\mathrm{ff}} \delta_{ij}, \tag{9}$$ {{eq:add_rec_fields}}

illustrated in Fig. 2c). In this network, the weighted and summed responses of neuron $i$'s nearest neighbors are thus already very close to neuron $i$'s target response. This is enabled by the specific exponential shape of the RFs. Only for the preferred input of a feature neuron, the response is too low. The neuron corrects for this by recruiting the missing input through a feedforward connection. Such an "explanatory gap" that is left by the recurrent inputs and can be filled by external input is important for cooperative coding, because the output must depend on external input.

### Spatial demand and metabolic cost

To compare the efficiency of the introduced implementations, we focus on two cost dimensions: the space needed to implement the network and the metabolic cost of generating the stationary dynamics. As measure for the space needed for the network we take the number of synaptic connections, or, in other words, the L0 norm of the synaptic weight matrix. In the feedforward network eq. (2), it increases linearly with the width $d$ of the RF (if small responses can be neglected). In the recurrent network eq. (5) three synapses per neuron suffice to generate the desired stationary response regardless of the RF size. We show in appendix S1 that the recurrent network eq. (5) therefore minimizes the L0 norm.

For the metabolic cost of generating the stationary dynamics, we can take a multiple of the L1 norm of the synaptic currents (see Discussion). Since in both networks all modeled synaptic currents are excitatory, the L1 norm of synaptic currents equals the total synaptic current. In the stationary state this current is the same in both implementations, because neurons have the same stationary activity; therefore also the cost is the same. We conclude that the metabolic cost for maintaining the stationary network state is the same in both the feedforward and the recurrent network implementation.

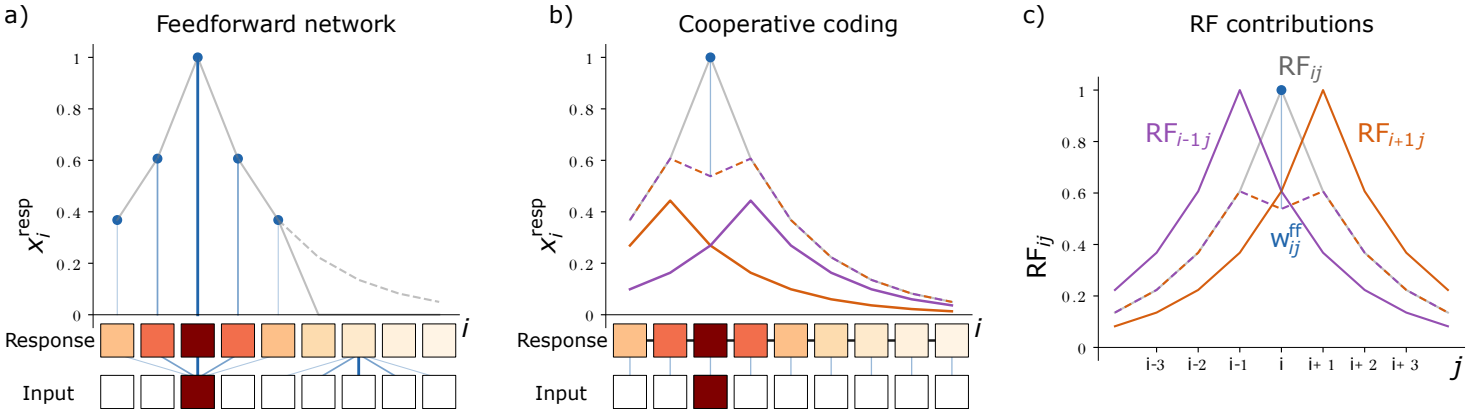


Fig 2: **Schematics of feedforward and cooperatively coding networks.** **a)** Top: In the feedfoward network, the response $x_i^{\text{resp}}$ (gray solid curve) to an isolated input is fully generated by the neurons' feedforward inputs (blue lines and dots). For the displayed RF width $d = 2$, five neurons receive feedforward input, so that the network response (gray solid curve) represents $\approx 63\%$ of the summed target response (gray dashed curve). Bottom: Feature and input neuron activities as in Fig. 1. Outgoing feedforward synapses from the active input neuron $j = 2$ and incoming feedforward synapses to feature neuron $i = 6$ are shown in blue. **b)** Top: In the cooperatively coding network model, the network response (gray solid curve) is the sum of feedforward input (blue line and dot) and recurrent input (brown-purple dashed curve). For the displayed case of an isolated input, only one neuron receives feedforward input, which induces a part of the stationary response of the most active feature neuron. The rest of the response and all other responses are induced by recurrent input from neighboring neurons. The total recurrent input that each feature neuron receives is the sum of recurrent input from the right (brown solid curve) and left neighbor (purple solid curve). Bottom: Each feature neuron receives one feedforward synapse (blue lines) and two recurrent synapses (black lines, all recurrent connections are bidirectional). **c)** The RF of feature neuron $i$ ($\text{RF}_{ij}$ for varying $j$, gray solid curve) is the weighted sum (brown-purple dashed curve) of the RFs of its left ($\text{RF}_{i-1\,j}$, purple) and right neighbors ($\text{RF}_{i+1\,j}$, brown) plus a contribution from feedforward input ($W_{ij}^{\text{ff}}$, blue line and dot). All shown RFs have width $d = 2$. {fig:2_schematics_FF_CoopCoding}

## Response speed

In the feedforward network, activity converges with the intrinsic time constant $\tau$, which we define as its response time. In the cooperatively coding network, the excitatory recurrent connectivity increases the response time: Fig. 3 shows the dynamics of the response formation. The L1-norm of the deviation of the response from the steady state,

$$L(t) = |x(t) - x^{\text{steady}}|_1, \tag{10}$$ {{eq:L1_loss}}

which we use as a loss measure, decays exponentially with time constant

$$\tau_{\text{resp}} \equiv \frac{\tau}{1 - w_{\text{sum}}^{\text{rec}}}, \tag{11}$$ {{eq:tresp_net_1D_definition}}

(see Fig. 4), where $w_{\text{sum}}^{\text{rec}} = \sum_j W_{ij}^{\text{rec}}$ is the sum of recurrent weights arriving at (or, equivalently, originating from) a neuron. We define $\tau_{\text{resp}}$ as the response time of the networks. It scales inversely with the difference of the largest eigenvalue of the network, which is equal to $w_{\text{sum}}^{\text{rec}}$ (cf. appendix S1), from 1. In particular, it depends only on the summed recurrent weights. Equation (11) holds generally, for networks of the type eq. (4) with purely excitatory circulant recurrent weight matrix and convergent dynamics.

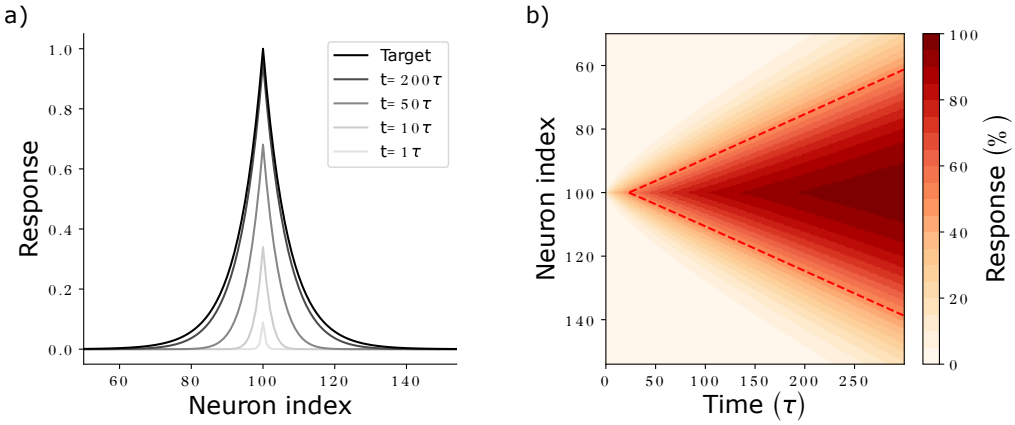


Fig 3: **Response formation and activity propagation. a)** Network activity at different times (shaded curves) after $r_{100}$ has been set from 0 to 1. For long times, network activity approaches the target response (black curve). **b)** Development of the activity of neurons (y axis) with time (x axis), measured relative to their target activities. The diagonal fronts of equal relative activities indicate propagation of activity with constant propagation speed. The time points at which neurons reach 50% of their final activity are connected by a red dashed line. Parameters: $w^{\text{rec}} = 0.5 \cdot 1/(1 - 1/100)$, such that $\tau_{\text{resp}} = 100\tau$ (see eq. (11)), $N = 200$ neurons. {fig:3_activity_propagation}

We now specialize the result to networks with nearest-neighbor coupling eq. (5) that generate the RFs eq. (1). Inserting $w^{\text{rec}} = 1/(\gamma + \gamma^{-1})$ and $\gamma = \exp(-1/d)$ relates the response time to the RF width. By approximating $\exp(\pm 1/d) \approx 1 \pm 1/d + (1/2)d^2$ for large $d$, we obtain $w^{\text{rec}} \approx 1/(2 + 1/d^2)$ and, inserting this into eq. (11),

$$\tau_{\text{resp}} \approx \frac{\tau}{1 - \frac{2}{2+\frac{1}{d^2}}} = (1 + 2d^2)\ \tau \approx 2d^2\tau \approx \frac{1}{2} n_{\text{RF}}^2 \tau. \tag{12}$$ {{eq:tresp_net_1D_vs_d}}

In the last part of the equation we used that $n_{\text{RF}} = 1 + 2d \approx 2d$ for large $d$. Equation (12) shows that wide RFs require long equilibration time. This is because they need strong recurrent weights with a largest eigenvalue close to 1. Further the equation reveals the trade-off between response time and number of employed synapses: The feedforward implementation eq. (2) needs $n_{\text{RF}}$ synapses and has a response time $\tau$. The recurrent implementation thus saves $n_{\text{RF}} - 3 \approx n_{\text{RF}}$ synapses per feature neuron. Equation (12) shows that the response time increases quadratically in the number of saved synapses, see also Fig. 4b).

The quadratic dependence of the response time on $d$ reflects that, as the RF becomes wider, not only does activity have to spread further, it also spreads more slowly: this is consistent with the idea that the settling of a neuron depends (indirectly) more on activity propagating back from more distant neurons that settle after it.

### Faster response with spike frequency adaptation

For activity to rapidly spread through the network, neurons need to be able to cause a large activity change in their neighbors within a short period of time. To achieve this, they need strong recurrent weights. However, recurrent weights are restricted to $w_{\text{sum}}^{\text{rec}} < 1$ to not cause runaway activity. We now show how spike frequency adaptation (SFA) can help ease this conflict and speed up network dynamics. SFA is typical for excitatory principal neurons and induces a reduction of their response to constant inputs in the long run [42, 49, 50].

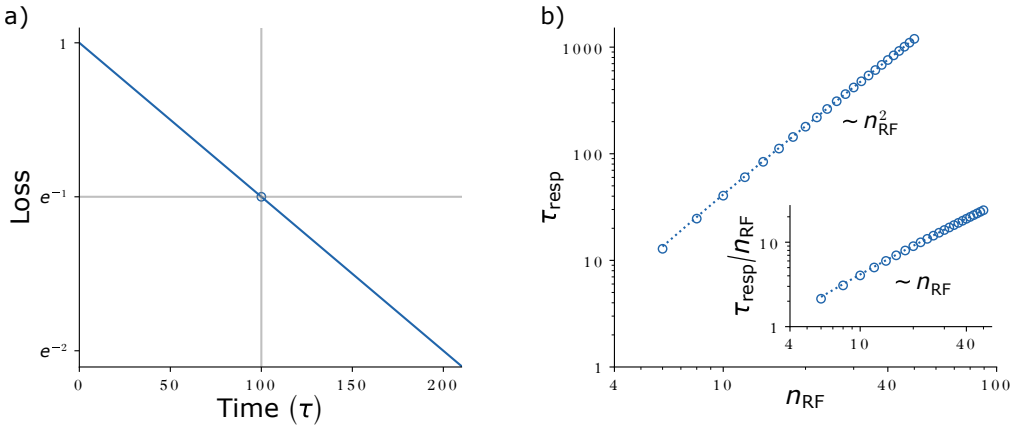


Fig 4: **Loss evolution and response speed - synapse number trade-off. a)** Exemplary loss evolution for a network with $w_{\text{sum}}^{\text{rec,net}} = 0.5\cdot(1-1/100)$ so that $\tau_{\text{resp}} = 100\tau$. Experimentally, $\tau_{\text{resp}}$ is determined as the time (gray vertical line) at which the loss drops to $\text{e}^{-1}$ (gray horizontal line, blue open circle). **b)** Response times $\tau_{\text{resp}}$ (circles: simulation results; dotted line: analytical solution eq. (12)) for target RFs of different widths $n_{\text{RF}}$. Data was created by scanning $n_{\text{RF}}$, setting $w_{\text{sum}}^{\text{rec,net}}$ to yield a RF of size $n_{\text{RF}}$ and determining $\tau_{\text{resp}}$ from the loss dynamics. {fig:4_tresp_1D}

We model SFA through a negative-feedback adaptation current $u(t)$, which is triggered by neuronal activity $x(t)$ and characterized by its scale $a_{\text{SFA}}$ and time constant $\tau_{\text{SFA}}$,

$$\tau\dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec}} x_j(t) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j(t) - a_{\text{SFA}} u_i(t), \tag{13}$$ {{eq:eom_SFA_dxdt}}

$$\tau_{\text{SFA}}\dot{u}_i(t) = -u_i(t) + x_i(t). \tag{14}$$ {{eq:eom_SFA_dudt}}

[51–53]. Setting $\dot{x}_i(t) = 0$ and $\dot{u}_i(t) = 0$ yields the steady state. We see immediately that it implies $u_i = x_i$. Inserting this into eq. (13) shows that in the stationary state the spike frequency adaptation results in a stronger leak current, $-(1 + a_{\text{SFA}})x_i$. Dividing by $1 + a_{\text{SFA}}$ yields

$$0 = -x_i + \sum_{j=1}^{N} \frac{W_{ij}^{\text{rec}}}{1 + a_{\text{SFA}}} x_j + \sum_{j=1}^{N} \frac{W_{ij}^{\text{ff}}}{1 + a_{\text{SFA}}} r_j. \tag{15}$$ {{eq:eom_SFA_dxdt_steady state}}

Consequently, in order to implement the same response as a network without SFA ($a_{\text{SFA}} = 0$, cf. eq. (4)), the recurrent and feedforward weights have to be scaled up by a factor of $1 + a_{\text{SFA}}$. The additional excitatory synaptic input compensates in the steady state the added inhibitory adaptation current.

To understand the network dynamics, it is instructive to consider the limit $\tau_{\text{SFA}} \to 0$ where $u_i(t) \to x_i(t)$ as in the steady state. Inserting this into eq. (13) and again dividing by $1+a_{\text{SFA}}$ yields an equation equivalent to eq. (4) with shortened neuronal time constant and increased weights,

$$\frac{\tau}{1 + a_{\text{SFA}}}\dot{x}_i(t) \overset{\tau_{\text{SFA}}\to 0}{=} -x_i(t) + \sum_{j=1}^{N} \frac{W_{ij}^{\text{rec}}}{1 + a_{\text{SFA}}} x_j(t) + \sum_{j=1}^{N} \frac{W_{ij}^{\text{ff}}}{1 + a_{\text{SFA}}} r_j(t). \tag{16}$$ {{eq:eom_SFA_dxdt_small_tSFA_limit}}

We see that a network with arbitrarily fast SFA and appropriately scaled weights has the same dynamics as a network without SFA, but with its time constant reduced by $1 + a_{\text{SFA}}$. This factor only depends on $a_{\text{SFA}}$ and is independent of the RF width that

the network implements. We might thus expect that introducing SFA with a given $a_{\mathrm{SFA}}$ and small $\tau_{\mathrm{SFA}}$ causes a constant speedup, but still results in a quadratic dependence of $\tau_{\mathrm{resp}}^{\mathrm{SFA}}$ on $n_{\mathrm{RF}}$ (see eq. (12)).

There is, however, an additional possibility: SFA might yield faster dynamics for finite, nonzero $\tau_{\mathrm{SFA}}$. This is because then $u_i(t)$ lags behind $x_i(t)$, which creates a temporal 'window of opportunity'. Within this window, the up-scaled weights can mediate strong interactions that are not yet cancelled by the retarded adaptation currents of the receiving neurons. In our networks, this leads to the following concept to exploit SFA: During the initial response phase, strong weights should cause a fast response while SFA keeps the steady state before and after an input change at the desired activity values as well as dynamically stable. In particular, the modified recurrent synaptic weights may then be (and to optimally exploit SFA: should be) so strong that without the SFA current the network dynamics are unstable.

This second possibility applies to our networks: Measuring the response time as a function of $\tau_{\mathrm{SFA}}$, we observe that it first decreases when increasing $\tau_{\mathrm{SFA}}$ from zero and reaches a minimum at a nonzero, optimal value of the SFA time scale (appendix S4). Increasing $\tau_{\mathrm{SFA}}$ further eventually causes diverging activity, because the retarded adaptation current $u(t)$ becomes so slow that it never compensates the stronger input due to the up-scaled weights. We note that also when keeping $\tau_{\mathrm{SFA}}$ at a fixed value, there is an optimal nonzero value of the inhibitory feedback strength $a_{\mathrm{SFA}}$. Importantly, we find that introducing SFA with finite $\tau_{\mathrm{SFA}}$ and optimal $a_{\mathrm{SFA}}$ improves the scaling of $\tau_{\mathrm{resp}}^{\mathrm{SFA}}$ with $n_{\mathrm{RF}}$ from quadratic as without or with arbitrarily fast SFA to linear.

To incorporate SFA in a cooperatively coding network, we modify the weights in eq. (5) as described above and add the SFA current. For the neuron activities, this yields the dynamical equation

$$\begin{aligned}\tau \dot{x_i}(t) = -x_i(t) &+ (1+a_{\mathrm{SFA}})w^{\mathrm{rec}}(x_{i+1}(t)+x_{i-1}(t)) \\ &+ (1+a_{\mathrm{SFA}})w^{\mathrm{ff}}r_i(t) - a_{\mathrm{SFA}}u_i(t);\end{aligned} \tag{17}$$

{{eq:eom_1D_SFA}}

where the adaptation current obeys eq. (14).

Concerning the use of resources, SFA does not require additional synaptic connections, so the spatial demand of the cooperatively coding network is the same as in the original model eq. (5). The increased weights, however, lead to stronger synaptic currents. Together with the added adaptation currents, this increases the energetic cost of maintaining the stationary state.

## Balanced networks

The networks we studied so far had only excitatory synapses, while biological neural networks also have recurrent inhibition, which balances the excitation [32, 54]. These are likely required for a range of reasons, such as ensuring network stability and maintaining irregular spiking activity [55–59]. Given their existence, we here show how inhibition can be used to speed up the network response, in an architecture that still relies on few synapses.

The individual excitatory and inhibitory currents can be much larger than their sum and precisely temporally balanced with a lag smaller than the neuronal time constant [60, 61]. Further, many inhibitory neurons are rather sharply tuned [61–63], sometimes similarly sharply as excitatory ones.

To incorporate inhibition consistent with these experimental findings, we add inhibitory interneurons to the existing network of excitatory feature neurons, which represent principle neurons. The interneurons derive their tuning from the feature neurons via specific recurrent inputs. For concreteness we assume that there are as many

interneurons as feature neurons and that each interneuron follows the activity of one feature neuron with a small delay, $\tau_{\text{lag}}$. Equation (4) thus becomes

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,E}} x_j(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,I}} x_j^{\text{I}}(t) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j(t) \tag{18}$$

$$= -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,E}} x_j(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,I}} x_j(t-\tau_{\text{lag}}) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j(t), \tag{19}$$ {{eq:eom_balanced_general_0}}

where $x_i^{\text{I}}(t)$ is the inhibitory activity, which equals the delayed excitatory feature neuron activity $x_i(t-\tau_{\text{lag}})$. $W_{ij}^{\text{rec,I}} \leq 0$ is the coupling from inhibitory neuron $j$ to feature neuron $i$. We now introduce the state change of feature neuron $i$ between $t-\tau_{\text{lag}}$ and $t$,

$$\Delta x_i(t) = x_i(t) - x_i(t-\tau_{\text{lag}}), \tag{20}$$

as well as the the sum of recurrent excitation and inhibition

$$W_{ij}^{\text{rec,net}} = W_{ij}^{\text{rec,E}} + W_{ij}^{\text{rec,I}}, \tag{21}$$ {{eq:def_Wrecnet}}

which we call net weights. These definitions allow to rewrite eq. (19) as

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,net}} x_j(t) - \sum_{j=1}^{N} W_{ij}^{\text{rec,I}} \Delta x_j(t) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j(t). \tag{22}$$ {{eq:eom_balanced_general}}

We now insert the values of the cooperatively coding network eq. (5) into the equation and further assume that the interneurons inhibit and balance the same sets of neurons that their driving feature neurons excited, i.e. $W_{ij}^{\text{rec,E}} = w^{\text{rec,E}} \left(\delta_{i+1,j} + \delta_{i-1,j}\right)$ and $W_{ij}^{\text{rec,I}} = w^{\text{rec,I}} \left(\delta_{i+1,j} + \delta_{i-1,j}\right)$. Equation (19) then becomes

$$\begin{aligned} \tau \dot{x}_i(t) = -x_i(t) &+ w^{\text{rec,E}} \big(x_{i+1}(t) + x_{i-1}(t)\big) \\ &+ w^{\text{rec,I}} \big(x_{i+1}(t-\tau_{\text{lag}}) + x_{i-1}(t-\tau_{\text{lag}})\big) + w^{\text{ff}} r_i(t) \end{aligned} \tag{23}$$ {{eq:eom_balanced_0}}

and eq. (22)

$$\begin{aligned} \tau \dot{x}_i(t) = -x_i(t) &+ w^{\text{rec,net}} \big(x_{i+1}(t) + x_{i-1}(t)\big) \\ &- w^{\text{rec,I}} \big(\Delta x_{i+1}(t) + \Delta x_{i-1}(t)\big) + w^{\text{ff}} r_i(t). \end{aligned} \tag{24}$$ {{eq:eom_balanced}}

The residual inhibitory interaction, i.e. the inhibitory interaction that is not included in $w^{\text{rec,net}}$, depends on $\Delta x_{i\pm1}(t)$ and therefore only acts when there are activity changes during the preceding brief E-I lag. A total change $\delta x_j$ in the activity of neuron $j$ causes an integrated postsynaptic activity change in neuron $i$ of $-\frac{\tau_{\text{lag}}}{\tau} W_{ij}^{\text{rec,I}}\, \delta x_j$ (see appendix S5).

To connect eq. (24) to our previous, unbalanced network eq. (5), we set

$$w^{\text{rec,net}} = w^{\text{rec}}. \tag{25}$$ {{eq:wrecnet wrec}}

The dynamical equations then agree if $w^{\text{rec,I}} = 0$ or $\Delta x_i(t) = 0$. The latter is satisfied in the steady state. The steady state is thus independent of the amount of inhibition (given that excitation covaries with inhibition such that $w^{\text{rec,E}} + w^{\text{rec,I}} = w^{\text{rec}}$ as implied by eq. (25)) and it is the same as in eq. (5); its stability can, however, change. We may thus think of $w^{\text{rec,net}}$ as defining the RF, and of $w^{\text{rec,I}}$ as affecting the dynamics. During the build-up of the response strong excitation ramps up slightly before the balancing inhibition. The analytical solution of eq. (24) and our stability analysis (see next section and Fig. 9) show that throughout this window of opportunity excitation may be up to approximately $\tau/\tau_{\text{lag}}$ times larger than the net interaction without destabilizing the network. This strong interaction allows a much quicker propagation of activity, convergence to the steady state and decay of the loss function.

### Spatial demand and metabolic cost in the balanced network

Compared to the unbalanced network, the balanced network requires three additional synapses per principle feature neuron, one E-to-I and two I-to-E synapses, i.e. a total of six synapses. This is again independent of the RF width, such that for large RFs, the balanced, cooperatively coding network still saves synapses compared to the feedforward network. It requires additional space for the inhibitory neurons, which may, however, be needed for other purposes anyways.

Also the metabolic maintenance cost increases, since there are more neurons. Further, there is an increased metabolic cost to sustain the synaptic currents in the stationary state: In this state, large parts of the excitatory and inhibitory currents cancel to give rise to a net current that equals the one in the purely excitatory recurrent network, see eq. (21) and eq. (25). In the L1 norm of the synaptic currents, the excitatory currents and the absolute inhibitory currents, however, add up. The metabolic cost thus increases by the amount of excitatory and inhibitory currents that cancel each other.

### Response speed in the balanced network

Under some additional assumptions, the evolution of the L1-loss eq. (10) can be analytically approximated as the solution of a linear delay differential equation, see appendix S6 for details. The resulting dynamics are those of a damped oscillator, see Fig. 5a): For weak inhibition, they are 'overdamped' in the sense that they are well described by the sum of two exponentials with different decay rates. At a specific intermediate inhibitory strength, the two decay rates agree and we have 'critical damping'. For stronger inhibition the dynamics are 'underdamped' in the sense that the loss behaves as the absolute value of an oscillation with exponentially decaying amplitude. Overly strong inhibition causes divergence of the dynamics.

In the overdamped regime, the smaller decay rate is the relevant one, as it dominates the speed of the decay for longer times. The larger decay rate rather describes how quickly faster dynamics, that may be present due to the initial conditions, are suppressed and the dynamics converge to the slower mode. The smaller decay rate increases when the strength of inhibition approaches its critical value. The same holds for the single decay rate in the oscillatory regime. At the critical inhibitory strength the overall decay of the loss is thus fastest, see Fig. 5b). We find analytically that its decay time constant is approximately proportional to the geometric mean of the response time in absence of inhibition and the inhibitory delay,

$$\tau_{\text{resp}}^{\text{bal},c} \approx \sqrt{\frac{\tau_{\text{resp}}\tau_{\text{lag}}}{2}}, \tag{26}$$ {{eq:trespc_approx}}

(cf. eq. (88) in appendix S6); the superscript "$c$" indicates that the result holds for critical inhibitory strength.

Importantly, this implies that the scaling of the response time with the receptive field width and size improves compared to the purely excitatory network. This is because $\tau_{\text{resp}}^{\text{bal},c} \sim \sqrt{\tau_{\text{resp}}}$. Inserting eq. (12) into eq. (26) for the 1D network, we obtain

$$\tau_{\text{resp}}^{\text{bal},c} \approx \sqrt{\frac{(1+2d^2)\ \tau\tau_{\text{lag}}}{2}} \approx d\sqrt{\tau\tau_{\text{lag}}} \approx \frac{1}{2}n_{\text{RF}}\sqrt{\tau\tau_{\text{lag}}}, \tag{27}$$ {{eq:s_trespc_vs_d_approx}}

which is only linear in the RF width $d$ and size $n_{\text{RF}}$, instead of quadratic as in the case of $w_{\text{sum}}^{\text{rec,I}} = 0$, compare eq. (27) with eq. (12) and in Fig. 6a) the red and blue dotted curves. As a consequence also the speedup gained through the inhibition, $\tau_{\text{resp}}/\tau_{\text{resp}}^{\text{bal},c}$, increases for wider RFs. We finally note that the interactions mediated by $W^{\text{rec,I}}$ can also be thought of as implementing an excitatory transmission of activity changes: an

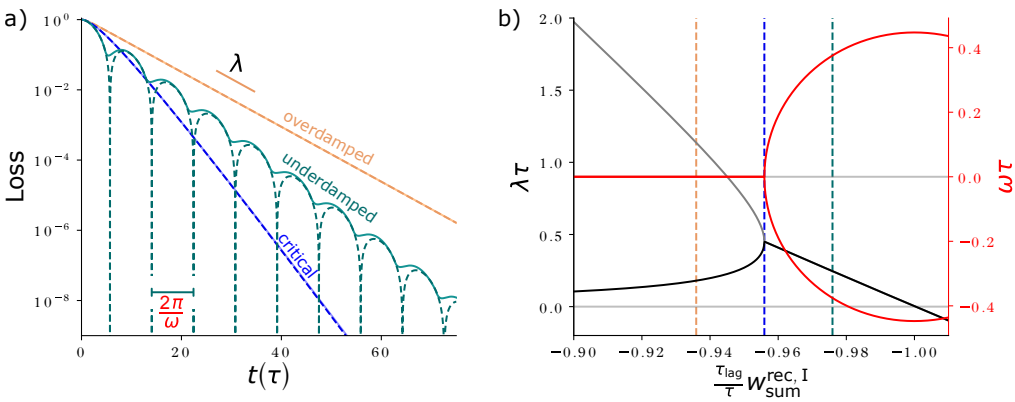


Fig 5: **Loss evolution for different inhibitory strengths. a)** Loss evolution (dashed: analytical approximation (cf. appendix S6, eqs. (64) and (74), partly occluded; solid: network simulation) for inhibitory strengths that are slightly weaker (orange), equal (blue) or slightly stronger (teal) than the critical strength, on a logarithmic scale. The slope of the decay is given by $\lambda$ (see b), explicitly highlighted for the overdamped dynamics. The oscillation period of the underdamped dynamics is $T_{\text{osci}} = 2\pi/\omega$. In case of oscillations, the analytic approximation briefly reaches zero loss once in a period (sharp dips in dashed curve). In the network simulation there is also a pronounced oscillation, but there always remains a finite error (solid, see appendix S6). **b)** Real part (decay rate $\lambda$, black/gray) and imaginary part (oscillation frequency $\omega$ times $\pm 1$, red) of the complex frequency of the exponential loss evolution, scaled by $\tau$. For weak inhibition there are two exponentially decaying modes ($\lambda$, black and gray curve). At the critical inhibitory strength (blue dashed vertical line) there is only a single decay rate and no oscillation. The decay rate (in the overdamped case: of the relevant slower-decaying mode) is maximized. For stronger inhibition, network activity begins to oscillate (nonzero $\omega$, red), and diverges for $\frac{\tau_{\text{lag}}}{\tau} w_{\text{sum}}^{\text{rec,I}} < -1$, where $\lambda$ becomes negative. Dashed vertical lines show the inhibitory strengths scaled by $\tau_{\text{lag}}/\tau$ for the curves in a) ($(\tau_{\text{lag}}/\tau) w_{\text{sum},c}^{\text{rec,I}} + 0.02$, $(\tau_{\text{lag}}/\tau) w_{\text{sum},c}^{\text{rec,I}}$, $(\tau_{\text{lag}}/\tau) w_{\text{sum},c}^{\text{rec,I}} - 0.02$). Parameters: $w_{\text{sum}}^{\text{rec,net}} = 0.99$, $\tau = 1$, $\tau_{\text{lag}} = 0.1$, and $N = 200$ for the network simulation. {fig:5_around_critical}

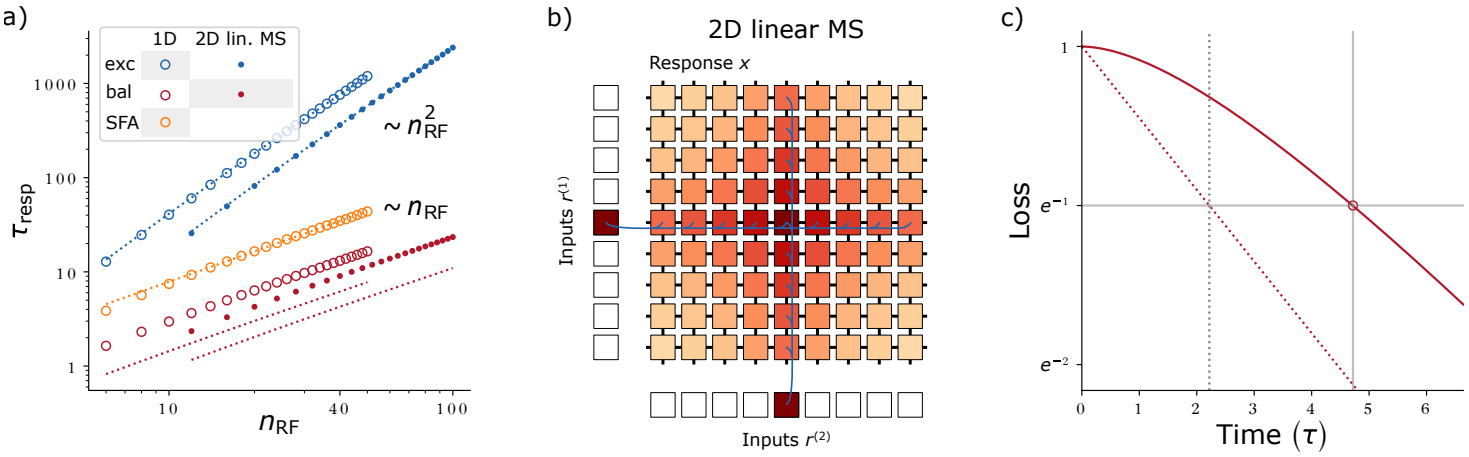


Fig 6: **Response speed of networks with inhibition and linear MS. a)** Response times for the 1D network and for the 2D linear MS network. The quadratic scaling of $\tau_{\mathrm{resp}}$ with $n_{\mathrm{RF}}$ for the excitatory networks (blue) can be improved to a linear dependence by introducing balancing, delayed inhibition (red) or SFA (orange). Open (1D network) and filled (MS network) circles display numerical results. Alike colored dotted (1D network) or continuous (MS network) curves show theoretical estimates (eqs. (12), (27), (33) and (34)) or, for the SFA network, fit results (monomial fit: $\tau_{\mathrm{resp}}^{\mathrm{SFA}}(n_{\mathrm{RF}}) = 0.66724(n_{\mathrm{RF}})^{1.07063}$). We use the slowest decaying eigenmode to theoretically estimate the response times (see c). Since the balanced networks are not initialized in this eigenmode (in contrast to the purely excitatory networks), the numerically measured response times (red markers) lie above the theoretical values (red lines). **b)** Schematic of a 2D network with linear MS. Feature neurons are arranged on a two-dimensional grid (labeled 'Response $x$'). Each receives feedforward input from two arrays of input neurons (labeled 'Inputs $r^{(1|2)}$') and four recurrent inputs. Feedforward and recurrent synapses are shown in blue (exemplarily) and black, respectively. Input and feature neuron activities are color-coded. The (linear) network response is the sum of the responses to input one and input two. **c)** Exemplary loss evolution for a 1D network with lagged inhibition. Due to the temporally constant initialization ($x_i(0) = 0$, $\Delta x_i(0) = 0$), the network activity (solid red curve) converges initially more slowly than the network's slowest eigenmode (dotted red line). The experimentally measured response time (continuous vertical gray line) is defined as the time when the loss has decayed by $1/e$ (red open circle, horizontal gray line, see also Fig. 4a). It is larger than that of the network's eigenmode (dotted gray line), which we use as analytical estimate of the response time. We created the data in a) by scanning $n_{\mathrm{RF}}$, setting $w_{\mathrm{sum}}^{\mathrm{rec,net}}$ to yield a RF of size $n_{\mathrm{RF}}$, setting $w_{\mathrm{sum}}^{\mathrm{rec,I}}$ to 0 or its critical value, and determining $\tau_{\mathrm{resp}}$ or $\tau_{\mathrm{resp}}^{\mathrm{bal}}$ from the loss dynamics. For the SFA network we set $\tau_{\mathrm{SFA}} = \tau$, scanned $a_{\mathrm{SFA}}$ and used the value that minimized the temporally integrated loss. {fig:6_tresp_1D2Dlin_with_scheme}

activity change in neuron $j$ adds an activity change with the same sign to neuron $i$, because $-W^{\mathrm{rec,I}}$ in eq. (22) is positive. Thus activity changes in different neurons in the network amplify each other. In the limit of small $\tau_{\mathrm{lag}}$, the temporal derivative is transmitted, see appendix S5.

## Linear mixed selectivity

Neurons often respond selectively to more than one stimulus or input feature [64]. This phenomenon is called mixed selectivity (MS). Here we consider linear MS [64, 65], where neuronal responses are linear functions of multiple stimuli. Concretely, neurons with activities $x_{ij}$, $i, j = 1, \cdots, N$ are arranged on a two-dimensional grid and respond with equal selectivity to two input features, represented by input neurons $r_k^{(1)}$ and $r_l^{(2)}$ with $k, l = 1, \cdots, N$,

$$x_{ij}^{\mathrm{resp}} = \sum_{k=1}^{N} \mathrm{RF}_{ijk}^{(1)} r_k^{(1)} + \sum_{l=1}^{N} \mathrm{RF}_{ijl}^{(2)} r_l^{(2)}. \tag{28}$$ {{eq:RF_sum_linear_mixed}}

Due to the linearity in the input representation and in the network, the total response is the sum of the responses to the single input features. We take the grid axes to be aligned with the stimulus dimensions, so that the first index in $x_{ij}$ determines its response to $r^{(1)}$ and the second that to $r^{(2)}$. We model this dependence as the same localized, exponentially decaying shape as for the 1D network (cf. Fig. 6b)),

$$\mathrm{RF}_{ijk}^{(1)} = \gamma^{|i-k|} \qquad \mathrm{RF}_{ijk}^{(2)} = \gamma^{|j-k|}. \tag{29}$$ {{eq:RF_linear_mixed}}

The desired network response can be generated as the steady state of a recurrent network that is equivalent to the 1D network eq. (5) in each dimension of the 2D grid (see next section),

$$\begin{aligned} \tau \dot{x}_{ij} = -x_{ij} &+ w^{\mathrm{rec,MS}}(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1}) \\ &+ w^{\mathrm{ff,MS}}(r_i^{(1)} + r_j^{(2)}), \end{aligned} \tag{30}$$ {{eq:eom_lin_mixed_selectivity}}

with the modified constants $w^{\mathrm{rec,MS}} = \frac{w^{\mathrm{rec}}}{1+2w^{\mathrm{rec}}} = \frac{\gamma}{(1+\gamma)^2}$ and $w^{\mathrm{ff,MS}} = \frac{w^{\mathrm{ff}}}{1+2w^{\mathrm{rec}}} = \frac{1-\gamma^2}{1+\gamma^2}$. Each neuron receives two external inputs and is connected to its nearest neighbors along each stimulus axis. The network has thus only six synapses per neuron, regardless of the RF width. Also a feedforward network where each dimension of the 2D grid is equal to the 1D network eq. (2) generates the desired response. This implementation requires $n_{\mathrm{RF}}^{\mathrm{MS}} = 2n_{\mathrm{RF}} = 2(2d+1)$ synapses per neuron, a number that increases linearly with the RF width.

### Mapping to a 1D system

In the following, we trace the network dynamics eq. (30) back to those of the 1D system eq. (5). Due to the linearity of eq. (30), network responses again superpose. It thus suffices to study the network in the case where only one input neuron is active: we choose $r_i^{(1)}$, which specifies a property of the first stimulus, to be nonzero. Since the input is independent of $j$, the dynamics eq. (30) are (for initial conditions homogeneous in $j$ such as $x_{ij}(0) = 0$) independent of $j$, $x_{ij}(t) = x_i(t)$. The recurrent inputs $w^{\mathrm{rec,MS}}(x_{i,j+1}(t) + x_{i,j-1}(t)) = 2w^{\mathrm{rec,MS}} x_i(t)$ then simply amount to a modification of the leak current to $-(1 - 2w^{\mathrm{rec,MS}})x_i$,

$$\tau \dot{x}_i = -(1 - 2w^{\mathrm{rec,MS}})x_i + w^{\mathrm{ff,MS}} r_i^{(1)} + w^{\mathrm{rec,MS}}(x_{i+1} + x_{i-1}) \tag{31}$$ {{eq:a_1D_correspondence_factor_0}}

After dividing by $(1-2w^{\mathrm{rec,MS}})$, the differential equation for $x_i$ becomes

$$a\tau\dot{x}_i = -x_i + aw^{\mathrm{rec,MS}}(x_{i+1}+x_{i-1}) + aw^{\mathrm{ff,MS}}r_i^{(1)}, \tag{32}$$ {{eq:linMS_mapping_to_1D}}

where we introduced $a = (1-2w^{\mathrm{rec,MS}})^{-1}$ for brevity. With the values of the constants $w^{\mathrm{rec,MS}}$ and $w^{\mathrm{ff,MS}}$ highlighted after eq. (30) this is equivalent to the one-dimensional network dynamics eq. (5) up to a different neuronal time constant $a\tau$ instead of $\tau$. (We note that we obtained the modified constants such that this holds. For example, equating the prefactors of the recurrent term in eq. (32) and eq. (5) gives $w^{\mathrm{rec}} = aw^{\mathrm{rec,MS}} = (1-2w^{\mathrm{rec,MS}})^{-1}w^{\mathrm{rec,MS}}$, which then can be solved for $w^{\mathrm{rec,MS}}$.) As a direct consequence, while the 1D network must have recurrent coupling strength of $w^{\mathrm{rec}} < 0.5$ for being stable, the 2D MS network must have $w^{\mathrm{rec,MS}} < 0.25$. This is because in the MS network, a neuron receives direct recurrent input from four nearest neighbors instead of two as in the 1D case.

Equation (32) means that the RF of the MS network, along one axis, has the same shape and width $d$ as the equivalent one-dimensional network. In particular, $d$ is related to the recurrent weight strength $aw^{\mathrm{rec,MS}}$ via $aw^{\mathrm{rec,MS}} = w^{\mathrm{rec}} = \frac{\gamma}{(1+\gamma)^2}$ and $\gamma = \exp(-1/d)$; the two RF components in eq. (29) are the same as the RFs in eq. (1), for example $\mathrm{RF}^{(1)}_{ijk} = \mathrm{RF}_{ik}$.

### Response speed

From the mapping of the MS to the 1D system, eq. (32), we see that the MS dynamics behaves in response to a single input like the 1D dynamics with the neuronal time constant $\tau$ enlarged by a factor of $a$. The response time is thus given by eq. (12), but with enlarged neuronal time constant, $\tau \to a\tau$. For sufficiently large $d$, we have $\gamma \approx 1$ (reflecting the spatially slow RF decay), $w^{\mathrm{rec}} \approx 1/2$, $w^{\mathrm{rec,MS}} \approx 1/4$ and thus $a \approx 2$. The scaling of the response time with the RF width $d$ and size $n^{\mathrm{MS}}_{\mathrm{RF}}$ is thus again quadratic,

$$\tau^{\mathrm{MS}}_{\mathrm{resp}} \approx (1+2d^2)\,2\tau \approx 4d^2\tau \approx \frac{1}{4}\left(n^{\mathrm{MS}}_{\mathrm{RF}}\right)^2\tau. \tag{33}$$ {{eq:tresp_net_MS_vs_d}}

In the last equation we used $n^{\mathrm{MS}}_{\mathrm{RF}} = 2(2d+1) \approx 4d$. Compared to the 1D case (eq. (12)), the response time as a function of $d$ is therefore larger by a factor $a \approx 2$. In contrast, it is smaller by a factor 1/2 as a function of the RF size, compare eq. (33) with eq. (12) and the blue continuous and dotted curves in Fig. 6a). In other words: the trade-off between response time and number of needed synapses improves for sufficiently large RFs by a constant factor of about 1/2 compared to the 1D network. This is because the MS network effectively implements two 1D RFs (eq. (28)).

### Balanced network

We now incorporate the effect of inhibitory neurons into the MS network. As in the 1D case, we assume that the generated inhibition precisely tracks excitation with a short time delay. We thus add to each recurrent excitatory connection an inhibitory one that is slightly delayed. This results in a delayed differential equation like eq. (24) for the balanced MS network dynamics. The parameters are given by those of the 1D balanced system up to a factor $a = 1 + w^{\mathrm{rec,net}}_{\mathrm{sum}}$, like in eq. (32). Further, it is again sufficient to study the response dynamics to a single input, which can be reduced to those of the 1D balanced network eq. (24) with adapted parameters. As in the purely excitatory case, for a fair comparison of response times, we consider MS and 1D networks with the same neuronal time constant $\tau$. The effective time constant of the MS dynamics are then $a\tau$. Therefore the response time of the MS network are given by those of the 1D network

eq. (27) with neuronal time constant $\tau$ replaced by $a\tau \approx 2\tau$,

$$\tau_{\text{resp}}^{\text{bal},c,\text{MS}} \approx \sqrt{(1+2d^2)\ \tau\tau_{\text{lag}}} \approx \sqrt{2}d\sqrt{\tau\tau_{\text{lag}}} \approx \frac{1}{2\sqrt{2}} n_{\text{RF}}^{\text{MS}} \sqrt{\tau\tau_{\text{lag}}}. \tag{34}$$ {{eq:tresp MS bal}}

At the critical inhibitory strength, the response time thus scales again with the square root of the response time of the network without inhibition (eq. (26)). Therefore it scales linearly with the receptive field width $d$ and size $n_{\text{RF}}^{\text{MS}}$. The response time is as a function of the RF size by a factor of about $1/\sqrt{2}$ smaller than that in the 1D case, eq. (27), see Fig. 6a), red continuous and dotted curves. In other words, the trade-off between response time and number of required synapses improves by a factor of $1/\sqrt{2}$. This is again because the MS network effectively implements two RFs in the MS case; the RF size doubles for the same width compared to the 1D network.

As for the 1D stimulus, the balanced networks have twice as many neurons and additional synapses: Each (excitatory) feature neuron drives one inhibitory neuron, which mirrors its activity. This inhibitory neuron in turn forms inhibitory synapses to the four nearest neighbors that its presynaptic feature neuron excites. In total, the balanced, cooperatively coding network thus requires eleven synapses per feature neuron, instead of six for the unbalanced network. This is independent of the RF width, such that the balanced, cooperatively coding network save synapses for sufficiently wide RFs.

### Higher-dimensional linear MS

We can straightforwardly extend the introduced scheme to networks that have MS with $P > 2$ stimuli. Neurons are then arranged on a hyper-grid with one grid axis per stimulus dimension, so that $N^P$ feature neurons respond to $PN$ input neurons. In the cooperatively coding network, each neuron receives $P$ feedforward and $2P$ recurrent inputs, requiring a total of $3P$ synapses per neuron. The feedforward network, in contrast, needs for each stimulus dimension $2d+1$ synapses, in total $P(2d+1)$ synapses per neuron. The number of saved synapses thus grows linearly with the number of encoded stimulus dimensions and the receptive field width.

## Encoding a 2D stimulus

We finally consider the encoding of a two-dimensional stimulus, with both input and feature neurons arranged on a two-dimensional grid, see Fig. 7a). Two-dimensional input appears for example in vision [36] or planar navigation tasks [66]. Each feature neuron responds to inputs that are close to its preferred input in both stimulus dimensions. The RFs of neighboring neurons thus overlap and neuronal responses tile the represented stimulus space. Concretely, the desired activity of feature neuron $ij$ is

$$x_{ij}^{\text{steady}} = \sum_{kl=1}^{N} \text{RF}_{ijkl} r_{kl}, \qquad \text{RF}_{ijkl} = \gamma^{|i-k|}\gamma^{|j-l|}. \tag{35}$$

The neuron thus responds to two input features represented by input neurons $r_{ij}$ with $i,j = 1,\cdots,N$. A purely excitatory cooperative coding network generating the desired activity as stationary state is given by

$$\tau\dot{x}_{ij} = -x_{ij} + w^{\text{rec,2D}}(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1}) + w^{\text{ff,2D}} r_{ij}. \tag{36}$$ {{eq:eom_nonlin_mixed_selectivity}}

The RF size depends on $w^{\text{rec,2D}}$ and the response amplitude on $w^{\text{ff,2D}}$. In contrast to the 1D and the MS case, these dependencies cannot easily be computed analytically; in Fig. 7 we thus determine the receptive field sizes resulting from the tried $w^{\text{rec,2D}}$ numerically. The network requires 5 synapses per neuron.

We also constructed a balanced network by introducing for each excitatory recurrent input a delayed inhibitory one, like in the balanced 1D and MS networks. A balanced implementation with explicit inhibitory interneurons requires twice as many neurons and synapses than the purely excitatory network: it requires additionally one interneuron per principle neuron, one inhibitory synapse for each excitatory recurrent synapse and one synapse from each principle neuron to its corresponding interneuron.

### Response speed

To estimate the dependence of the response speed on the RF size, we first need to appropriately adapt the definition of the RF size, which we introduced after eq. (3). For the one-dimensional network, this definition can be reformulated as follows: we count the number of synapses that are necessary to generate the largest (around the center) RF responses such that these responses summed together amount to a fraction of about $1 - e^{-1} \approx 63\%$ of the summed non-truncated RF. Accordingly, for the 2D network at hand we define the RF size as the number of feedforward synapses that are necessary to implement the largest RF entries, such that together they account for a fraction of approximately 63% of the summed nontruncated RF. We denote the so-defined receptive field sizes by $\hat{n}_{\mathrm{RF}}$.

As for the 1D and linear MS networks, the response time with or without lagged inhibition depends only on the summed excitatory weights or on the summed net and inhibitory weights. It is thus given by eq. (11) or by eq. (26) in terms of $w_{\mathrm{sum}}^{\mathrm{rec}} = 4w^{\mathrm{rec,2D}}$ or in terms of the (alike obtained) $w_{\mathrm{sum}}^{\mathrm{rec,net}}$ and $w_{\mathrm{sum}}^{\mathrm{rec,I}}$. Fig. 7b) shows that the scaling of the response time with the RF size is linear for unbalanced and square-root-like for balanced networks. We give a geometric argument for this general scaling in the Discussion (subsection 'Multi-dimensional stimuli'). The scaling is more economical than for the 1D and 2D linear MS networks, cf. Fig. 6.

## Discussion

In this work, we have studied networks that encode continuous variables with neurons that have overlapping response properties. We developed a cooperative coding scheme that enables them to share and distribute computations among similarly-tuned neurons, crucially using (net) excitatory connections. For the simplest considered networks this sharing minimizes the number of required synapses while the total amount of synaptic current remains the same as in a purely feedforward implementation. For networks of neurons that represent higher-dimensional stimuli, the number of saved synapses is especially large. The saving of synapses comes at the cost of longer response times. We find, however, that neurons with spike frequency adaptation and neurons in networks in which excitation is largely balanced by delayed inhibition can use the window of opportunity between the arrival of excitation and inhibition to significantly speed up their convergence to the steady state response, decreasing response times by orders of magnitude and improving their scaling with RF size.

**RF shape and more complex networks** The exponentially decaying RFs that we consider act as tractable models for experimentally encountered localized, overlapping and broadening RFs [2, 38, 40]. They allow to elegantly illustrate how neurons can use recurrent interactions to cooperatively share feedforward information and shape the network response. However, experimentally measured RFs have different and often more complex shapes [7–9, 36]. We are optimistic that these can still be approximated by the steady state of a cooperatively coding recurrent network with sparse connectivity,

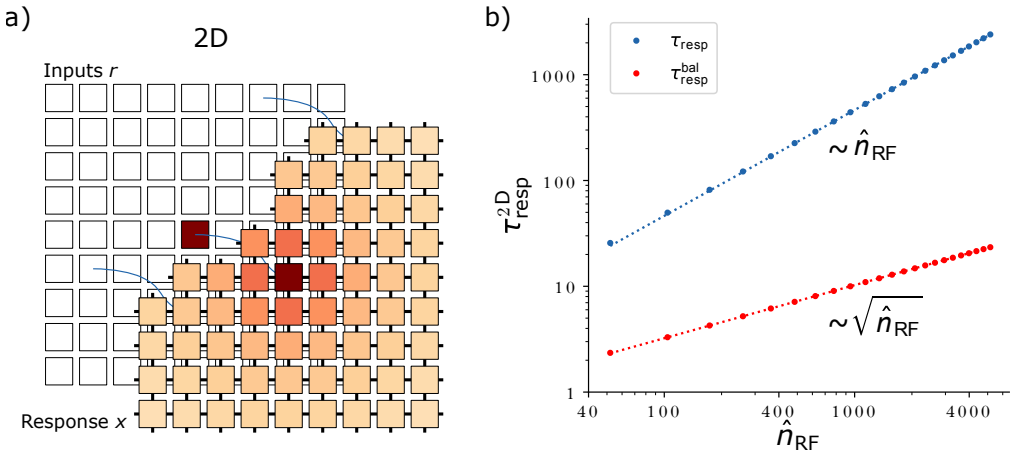


Fig 7: **2D network schematic and response times versus RF size. a)** Schematic of a two-dimensional network responding to a two-dimensional stimulus. Feature neurons (labeled 'Response $x$') and input neurons (labeled 'Inputs $r$') are arranged on two-dimensional grids. In the cooperatively coding network each feature neuron receives one feedforward and four recurrent inputs; activities and shown connections are color-coded as in Fig. 2. **b)** Response times in the 2D network increase linearly (without inhibition, blue. Monomial fit: $\tau_{\text{resp}}^{\text{2D}} = 0.46115\ (\hat{n}_{\text{RF}})^{1.00085}$) or square-root-like (with inhibition, red. Monomial fit: $\tau_{\text{resp}}^{\text{bal,2D}} = 0.31995\ (\hat{n}_{\text{RF}})^{0.50202}$) with the RF size $\hat{n}_{\text{RF}}$. Dotted lines represent the monomial fits. Data was created by scanning $w_{\text{sum}}^{\text{rec,net}}$, setting $w_{\text{sum}}^{\text{rec,I}}$ to 0 or its critical value, and determining $\hat{n}_{\text{RF}}$ and $\tau_{\text{resp}}$ or $\tau_{\text{resp}}^{\text{bal}}$, respectively, from the response curves after network activity converged. {fig:7_tresp_2D_nonlinearMS}

although more synapses will be required. Determining the necessary network parameters might involve minimizing a loss with L0 regularization, which is challenging.

**Related work** Conceptually, our 1D model is a ring model. Models of this type have been proposed to model orientation selectivity in the visual cortex [23, 52] (see also [67]), head direction cells [68] and spatial memory [69]. Similarly, our 2D and 2D mixed selectivity models have a toroidal or, when removing the periodicity, a planar structure. Such networks may be important for spatial navigation [66]. With appropriate coupling, ring-like networks can have two different dynamical regimes [42]: (i) an input-driven regime where there is a single, homogeneous ground state, which is assumed in absence of input and (ii) a regime of bump attractors, where there is spatially localized, persistent activity in absence of input. Our networks are linear, therefore without input there cannot be multiple stationary activity patterns whose amplitudes are asymptotically stable. This is because each multiple of a stationary solution is a solution as well. We thus work in the input-driven regime, with a single stable zero ground state. Previous models have broad coupling fields or ranges of coupling probabilities, equivalent to many recurrent synaptic connections that extend over neurons with quite different preferred stimuli [23, 42, 52, 66, 68, 69]. In contrast, in cooperative coding networks, we have very sparse synaptic connections between neurons with highly similar tuning.

In this work we considered the encoding of continuous variables in a scheme that minimizes the total number of required synapses. Relatedly, ref. [70] investigates the problem of classifying discrete patterns in large networks of neurons with limited and fixed in-degrees. It solves the problem by introducing an intermediate layer consisting of interconnected perceptrons, which each receive only part of the overall input. This intermediate layer may be seen as analogous to our feature layer. In the mainly studied

binary classification task, the intermediate layer is equipped with excitatory connections. The assumed connectivity is untuned except for being overcritically excitatory such that it drives neurons into saturation. In the stationary response to a pattern then all (more precisely: nearly all [71]) neurons in the network assume the same output, either positive or negative one. This allows a readout neuron to perform binary classification with sparse readout weights. A refined architecture combines several such intermediate layers in parallel. This yields an intermediate layer with several subpopulations, which code in binary manner, allowing for classification with a binary vector. Another way highlighted to achieve this is by using as intermediate layer a Hopfield network with recurrent connections that are randomly realized with fixed probability. Since the binary coding network and the network with several subpopulations have specific like-to-like connectivity to save feedforward connections, they realize cooperative coding in our sense, in a binary manner. By choosing the recurrent coupling probability of the Hopfield network intermediate layer such that stronger positive and negative weights (which connect similarly and oppositely tuned neurons, respectively) are realized with larger probability, one could implement cooperative coding in them as well. Ref. [72] find that local recurrent connectivity in Hebbian assemblies of spiking neurons can reduce the number of feedforward connections between assemblies required for memory replay. The total number of synapses in their model is, however, minimized by a purely feedforward architecture. They argue that one benefit of sparse feedforward connectivity, augmented through local recurrence, might be enabling one-shot learning.

The signature of cooperative coding in the networks that we investigated and in more general ones is that the network trades feedforward and less specific recurrent synapses for recurrent synapses mediating on average net-excitatory recurrent interactions between similarly tuned neurons (and net-inhibitory interactions between highly anti-tuned neurons). Consistent with this, intermediate-depth ML networks featuring recurrent and feedback connections can match the performance of much deeper feedforward networks while requiring less units and parameters [73]. It would be interesting to investigate whether the recurrent connectivity in such networks is also like-to-like. If yes, this would indicate that cooperative coding naturally appears also in ML networks. It may be helpful in particular in convolutional networks to save feedforward connections and rely on very sparse recurrent connectivity instead.

**Response speed in excitatory 1D networks** In our most simple, purely excitatory cooperatively coding networks the response is slowed down compared to that of single neurons due to recurrent excitation, which implements a positive feedback loop. This feedback loop increases the eigenvalues of the matrix governing the differential equation of the dynamics and increases the duration of the response to single pulses. The prolonged response to a single short pulse adds up for prolonged inputs and leads to their amplification as well. This type of amplification has been termed "Hebbian amplification" in [21].

**Networks with SFA** To speed up the response, we first introduced SFA, a typical feature of excitatory principal neurons [42, 49, 50]. Our model is a slightly simplified version of that in [51] and the same as in [52, 53].

In the networks with SFA excitation still dominates, i.e. we have on the one hand Hebbian amplification as well. On the other hand, we have "balanced amplification" [21]: Each neuron is a 2-dimensional dynamical system, the network consists of $N$ of them; they are excitatorily coupled. Without the inhibitory feedback from the adaptation the dynamics of the purely excitatory neuron population would be unstable. When a short input arrives to the neuron population, its activity therefore increases rapidly, using the window of opportunity before the inhibitory feedback current becomes prominent.

Inhibition then takes over and largely terminates the excursion leading to a shorter duration impulse response of the network. For prolonged input, these impulses basically add up, which leads to the fast convergence to new stationary activity, amplifying the input. The complex Schur decomposition of the matrix governing the network dynamics reveals a strong feedforward coupling from an oscillatory difference to an oscillatory sum mode. This is similar to the networks in [21], which, however, have mostly real, non-positive eigenvalues and thus non-oscillatory modes without Hebbian amplification.

**Balanced networks** The balanced versions of our networks incorporate inhibition that closely tracks excitation with a short delay, as observed in experiments [60, 74, 75]. For stationary dynamics, excitatory and inhibitory currents largely cancel, resulting in a relatively weak net excitatory interaction between similarly-tuned neurons [18, 19]. However, excitation reacts slightly faster than inhibition to dynamic activity changes, allowing strong interactions during the window of opportunity where the change in excitation is not yet balanced by the delayed inhibition. This mechanism allows activity changes to propagate quickly through the network, significantly decreasing its response time.

The effective lag of inhibitory feedback originates in our SFA and in other balanced amplification [21] networks from the fact that the inhibitory currents are evoked by a low-pass filtered version of the excitatory activity. In contrast, in our balanced networks, there is an explicit, fixed lag between excitatory and inhibitory activity. We thus have an infinite dimensional dynamical system governed by a delay differential equation. However, the basic mechanism of shortening the impulse response and speeding up the reaction to inputs is the same: Excitation without feedback inhibition is unstable. When an input arrives, inhibition is insufficient to balance it. The excitatory activity therefore increases strongly during the temporal window of opportunity. Then inhibition sets in, largely balances excitation and limits it to its stationary value. We note that still excitation dominates our balanced networks. We have numerically and analytically studied the resulting convergence to stationary activity. This revealed qualitatively different types of dynamics, which are familiar from the harmonic oscillator, namely overdamped, critical and underdamped dynamics. Their occurrence depends on the strength of the inhibitory feedback and the size of the time lag between excitation and inhibition.

The increased response speed comes at the cost of requiring additional synapses and interneurons to implement the inhibition. In our model this roughly doubles the number of required synapses compared to the unbalanced cooperatively coding networks. As the number of needed synapses is still independent of the RF size, the balanced networks still save synapses for larger RFs. Furthermore, inhibitory neurons may be required anyways for purposes such as maintaining irregular spiking activity [55–58]. Alongside the increased synaptic requirements, the cancellation of currents means an increased metabolic cost.

We covered inhibition in a simplified, effective manner without explicitly incorporating inhibitory neurons. As argued above, we are optimistic that our qualitative findings will hold in more complex networks, for example if they are optimized for sparseness. In particular we expect that such networks will still show characteristics of cooperative coding and that inhibition will allow to reduce response times.

From our analysis, the brain might use cooperative coding to save synapses and space compared to a purely feedforward or more wasteful recurrent implementation, but might invest some synapses, neurons, space and energy in balancing inhibition to retain a reasonable response speed.

**Mixed selectivity** We also considered networks with two-dimensional mixed selectivity, where neurons are not only selective to a single stimulus but to two stimuli. Since our networks are linear, we consider linear mixed selectivity. This is a simplification compared to the nonlinear mixed selectivity that is ubiquitous in the brain [64, 65]. We observe that the tradeoff between saved synapses and speed is for networks using linear mixed selectivity improved by a constant factor compared to networks where each neuron codes for a single stimulus.

**Multi-dimensional stimuli** We considered networks encoding one- and two-dimensional stimuli. We found that the trade-off between saved synapses and increased response time is much more favorable for neurons that encode 2D stimuli.

For $P \geq 2$-dimensional stimuli, the fact that activity simultaneously propagates along all dimensions suggests that the scaling of the response time with $d$, the characteristic RF width along one of the dimensions, does not change with the number of dimensions. However, the RF size $n_{\mathrm{RF}}$, the number of synapses needed in a purely feedforward implementation, can be assumed to scale as $n_{\mathrm{RF}} \sim (2d+1)^P$, with a prefactor that depends on geometry. This reasoning suggests that the response time of networks encoding higher-dimensional stimuli scales like $\tau_{\mathrm{resp}} \sim n_{\mathrm{RF}}^{2/P}$ and $\tau_{\mathrm{resp}}^{\mathrm{bal}} \sim n_{\mathrm{RF}}^{1/P}$, which we verified for $P = 1, 2$. The number of synapses required in a cooperatively coding network will also in higher dimensions be negligible compared to $n_{\mathrm{RF}}$. Therefore the number of saved synapses in the cooperatively coding network is still approximately $n_{\mathrm{RF}}$. For higher-dimensional stimuli, the trade-off between response time and saved synapses would thus become highly beneficial: the response time $\tau_{\mathrm{resp}}$ or $\tau_{\mathrm{resp}}^{\mathrm{bal}}$ would grow only slowly with the number of saved synapses $n_{\mathrm{RF}}$ due to the strongly sublinear relationship for larger $P$.

**Properties of connectivity** Our cooperative coding scheme relies on the presence of few strong recurrent excitatory connections between similarly-tuned cells. Excitation can be balanced by inhibition, while interactions between similarly-tuned cells remain net excitatory. Indeed, in layer 2/3 of mouse visual cortex, pyramidal neurons with the same orientation preference connect at higher rates and form more bidirectional connections [15]. This pattern of increased connection probability between neurons with highly similar tuning extends across layers and visual areas, including feedforward and feedback connections [14]. Furthermore, synapses between neurons with similar spatial RFs are markedly stronger such that neurons receive the majority of their local excitation from few similarly-tuned cells [16]. This strong, sparse local excitation matches the RF structure of the receiving neurons [16]. Recurrent connections are generally sparse in the cortex [76–79].

Studies of recurrent functional connectivity found net excitation between (spatially close) neurons with similar tuning [19] and most correlated responses [18], consistent with our model. Ref. [19] showed specifically that when optogenetically stimulating spatially compact ensembles of co-tuned neurons, similarly-tuned neurons were excited while differently-tuned neurons were inhibited. In a 1D-model, ref. [18] had to incorporate strong nearest-neighbor-like excitatory interactions to match the experimental data, arguing that they stabilize network responses in the presence of input noise. Our interpretation is that they not only stabilize, but actively form the RF. Refs. [18, 19] also show an inhibitory effect on largely differently tuned neurons. Ref. [18] found net inhibition between rather similarly tuned neurons as well. This is assumed to implement feature competition, which we did not include in our model.

A particular benefit of our cooperative coding scheme is that it allows feedforward connections to be sparse. This fits for example experimental observations in the primary visual cortex, where the vast majority of inputs are local recurrent ones, while only a

few percent are feedforward inputs [80, 81]. Ref. [82] estimated based on experimental studies [83] that a single hypercolumn in primate V1 receives only 10-30 feedforward inputs from the magnocellular layer of dorsal LGN mediating retinal input, with single cells in L4$\alpha$ receiving as little as 0 - 6 inputs. Also in the hippocampal region CA3, where place fields are enlarged compared to the upstream dentate gyrus [40], the recurrent connectivity is high [78, 84].

Experiments that aim to disentangle feedforward from recurrent contributions to orientation selectivity resulted in mixed findings. Ref. [85] showed that the input to simple cells in L4 of cat primary visual cortex still exhibit tuned, modulated responses to drifting gratings after cortical activity was suppressed by cooling. In line with this, ref. [17] found that thalamic and cortical contributions to the first harmonic of the response curve (F1) were co-tuned. However, the temporally averaged response (F0) to drifting gratings was tuned only in cortical but not in thalamic inputs. A recent study, ref. [86], suggests that total input current from L4 of mouse primary visual cortex to L2/3 may lack orientation tuning and that orientation selectivity is determined by recurrent inputs from within L2/3.

**Optimality** By distributing and reusing computations via excitatory connections, cooperatively coding networks minimize or strongly reduce the number of synapses required to generate their responses. This corresponds to minimizing or strongly reducing the L0 norm of the weights (or synaptic currents).

In the following we compare the metabolic cost of the feedforward and the simplest, purely excitatory cooperative coding networks in the stationary state. We consider the three main contributions of the metabolic cost [26, 28]: the cost of keeping up the resting network, of generating the required activity and of the synaptic transmissions. Keeping up the resting network (housekeeping and maintaining the resting potentials) requires the same energy expenditure in both implementations, as the number of neurons is the same. Also the energy required to generate the stationary output activity is the same, as corresponding neurons generate the same activity in both implementations.

The cost of synaptic transmission is dominated by the metabolic cost caused by the postsynaptic currents [28]. We assume that this cost is characterized by the sum of the absolute current strengths at individual synapses, i.e. by the L1 norm of the synaptic currents [87]. This is the same in both implementations. Finally, a small part of the overall cost (less than 10% [28]) arises due to presynaptic calcium influx and usage of neurotransmitter during synaptic transmission. A comparison of these contributions between the network implementations is more difficult. The surface of the active zone [88] increases linearly with the synaptic strength [89]. Assuming that the area where calcium influx happens increases linearly with the active zone size, the amount of influx during a single synaptic transmission depends linearly on the synaptic strength, consistent with [90]. The same should then hold for the related cost. Concerning the amount of neurotransmitter used, experiments observe that the number of directly releasable vesicles is linearly related to the synaptic strength, while the impact of a single vesicle and its release probability are independent of it [89, 91] (this may differ for different connected neuron types [92]). The finding suggests a linear dependence between the synaptic strength and the used neurotransmitter and thus the related cost at a single transmission. The total cost due to presynaptic calcium influx and neurotransmitter usage is therefore proportional to the absolute synaptic strength times the number of transmissions (the presynaptic activity). This, in turn, is proportional to the absolute value of the induced postsynaptic current, which is identical in both implementations.

We conclude that the metabolic cost of generating the stationary state is the same in the feedforward and the simplest, purely excitatory cooperative coding implementation. Since the latter takes longer to reach the stationary state, however, it also requires

more current and energy before generating a useful result. Due to the cancellation of excitatory and inhibitory currents, the networks with SFA and the balanced networks also use more current and thus more energy in the stationary state.

Previous studies often minimized the number of spikes or, more generally, the neuronal activity needed to represent encoded features. Refs. [30, 32, 34] follow this approach and suggest that tight EI-balance may be a signature of a highly coordinated and competitive code that, despite the irregular firing, is orders of magnitude more precise than a Poisson rate code. This spike-code depends on an extremely structured, dense connectivity, through which similarly-coding neurons quickly inhibit each other to prevent redundant spiking. From this standpoint the findings of excitatory functional connectivity between similarly-tuned neurons [15, 16, 18, 19] seem counter-intuitive.

Although the energetic costs of spike generation and synaptic transmission very likely play a role in shaping cortical networks, our work suggests that space constraints and the number of required synapses may be another main factor.

**Conclusion** To conclude, net excitatory connectivity between similarly tuned neurons is compatible with a cooperative coding scheme that generates network responses with a minimal number of synapses. This suggests space constraints as an important factor in shaping neural networks. The window of opportunity between excitation and balancing, delayed adaptation or inhibition may be harnessed to rapidly propagate activity changes through the network, speeding up equilibration times by orders of magnitude.

# Methods

All simulations have periodic boundary conditions. Fixed network parameters are the number of neurons $N$ for 1D and $N^2$ for 2D networks, the neuronal time-constant $\tau$ and, in networks with inhibition, the EI-lag $\tau_{\text{lag}}$. We set $N = 200$, $\tau = 1$ and $\tau_{\text{lag}} = 0.1$. In the networks with SFA we use a fixed value of $\tau_{\text{SFA}} = \tau = 1$ and, for each RF size, obtain the value of $a_{\text{SFA}}$ that minimizes the temporal mean of the normalized L1-loss, $(1/T)\int_0^T \mathrm{d}t\ |x(t) - x^*(t)|_1/|x^*|_1$, through a linear grid search. Here $T = 500\tau$ is the length of a trial as described in Fig. 8a) and $x^*(t)$ the target corresponding to the present input. Fig. 8e) and c) show the scans over $a_{\text{SFA}}$ and the individual loss curves for the optimal $a_{\text{SFA}}$ values. In all networks with inhibition, we set $w_{\text{sum}}^{\text{rec,I}}$ to its critical value given by eq. (76).

We simulate our networks with SFA using the Euler method and all other networks using the midpoint method with stepsize $\mathrm{d}t = 0.01$. To simulate the networks with delayed inhibition, we also need midpoint values of the delayed activity. We obtain them by copying the midpoint values of the non-delayed activity $\tau_{\text{lag}}$ ($\tau_{\text{lag}}/\,\mathrm{d}t$ simulation steps) before.

For the data in Figs. 4 and 6, we obtain RFs with different sizes by setting $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$ to appropriate values. In the case of 1D networks with and without SFA and in the case of 2D linear MS networks, we have analytical expressions for the RF sizes as a function of $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$. We thus chose $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$ such that the RF sizes are sampled linearly from $n_{\text{RF}}^{\text{1D}} = 6$ to $n_{\text{RF}}^{=}50$ in steps of two. For the 2D network, we do not have analytical expressions of the RF sizes as a function of $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$. Therefore we simulate networks with 20 different values of $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$ and measure the RF sizes that the networks generate after convergence. We obtain $w_{\text{sum}}^{\text{rec}}$ or $w_{\text{sum}}^{\text{rec,net}}$ as $w_{\text{sum}}^{\text{rec}}(\text{or } w_{\text{sum}}^{\text{rec,net}}) = 1 - \tau/\tau_{\text{resp}}$ by varying $\tau_{\text{resp}}$ from 10 to $1,000$ with equal spacing on a logarithmic scale.

To numerically determine a network's response time, we first simulate the network for a long time, clearly longer than the convergence time, and define the resulting state as the final, target state $x^*$. The loss is the L1 norm of the difference between $x^*$ and

the current state. We then simulate the network for a second time. We obtain $\tau_{\text{resp}}$ or $\tau_{\text{resp}}^{\text{bal}}$ as the earliest time at which the loss drops and stays below $e^{-1}$ times the initial loss.

## Acknowledgments

We thank Felipe Kalle Kossio for fruitful discussions.

## References

1. Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Computational Neuroscience. Massachusetts Institute of Technology Press; 2001.

2. Trappenberg T. Fundamentals of Computational Neuroscience. OUP Oxford; 2010. Available from: `https://books.google.de/books?id=4PDsA1EVCx0C`.

3. Hubel DH, Wiesel TN. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. The Journal of Physiology. 1962;160(1):106–154. doi:10.1113/jphysiol.1962.sp006837.

4. Nauhaus I, Nielsen KJ, Callaway EM. Efficient Receptive Field Tiling in Primate V1. Neuron. 2016;91(4):893–904. doi:10.1016/j.neuron.2016.07.015.

5. Taube S. Head Direction Cells Recorded in the Anterior Thalamic Nuclei of Freely Moving Rats. The Journal of Neuroscience. 1995;.

6. DiCarlo JJ, Johnson KO, Hsiao SS. Structure of Receptive Fields in Area 3b of Primary Somatosensory Cortex in the Alert Monkey. The Journal of Neuroscience. 1998;18(7):2626–2645. doi:10.1523/JNEUROSCI.18-07-02626.1998.

7. O'Keefe J, Dostrovsky J. The Hippocampus as a Spatial Map. Preliminary Evidence from Unit Activity in the Freely-Moving Rat. Brain Research. 1971;34(1):171–175. doi:10.1016/0006-8993(71)90358-1.

8. Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. Microstructure of a Spatial Map in the Entorhinal Cortex. Nature. 2005;436(7052):801–806. doi:10.1038/nature03721.

9. Stensola H, Stensola T, Solstad T, Frøland K, Moser MB, Moser EI. The Entorhinal Grid Map Is Discretized. Nature. 2012;492(7427):72–78. doi:10.1038/nature11649.

10. Fukushima K. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics. 1980;36(4):193–202. doi:10.1007/BF00344251.

11. Olshausen BA, Field DJ. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Nature. 1996;381(6583):607–609. doi:10.1038/381607a0.

12. Banino A, Barry C, Uria B, Blundell C, Lillicrap T, Mirowski P, et al. Vector-Based Navigation Using Grid-like Representations in Artificial Agents. Nature. 2018;557(7705):429–433. doi:10.1038/s41586-018-0102-6.

13. Cueva CJ, Wei XX. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In: International Conference on Learning Representations; 2018.Available from: `https://openreview.net/forum?id=B17JTOe0-`.

14. Ding Z, Fahey PG, Papadopoulos S, Wang EY, Celii B, Papadopoulos C, et al. Functional Connectomics Reveals General Wiring Rule in Mouse Visual Cortex. bioRxiv. 2023;doi:10.1101/2023.03.13.531369.

15. Ko H, Hofer SB, Pichler B, Buchanan KA, Sjöström PJ, Mrsic-Flogel TD. Functional Specificity of Local Synaptic Connections in Neocortical Networks. Nature. 2011;473(7345):87–91. doi:10.1038/nature09880.

16. Cossell L, Iacaruso MF, Muir DR, Houlton R, Sader EN, Ko H, et al. Functional Organization of Excitatory Synaptic Strength in Primary Visual Cortex. Nature. 2015;518(7539):399–403. doi:10.1038/nature14182.

17. Lien AD, Scanziani M. Tuned Thalamic Excitation Is Amplified by Visual Cortical Circuits. Nature Neuroscience. 2013;16(9):1315–1323. doi:10.1038/nn.3488.

18. Chettih SN, Harvey CD. Single-Neuron Perturbations Reveal Feature-Specific Competition in V1. Nature. 2019;567(7748):334–340. doi:10.1038/s41586-019-0997-6.

19. Oldenburg IA, Hendricks WD, Handy G, Shamardani K, Bounds HA, Doiron B, et al. The Logic of Recurrent Circuits in the Primary Visual Cortex. Nature Neuroscience. 2024;27(1):137–147. doi:10.1038/s41593-023-01510-5.

20. Valente M, Pica G, Bondanelli G, Moroni M, Runyan CA, Morcos AS, et al. Correlations Enhance the Behavioral Readout of Neural Population Activity in Association Cortex. Nature Neuroscience. 2021;24(7):975–986. doi:10.1038/s41593-021-00845-1.

21. Murphy BK, Miller KD. Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. Neuron. 2009;61(4):635–648. doi:10.1016/j.neuron.2009.02.005.

22. Lim S, Goldman MS. Balanced Cortical Microcircuitry for Maintaining Information in Working Memory. Nature Neuroscience. 2013;16(9):1306–1314. doi:10.1038/nn.3492.

23. Ben-Yishai R, Bar-Or RL, Sompolinsky H. Theory of Orientation Tuning in Visual Cortex. Proceedings of the National Academy of Sciences. 1995;92(9):3844–3848. doi:10.1073/pnas.92.9.3844.

24. Miller P. A Recurrent Network Model of Somatosensory Parametric Working Memory in the Prefrontal Cortex. Cerebral Cortex. 2003;13(11):1208–1218. doi:10.1093/cercor/bhg101.

25. Zhang WH, Wu S, Josić K, Doiron B. Sampling-Based Bayesian Inference in Recurrent Circuits of Stochastic Spiking Neurons. Nature Communications. 2023;14(1):7074. doi:10.1038/s41467-023-41743-3.

26. Attwell D, Laughlin SB. An Energy Budget for Signaling in the Grey Matter of the Brain. Journal of Cerebral Blood Flow & Metabolism. 2001;21(10):1133–1145. doi:10.1097/00004647-200110000-00001.

27. Zalc B. The Acquisition of Myelin: A Success Story. In: Chadwick DJ, Goode J, editors. Novartis Foundation Symposia. vol. 276. 1st ed. Wiley; 2006. p. 15–25. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9780470032244.ch3.

28. Howarth C, Gleeson P, Attwell D. Updated Energy Budgets for Neural Computation in the Neocortex and Cerebellum. Journal of Cerebral Blood Flow & Metabolism. 2012;32(7):1222–1232. doi:10.1038/jcbfm.2012.35.

29. Garcia KE, Kroenke CD, Bayly PV. Mechanics of Cortical Folding: Stress, Growth and Stability. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018;373(1759):20170321. doi:10.1098/rstb.2017.0321.

30. Boerlin M, Machens CK, Denève S. Predictive Coding of Dynamical Variables in Balanced Spiking Networks. PLoS Computational Biology. 2013;9(11):e1003258. doi:10.1371/journal.pcbi.1003258.

31. Thalmeier D, Uhlmann M, Kappen HJ, Memmesheimer RM. Learning universal computations with spikes. PLOS Comput Biol. 2016;12:e1004895.

32. Denève S, Machens CK. Efficient Codes and Balanced Networks. Nature Neuroscience. 2016;19(3):375–382. doi:10.1038/nn.4243.

33. Abbott LF, DePasquale B, Memmesheimer RM. Building Functional Networks of Spiking Model Neurons. Nature Neuroscience. 2016;19:350–355.

34. Kadmon J, Timcheck J, Ganguli S. Predictive Coding in Balanced Neural Networks with Noise, Chaos and Delays. NeurIPS Proceedings. 2020; p. 12.

35. Pulido C, Ryan TA. Synaptic Vesicle Pools Are a Major Hidden Resting Metabolic Burden of Nerve Terminals. SCIENCE ADVANCES. 2021;.

36. Jones JP, Palmer LA. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. Journal of Neurophysiology. 1987;58(6):1233–1258. doi:10.1152/jn.1987.58.6.1233.

37. Dumoulin SO, Wandell BA. Population Receptive Field Estimates in Human Visual Cortex. NeuroImage. 2008;39(2):647–660. doi:10.1016/j.neuroimage.2007.09.034.

38. Wandell BA, Winawer J. Computational Neuroimaging and Population Receptive Fields. Trends in Cognitive Sciences. 2015;19(6):349–357. doi:10.1016/j.tics.2015.03.009.

39. Serre T. In: Jaeger D, Jung R, editors. Hierarchical Models of the Visual System. New York, NY: Springer New York; 2013. p. 1–12. Available from: https://doi.org/10.1007/978-1-4614-7320-6_345-1.

40. Neher T, Azizi AH, Cheng S. From grid cells to place cells with realistic field sizes. PLoS One. 2017;12(7):e0181618.

41. Dayan P, Abbott L. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge: MIT Press; 2001.

42. Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal Dynamics - From single neurons to networks and models of cognition. Cambridge: Cambridge University Press; 2014.

43. Jetschke G. Mathematik der Selbstorganisation. Frankfurt am Main: Harri Deutsch; 2009.

44. Hirsch MW, Smale S. Differential equations, dynamical systems, and linear algebra. No. 60 in Pure and applied mathematics. San Diego [u.a.]: Acad. Press; 1974. Available from: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+022705937&sourceid=fbw_bibsonomy.

45. Pernice V, Staude B, Cardanobile S, Rotter S. How structure determines correlations in neuronal networks. PLOS Comput Biol. 2011;7:e1002059.

46. Kalle Kossio YF, Goedeke S, van den Akker B, Ibarz B, Memmesheimer RM. Growing Critical: Self-Organized Criticality in a Developing Neural System. Physical Review Letters. 2018;121:058301. doi:10.1103/PhysRevLett.121.058301.

47. White OL, Lee DD, Sompolinsky H. Short-term memory in orthogonal neural networks. Phys Rev Lett. 2004;92:148102.

48. Ganguli S, Huh D, Sompolinsky H. Memory traces in dynamical systems. Proc Natl Acad Sci U S A. 2008;105:18970–18975. doi:10.1073/pnas.0804451105.

49. Izhikevich EM. Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting. Cambridge: MIT Press; 2007.

50. Gutkin B, Zeldenrust F. Spike Frequency Adaptation. Scholarpedia. 2014;9(2):30643. doi:10.4249/scholarpedia.30643.

51. Treves A. Mean-field analysis of neuronal spike dynamics. Network: Computation in Neural Systems. 1993;4(3):259–284. doi:10.1088/0954-898x_4_3_002.

52. Hansel D, Sompolinsky H. Modeling feature selectivity in local cortical circuits. In: Koch C, Segev I, editors. Methods in Neuronal Modeling, 2nd ed. Cambridge, MA.: MIT press, Cambridge, MA.; 1998. p. 499–566.

53. Fuhrmann G, Markram H, Tsodyks M. Spike Frequency Adaptation and Neocortical Rhythms. Journal of Neurophysiology. 2002;88(2):761–770. doi:10.1152/jn.2002.88.2.761.

54. Dehghani N, Peyrache A, Telenczuk B, Le Van Quyen M, Halgren E, Cash SS, et al. Dynamic Balance of Excitation and Inhibition in Human and Monkey Neocortex. Scientific Reports. 2016;6(1):23176. doi:10.1038/srep23176.

55. Gerstein G, Mandelbrot B. Random Walk Models for the Spike Activity of a Single Neuron. Biophys J. 1964;4:41–68.

56. Shadlen M, Newsome W. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. J Neurosci. 1998;18:3870–3896.

57. Van Vreeswijk C, Sompolinsky H. Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. Science. 1996;274(5293):1724–1726. doi:10.1126/science.274.5293.1724.

58. Renart A, de la Rocha J, Bartho P, Hollender L, Parga N, Reyes A, et al. The Asynchronous State in Cortical Circuits. Science. 2010;327:587–590.

59. Herstel LJ, Wierenga CJ. Network control through coordinated inhibition. Current Opinion in Neurobiology. 2021;67:34–41. doi:10.1016/j.conb.2020.08.001.

60. Okun M, Lampl I. Instantaneous Correlation of Excitation and Inhibition during Ongoing and Sensory-Evoked Activities. Nature Neuroscience. 2008;11(5):535–537. doi:10.1038/nn.2105.

61. Znamenskiy P, Kim MH, Muir DR, Iacaruso MF, Hofer SB, Mrsic-Flogel TD. Functional Specificity of Recurrent Inhibition in Visual Cortex. Neuron. 2024; p. S0896627323009728. doi:10.1016/j.neuron.2023.12.013.

62. Runyan CA, Schummers J, Van Wart A, Kuhlman SJ, Wilson NR, Huang ZJ, et al. Response features of parvalbumin-expressing interneurons suggest precise roles for subtypes of inhibition in visual cortex. Neuron. 2010;67(5):847–857.

63. Najafi F, Elsayed GF, Cao R, Pnevmatikakis E, Latham PE, Cunningham JP, et al. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. Neuron. 2020;105(1):165–179.e8.

64. Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, et al. The Importance of Mixed Selectivity in Complex Cognitive Tasks. Nature. 2013;497(7451):585–590. doi:10.1038/nature12160.

65. Fusi S, Miller EK, Rigotti M. Why Neurons Mix: High Dimensionality for Higher Cognition. Current Opinion in Neurobiology. 2016;37:66–74. doi:10.1016/j.conb.2016.01.010.

66. McNaughton BL, Battaglia FP, Jensen O, Moser EI, Moser MB. Path integration and the neural basis of the 'cognitive map'. Nat Rev Neurosci. 2006;7(8):663–678.

67. von der Malsburg C. Self-organization of orientation sensitive cells in the striate cortex. Kybernetik. 1973;14(2):85–100.

68. Redish AD, Elga AN, Touretzky DS. A coupled attractor model of the rodent head direction system. Network. 1996;7(4):671–685.

69. Renart A, Song P, Wang XJ. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron. 2003;38:473–485.

70. Kushnir L, Fusi S. Neural Classifiers with Limited Connectivity and Recurrent Readouts. The Journal of Neuroscience. 2018;38(46):9900–9924. doi:10.1523/JNEUROSCI.3506-17.2018.

71. Turcu D, Abbott LF. Sparse RNNs Can Support High-Capacity Classification. PLOS Computational Biology. 2022;18(12):e1010759. doi:10.1371/journal.pcbi.1010759.

72. Chenkov N, Sprekeler H, Kempter R. Memory Replay in Balanced Recurrent Networks. PLOS Computational Biology. 2017;13(1):e1005359. doi:10.1371/journal.pcbi.1005359.

73. Nayebi A, Sagastuy-Brena J, Bear DM, Kar K, Kubilius J, Ganguli S, et al. Recurrent Connections in the Primate Ventral Visual Stream Mediate a Trade-Off Between Task Performance and Network Size During Core Object Recognition. Neural Computation. 2022;34(8):1652–1675. doi:10.1162/neco_a_01506.

74. Graupner M, Reyes AD. Synaptic Input Correlations Leading to Membrane Potential Decorrelation of Spontaneous Activity in Cortex. The Journal of Neuroscience. 2013;33(38):15075–15085. doi:10.1523/JNEUROSCI.0347-13.2013.

75. Barral J, Reyes AD. Synaptic Scaling Rule Preserves Excitatory–Inhibitory Balance and Salient Neuronal Network Dynamics. Nature Neuroscience. 2016;19(12):1690–1696. doi:10.1038/nn.4415.

76. Braitenberg V, Schüz A. Cortex: Statistics and geometry of neuronal connectivity. Springer Berlin; 1998.

77. Holmgren C, Harkany T, Svennenfors B, Zilberter Y. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. J Physiol. 2003;551.1:139–153.

78. Cutsuridis V, Graham BP, Cobb S, Vida I, editors. Hippocampal microcircuits. 2nd ed. Springer Series in Computational Neuroscience. Cham, Switzerland: Springer International Publishing; 2018.

79. Seeman SC, Campagnola L, Davoudian PA, Hoggarth A, Hage TA, Bosma-Moody A, et al. Sparse recurrent excitatory connectivity in the microcircuit of the adult mouse and human cortex. Elife. 2018;7.

80. da Costa NM, Martin KAC. The proportion of synapses formed by the axons of the lateral geniculate nucleus in layer 4 of area 17 of the cat. J Comp Neurol. 2009;516(4):264–276.

81. Markov NT, Misery P, Falchier A, Lamy C, Vezoli J, Quilodran R, et al. Weight consistency specifies regularities of macaque cortical networks. Cereb Cortex. 2011;21(6):1254–1272.

82. Chariker L, Shapley R, Young LS. Orientation Selectivity from Very Sparse LGN Inputs in a Comprehensive Model of Macaque V1 Cortex. The Journal of Neuroscience. 2016;36(49):12368–12384. doi:10.1523/JNEUROSCI.2603-16.2016.

83. Angelucci A, Sainsbury K. Contribution of Feedforward Thalamic Afferents and Corticogeniculate Feedback to the Spatial Summation Area of Macaque V1 and LGN. Journal of Comparative Neurology. 2006;498(3):330–351. doi:10.1002/cne.21060.

84. Sammons RP, Vezir M, Moreno-Velasquez L, Cano G, Orlando M, Sievers M, et al. Structure and function of the hippocampal CA3 module. Proceedings of the National Academy of Sciences. 2024;121(6). doi:10.1073/pnas.2312281120.

85. Ferster D, Chung S, Wheat H. Orientation Selectivity of Thalamic Input to Simple Cells of Cat Visual Cortex. Nature. 1996;380(6571):249–252. doi:10.1038/380249a0.

86. Biggiogera J, Sanzeni A. Feature Tuning and Network Dynamics in Mouse Visual Cortex: Insights from Connectomics; 2023. Available from: `http://doi.org/10.12751/nncn.bc2023.217`.

87. Sacramento J, Wichert A, Van Rossum MCW. Energy Efficient Sparse Connectivity from Imbalanced Synaptic Plasticity Rules. PLOS Computational Biology. 2015;11(6):e1004265. doi:10.1371/journal.pcbi.1004265.

88. Welzel O, Henkel AW, Stroebel AM, Jung J, Tischbirek CH, Ebert K, et al. Systematic Heterogeneity of Fractional Vesicle Pool Sizes and Release Rates of Hippocampal Synapses. Biophysical Journal. 2011;100(3):593–601. doi:10.1016/j.bpj.2010.12.3706.

89. Holler S, Köstinger G, Martin KAC, Schuhknecht GFP, Stratford KJ. Structure and Function of a Neocortical Synapse. Nature. 2021;591(7848):111–116. doi:10.1038/s41586-020-03134-2.

90. Kirischuk S, Grantyn R. Inter-Bouton Variability of Synaptic Strength Correlates With Heterogeneity of Presynaptic Ca $^{2+}$ Signals. Journal of Neurophysiology. 2002;88(4):2172–2176. doi:10.1152/jn.2002.88.4.2172.

91. Loebel A. Multiquantal Release Underlies the Distribution of Synaptic Efficacies in the Neocortex. Frontiers in Computational Neuroscience. 2009;3. doi:10.3389/neuro.10.027.2009.

92. Silver RA, Lübke J, Sakmann B, Feldmeyer D. High-Probability Uniquantal Transmission at Excitatory Synapses in Barrel Cortex. Science. 2003;302(5652):1981–1984. doi:10.1126/science.1087160.

93. Gray RM. Toeplitz and Circulant Matrices: A Review. Foundations and Trends in Communications and Information Theory. 2006;2(3):155–239.

94. Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE. On the LambertW Function. Advances in Computational Mathematics. 1996;5(1):329–359. doi:10.1007/BF02124750.

# Supporting information

## S1 Dynamics of excitatory networks

{sec:S_network_dynamics}

The dynamics of our networks without SFA and explicit inhibition read for constant input $r_i(t) = r_i$

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_j W_{ij}^{\mathrm{rec}} x_j(t) + \sum_j W_{ij}^{\mathrm{ff}} r_j, \tag{37}$$

{{eq:S_network_dynamics}}

see eq. (4). In the following, we discuss stationary states and time-dependent behaviors of these dynamics.

We first verify that the recurrent network eq. (5), where $W_{ij}^{\mathrm{rec}} = (\delta_{i+1,j} + \delta_{i-1,j})/(\gamma + \gamma^{-1})$ and $W_{ij}^{\mathrm{ff}} = \delta_{ij}\left(1 - 2\gamma/(\gamma + \gamma^{-1})\right)$, has the desired stationary state $x_i^{\mathrm{resp}} =$

$\sum_j \gamma^{|i-j|} r_j$. For this we insert eq. (1) into eq. (5),

$$\begin{aligned}
\tau \dot{x}_i &= -\sum_j \gamma^{|i-j|} r_j + \frac{1}{\gamma+\gamma^{-1}} \left( \sum_j \gamma^{|i+1-j|} r_j + \sum_j \gamma^{|i-1-j|} r_j \right) \\
&\quad + \left(1 - \frac{2\gamma}{\gamma+\gamma^{-1}}\right) r_i && (38) \\
&= -\sum_j \gamma^{|i-j|} r_j + \frac{1}{\gamma+\gamma^{-1}} \left( \sum_{j<i+1} \gamma^{i+1-j} r_j + \sum_{j\geq i+1} \gamma^{j-i-1} r_j \right. \\
&\quad \left. + \sum_{j\leq i-1} \gamma^{i-1-j} r_j + \sum_{j>i-1} \gamma^{j-i+1} r_j \right) + \left(1 - \frac{2\gamma}{\gamma+\gamma^{-1}}\right) r_i && (39) \\
&= -\sum_j \gamma^{|i-j|} r_j + \frac{1}{\gamma+\gamma^{-1}} \left( \gamma \sum_{j\leq i} \gamma^{i-j} r_j + \gamma^{-1} \sum_{j>i} \gamma^{j-i} r_j \right. \\
&\quad \left. + \gamma^{-1} \sum_{j<i} \gamma^{i-j} r_j + \gamma \sum_{j\geq i} \gamma^{j-i} r_j \right) + \left(1 - \frac{2\gamma}{\gamma+\gamma^{-1}}\right) r_i && (40) \\
&= -\sum_j \gamma^{|i-j|} r_j + \frac{1}{\gamma+\gamma^{-1}} \left( \gamma \sum_j \gamma^{|i-j|} r_j + \gamma r_i \right. \\
&\quad \left. + \gamma^{-1} \sum_j \gamma^{|i-j|} r_j - \gamma^{-1} r_i \right) + \frac{\gamma+\gamma^{-1}-2\gamma}{\gamma+\gamma^{-1}} r_i && (41) \quad \texttt{\{\{eq:eom_1D_steady_state_line_4\}\}} \\
&= \frac{\gamma-\gamma^{-1}}{\gamma+\gamma^{-1}} r_i + \frac{\gamma^{-1}-\gamma}{\gamma+\gamma^{-1}} r_i = 0 && (42) \quad \texttt{\{\{eq:eom_1D_steady_state\}\}}
\end{aligned}$$

The computation shows that the recurrent input from neurons $i+1$ and $i-1$ add, in the steady state, up to nearly generate the desired response of neuron $i$: the first and third summand in the bracket in eq. (41) give the desired response. The input that is missing, $\frac{\gamma-\gamma^{-1}}{\gamma+\gamma^{-1}} r_i < 0$, is contributed by the feedforward input (last summand in each line, which cancels the missing input in eq. (42)).

We now turn to non-stationary solutions. The recurrent weight matrices in our recurrent networks eq. (5) are real symmetric matrices and therefore diagonalizable with real eigenvalues and orthogonal eigenvectors. We denote by $\hat{x}^\mu$, $\mu = 0, \cdots, N-1$, the $\mu$th normalized eigenvector of $W^{\text{rec}}$ with eigenvalue $\alpha^\mu$, and by $x^\mu(t) = \hat{x}^{\mu T} x(t)$ the projection of the network activity on this eigenvector (a scalar). Any activity $x(t)$ can be expressed as a linear combination of the orthonormal eigenvectors $\hat{x}^\mu$ with coefficients $x^\mu(t)$. $\hat{x}^\mu$ is also an eigenvector of $W^{\text{rec}} - \mathbb{1}$, with eigenvalue $\alpha^\mu - 1$. Multiplying eq. (37) with $\hat{x}^{\mu T}$ from the left shows that the evolution of network activity can be separated into the evolution of $N$ individual components,

$$\tau \dot{x}^\mu(t) = -x^\mu(t) + \alpha^\mu x^\mu(t) + \sum_{i,j} \hat{x}_i^\mu W_{ij}^{\text{ff}} r_j, \tag{43}$$

$$\dot{x}^\mu(t) = -\lambda^\mu x^\mu(t) + \frac{1}{\tau} \sum_{i,j} \hat{x}_i^\mu W_{ij}^{\text{ff}} r_j, \tag{44}$$ {{eq:S_network_dynamics_components}}

see, e.g., [41]. Here we defined

$$\lambda^\mu = \frac{1-\alpha^\mu}{\tau} \tag{45}$$ {{eq:lambda mu}}

as the decay rate of network activity in the $\mu$th eigenmode. The response time of the network equals $1/\lambda^\mu$, if it is initialized in the $\mu$th eigenmode. When initialized by $x(0)$, the state decays to a stationary value in the different components,

$$x^\mu(t) = \mathrm{e}^{-\lambda^\mu t} x^\mu(0) + (1 - \mathrm{e}^{-\lambda^\mu t}) x^{*\mu}. \tag{46}$$ {{eq:S_free_evolution_eigstate}}

Here $x^\mu(0) = \hat{x}^{\mu T} x(0)$ are the initial and $x^{*\mu} = \hat{x}^{\mu T} x^* = \frac{1}{\lambda^\mu \tau} \sum_{i,j} \hat{x}_i^\mu W_{ij}^{\mathrm{ff}} r_j$ the final stationary values of the $\mu$th component in the eigenbasis. The vector of the stationary dynamics $x^*$ has the entries $x_i^* = \sum_{j=1}^N \mathrm{RF}_{ij} r_j$, if the network generates the desired stationary dynamics. In the absence of recurrent input we have $W^{\mathrm{rec}} = 0$ and thus $\alpha^\mu = 0$, such that the activity decays at a rate of $\lambda = 1/\tau$ to its target. Positive eigenvalues $\alpha^\mu > 0$ mean a slower decay. At $\alpha^\mu = 1$ there is no decay at all, and for $\alpha^\mu > 1$ activity diverges. Stable activity thus requires the largest eigenvalue of the recurrent weight matrix to be smaller than one, such that together with the individual intrinsic decay of each neuron the dynamics are a contraction.

Because the system is linear, the different eigenmodes decay independently from each other at different rates, following eq. (46). Thus with time the faster-decaying modes connected to smaller eigenvalues become exponentially suppressed relative to the dominant, slowest-decaying mode.

The recurrent weight matrices in our recurrent networks eq. (5) furthermore have the property that they are circulant matrices, i.e. each row is equal to the row before rotated one element to the right. The eigenvalues of such matrices are given by the explicit formula [93]

$$\alpha^\mu = \sum_{j=1}^N W_{1j}^{\mathrm{rec}} \cdot \mathrm{e}^{\frac{2\pi I}{N}(j-1)\mu}, \tag{47}$$ {{eq:S_circulant_eigvals}}

where $I$ is the imaginary unit. For purely excitatory networks, we have $W_{1j}^{\mathrm{rec}} \geq 0$, such that $\alpha^\mu$ is maximal if $\mu = 0$, as otherwise the real part of each addend in eq. (47) is smaller or equal. The corresponding eigenvector of $W^{\mathrm{rec}}$ is

$$\hat{x}^0 = \frac{1}{\sqrt{N}} (1, 1, 1, ..., 1)^T. \tag{48}$$ {{eq:hat x 0}}

Therefore, we obtain the slowest convergence for

$$\alpha^{\max} = \alpha^0 = w_{\mathrm{sum}}^{\mathrm{rec}} = \sum_{j=1}^N W_{ij}^{\mathrm{rec}}, \tag{49}$$

where $w_{\mathrm{sum}}^{\mathrm{rec}}$ is the row sum of the recurrent coupling matrix, which is independent of $i$ as the matrix is circulant. The slowest exponential decay dominates the behavior for longer times. Inserting $\alpha^{\max}$ into eq. (45) thus yields $\tau_{\mathrm{resp}}$, the generic time scale of convergence of the network dynamics eq. (5) to the target state $x_i = \sum_j \mathrm{RF}_{ij} r_j$,

$$\tau_{\mathrm{resp}} = \frac{\tau}{1 - \alpha^0} = \frac{\tau}{1 - w_{\mathrm{sum}}^{\mathrm{rec}}}. \tag{50}$$ {{eq:tau_resp}}

## S2 Loss evolution of excitatory networks

{sec:S_evolution_of_loss_net}

In this section, we compute the time evolution of the network loss for our networks without SFA and explicit inhibition. We assume that the networks receive constant input and have initial state $x_i(0) = 0$. All convergence time scales that are present in the network (cf. eqs. (45) and (47)) could in principle contribute to this time evolution. We

will however see, that the decay time of the loss is equal to that of the slowest-decaying mode.

We define the loss as the 1-norm of the deviation of the network activity from the target activity,

$$L(t) = \frac{1}{N}\sum_{i=1}^{N} |x_i(t) - x_i^*|. \tag{51}$$

Under the assumption that network activity does not 'overshoot', i.e. that it is always lower than or equal to the target activity, $x_i(t) \leq x_i^* \ \forall t$, we can replace the loss function by a linear loss function

$$L^{\text{lin}}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i^* - x_i(t). \tag{52}$$ {{eq:S_L_lin_definition}}

We can express the linearized loss as the (scaled) projection of the deviation of the activity from its target, $x^* - x(t)$, onto the eigenvector $\hat{x}^0$ eq. (48),

$$L^{\text{lin}}(t) = \frac{1}{\sqrt{N}}\hat{x}^{0T}\left(x^* - x(t)\right) = \frac{1}{\sqrt{N}}\left(x^{*0} - x^0(t)\right). \tag{53}$$ {{eq:S_L_lin_components}}

The linear loss thus has a single exponential decay. This is a consequence of the fact that $(1, 1, 1, ...1)^T$ is an eigenvector of the dynamics, which, in turn, holds because the row sum of a circulant matrix is the same for each row. The linear loss therefore decays with the time constant $\tau_{\text{resp}}$ obtained in eq. (50) to its stationary value, 0.

## S3 Cooperative coding minimizes the number of required synapses in 1D excitatory networks

{sec:S_least_synapses_proof}

The cooperatively coding recurrent networks eq. (5) and, with SFA, eq. (17) generate the required stationary dynamics eq. (1) with three synapses per neuron. Here we show that this is the minimal number of synapses. Specifically, we show that it is impossible to construct the RF of one neuron as the sum of only one or two other RFs and/or feedforward inputs.

Any network of the form eq. (4) that correctly implements the desired target response has RFs that satisfy eq. (8),

$$\text{RF}_{ij} = W_{ij}^{\text{ff}} + \sum_k W_{ik}^{\text{rec}}\text{RF}_{kj}, \tag{54}$$

i.e. $\text{RF}_{i\cdot}$ is a sum of localized peaks from feedforward inputs and extended two-sided exponentials from recurrent inputs (see Fig. 2c) for an example). First, there needs to be at least one (nonzero) feedforward synapse, as otherwise the network could not respond to input neuron activity. Clearly two feedforward synapses cannot solve the problem for RFs with $n_{\text{RF}} > 2$. But the sum of a localized peak from a feedforward input and the extended response from a recurrent input (with nonzero coefficients) cannot match the target shape (because the difference between the target RF and another RF is either zero or has extended support). Thus every neuron needs at least three synapses to implement the target network response.

## S4 Spike frequency adaptation

{sec:S_spike_frequency_adaptation}

Here we study the dynamics of a 1D network with SFA, which are given by eq. (17). Fig. 8a) shows the dynamic response of neuron $j_0$, whose preferred input is presented

as isolated unit input. The initial dynamics are faster than for a network without SFA. This is because the adaptation current effectively reduces the membrane time constant and because it does not yet fully compensate for the stronger recurrent and feedforward weights, as it is slow (see eqs. (14) and (16)).

To quantify the change in response speed, we compute the integrated loss, as a fast convergence implies that the integrated loss is small. We normalize the loss, $\mathrm{Loss}(t) = |x(t) - x^*|_1/|x^*|_1$ to render it comparable for different RF widths. We define the integrated loss as the temporal mean of the normalized loss. Fig. 8d) shows the dependence of the integrated loss on the SFA time constant. The left-most data point at $\tau_{\mathrm{SFA}} = 0$ corresponds to the zero-lag limit where $u_i(t) = x_i(t)$, which effectively modifies the leak current to $-(1 + a_{\mathrm{SFA}})x_i(t)$ (eq. (16)), resulting in a reduced neuronal time constant $\tau \to \tau/(1 + a_{\mathrm{SFA}})$. Consequently, in this limit and for $a_{\mathrm{SFA}} = 1$, the response time of the SFA network is half as large as that of a network without SFA, which is $\tau_{\mathrm{resp}} = \tau/(1 - w^{\mathrm{rec}}_{\mathrm{sum}}) \approx 201$, eq. (11). Also the integrated loss is smaller; without SFA it would be $\approx 0.413$. The negative slope at $\tau_{\mathrm{SFA}} = 0$ and the clear minimum at $\tau_{\mathrm{SFA}} \approx 0.87$ show that the network can increase its response speed beyond this effective reduction of $\tau$ by utilizing the larger recurrent and feedforward weights, eq. (15), which are not yet fully compensated during equilibration. For oscillating dynamics, $\tau_{\mathrm{resp}}$ can jump discontinuously. This happens if the time point at which all later losses are below $\mathrm{e}^{-1}$ jumps from one oscillation peak to the next. Therefore we use the integrated loss here as the minimization target.

Fig. 8f) shows the integrated loss as a function of $a_{\mathrm{SFA}}$ and $w^{\mathrm{rec,I}}_{\mathrm{sum}}$ in networks that combine SFA and balancing inhibition. The integrated loss is minimized in networks with critical inhibitory strength and without SFA. Larger adaptation can partly compensate weaker inhibition; the data suggest that the optimal values of the adaptation and inhibition parameters satisfy the relation $a_{\mathrm{SFA}}\tau_{\mathrm{SFA}} + w^{\mathrm{rec,I}}_{\mathrm{sum}}/w^{\mathrm{rec,I}}_{\mathrm{sum},c} \approx 1$.

## S5 Effective residual inhibitory interaction strength

{sec:S effective interaction strength}

The residual inhibitory interaction term in the dynamics eq. (22) reads

$$\tau\dot{x}_i = \ldots - \sum_{j=1}^{N} W^{\mathrm{rec,I}}_{ij}\Delta x_j(t) + \ldots. \tag{55}$$

{{eq:residual interaction}}

In the following we compute the effective strength of the interaction mediated by this term, which we define as the total, integrated contribution to the state change of feature neuron $i$ that it causes. For this we assume that the presynaptic activity $x_j$ of neuron $j$ changes only for a limited amount of time, i.e. its derivative has limited support. Concretely we assume that $\dot{x}_j(t) = 0$ for $|t| > T - \tau_{\mathrm{lag}}$ for some finite time $T$. The component of the overall state change $\delta x_i = x_i(T) - x_i(-T)$ in neuron $i$ between $-T$ and $T$ that is caused by the residual inhibitory term due to changes in neuron $j$ is then

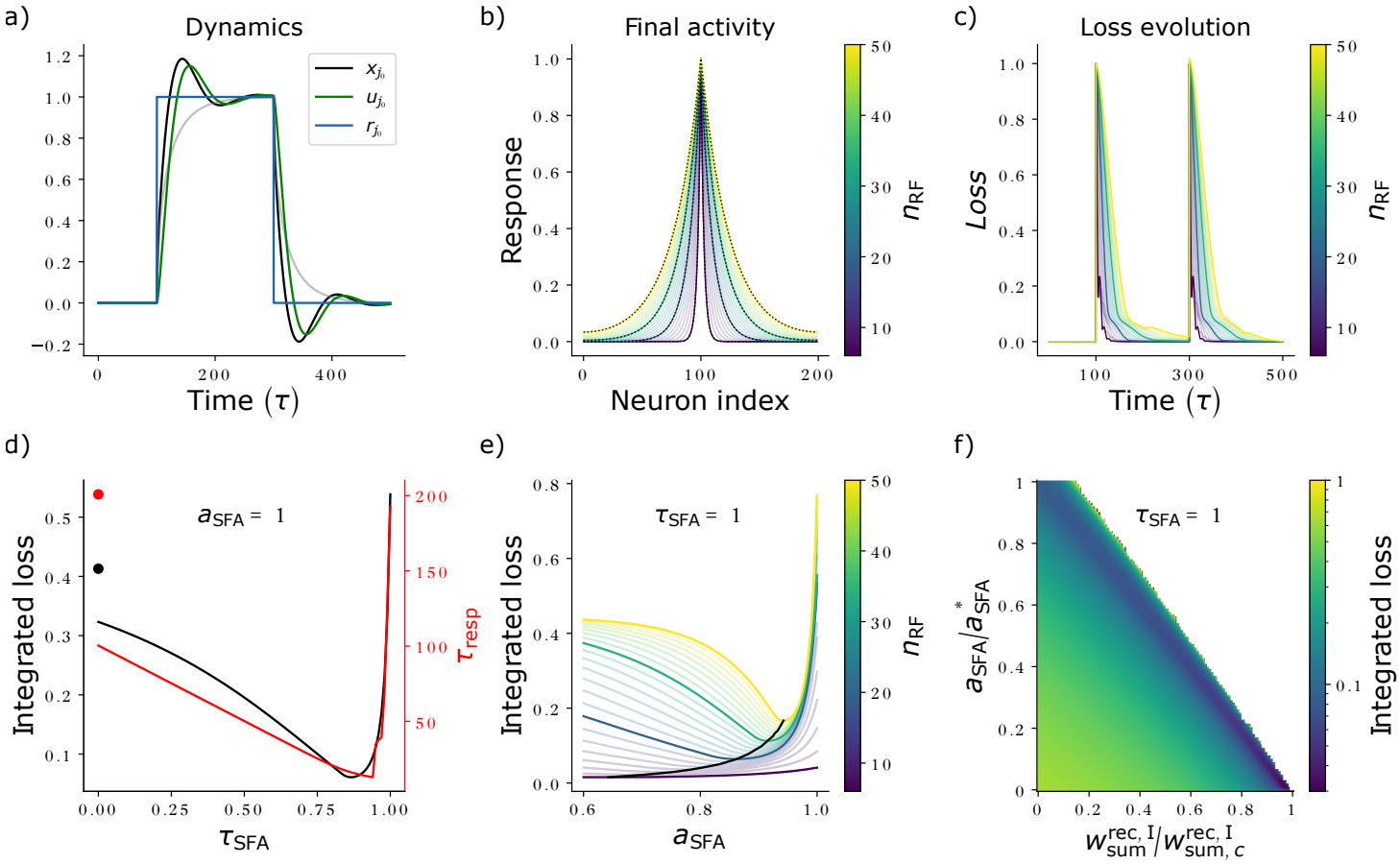


Fig 8: **Network and loss dynamics in a 1D network with SFA. a)** Dynamics of activity ($x_{j_0}(t)$, black), adaptation variable ($u_{j_0}(t)$, green) and input ($r_{j_0}(t)$, blue) for the $j_0 = 100$th neuron, which receives its preferred input as isolated unit input that is switched on at $t = 100\tau$ and off at $t = 300\tau$ ($r_j(t) = \delta_{jj_0}$ between these times). The network implements an RF with $d = 4.5$ ($n_{\mathrm{RF}} = 10$) and has, for better illustration, a rather slow $\tau_{\mathrm{SFA}} = 10\tau$ and slightly stronger-than-optimal $a_{\mathrm{SFA}} = 0.09$, resulting in a visible lag between $x(t)$ and $u(t)$ and oscillatory dynamics. During the initial rising phase ($t \gtrsim 100\tau$), activity rises faster than for a network without SFA (gray line). **b)** Stationary activity of SFA networks with different receptive field width. The networks receive an isolated input (parameters are $\tau_{\mathrm{SFA}} = \tau$ and optimal $a_{\mathrm{SFA}} = a^*_{\mathrm{SFA}}$, see e)). The stationary activity matches its target for different $n_{\mathrm{RF}}$ (color-coded). To show this, target (black dashed) and final activity (color-coded) of four networks with $n_{\mathrm{RF}} \in 6, 20, 34, 50$ are highlighted. **c)** Evolution of the L1-loss of the deviation of network activity from its target, normalized by the L1-norm of the target response for present input, under the stimulus protocol as described in a) for the same networks as in b), with $a_{\mathrm{SFA}}$ determined as described in e). **d)** Integrated loss (black), determined as the temporal mean of the normalized L1-loss shown in c), and response time (red), determined as the earliest time after the onset of input for which the normalized L1-loss drops and stays below $\mathrm{e}^{-1}$, for a network with $d_{\mathrm{RF}} = 10$ ($n_{\mathrm{RF}} = 21$), $a_{\mathrm{SFA}} = 1$ and 101 values of $\tau_{\mathrm{SFA}}$ scanned between 0 and 1. The integrated loss and response time of a network without SFA are shown by colored dots. **e)** Integrated loss for different $n_{\mathrm{RF}}$ (curves are color coded as in b)) as a function of $a_{\mathrm{SFA}}$ (for $\tau_{\mathrm{SFA}} = 1$). The minima, determining $a^*_{\mathrm{SFA}}$ used in b), c) and Fig. 6a), are connected by a black curve, showing that the optimal adaptation strength increases for larger RFs. We note that the data in a),d) and e) suggest that the optimal adaptation parameters fulfill the relation $a_{\mathrm{SFA}}\tau_{\mathrm{SFA}} \lesssim 1$. *(Caption continued on next page)*

{fig:S_SFA}

Fig 8: *(Continued)* **f)** Grid scan of the integrated loss for $a_{\mathrm{SFA}}$ linearly scanned between 0 and $a^*_{\mathrm{SFA}}$ and $w^{\mathrm{rec,I}}_{\mathrm{sum}}$ linearly scanned between 0 and $w^{\mathrm{rec,I}}_{\mathrm{sum},c}$ (101 values each, $d_{\mathrm{RF}} = 10$ ($n_{\mathrm{RF}} = 21$), $\tau_{\mathrm{SFA}} = 1$). The white region indicates parameters for which network activity diverges, identified by an integrated loss larger than 1, cf. also e). Here the loss was computed for the time window with present input only, more precisely for $t \in [100\tau - \mathrm{d}t, 300\tau]$, different from a)-e). For all simulations, we created data by simulating 1D networks with $N = 200$, $\tau = 1$ using the Euler method with step size $\mathrm{d}t = 0.01$; $n_{\mathrm{RF}}$, $\tau_{\mathrm{SFA}}$, $a_{\mathrm{SFA}}$ and $w^{\mathrm{rec,I}}_{\mathrm{sum}}$ varied as described above.

given by

$$\begin{aligned}
\int_{-T}^{T} \mathrm{d}t\ \frac{-W^{\mathrm{rec,I}}_{ij}}{\tau} \Delta x_j(t) &= \frac{-W^{\mathrm{rec,I}}_{ij}}{\tau} \left( \int_{-T}^{T} \mathrm{d}t\ x_j(t) - \int_{-T}^{T} \mathrm{d}t\ x_j(t - \tau_{\mathrm{lag}}) \right) \\
&= \frac{-W^{\mathrm{rec,I}}_{ij}}{\tau} \left( \int_{-T}^{T} \mathrm{d}t\ x_j(t) - \int_{-T-\tau_{\mathrm{lag}}}^{T-\tau_{\mathrm{lag}}} \mathrm{d}\tilde{t}\ x_j(\tilde{t}) \right) \\
&= \frac{-W^{\mathrm{rec,I}}_{ij}}{\tau} \left( \int_{T-\tau_{\mathrm{lag}}}^{T} \mathrm{d}t\ x_j(t) - \int_{-T-\tau_{\mathrm{lag}}}^{-T} \mathrm{d}t\ x_j(t) \right) \\
&= \frac{-W^{\mathrm{rec,I}}_{ij}}{\tau} \Big( \tau_{\mathrm{lag}} x_j(T) - \tau_{\mathrm{lag}} x_j(-T) \Big) \\
&= -\frac{\tau_{\mathrm{lag}}}{\tau} W^{\mathrm{rec,I}}_{ij}\ \delta x_j. \qquad (56)
\end{aligned}$$

{{eq:S_rescaling_of_wri}}

Here we substituted $\tilde{t} = t - \tau_{\mathrm{lag}}$ in the second line; in the third line we used that large parts of the two integrals cancel and in the fourth line that $x_j(t)$ is constant between $T - \tau_{\mathrm{lag}}$ and $T$ as well as between $-T - \tau_{\mathrm{lag}}$ and $-T$. $\delta x_j = x_j(T) - x_j(-T)$ is the total change in presynaptic activity. The result states that a change $\delta x_j$ in presynaptic activity causes a total postsynaptic state change of $-\frac{\tau_{\mathrm{lag}}}{\tau} W^{\mathrm{rec,I}}_{ij}\ \delta x_j$. This may also be intuitively understood as follows: A step-like activity change in $x_j(t_0)$ by $\delta x_j(t_0)$ at time $t_0$ increases $\Delta x_j(t)$ by $\delta x_j(t_0)$ for all $t$ with $t_0 \leq t \leq t_0 + \tau_{\mathrm{lag}}$. According to eq. (55) it thus changes $\dot{x}_i(t)$ by $-1/\tau W^{\mathrm{rec,I}}_{ij} \delta x_j(t_0)$ for a duration of $\tau_{\mathrm{lag}}$. The integrated effect is thus $-\tau_{\mathrm{lag}}/\tau W^{\mathrm{rec,I}}_{ij} \delta x_j(t_0)$. A continuous change of $x_j(t)$ may be seen as assembled of many small step-like ones, which together sum to $\delta x_j$ and thus have the integrated effect eq. (56).

In the limit of small $\tau_{\mathrm{lag}}$, the residual inhibitory interaction term transmits the temporal derivative of the neuronal dynamics in excitatory manner, adding a contribution with the same sign to the temporal derivative of the postsynaptic neuron. To see this, we scale the inhibitory weights with $\tau_{\mathrm{lag}}$ in such a way that the total, integrated effect of the residual inhibition due to a presynaptic activity change at $t$ is independent of the lag between excitatory and inhibitory activity, $\tau_{\mathrm{lag}}$, i.e. we set $W^{\mathrm{rec,I}}_{ij} = \frac{\tau}{\tau_{\mathrm{lag}}} c_{ij}$ with constant $c_{ij}$, compare eq. (56). In the limit of short lag the delayed interaction term then becomes $\lim_{\tau_{\mathrm{lag}} \to 0} -W^{\mathrm{rec,I}}_{ij} \Delta x_j(t) = \lim_{\tau_{\mathrm{lag}} \to 0} -\tau c_{ij} \Delta x_j(t)/\tau_{\mathrm{lag}} = -\tau c_{ij} \dot{x}_j(t)$. The prefactor $-\tau c_{ij} > 0$ is positive, which renders the coupling excitatory.

## S6 Loss evolution of balanced networks

{sec:S_evolution_of_loss_full}

To analytically estimate the evolution of the loss in our networks with balancing, delayed inhibition, we use again the linearized L1 loss eq. (52)), i.e. we assume again that the activity is always lower than or equal to the target activity. In the purely excitatory network, the linearized loss satisfied some simple dynamical equations, eq. (53) and

eq. (44). We will see in the following that this also holds in our balanced networks. To obtain the dynamical equation we compute the temporal derivative of the loss, using the linear time evolution of the activityeqs. (22) and (24). We assume that the input is constant, $r_j(t) = r_j$, and use the knowledge that for such input the dynamics converge to a stationary target state $x_i^*$. This yields

$$\begin{aligned}\tau \dot{L}^{\text{lin}}(t) &= \frac{\tau}{N} \sum_{i=1}^{N} \frac{\mathrm{d}}{\mathrm{d}t} \left(x_i^* - x_i(t)\right) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \Big( - x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec,net}} x_j(t) \\ &\qquad - \sum_{j=1}^{N} W_{ij}^{\text{rec,I}} \Delta x_j(t) + \sum_{j=1}^{N} W_{ij}^{\text{ff}} r_j \Big). \end{aligned} \tag{57}$$

{{eq:S_loss_dynamics_general_0}}

We now specialize the computation further by taking into account that in the cases of interest for us, $W^{\text{rec}}$, $W^{\text{rec,I}}$ and thus also $W^{\text{rec,net}}$, as well as $W^{\text{ff}}$ are circulant matrices. This implies that the column sums $w_{\text{sum}}^{\text{rec,E}} = \sum_{i=1}^{N} W_{ij}^{\text{rec,E}}$, $w_{\text{sum}}^{\text{rec,I}} = \sum_{i=1}^{N} W_{ij}^{\text{rec,I}}$, $w_{\text{sum}}^{\text{rec,net}} = \sum_{i=1}^{N} W_{ij}^{\text{rec,net}}$ and $w_{\text{sum}}^{\text{ff}} = \sum_{i=1}^{N} W_{ij}^{\text{ff}}$ are independent of the column $j$. Equation (57) thus simplifies to

$$\begin{aligned}\tau \dot{L}^{\text{lin}}(t) &= \frac{1}{N} \sum_{i=1}^{N} x_i(t) - w_{\text{sum}}^{\text{rec,net}} \frac{1}{N} \sum_{j=1}^{N} x_j(t) \\ &\quad + w_{\text{sum}}^{\text{rec,I}} \frac{1}{N} \sum_{j=1}^{N} \Delta x_j(t) - w_{\text{sum}}^{\text{ff}} \frac{1}{N} \sum_{j=1}^{N} r_j. \end{aligned} \tag{58}$$

{{eq:S_loss_dynamics_general_1}}

The networks that we want to track analytically start with $x_i(0) = 0$, such that the initial linear loss is $L^{\text{lin}}(0) = \sum_{i=1}^{N} x_i^* - 0$. To describe the network loss dynamics, we can thus use

$$\frac{1}{N} \sum_{i=1}^{N} x_i(t) = \frac{1}{N} \sum_{i=1}^{N} x_i^* - 0 - \big(x_i^* - x_i(t)\big) = L^{\text{lin}}(0) - L^{\text{lin}}(t). \tag{59}$$

{{eq:ensemble mean of states 1}}

An alike equation holds for $\frac{1}{N} \sum_{i=1}^{N} x_i(t - \tau_{\text{lag}})$, such that

$$\begin{aligned}\frac{1}{N} \sum_{i=1}^{N} \Delta x_i(t) &= \frac{1}{N} \sum_{i=1}^{N} x_i(t) - x_i(t - \tau_{\text{lag}}) \\ &= L^{\text{lin}}(0) - L^{\text{lin}}(t) - \big(L^{\text{lin}}(0) - L^{\text{lin}}(t - \tau_{\text{lag}})\big) \\ &= -\big(L^{\text{lin}}(t) - L^{\text{lin}}(t - \tau_{\text{lag}})\big) = -\Delta L^{\text{lin}}(t), \end{aligned} \tag{60}$$

{{eq:ensemble mean of states 2}}

where we introduced the abbreviation

$$\Delta L(t) = L^{\text{lin}}(t) - L^{\text{lin}}(t - \tau_{\text{lag}}) \tag{61}$$

{{eq:Delta L definition}}

for the difference between the current and the delayed loss. Inserting eq. (59) and eq. (60) into eq. (58) gives

$$\tau \dot{L}^{\text{lin}}(t) = \big(1 - w_{\text{sum}}^{\text{rec,net}}\big)\big(L^{\text{lin}}(0) - L^{\text{lin}}(t)\big) - w_{\text{sum}}^{\text{rec,I}} \Delta L^{\text{lin}}(t) - w_{\text{sum}}^{\text{ff}} \frac{1}{N} \sum_{j=1}^{N} r_j. \tag{62}$$

{{eq:S_loss_dynamics_2}}

To eliminate the explicit occurrence of the inputs, we use that in the stationary state, which is reached for $t \to \infty$, we have $L^{\mathrm{lin}}(\infty) = 0$, $\dot{L}^{\mathrm{lin}}(\infty) = 0$ and $\Delta L^{\mathrm{lin}}(\infty) = 0$. For $t \to \infty$, eq. (62) thus shows that

$$0 = (1 - w_{\mathrm{sum}}^{\mathrm{rec,net}}) L^{\mathrm{lin}}(0) - w_{\mathrm{sum}}^{\mathrm{ff}} \frac{1}{N} \sum_{j=1}^{N} r_j. \tag{63}$$

Employing this in eq. (62) gives our final dynamical equation for the linearized loss in terms of the variable $L^{\mathrm{lin}}$ only,

$$\tau \dot{L}^{\mathrm{lin}}(t) = -\big(1 - w_{\mathrm{sum}}^{\mathrm{rec,net}}\big) L^{\mathrm{lin}}(t) - w_{\mathrm{sum}}^{\mathrm{rec,I}} \Delta L^{\mathrm{lin}}(t). \tag{64}$$ {{eq:S_loss_dynamics}}

We solve this delay differential equation with the ansatz

$$L^{\mathrm{lin}}(t) = L^{\mathrm{lin}}(0) \exp(-\Omega t). \tag{65}$$ {{eq:S_loss_ansatz}}

We note that

$$\Omega = \lambda + \mathrm{i}\omega \tag{66}$$ {{eq:Omega}}

is generally complex. The ansatz implies

$$\begin{aligned} \dot{L}^{\mathrm{lin}}(t) &= -\Omega L^{\mathrm{lin}}(t), \\ \Delta L^{\mathrm{lin}}(t) &= L^{\mathrm{lin}}(t) - L^{\mathrm{lin}}(t - \tau_{\mathrm{lag}}) = (1 - \exp(\Omega \tau_{\mathrm{lag}}))\, L^{\mathrm{lin}}(t) \\ &= -\left(\exp(\Omega \tau_{\mathrm{lag}}) - 1\right) L^{\mathrm{lin}}(t). \end{aligned} \tag{67}$$ {{eq:S_loss_ansatz_DL}}

Inserting eq. (67) into eq. (64) and dividing by $-\tau L^{\mathrm{lin}}(t)$ yields

$$\begin{aligned} -\tau \Omega L^{\mathrm{lin}}(t) &= -(1 - w_{\mathrm{sum}}^{\mathrm{rec,net}}) L^{\mathrm{lin}}(t) + w_{\mathrm{sum}}^{\mathrm{rec,I}} \big(\exp(\Omega \tau_{\mathrm{lag}}) - 1\big) L^{\mathrm{lin}}(t), \\ \Omega &= \frac{1 - w_{\mathrm{sum}}^{\mathrm{rec,net}}}{\tau} - \frac{w_{\mathrm{sum}}^{\mathrm{rec,I}}}{\tau} \big(\exp(\Omega \tau_{\mathrm{lag}}) - 1\big). \end{aligned} \tag{68}$$ {{eq:S_loss_complex_freq}}

We see immediately that in the absence of inhibition, $w_{\mathrm{sum}}^{\mathrm{rec,I}} = 0$, we have $\Omega = \lambda = \frac{1 - w_{\mathrm{sum}}^{\mathrm{rec,net}}}{\tau} = \frac{1}{\tau_{\mathrm{resp}}}$ (cf. eq. (11)), such that the decay rate of the purely excitatory network eq. (4) is recovered, as it has to be. To solve eq. (68) in presence of inhibition, we rewrite it as

$$\Omega \tau_{\mathrm{lag}} = \frac{\tau_{\mathrm{lag}}}{\tau_{\mathrm{resp}}} - \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}} \big(\exp(\Omega \tau_{\mathrm{lag}}) - 1\big), \tag{69}$$

$$\left(\Omega \tau_{\mathrm{lag}} - \frac{\tau_{\mathrm{lag}}}{\tau_{\mathrm{resp}}} - \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}}\right) \exp(-\Omega \tau_{\mathrm{lag}}) = -\frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}}. \tag{70}$$

Substituting

$$\Omega \tau_{\mathrm{lag}} - \frac{\tau_{\mathrm{lag}}}{\tau_{\mathrm{resp}}} - \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}} = -z, \tag{71}$$ {{eq:z definition}}

we obtain

$$z \exp(z) = \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}} \exp\left(\frac{\tau_{\mathrm{lag}}}{\tau_{\mathrm{resp}}} + \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}}\right). \tag{72}$$ {{eq:S_Omega_Equation}}

The branches of the Lambert W function solve $z \exp(z) = \mathrm{RHS}$ for $z$. Applying them to eq. (72) yields $z = W_k(\mathrm{RHS})$, where $W_k$ denotes the $k$th branch and RHS the right hand side of eq. (72),

$$z = W_k\left(\frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}} \exp\left(\frac{\tau_{\mathrm{lag}}}{\tau_{\mathrm{resp}}} + \frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}}\right)\right). \tag{73}$$ {{eq:S_determining_z}}

Resubstituting for $z$ then gives the decay rate $\lambda$ (cf. eq. (66)) and, if an oscillation is present, the oscillation frequency $\omega$ of the linearized loss,

$$\Omega\tau_{\text{lag}} = \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + \frac{\tau_{\text{lag}}}{\tau} w_{\text{sum}}^{\text{rec,I}} - W_k\left(\frac{\tau_{\text{lag}}}{\tau} w_{\text{sum}}^{\text{rec,I}} \exp\left(\frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + \frac{\tau_{\text{lag}}}{\tau} w_{\text{sum}}^{\text{rec,I}}\right)\right). \tag{74}$$ {{eq:S_Omega_LambertW_solution}}

Figs. 5 and 9 show the relevant solutions, obtained from the branches $k = 0$ and $k = -1$.

Similarly to the dynamics of a dampened harmonic oscillator, there is a transition from an exponentially-decaying ('overdamped') regime with two decay rates to an oscillating ('underdamped') regime with a single one. At the transition point we have critical inhibitory strength ('critical damping'). In the oscillatory regime it is the amplitude of the oscillation that decays exponentially. Because the actual error is always smaller or equal to the amplitude, having slightly larger than critical inhibition may be optimal, see Fig. 5a). The solutions corresponding to branches other than $k = 0, -1$ have markedly higher decay rates and are thus of little relevance. They also have high oscillation frequencies with periods smaller than the lag.

We note that in the case of oscillations ($\omega \neq 0$), the linearized loss eq. (65) periodically reaches zero. This corresponds to $x_i(t) = x_i^*$ for all neurons, meaning that all neurons simultaneously go from positive to negative deviations from their target activities, or vice-versa. In the full network model, however, these transitions will occur at different times, so that when one neuron matches its target activity there are others that don't. Therefore, the minima of the oscillation are not at zero but at a final error, see Fig. 5a).

We now determine the critical inhibitory strength and the critical decay rate. These are determined by $z = W_k(\text{RHS})$ (eq. (73), with RHS the right hand side of eq. (72)) having only one real solution $z$. This happens at the branch point $\text{RHS} = -\mathrm{e}^{-1}$ where the 0th and $-1$st branch agree, $z = W_0(-\mathrm{e}^{-1}) = W_{-1}(-\mathrm{e}^{-1}) = -1$ [94]. We first set $\text{RHS} = z\mathrm{e}^z = -\mathrm{e}^{-1}$ to obtain $w_{\text{sum},c}^{\text{rec,I}}$,

$$\frac{\tau_{\text{lag}}}{\tau} w_{\text{sum},c}^{\text{rec,I}} \exp\left(\frac{\tau_{\text{lag}}}{\tau} w_{\text{sum},c}^{\text{rec,I}}\right) = -\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right), \tag{75}$$

$$\frac{\tau_{\text{lag}}}{\tau} w_{\text{sum},c}^{\text{rec,I}} = W_k\left(-\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\right), \tag{76}$$ {{eq:S_wrisc}}

where we have again used a (yet unspecified) branch of the Lambert W function to solve for $w_{\text{sum},c}^{\text{rec,I}}$. We know that the inhibitory strength must be real and negative; this can only hold for $k = 0$ or $k = -1$. It might seem surprising that we obtain two solutions here. The reason is that there are actually two critical points, seeFig. 9a): one with a positive decay rate, and another with a negative decay rate, which thus describes exponential growth. We are interested only in the converging dynamics and thus choose $k = 0$.

Next we resubstitute $\lambda^c\tau_{\text{lag}} = \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + \frac{\tau_{\text{lag}}}{\tau} w_{\text{sum},c}^{\text{rec,I}} - z$ (eq. (71) with $\Omega$ replaced by the critical, real decay rate $\lambda^c$), use $z = -1$ and insert the critical inhibitory strength eq. (76) to obtain the critical decay rate,

$$\lambda^c\tau_{\text{lag}} = \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + \frac{\tau_{\text{lag}}}{\tau} w_{\text{sum},c}^{\text{rec,I}} + 1 \tag{77}$$

$$= 1 + \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + W_0\left(-\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\right). \tag{78}$$ {{eq:S_lambda_crit_exact}}

The response time of the balanced network is the inverse of $\lambda^c$,

$$\tau_{\text{resp}}^{\text{bal},c} = \frac{1}{\frac{1}{\tau_{\text{lag}}} + \frac{1}{\tau_{\text{resp}}} + \frac{1}{\tau_{\text{lag}}} W_0\left(-\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\right)}. \tag{79}$$ {{eq:S_tresp_bal_crit_exact}}

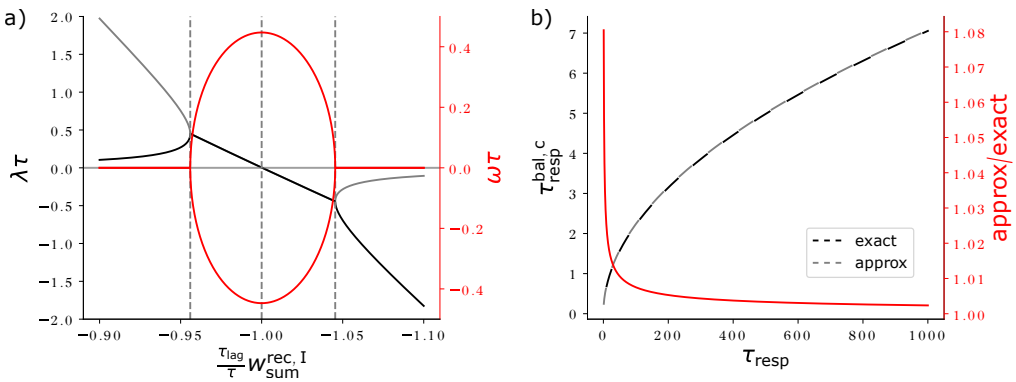


Fig 9: **Critical points and approximation of the analytical solution for $\tau_{\text{resp}}^{\text{bal,c}}$.** **a)** Real part (decay rate $\lambda$, black/gray) and imaginary part (oscillation frequency $\omega$ times $\pm 1$, red) of the complex frequency of the exponential loss evolution, scaled by $\tau$ (compare Fig. 5). The leftmost dashed gray vertical line marks the first critical inhibitory strength at which there is only a single, positive decay rate. For $(\tau_{\text{lag}}/\tau) w_{\text{sum}}^{\text{rec,I}} = -1$, marked by the middle dashed gray vertical line, the decay rate is zero and transitions from positive (decaying) to negative (exponentially growing). There is a second critical inhibitory strength, marked by the rightmost dashed gray vertical line, at which there is a single, negative decay rate. The critical point with the decaying dynamics corresponds to the solution of eq. (76) with $k = 0$, the one with the exponentially growing dynamics to that with $k = -1$. **b)** Exact (black dashed, cf. eq. (79)) and approximate (gray dashed, cf. eq. (88)) values of the critical decay time of the balanced network, $\tau_{\text{resp}}^{\text{bal,c}}$, as a function of that of the excitatory network. Approximation and exact solution agree well. Their ratio (red) is close to one, and approaches one for $\tau_{\text{resp}} \gg \tau_{\text{lag}}$. Parameters: $\tau = 1$, $\tau_{\text{lag}} = 0.1$ and, in a), $w_{\text{sum}}^{\text{rec,net}} = 0.99$.

{fig:S_1_critical_point_approximation}

We now want to derive a more easily interpretable approximation for $\lambda^c$ and $\tau_{\text{resp}}^{\text{bal},c}$ for small $\tau_{\text{lag}}/\tau_{\text{resp}}$. To this end, we modify eq. (78) to

$$\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} = W_0\left(-\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\right) \tag{80}$$

$$\Rightarrow \quad \left(\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\exp\left(\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right) = -\exp\left(-1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right) \tag{81}$$

$$\Leftrightarrow \quad \left(\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\exp\left(\lambda^c \tau_{\text{lag}}\right) = -1. \tag{82}$$

Here we have applied the inverse of $W_0$, $z \to z\mathrm{e}^z$, to both sides, and multiplied with $\exp(-1 - \tau_{\text{lag}}/\tau_{\text{resp}})$. Next we assume that $\lambda^c \tau_{\text{lag}}$ is small, meaning that the response time is much larger than the E-I lag, expand the exponential up to the second power in $\lambda^c$ and $\tau_{\text{lag}}$, and solve for $\lambda^c \tau_{\text{lag}}$,

$$\left(\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)\left(1 + \lambda^c \tau_{\text{lag}} + \frac{1}{2}(\lambda^c \tau_{\text{lag}})^2 + \dots\right) = -1 \tag{83}$$

$$\lambda^c \tau_{\text{lag}} - 1 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} + (\lambda^c \tau_{\text{lag}})^2 - \lambda^c \tau_{\text{lag}} - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\lambda^c \tau_{\text{lag}} - \frac{1}{2}(\lambda^c \tau_{\text{lag}})^2 \approx -1 \tag{84}$$

$$\frac{1}{2}(\lambda^c \tau_{\text{lag}})^2 - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}(\lambda^c \tau_{\text{lag}}) - \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} \approx 0 \tag{85}$$

$$\lambda_{\pm}^c \tau_{\text{lag}} \approx \frac{\tau_{\text{lag}}}{\tau_{\text{resp}}} \pm \sqrt{\left(\frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}\right)^2 + 2\frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}}. \tag{86}$$ {{eq:lambda_taulag_0}}

Here only the positive solution makes sense. Since the delayed inhibition speeds up responses, our assumption that $\lambda^c \tau_{\text{lag}}$ is small implies that also $\tau_{\text{lag}}/\tau_{\text{resp}}$ with the response time of the network without delayed inhibition is small. We can thus neglect in the radicand of eq. (86)'s RHS the quadratic term compared to the linear one. Compared to the resulting square root term we can neglect the first, linear RHS term, such that we obtain

$$\lambda^c \tau_{\text{lag}} \approx \sqrt{2\frac{\tau_{\text{lag}}}{\tau_{\text{resp}}}}. \tag{87}$$ {{eq:S_lambda_crit_approx}}

The response time of the balanced network follows as the inverse of $\lambda^c$,

$$\tau_{\text{resp}}^{\text{bal},c} \approx \sqrt{\frac{\tau_{\text{resp}}\tau_{\text{lag}}}{2}}. \tag{88}$$ {{eq:S_trespc_approx}}

Fig. 9 shows that, despite the simple formula, the quality of the approximation is very good.

## S7 Initial response of balanced networks

{sec:S5_initial_response}

In the following we explain the discrepancy between the analytical and numerical linearized loss evolution, which results in different response times, see Fig. 6a), c). For this, we first note that an exponential function eq. (65), $L^{\text{lin}}(t) = L^{\text{lin}}(0)\exp(-\Omega t)$, with $\Omega$ solving eq. (68) and arbitrary amplitude $L^{\text{lin}}(0)$, is an eigenmode of the time evolution of the linearized loss eq. (64), as the time evolution preserves its functional form. The results of the last section describe the loss evolution if the system is in such an eigenmode. However, in the main text simulations, networks are not initialized in an eigenmode: They are initialized with zero activity and respond to a sudden jump in input. This corresponds to an initial state with $x(t) = 0$ for $t \leq 0$; in particular $\Delta x(0) = 0$. Therefore

both the initial and earlier losses equal the maximal loss, $L(t) = L(-\tau_{\text{lag}}) = |x^*|_1/N$ for $t \leq 0$ and $\Delta L(0) = 0$, see eqs. (60) and (61) and Fig. 10b). In contrast, for the eigenmodes we have $\Delta L^{\text{lin}}(t) = -(\exp(\Omega\tau_{\text{lag}})-1)L^{\text{lin}}(t)$ for $t \leq 0$ (from eqs. (61) and (65)); especially $\Delta L^{\text{lin}}(0) < 0$. The loss is thus not initialized in an eigenmode, but the decay rate of the loss converges to that of the slowliest decaying eigenmode over time.

For the chosen initial conditions the loss decay speeds up during this process, see Fig. 6c and Fig. 10a),b). This can be understood as follows: In our network simulations, activity initially increases, such that $\Delta x_i(t)$ is positive. The amplifying current term $-W^{\text{rec,I}}\Delta x(t)$ ($-W_{ij}^{\text{rec,I}} \geq 0$) in eq. (57) is thus positive, which causes a dynamical speed-up in the activity. Its sum is proportional to $-\Delta L(t)$, which is initially zero. The term $-w_{\text{sum}}^{\text{rec,I}}\Delta L^{\text{lin}}(t) \leq 0$ introduces the dynamical speed up evoked by $\Delta x_i(t)$ into the equation of the loss dynamics, where it causes a faster decay of the loss. As both $\Delta x_i(t)$ and $|\Delta L(t)|$ increase, the terms that generate the speed up in the activity and in the loss dynamics increase in absolute value. As a consequence, the activity increases faster and the loss decays faster. The inset in Fig. 10b) shows that $|\Delta L|$ increases (initially in absolute, later in relative terms) and approaches the same proportionality to $L$ as for the slowliest decaying eigenmode, which is non-oscillatory in the displayed example. This means that the system settles in the eigenmode. The loss curves then become parallel lines on the logarithmic axes (extrapolation of Fig. 10a)).

This different initialization and the resulting initially lower impact of lagged interactions explain why the red data points in Fig. 6 show longer response times than expected from the theory.

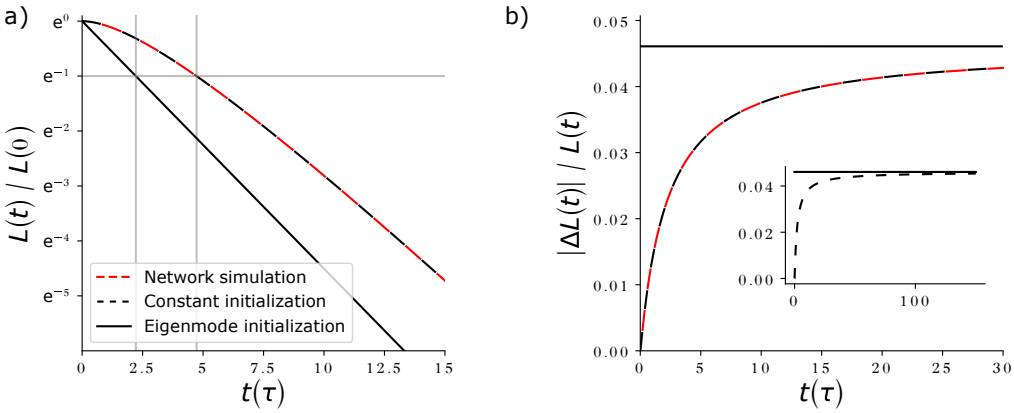


Fig 10: **Effects of initial conditions on loss evolution. a)** Normalized loss evolution obtained by solving eq. (64) (black curves) for critical inhibitory strength. The loss is either initialized in the slowest eigenmode (solid, analytical solution eqs. (65) and (68)), like in our analytical computations, or it is initialized as a constant function (dashed, $L(t) = L(0)$ for $t \leq 0$, numerical solution), like in our network simulations. Vertical lines indicate the time at which the error drops below $\mathrm{e}^{-1}$, which experimentally defines the response time. The theory with constant initialization describes the loss evolution of a numerically simulated network (red dashed, $N = 200$) well. With time, the loss decay rate approaches the same value, independent of initialization; the curves become parallel. **b)** In the for the chosen parameters non-oscillatory slowliest decaying eigenmode, $\Delta L(t)$ is always proportional to $L(t)$ (solid horizontal line at the proportionality factor $1 - \mathrm{e}^{\lambda^c \tau_{\mathrm{lag}}} \approx -0.046$). With constant initialization (dashed curves), $\Delta L(t)$ is initially zero, such that the impact of the interactions mediated by $W^{\mathrm{rec,I}}$ (see eq. (24)) is initially small. $\Delta L(t)$ then tends to the same proportionality to $L(t)$ as for the eigenmode initialization. Inset: loss evolution for long times. Parameters: $w_{\mathrm{sum}}^{\mathrm{rec,net}} = 0.99$, $\tau = 1$, $\tau_{\mathrm{lag}} = 0.1$, $w^{\mathrm{rec,I}} = 0.5 w_{\mathrm{sum},c}^{\mathrm{rec,I}}$; we thus have $\frac{\tau_{\mathrm{lag}}}{\tau} w_{\mathrm{sum}}^{\mathrm{rec,I}} \approx -0.95594$; the displayed slowliest decaying eigenmode is non-oscillatory, cf. Fig. 9. {fig:S_initial_loss_evolution}