Supplementary Math Note

1. Analytic expression of changes of firing rates with learning and application to the covariance learning rule

In the main text, we showed the relation between the synaptic plasticity rule and changes in input currents and network activity with learning (Eq. (1)-(3)). Based on such a relation, here we derive an analytic expression that describes how network activity changes with learning when the transfer functions and the synaptic plasticity rule are known. Then, we apply this expression to an example Hebbian learning rule that cannot reproduce the key feature of changes of activities observed during visual learning (Fig. 2a).

The expression for the changes of neuronal response Δr_i can be found using the following relations

$$\Delta h_i \sim \sum_j \Delta W(r_i, r_j) r_j + \sum_j W_{ij} \Delta r_j,$$
$$\Delta h_i = \Phi^{-1}(r_i + \Delta r_i) - \Phi^{-1}(r_i) \sim \Phi^{-1\prime}(r_i) \Delta r_i.$$

The first equation is Eq. (3) in the main text and the second is obtained from the definition of the transfer function given in Eqs. (1)-(2). If we assume that W_{ij} is either uniform ($W_{ij} = w/N$ where N is the network size), or is random with a mean w/N and independent of Δr_j, then we can obtain an analytic expression for Δr_i,

$$\Phi^{-1\prime}(r_i) \Delta r_i \sim \sum_j \Delta W(r_i, r_j) r_j + \sum_j W_{ij} \Delta r_j \Rightarrow$$

$$\Delta r_i \sim a_i \left(\sum_j \Delta W(r_i, r_j) r_j + \sum_k \frac{a_k w/N}{1 - w/N \sum_l a_l} \sum_j \Delta W(r_k, r_j) r_j \right) \text{ with } a_i = 1/\Phi^{-1\prime}(r_i) \quad \text{(S1)}$$

Next, to build intuition about how a synaptic plasticity rule affects the distribution of firing rates, we consider a linear transfer function $\Phi_i(x) = x$, a uniform initial connectivity matrix $W_{ij} = w/N$ and the `covariance rule'[37] $\Delta W(r_i, r_j) = \alpha(r_i - \text{mean}(r))(r_j - \text{mean}(r))$, where the mean is an average response across the population. Note that this population average is equal to the activity averaged in time over multiple stimuli in a homogeneous network. Then, with a linear transfer function and the covariance learning rule, Eq. (S1) becomes

$$\Delta r_i \sim \left(\sum_j \alpha \left(r_i - m(r) \right) \left(r_j - m(r) \right) r_j + \frac{w/N}{1-w} \sum_{k,j} \alpha \left(r_k - m(r) \right) \left(r_j - m(r) \right) r_j \right)$$
$$= N \alpha \left(r_i - m(r) \right) \text{var(r)}$$

so that the average rate change over the whole population is zero. The above expression shows that initially strong responses are strengthened by this learning rule, while initially weak responses

are weakened. Hence, the covariance rule broadens the distribution of firing rates, leading to an increase in variance. Such a broadening of the distribution of firing rates is expected for any Hebbian synaptic plasticity rule. While this learning rule reproduces qualitatively some aspects of the changes of the distributions of firing rates between familiar and novel stimuli such as the broadening of the distribution, it fails to reproduce another important aspect - the average decrease in rates with familiarity.

2. Proof that the rank preservation assumption minimizes $\sum(\Delta r_i)^2$ among all possible set of Δr_i

We assume that the numbers of novel and familiar stimuli are equal, denoted as *n*, and the firing rates in response to novel and familiar stimuli are denoted as r_i^{nov} and r_i^{fam}, respectively, for $i = 1$ to n. We first show that the rank preservation assumption minimizes $\sum_i \left(r_i^{fam} - r_i^{nov} \right)^2$ for $n = 2$ as follows: we denote the lower firing rates for novel and familiar stimuli as r^{nov} and r^{fam}, respectively, and the higher firing rates as $r^{nov}+\Delta r^{nov}$ and $r^{fam}+\Delta r^{fam}$ with an increment from the lower rates denoted as Δr^{nov} and Δr^{fam}. Then, there are two combinations to match the firing rates, either i) matching the novel and familiar stimuli at the same rank or ii) switching the ranks as matching r^{nov} with $r^{fam}+\Delta r^{fam}$, and $r^{nov}+\Delta r^{nov}$ with r^{fam}. In each case, $\sum_i \left(r_i^{fam} - r_i^{nov} \right)^2 = \sum_i \left(\Delta r_i \right)^2$ becomes

$$\text{i)} \quad \left(r^{fam} - r^{nov} \right)^2 + \left(r^{fam} + \Delta r^{fam} - r^{nov} - \Delta r^{nov} \right)^2$$
$$= 2\left(r^{fam} - r^{nov} \right)^2 + \left(r^{fam} - r^{nov} \right)\left(\Delta r^{fam} - \Delta r^{nov} \right) + \left(\Delta r^{fam} - \Delta r^{nov} \right)^2$$
$$\text{ii)} \quad \left(r^{fam} + \Delta r^{fam} - r^{nov} \right)^2 + \left(r^{fam} - r^{nov} - \Delta r^{nov} \right)^2$$
$$= 2\left(r^{fam} - r^{nov} \right)^2 + \left(r^{fam} - r^{nov} \right)\left(\Delta r^{fam} - \Delta r^{nov} \right) + \left(\Delta r^{fam} \right)^2 + \left(\Delta r^{nov} \right)^2$$

Since the second case has larger $\sum(\Delta r_i)^2$ than the first one, $\sum(\Delta r_i)^2$ is minimized when the rank is preserved for n=2.

Using the case for $n = 2$, we can show that the statement holds for any *n*: we denote the rank of the familiar stimuli matching the novel stimuli with the rank *i* as *k(i)*. Then, if the rank is not preserved, there is some *i* where $k(i)>k(i+1)$. According to the case for $n = 2$, $\sum(\Delta r_i)^2$ gets smaller by changing $k(i)$ and $k(i+1)$. We can continue such a procedure to lower $\sum(\Delta r_i)^2$ until $k(i)$ is an increasing function, that is, $k(i) = i$, where the rank is preserved with learning. Thus, the rank preservation assumption minimizes $\sum(\Delta r_i)^2$ among all possible sets of Δr_i for arbitrary *n*.

3. Comparison of the effects of synaptic plasticity in different recurrent connections

In this section, we explore what types of plasticity in recurrent connections can reproduce experimental observations shown in Figs. 3 and 4. In the networks composed of excitatory (*E*) and inhibitory (*I*) neurons, there are 4 types of connections, *E*-to-*E*, *E*-to-*I*, *I*-to-*E* and *I*-to-*I* connections. Here, we show that to reproduce the increase of maximal response of excitatory neurons, plasticity in connections onto excitatory neurons (*E*-to-*E*, and/or *I*-to-*E* connections) is required, and further find the functional form of plasticity in connections onto excitatory neurons that reproduce changes of firing rates with learning.

First, we claim that plasticity only in *E*-to-*I* or *I*-to-*I* connections cannot explain the input changes observed in excitatory neurons (Fig. 4c). This is because without plasticity in connections onto excitatory neurons, input changes in excitatory neurons arise only from changes of firing rates $\Delta h^E \sim W^{EE}\Delta r^E - W^{EI}\Delta r^I$, which are independent of the post-synaptic firing rates of excitatory neurons when the connectivity matrix W is uniform or random with independent and identically distributed entries. Such constant input changes in excitatory neurons are inconsistent with the experimental observations. On the other hand, input changes observed in inhibitory neurons that are almost constant regardless of the post-synaptic rates (Fig. 4d) can be reproduced without plasticity onto inhibitory neurons but with plasticity onto excitatory neurons, *E*-to-*E* or *I*-to-*E* connections.

Next, we determine the functional forms of the learning rules in the *E*-to-*E*, and *I*-to-*E* connections that are consistent with data. As explained in equation (4) in the main paper, with plasticity in the *E*-to-*E* connections and under the assumption of its separable form as $\Delta W^{EE} = f^{EE}_{post}(r^E_i) f^{EE}_{pre}(r^E_j)$, the post-synaptic dependence can be found as

$$\Delta h^E \sim \Delta W^{EE} r^E + W^{EE}\Delta r^E - W^{EI}\Delta r^I$$
$$\Rightarrow f^{EE}_{post}(r^E_i) = \left(\Delta h^E - W^{EE}\Delta r^E + W^{EI}\Delta r^I\right) \Big/ \left(\sum_j f^{EE}_{pre}\left(r^E_j\right) r^E_j\right)$$

Thus, $f^{EE}_{post}(r^E_i)$ has a similar form to $\Delta h^E(r^E_i)$, showing depression for low rates and potentiation for high rates.

Similarly, in the case with the plasticity in *I*-to-*E* connections having a product form $\Delta W^{EI} = f^{EI}_{post}(r^E_i) f^{EI}_{pre}(r^I_j)$, the post-synaptic dependence can be expressed as

$$\Delta h^E \sim -\Delta W^{EI} r^I + W^{EE}\Delta r^E - W^{EI}\Delta r^I$$
$$\Rightarrow f^{EI}_{post}(r^E_i) = -\left(\Delta h^E - W^{EE}\Delta r^E + W^{EI}\Delta r^I\right) \Big/ \left(\sum_j f^{EI}_{pre}\left(r^I_j\right) r^I_j\right)$$

Thus, unless $\sum_j f_{pre}^{EI}\left(r_j^I\right) r_j^I$ is negative, $f_{post}^{EI}(r_i^E)$ has a similar form to $-\Delta h^E(r_i^E)$, showing potentiation for low rates and depression for high rates, that is, an anti-Hebbian learning rule.

However, the form of plasticity of inhibitory to excitatory connections suggested theoretically (Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569-1573 (2011)) and experimentally (Vogels, T.P., *et al.* Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Frontiers in neural circuits* **7**, 119 (2013); D'Amour J, A. & Froemke, R.C. Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* **86**, 514-528 (2015)) is Hebbian, which would lead to exactly the opposite changes in firing rates from those observed in the data (increase in rates for initially low rate responses, and decrease for initially high rate responses), consistent with their hypothesized homeostatic role. Our results therefore suggest we can rule out a dominant role of Hebbian plasticity from *I*-to-*E* connections. Therefore plasticity in *E*-to-*E* connections is the simplest, yet biological plausible scenario that is consistent with experimental observations.

4. Effects of bounds and other constraints on synaptic weights

In the derivation of the dependence of a learning rule on the post-synaptic firing rates (Eq. (1)-(4)), the bounds on the synaptic weights were not taken into consideration. Ignoring such bounds may lead to unrealistic situations after learning a stream of novel stimuli, such as negative synaptic weights or unbounded growth of synaptic weights. Thus, we need to introduce bounds and/or other constraints on synaptic weights to maintain synaptic strengths in a proper range. Consequently, we also need to examine their effects on the learning rule and the changes of firing rates.

We consider two types of constraints on the synaptic weights under which either the total sum of synaptic weights onto post-synaptic neurons (a constant total presynaptic weight constraint) or the total sum of synaptic weights from pre-synaptic neurons (a constant total postsynaptic weight constraint) is preserved. To maintain the total sum, for each learning rule, we subtract mean of synaptic changes[34]. Combined with the assumption that the learning rule is a separable function of pre- and post-synaptic rates as $\Delta W(r_i, r_j) = f_{post}(r_i) f_{pre}(r_j)$, each constraint leads to the modification of synaptic learning rule as

$$f_{post}(r_i) f_{pre}(r_j) - \text{mean}_k\left(f_{post}(r_i) f_{pre}(r_k)\right) = f_{post}(r_i)\left(f_{pre}(r_j) - \text{mean}_k\left(f_{pre}(r_k)\right)\right) \text{ for pre-constraint.}$$

$$f_{post}(r_i) f_{pre}(r_j) - \text{mean}_k\left(f_{post}(r_k) f_{pre}(r_j)\right) = \left(f_{post}(r_i) - \text{mean}_k\left(f_{post}(r_k)\right)\right) f_{pre}(r_j) \text{ for post-constraint.}$$

Under the constant total presynaptic weight constraint, while the pre-synaptic dependence

is replaced with $f_{pre}(r_j) - \text{mean}_k\left(f_{pre}(r_k)\right)$, $f_{post}(r_i)$ derived in Eq. (4) is unchanged, and the experimental data can be reproduced using a suitable function $f_{post}(r_i)$ unless mean of synaptic weights is too close to the bounds. On the other hand, the constant total postsynaptic weight constraint affects the post-synaptic dependence without changing the pre-synaptic dependence. Such a modification of the post-synaptic dependence lead to inconsistencies with data – in particular, it prevents the decrease in mean rates – for instance, if the system is linear with transfer function $\Phi_i(x) = x$, then mean changes of firing rates can be obtained by taking the mean of Eq. (S1), given as

$$\begin{aligned}\text{mean}\left(\Delta r_i\right) &\sim \frac{1}{N}\sum_i\left(\sum_j \Delta W(r_i, r_j) r_j + \frac{w/N}{1-w}\sum_{k.j}\Delta W(r_k, r_j) r_j\right) \\ &= 0 \quad \text{with } \sum_i \Delta W(r_i, r_j) = 0\end{aligned}$$

Thus, in contrast to the constant total postsynaptic weight constraint that modifies the post-synaptic dependence obtained from the experimental data, the constant total presynaptic weight constraint only affects the pre-synaptic dependence which is undetermined from the data, and thus, is able to reproduce the experimental data. Note that the constant total presynaptic weight constraint is consistent with the experimental observation that the summed synaptic surface area is maintained following plasticity induction[35].