# Hebbian Wiring Plasticity Generates Efficient Network Structures for Robust Inference with Synaptic Weight Plasticity

*Running title: Hebbian wiring plasticity*

Naoki Hiratani[1,2*], Tomoki Fukai[2*]

[1]Department of Complexity Science and Engineering, The University of Tokyo, Kashiwa, Chiba, Japan

[2]Laboratory for Neural Circuit Theory, RIKEN Brain Science Institute, Wako, Saitama, Japan

[*]Correspondance:

N.Hiratani@gmail.com (N.H), tfukai@riken.jp (T.F.)

Keywords: synaptic plasticity, synaptogenesis, neural decoding, computational model, connectomics

Number of figures: 8 main figures + 1 supplementary figure

## Abstract

**In the adult mammalian cortex, a small fraction of spines are created and eliminated every day, and the resultant synaptic connection structure is highly nonrandom, even in local circuits. However, it remains unknown whether a particular synaptic connection structure is functionally advantageous in local circuits, and why creation and elimination of synaptic connections is necessary in addition to rich synaptic weight plasticity. To answer these questions, we studied an inference task model through theoretical and numerical analyses. We demonstrate that a robustly beneficial network structure naturally emerges by combining Hebbian-type synaptic weight plasticity and wiring plasticity. Especially in a sparsely connected network, wiring plasticity achieves reliable computation by enabling efficient information transmission. Furthermore, the proposed rule reproduces experimental observed correlation between spine dynamics and task performance.**

## Introduction

The amplitude of excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs), often referred to as synaptic weight, is considered a fundamental variable in neural computation(Bliss and Collingridge, 1993)(Dayan and Abbott, 2005). In the mammalian cortex, excitatory synapses often show large variations in EPSP amplitudes(Song et al., 2005)(Ikegaya et al., 2013)(Buzsáki and

Mizuseki, 2014), and the amplitude of a synapse can be stable over trials(Lefort et al., 2009) and time(Yasumatsu et al., 2008), enabling rich information capacity compared with that at binary synapses(Brunel et al., 2004)(Hiratani et al., 2013). In addition, synaptic weight shows a wide variety of plasticity which depend primarily on the activity of presynaptic and postsynaptic neurons(Caporale and Dan, 2008)(Feldman, 2009). Correspondingly, previous theoretical results suggest that under appropriate synaptic plasticity, a randomly connected network is computationally sufficient for various tasks(Maass et al., 2002)(Ganguli and Sompolinsky, 2012).

On the other hand, it is also known that synaptic wiring plasticity and the resultant synaptic connection structure are crucial for computation in the brain(Chklovskii et al., 2004)(Holtmaat and Svoboda, 2009). Elimination and creation of dendritic spines are active even in the brain of adult mammalians. In rodents, the spine turnover rate is up to 15% per day in sensory cortex(Holtmaat et al., 2005) and 5% per day in motor cortex(Zuo et al., 2005). Recent studies further revealed that spine dynamics are tightly correlated with the performance of motor-related tasks(Yang et al., 2009)(Xu et al., 2009). Previous modeling studies suggest that wiring plasticity helps memory storage (Poirazi and Mel, 2001)(Stepanyants et al., 2002)(Knoblauch et al., 2010). However, in those studies, EPSP amplitude was often assumed to be a binary variable, and wiring plasticity was performed in a heuristic manner. Thus it remains unknown what should be encoded by synaptic connection structure when synaptic weights have a rich capacity for representation, and how such a connection structure can be achieved through a local spine elimination and creation mechanism, which is arguably noisy and stochastic (Kasai et al., 2010).

To answer these questions, we constructed a theoretical model of an inference task. We first studied how sparse connectivity affects the performance of the network by analytic consideration and information theoretic evaluations. Then, we investigated how synaptic weights and connectivity should be organized to perform robust inference, especially under the presence of variability in the input structure. Based on these insights, we proposed a local unsupervised rule for wiring and synaptic weight plasticity. In addition, we demonstrated that connection structure and synaptic weight learn different components under a dynamic environment, enabling robust computation. Lastly, we investigated whether the model is consistent with various experimental results on spine dynamics.

## Results

*Connection structure reduces signal variability in sparsely connected networks*

What should be represented by synaptic connections and their weights, and how are those representations acquired? To explore the answers to these questions, we studied a hidden variable estimation task (**Fig. 1A**), which appears in various stages of neural information processing(Beck et

al., 2008)(Lochmann and Deneve, 2011). In the task, at every time $t$, one hidden state is sampled with equal probability from $p$ number of external states $s^t = \{0,1,\ldots,p\text{-}1\}$. Neurons in the input layer show independent stochastic responses $r_{X,j}^{\ t} \sim N(\theta_{j\mu}, \sigma_X)$ due to various noises (**Fig. 1B** middle), where $\theta_{j\mu}$ is the average firing rate of neuron $j$ to the stimulus $\mu$, and $\sigma_X$ is the constant noise amplitude. Although, we used Gaussian noise for analytical purposes, the following argument is applicable for any stochastic response that follows a general exponential family, including Poisson firing (**Supplementary Fig. 1**). Neurons in the output layer estimate the hidden variable from input neuron activity and represent the variable with population firing $\{r_{Y,i}\}$. This task is computationally difficult because most input neurons have mixed selectivity for several hidden inputs, and the responses of the input neurons are highly stochastic (**Fig. 1C**). Let us assume that the dynamics of output neurons are written as follows:

$$r_{Y,i}^t = r_Y^o \exp\left[\sum_{j=1}^M c_{ij}\left(w_{ij} r_{X,j}^t - h_w\right) - I_{inh}^t\right], \quad I_{inh}^t = \log\left[\sum_{i=1}^N \exp\left(\sum_{j=1}^M c_{ij}\left[w_{ij} r_{X,j}^t - h_w\right]\right)\right], \quad (1)$$

where $c_{ij}$ (= 0 or 1) represents connectivity from input neuron $j$ to output neuron $i$, $w_{ij}$ is its synaptic weight (EPSP size), and $h_w$ is the threshold. $M$ and $N$ are population sizes of the input and output layers, respectively. In the model, all feedforward connections are excitatory, and the inhibitory input is provided as the global inhibition $I_{inh}^{\ t}$.

If the feedforward connection is all-to-all (i.e., $c_{ij}$ = 1 for all $i,j$ pairs), by setting the weights as $w_{ij} = q_{j\mu} \equiv \theta_{j\mu}/\sigma_X^2$ for output neuron $i$ that represents external state $\mu$, the network gives an optimal inference from the given firing rate vector $r_X^{\ t}$, because the value $q_{j\mu}$ represents how much evidence the firing rate of neuron $j$ provides for a particular external state $\mu$. (For details, see *Methods 1.1*). However, if the connectivity between the two layers is sparse, as in most regions of the brain(Potjans and Diesmann, 2014), optimal inference is generally unattainable because each output neuron can obtain a limited set of information from the input layer. How should one choose connection structure and synaptic weights in such a case? Intuitively, we could expect that if we randomly eliminate connections while keeping the synaptic weights of output neuron $i$ that represents external state $\mu$ as $w_{ij} \propto q_{j\mu}$ (below, we call it as weight coding), the network still works at a near-optimal accuracy. On the other hand, even if the synaptic weight is a constant value, if the connection probability is kept at $\rho_{ij} \propto q_{j\mu}$ (i.e. connectivity coding; see *Methods 1.2* for details of coding strategies), the network is expected to achieve near-optimal performance. **Figure 2A** describes the connection matrices between input/output layers in two strategies. In the weight coding, if we sort input neurons with their preferred external states, the diagonal components of the connection matrix show high synaptic weights, whereas in the connectivity coding, the diagonal components show dense connection (**Fig. 2A**). Both of realizations asymptotically converge to optimal solution when the number of neurons in the middle layer is sufficiently large, though in a

finite network, not strictly optimal under given constraints. In addition, both of them are obtainable through biologically plausible local Hebbian learning rules as we demonstrate in subsequent sections.

We evaluated the accuracy of the external state estimation using a bootstrap method (*Methods 3.2*) for both coding strategies. Under intermediate connectivity, both strategies showed reasonably good performance (as in **Fig. 1B** bottom). Intriguingly, in sparsely connected networks, the connectivity coding outperformed the weight coding, despite its binary representation (**Fig. 2B** cyan/orange lines). The analytical results confirmed this tendency (**Fig. 2B** red/blue lines; see *Methods 2.1* for the details) and indicated that the firing rates of output neurons selective for the given external state show less variability in connectivity coding than in the weight coding, enabling more reliable information transmission (**Fig. 2C**). To further understand this phenomenon, we evaluated the maximum transfer entropy of the feed forward connections: $T_E = \left\langle H\left(s^t\right) - H\left(s^t \mid r_X^t, C\right)\right\rangle_t$. Because of limited connectivity, each output neuron obtains information only from the connected input neurons. Thus, the transfer entropy was typically lower under sparse than under dense connections in both strategies (**Fig. 2D**). However, in the connectivity coding scheme, because each output neuron can get information from relevant input neurons, the transfer entropy became relatively large compared to the weight coding (orange line in **Fig. 2D**). Therefore, analyses from both statistical and information theory-based perspectives confirm the advantage of connectivity coding over the weight coding in the sparse regions.

The result above can also be extended to arbitrary feedforward network as below. For a feedforward network of $M$ times $N$ neurons with connection probability $\rho$, information capacity of connections is given as $I_C(\rho) \equiv \log {}_{MN}C_{\rho MN} \approx MN \cdot H(\rho)$, where $H$ represents the entropy function $H(\rho) \equiv -\rho \log \rho - (1-\rho)\log(1-\rho)$. Similarly, for a given connections between two layers, information capacity of synaptic weights is written as $I_w(\rho) \equiv \rho MN \log b$, where $b$ is the number of distinctive synaptic states (Varshney et al., 2006). Therefore, when the connection probability $\rho$ satisfies $b = \exp\left[H(\rho)/\rho\right]$, synaptic connections and weights have the same information capacities. This means that, as depicted in **Figure 2E**, in a sparsely connected network, synaptic connections tend to have larger relative information capacity, compared to a dense network with the same $b$. This result is consistent with the model above, because stochastic firing of presynaptic neuron can be translated as synaptic noise. Furthermore, in the CA3-to-CA1 connection of mice, connection probability is estimated to be around 6% (Sayer et al., 1990), and information capacity of synaptic weight is around 4.7 bits (Bartol et al., 2015), thus the connection structure should also play an active role in neural coding in the real brain (data point in **Fig. 2E**).

*Dual coding by synaptic weights and connections enables robust inference*

In the section above, we demonstrated that a random connection structure highly degrades information transmission in a sparse regime to the degree that weight coding with random connection fell behind connectivity coding with a fixed weight. Therefore, in a sparse regime, it is necessary to integrate representations by synaptic weights and connections, but how should we achieve such a representation? Theoretically speaking, we should choose a connection structure that minimizes the loss of information due to sparse connectivity. This can be achieved by minimizing the KL-divergence between the distribution of the external states estimated from the all-to-all network, and the distribution estimated from a given connection structure (i.e. $\underset{\|C\|_0=\rho MN}{\operatorname{argmin}}\left\langle D_{KL}\left[p(s^t \mid r_X, C_{all}) \,\|\, p(s^t \mid r_X, C)\right]\right\rangle_{r_X}$, see *Methods 2.2* for details). However, this calculation requires combinatorial optimization, and local approximation is generally difficult(Donoho, 2006), thus expectedly the brain employs some heuristic alternatives. Experimental results indicate that synaptic connections and weights are often representing similar features. For example, the EPSP size of a connection in a clustered network is typically larger than the average EPSP size(Lefort et al., 2009)(Perin et al., 2011), and a similar property is suggested to hold for interlayer connections(Yoshimura et al., 2005) (Ryan et al., 2015). Therefore, we could expect that by simply combining the weight coding and connectivity coding in the previous section, low performance at the sparse regime can be avoided. On the other hand, in the previous modeling studies, synaptic rewiring and resultant connection structure were often generated by cut-off algorithm in which a synapse is eliminated if the weight is smaller than the given criteria (Chechik et al., 1998)(Navlakha et al., 2015). Thus, let us next compare the representation by combining the weight coding and connectivity coding (we call it as the dual coding below), with the cut-off coding strategy.

**Figure 3A** describes the synaptic weight distributions in the two strategies, as well as in random connection (see *Methods 1.3* for details of the implementation). When connectivity coding and weight coding are combined (i.e. in the dual coding), connection probability becomes larger in proportion to its synaptic weight (**Fig. 3A** middle), and the resultant distribution exhibits a broad distribution as observed in the experiments (Song et al., 2005)(Ikegaya et al., 2013), whereas in the cut-off strategy, the weight distribution is concentrated at a non-zero value (**Fig. 3A** right). Intuitively, the cut-off strategy seems more selective and beneficial for inference. Indeed, in the original task, the cut-off strategy enabled near-optimal performance, though the dual coding also improved the performance compared to a randomly connected network(**Fig. 3C**). However, under the presence of variability in the input layer, cut-off strategy is no longer advantageous. For instance, let us consider the case when noise amplitude $\sigma_X$ is not constant but pre-neuron dependent. If the

firing rate variability of input neuron $j$ is given by $\sigma_{X,j} \equiv \sigma_X \exp\left(2\zeta_j \log\sigma_r\right)/\sigma_r$, where $\zeta_j$ is a random variable uniformly sampled from [0, 1), and $\sigma_r$ is the degree of variability, in an all-to-all network, optimal inference is still achieved by setting synaptic weights as $w_{ij} = q_{j\mu} \equiv \theta_{j\mu}/\sigma^2_{X,j}$. On the contrary, in the sparse region, the performance is disrupted especially in the cut-off strategy, so that the dual coding outperformed the cut-off strategy (**Fig. 3D**).

To further illustrate this phenomenon, let us next consider a case when a quarter of input neurons show a constant high response for all of the external states as $\tilde{\theta}_{j\mu} = \theta_{const}$, and the rest of input neurons show high response for randomly selected half of external states (i.e. $\Pr\left[\tilde{\theta}_{j\mu} = \theta_{high}\right] = \Pr\left[\tilde{\theta}_{j\mu} = \theta_{low}\right] = \frac{1}{2}$), where $\theta_{low} < \theta_{high} < \theta_{const}$, and $\theta_{j\mu} = \tilde{\theta}_{j\mu}/Z_\mu$ with the normalization factor $Z_\mu = r^o_X \Big/ \sqrt{\sum_{j=1}^{M} \tilde{\theta}_{j\mu}/M}$. Even in this case, $w_{ij} = q_{j\mu} \equiv \theta_{j\mu}/\sigma^2_X$ is the optimal synaptic weights configuration in the all-to-all network, but if we create a sparse network with cut-off algorithm, the performance drops dramatically at certain connectivity, whereas in the dual coding, the accuracy is kept at some high levels even in the sparse connectivity (**Fig. 3E**).

To get insights on why the dual coding is more robust against variability in the input layer, for three input configurations described above, we calculated the relationship between synaptic weight $w_{ij}$ and the information gained by a single synaptic connection $\Delta I_{ij}$. Here, we defined the information gain $\Delta I_{ij}$ by the mean reduction in the KL divergence $\left\langle D_{KL}\left[p(s^t \mid r_X, C_{all}) \,\|\, p(s^t \mid r_X, C)\right]\right\rangle_{r_X}$, achieved by adding one synaptic connection $c_{ij}$ to a randomly connected network $C$ (see *Method 2.2* for details). In the original model, $\Delta I_{ij}$ has nearly a linear relationship with the synaptic weight $w_{ij}$ (gray points in **Fig. 3B**), thus by simply removing the connections with small synaptic weights, a near-optimal connection structure was acquired (**Fig. 3C**). On the other hand, when the input layer is not homogeneous, large synapses tend to have negative (black circles in **Fig. 3B**) or zero (black points in **Fig. 3B**) gains, as a result, the linear relationship between the weight and the information gain was lost. Thus, in these cases, the dual coding is less likely to be disrupted by non-beneficial connections.

Although our consideration here is limited to a specific realization of synaptic weights, in general, it is difficult to represent the information gain by locally acquired synaptic weight, so we could expect that the cut-off strategy is not the optimal connectivity organization in many cases.

*Local Hebbian learning of the dual coding*

The argument in the previous section suggest that, by combining the weight coding and connectivity coding, the network can robustly perform inference especially in sparsely connected regions. However, in the previous sections, a specific connection and weight structure were given a priori, although structures in local neural circuits are expected to be obtained with local weight plasticity and wiring plasticity. Thus, we next investigate whether dual coding can be achieved through a local unsupervised synaptic plasticity rule.

Let us first consider learning of synaptic weights. In order to achieve the weight coding, synaptic weight $w_{ij}$ should converge to $w_{ij} = q_{j\mu}/\sigma_X^2\bar{\rho} = \left\langle r_{X,j}^t r_{Y,i}^t / \left(\sigma_X^2 \bar{\rho} r_{Y,i}^t\right)\right\rangle$ when output neuron $i$ represents external state $\mu$, and $\bar{\rho}$ represents the mean connectivity of the network. Thus, synaptic weight change $\Delta w_{ij} = w_{ij}^{t+1} - w_{ij}^t$ is given as:

$$\Delta w_{ij} = \left(\eta_X/\gamma\right)\left(r_{Y,i}^t\left[r_{X,j}^t - \sigma_X^2\bar{\rho}w_{ij}\right] + b_h\left[r_Y^o/N - r_{Y,i}^t\right]\right). \quad (2)$$

The second term is the homeostatic term heuristically added to constrain the average firing rates of output neurons(Turrigiano and Nelson, 2004). Note that the first term corresponds to stochastic gradient descending on $D_{KL}\left[p^*(r_X^t) \| p(r_X^t \mid C,W)\right]$, because the weight coding approximates the optimal representation by synaptic weights (Nessler et al., 2013)(see *Methods 1.4* for details). We performed this unsupervised synaptic weight learning on a randomly connected network. When the connectivity is sufficiently dense, the network successfully acquired a suitable representation (**Fig. 4A**). Especially under a sufficient level of homeostatic plasticity (**Fig. 4B**), the average firing rate showed a narrow unimodal distribution (**Fig. 4C** top), and most of the output neurons acquired selectivity for one of external states (**Fig. 4C** bottom).

We next investigated the learning of connection structures by wiring plasticity. Unlike synaptic weight plasticity, it is not yet well understood how we can achieve functional connection structure with local wiring plasticity. In particular, rapid rewiring may disrupt the network structure, and possibly worsen the performance (Chechik et al., 1998). Thus, let us first consider a simple rewiring rule, and discuss the biological correspondence later. Here, we introduced a variable $\rho_{ij}$, for each combination (i,j) of presynaptic neuron $j$ and postsynaptic neuron $i$, which represents the connection probability. If we randomly create a synaptic connection between neuron (i,j) with probability $\rho_{ij}/\tau_c$ and eliminate it with probability $(1\text{-}\rho_{ij})/\tau_c$, on average there is a connection between neuron (i,j) with probability $\rho_{ij}$, when the maximum number of synaptic connections is bounded by 1. In this way, the total number of synaptic connections is kept constant on average, without any global regulation mechanism.

From a similar argument done for synaptic weights, the learning rule for connection probability $\rho_{ij}$ is derived as:

$$\Delta\rho_{ij} = \eta_\rho r^t_{Y,i}\left[r^t_{X,j} - \sigma^2_X \rho_{ij} w_o\right], \qquad (3)$$

where $w_o$ is the expected mean synaptic weight (*Methods 1.5*). Under this rule, the connection probabilities converge to the connectivity coding. Moreover, although this rule does not maximize the transfer entropy of the connections, direction of learning is on average close to the direction of the stochastic gradient on transfer entropy. Therefore, the above rule does not reduce the transfer entropy of the connection on average (see *Methods 1.6*).

**Figure 5A** shows the typical behavior of $\rho_{ij}$ and $w_{ij}$ under combination of this wiring rule (equation (3)) and the weight plasticity rule described in equation (2) (we call this combination as the dual Hebbian rule because both equations (2) and (3) have Hebbian forms). When the connection probability is low, connections between two neurons are rare, and, even when a spine is created due to probabilistic creation, the spine is rapidly eliminated (**Fig. 5A** top). In the moderate connection probability, spine creation is more frequent, and the created spine survives longer (**Fig. 5A** middle). When the connection probability is high enough, there is almost always a connection between two neurons, and the synaptic weight of the connection is large because synaptic weight dynamics also follow a similar Hebbian rule (**Fig. 5A** bottom).

We implemented the dual Hebbian rule in our model and compared the performance of the model with that of synaptic weight plasticity on a fixed random synaptic connection structure. Because spine creation and elimination are naturally balanced in the proposed rule (**Fig. 5B** top), the total number of synaptic connections was nearly unchanged throughout the learning process (**Fig. 5B** bottom). As expected, the dual Hebbian rule yielded better performance (**Fig. 5C,D**) and higher estimated transfer entropy than the corresponding weight plasticity only model (**Fig. 5E**). This improvement was particularly significant when the frequency of rewiring was in an intermediate range (**Fig. 5F**). When rewiring was too slow, the model showed essentially the same behavior as that in the weight plasticity only model, whereas excessively frequent probabilistic rewiring disturbed the connection structure. Although a direct comparison with experimental results is difficult, the optimal rewiring timescale occurred within hours to days, under the assumption that firing rate dynamics (equation (1)) are updated every 10 to 100 ms. Initially, both connectivity and weights were random (**Fig. 5G** left), but after the learning process, the diagonal components of the weight matrix developed relatively larger synaptic weights, and, at the same time, denser connectivity than the off-diagonal components (**Fig. 5G** right). Thus, through dual Hebbian learning, the network can indeed acquire a connection structure that enables efficient information transmission between two layers; as a result, the performance improves when the connectivity is moderately

sparse (**Fig. 5D, E**). Although the performance was slightly worse than that of a fully-connected network, synaptic transmission consumes a large amount of energy(Sengupta et al., 2013), and synaptic connection is a major source of noise(Faisal et al., 2008). Therefore, it is beneficial for the brain to achieve a similar level of performance using a network with fewer connections.

*Connection structure can acquire constant components of stimuli and enable rapid learning*

We have shown that the dual coding by synaptic weights and connections robustly helps computation in a sparsely connected network, and the desirable weight and connectivity structures are naturally acquired through the dual Hebbian rule. Although we were primary focused on sparse regions, the rule potentially provides some beneficial effects even in densely connected networks. To consider this issue, we extended the previous static external model to a dynamic one, in which at every interval $T_2$, response probabilities of input neurons partly change. If we define the constant component as $\theta_{const}$ and the variable component as $\theta_{var}$, then the total model becomes $\theta_{j\mu} = \frac{1}{Z}\left[\kappa_m \theta_{j\mu}^{const} + (1-\kappa_m)\theta_{j\mu}^{\mathrm{var}}\right]$, where the normalization term is given as $\frac{1}{MZ^2}\sum_{j=1}^{M}\left[\kappa_m \theta_{j\mu}^{const} + (1-\kappa_m)\theta_{j\mu}^{\mathrm{var}}\right]^2 = \left(r_X^o\right)^2$ (**Fig. 6A**). In this case, when the learning was performed only with synaptic weights based on fixed random connections, although the performance rapidly improved, every time a part of the model changed, the performance dropped dramatically and only gradually returned to a higher level (cyan line in **Fig. 6B**). By contrast, under the dual Hebbian learning rule, the performance immediately after the model shift (i.e., the performance at the trough of the oscillation) gradually increased, and convergence became faster (**Fig. 6B,C**), although the total connectivity stayed nearly the same (**Fig. 6D**). After learning, the synaptic connection structure showed a higher correlation with the constant component than with the variable component (**Fig. 6E**; see *Methods 3.3*). By contrast, at every session, synaptic weight structure learned the variable component better than it learned the constant component (**Fig. 6F**). The timescale for synaptic rewiring needed to be long enough to be comparable with the timescale of the external variability $T_2$ to capture the constant component. Otherwise, connectivity was also strongly modulated by the variable component of the external model (**Fig. 6G**). After sufficient learning, the synaptic weight $w$ and the corresponding connection probability $\rho$ roughly followed a linear relationship (**Fig. 6H**). Remarkably, some synapses developed connection probability $\rho = 1$, meaning that these synapses were almost permanently stable because the elimination probability $(1-\rho)/\tau_c$ became nearly zero.

*Approximated dual Hebbian learning rule reconciles with experimentally observed spine dynamics*

Our results up to this point have revealed functional advantages of dual Hebbian learning. In this last

section, we investigated the correspondence between the experimentally observed spine dynamics and the proposed rule. To this end, we first studied whether a realistic spine dynamics rule approximates the proposed rule, and then examined if the rule explains the experimentally known relationship between synaptic rewiring and motor learning (Yang et al., 2009)(Xu et al., 2009).

Previous experimental results suggest that a small spine is more likely to be eliminated(Yasumatsu et al., 2008)(Kasai et al., 2010), and spine size often increases or decreases in response to LTP or LTD respectively, with a certain delay (Matsuzaki et al., 2004)(Wiegert and Oertner, 2013). In addition, though spine creation is to some extent influenced by postsynaptic activity (Knott et al., 2006)(Yang et al., 2014), the creation is expected to be more or less a random process (Holtmaat and Svoboda, 2009). Thus, changes in the connection probability can be described as

$$\rho_{ij}^{t} = \begin{cases} \rho_{ij}^{t-1} + \eta_{\rho}\left[\gamma^{2} w_{ij} - \rho_{ij}^{t-1}\right] & \left(\text{if } c_{ij} = 1\right) \\ \gamma^{2} w_{o} & \left(\text{if } c_{ij} = 0\right). \end{cases} \tag{4}$$

By combining this rule and the Hebbian weight plasticity described in equation (2), the dynamics of connection probability well replicated the experimentally observed spine dynamics (Yasumatsu et al., 2008)(Kasai et al., 2010) (**Fig. 7A-C**). Moreover, the rule outperformed the synaptic weight only model in the inference task, although the rule performed poorly compared to the dual Hebbian rule due to the lack of activity dependence in spine creation (magenta line in **Fig. 6I**). This result suggests that plasticity rule by equations (2) and (4) well approximates the dual Hebbian rule (equations (2)+(3)). This is because, even if the changes in the connection probability are given as a function of synaptic weight as in equation (4), as long as the weight plasticity rule follows equation (2), wiring plasticity indirectly shows a Hebbian dependency for pre- and postsynaptic activities as in the original dual Hebbian rule (equation (3)). As a result, the approximated rule gives a good approximation of the original dual Hebbian rule.

We next applied this approximated learning rule to motor learning tasks. The primary motor cortex has to adequately read-out motor commands based on inputs from pre-motor regions(Salinas and Romo, 1998)(Sul et al., 2011). In addition, the connection from layer 2/3 to layer 5 is considered to be a major pathway in motor learning(Masamizu et al., 2014). Thus we hypothesized that the input and output layers of our model can represent layers 2/3 and 5 in the motor cortex. We first studied the influence of training on spine survival(Xu et al., 2009) (**Fig. 8A**). To compare with experimental results, below we regarded $10^5$ time steps as one day, and described the training and control phases as two independent external models $\theta_{ctrl}$ and $\theta_{train}$. In both training and control cases, newly created spines were less stable than pre-existing spines (solid lines vs. dotted lines in **Fig. 8B**), because older spines tended to have a larger connection probability (**Fig.**

**7B**). In addition, continuous training turned pre-existed spines less stable and new spines more stable than their respective counterparts in the control case (red lines vs. lime lines in **Fig. 8B**). The 5-day survival rate of a spine was higher for spines created within a couple of days from the beginning of training compared with spines in the control case, whereas the survival rate converged to the control level after several days of training (**Fig. 8C**). We next considered the relationship between spine dynamics and task performance(Yang et al., 2009). For this purpose, we compared task performance at the beginning of the test period among simulations with various training lengths (**Fig. 8D**). Here, we assumed that spine elimination was enhanced during continuous training, as is observed in experiments(Yang et al., 2009)(Xu et al., 2009). The performance was positively correlated with both the survival rate at day 7 of new spines formed during the first 2 days, and the elimination rate of existing spines (left and right panels of **Fig. 8E**). By contrast, the performance was independent from the total ratio of newly formed spines from day 0 to 6 (middle panel of **Fig. 8E**). These results demonstrate that complex spine dynamics are well described by the approximated dual Hebbian rule, suggesting that the brain uses a dual learning mechanism.

## Discussion

In this study, we first analyzed how random connection structures impair performance in sparsely connected networks by analyzing the change in signal variability and the transfer entropy in the weight coding and the connectivity coding strategies (**Fig. 2**). Subsequently, we showed that connection structures created by the cut-off strategy are not beneficial under the presence of input variability, due to lack of positive correlation between the information gain and weight of synaptic connections (**Fig. 3**). Based on these insights, we proposed that the dual coding by weight and connectivity structures as a robust representation strategy, then demonstrated that the dual coding is naturally achieved through dual Hebbian learning by synaptic weight plasticity and wiring plasticity (**Fig. 4, 5**). We also revealed that, even in a densely connected network in which synaptic weight plasticity is sufficient in terms of performance, by encoding the time-invariant components with synaptic connection structure, the network can achieve rapid learning and robust performance (**Fig. 6**). Even if spine creation is random, the proposed framework still works effectively, and the approximated model with random spine creation is indeed sufficient to reproduce various experimental results (**Fig. 7, 8**).

### *Model evaluation*

Spine dynamics depend on the age of the animal(Holtmaat et al., 2005), the brain region(Zuo et al., 2005), and many molecules play crucial roles(Kasai et al., 2010)(Caroni et al., 2012), making it difficult for any theoretical models to fully capture the complexity. Nevertheless, our simple

mathematical model replicated many key features(Yasumatsu et al., 2008)(Yang et al., 2009)(Xu et al., 2009)(Kasai et al., 2010). For instance, small spines often show enlargement, while large spines are more likely to show shrinkage (**Fig. 7A**). Older spines tend to have a large connection probability, which is proportional to spine size (**Fig. 7B**), and they are more stable (**Fig. 7C**). In addition, training enhances the stability of newly created spines, whereas it degrades the stability of older spines (**Fig. 8B**).

*Experimental prediction*

In the developmental stage, both axon guidance(Munz et al., 2014) and dendritic extension(Matsui et al., 2013) show Hebbian-type activity dependence, but in the adult cortex, both axons and dendrites seldom change their structures(Holtmaat and Svoboda, 2009). Thus, although recent experimental results suggest some activity dependence for spine creation(Knott et al., 2006)(Yang et al., 2014), it is still unclear to what extent spine creation depends on the activity of presynaptic and postsynaptic neurons. Our model indicates that in terms of performance, spine creation should fully depend on both presynaptic and postsynaptic activity (**Fig. 6I**). However, we also showed that it is possible to replicate a wide range of experimental results on spine dynamics without activity-dependent spine creation (**Fig. 8**).

Furthermore, whether or not spine survival rate increases through training is controversial(Yang et al., 2009)(Xu et al., 2009). Our model predicts that the stability of new spines highly depends on the similarity between the new task and control behavior (**Fig. 8F**). When the similarity is low, new spines created in the new task are expected to be more stable than those created in the control case, because the synaptic connection structure would need to be reorganized. By contrast, when the similarity is high, the stability of the new spines would be comparable to that of the control. In addition, our model replicates the effect of varying training duration on spine stability(Yang et al., 2009). When training was rapidly terminated, newly formed spines became less stable than those undergoing continuous training (**Fig. 8G**).

*Related studies*

Previous theoretical studies revealed candidate rules for spine creation and elimination(Deger et al., 2012)(Zheng et al., 2013)(Fauth et al., 2015), yet their functional benefits were not fully clarified in those studies. Some modeling studies considered the functional implications of synaptic rewiring (Poirazi and Mel, 2001) or optimality in regard to benefit and wiring cost (Chen et al., 2006), but the functional significance of synaptic plasticity and the variability of EPSP size were not considered in those models. In comparison, our study revealed functional roles of wiring plasticity that cooperates with synaptic weight plasticity and obeys local unsupervised rewiring rules. In addition, we extended

the previous results on single-spine information storage and synaptic noise (Varshney et al., 2006) into a network, and provided a comparison with experimental results (**Fig. 2E**).

Previous studies on associative memory models found the cut-off coding as the optimal strategy for maximizing the information capacity per synapse (Chechik et al., 1998)(Knoblauch et al., 2010). Our results suggest that the above result is the outcome of the tight positive correlation between the information gain and synaptic weight in associative memory systems, and not generally applicable to other paradigms (**Fig. 3BC**). In addition, although cut-off strategy did not yield biologically plausible synaptic weight distributions in our task setting (**Fig. 3A** right), in perceptron-based models, this unrealistic situation can be avoided by tuning the threshold of neural dynamics (Brunel et al., 2004)(Sacramento et al., 2015). Especially, cut-off strategy may provide a good approximation for developmental wiring plasticity(Ko et al., 2013), though the algorithm is not fully consistent with wiring plasticity in the adult animals.

Finally, our model provides a biologically plausible interpretation for multi-timescale learning processes. It was previously shown that learning with two synaptic variables on different timescales is beneficial under a dynamically changing environment(Fusi et al., 2007). In our model, both fast and slow variables played important roles, whereas in previous studies, only one variable was usually more effective than others, depending on the task context.

## Methods

### 1. Model

*1.1 Model dynamics*

We first define the model and the learning rule for general exponential family, and derive equations for two examples (Gaussian and Poisson). In the task, at every time $t$, one hidden state $s^t$ is sampled from prior distribution $p(s)$. Neurons in the input layer show stochastic response $r_{X,j}{}^t$ that follows probabilistic distribution $f(r_{X,j} \mid s^t)$:

$$f(r_{X,j} \mid \mu) \equiv \exp\left[h(\theta_{j\mu})g(r_{X,j}) - A(\theta_{j\mu}) + B(r_{X,j})\right]. \qquad (5)$$

From these input neuron activities, neurons in output layer estimate the hidden variables. Here we assume maximum likelihood estimation for decision making unit, as the external state is a discrete variable. In this framework, in order to detect the hidden signal, firing rate of neuron $i$ should be proportional to posterior

$$r_{Y,i}^t \propto \Pr\left[s^t = \sigma_i \mid r_X^t\right]. \qquad (6)$$

where $\sigma_i$ represents the index of the hidden variable preferred by output neuron $i$ (Beck et al., 2008)(Lochmann and Deneve, 2011). Note that $\{r_{X,j}\}$ represent firing rates of input neurons, whereas

437 $\{r_{Y,i}\}$ represent the rates of output neurons. Due to Bayes rule, estimation of $s^t$ is given by,

438
$$\begin{aligned}\log p(s^t=\mu \mid r_X^t) &= \sum_{j=1}^{M} \log p(r_{X,j}^t \mid s^t=\mu) + \log p(s^t=\mu) - \log p(r_X^t) \\ &= \sum_{j=1}^{M}\left[q_{\mu j} g(r_{X,j}^t) - \alpha(q_{\mu j}) + B(r_{X,j}^t)\right] + \log p(s^t=\mu) - \log p(r_X^t),\end{aligned} \tag{7}$$

439 where $q_{j\mu} \equiv h(\theta_{j\mu})$, $\alpha(q_{j\mu}) \equiv A\left(h^{-1}(q_{j\mu})\right)$. If we assume the uniformity of hidden states as

440 $\log p(s^t=\mu): const$, and $\frac{1}{M}\sum_{j=1}^{M}\alpha(q_{j\mu}) = \alpha_o$, the equation above becomes

441
$$\log p(s^t=\mu \mid r_X^t) = \sum_{j=1}^{M}\left[q_{\mu j} g(r_{X,j}^t) + B(r_{X,j}^t)\right] - \log p(r_X^t) + \text{const.}$$

442 To achieve neural implementation of this inference problem, let us consider a neural dynamics in
443 which the firing rates of output neurons follow,

444
$$r_{Y,i}^t = r_Y^o \exp\left[\sum_{j=1}^{M} c_{ij}\left(w_{ij} g(r_{X,j}^t) - h_w\right) - I_{inh}^t\right], \tag{8}$$

445 where,

446
$$I_{inh}^t \equiv \log\left[\sum_{i=1}^{N}\exp\left(\sum_{j=1}^{M} c_{ij}\left[w_{ij} g(r_{X,j}^t) - h_w\right]\right)\right],$$

447 and $h_w$ is the threshold. If connection is all-to-all, $w_{ij} = q_{j\mu}$ gives optimal inference, because

448
$$\frac{r_{Y,i}^t}{r_Y^o} = \frac{\exp\left[\sum_j q_{j\mu} g(r_{X,j}^t)\right]}{\sum_\nu \exp\left[\sum_j q_{j\nu} g(r_{X,j}^t)\right]} = p(s^t=\mu \mid r_X^t) \tag{9}$$

449 Note that $h_w$ is not necessary to achieve optimal inference, however, under a sparse connection, $h_w$ is
450 important for reducing the effect of connection variability. In this formalization, even in
451 non-all-to-all network, if the sparseness of connectivity stays in reasonable range, near-optimal
452 inference can be performed for arbitrary feedforward connectivity by adjusting synaptic weight to

453 $w_{ij} = w_{\mu j} \equiv q_{j\mu}/\rho_{\mu j}$ where $\rho_{\mu j} = \frac{1}{|\Omega_\mu|}\sum_{i\in\Omega_\mu} c_{ij}$.

454

455 *1.2. Weight coding and connectivity coding*

456 Let us first consider the case when the connection probability is constant (i.e. $\rho_{ij}=\rho$). By substituting

457 $\rho_{ij}=\rho$ into the above equations, $c$ and $w$ are given with $\Pr[c_{ij}=1]=\rho$ and $w_{ij} = w_{\mu j} = q_{j\mu}/\rho$,

where the mean connectivity is given as $\rho = \gamma \bar{q}$, and $\bar{q}$ is the average of the normalized mean response $q_{j\mu}$ (i.e., $\bar{q} = \frac{1}{Mp}\sum_j \sum_\mu q_{j\mu}$). Parameter $\gamma$ is introduced to control the sparseness of connections, and here we assumed that neuron $i$ represents the external state $\mu = floor\left(\frac{p \times i}{N}\right)$ (i.e., if $\frac{\mu N}{p} < i \le \frac{(\mu+1)N}{p}$, output neuron $i$ represents the state $\mu$). Under this configuration, the representation is solely achieved by the synaptic weights, thus we call this coding strategy as the weight coding.

On the other hand, if the synaptic weight is kept at a constant value, the representation is realized by synaptic connection structure (i.e. connectivity coding). In this case, the model is given by $\Pr\left[c_{ij} = 1\right] = \rho_{\mu j}$ and $w_{ij} = w_{\mu j} = 1/\gamma$, where $\rho_{\mu j} = \min\left(\gamma q_{j\mu}, 1\right)$.

*1.3 Dual coding and cut-off coding*

By combining the weight coding and connectivity coding described above, the dual coding is given as $w_{ij} = w_{\mu j} = q_{j\mu}/\rho$, $\Pr\left[c_{ij} = 1\right] = \rho_{\mu j}$, $\rho_{\mu j} = \min\left(\gamma q_{j\mu}, 1\right)$, where $\rho$ was defined by $\rho = \gamma \bar{q}$, $\bar{q} = \frac{1}{Mp}\sum_j \sum_\mu q_{j\mu}$, as in the weight coding. For the cut-off coding strategy, the synaptic weight was chosen as $w_{ij} = w_{\mu j} = q_{j\mu}/\rho_o$ where $\rho_o$ is the mean connection probability. Based on these synaptic weights, for each output neuron, we selected $M\rho_o$ largest synaptic connections, and eliminated all other connections. Thus, connection matrix $C$ was given as $c_{ij} = \left[\sum_{j'}\left[w_{ij} \le w_{ij'}\right]_+ \le M\rho_o\right]_+$, where $[\text{true}]_+ = 1$, $[\text{false}]_+ = 0$. When multiple connections have the same weight, we randomly selected the connections so that the total number of inbound connections becomes $M\rho_o$. Finally, in the random connection strategy, synaptic weights and connections were determined as $w_{ij} = w_{\mu j} = q_{j\mu}/\rho_o$, $\Pr\left[c_{ij} = 1\right] = \rho_o$.

*1.4 Synaptic weight learning*

To perform maximum likelihood estimation from output neuron activity, synaptic weight matrix between input neurons and output neurons should provide a reverse model of input neuron activity. If the reverse model is faithful, KL-divergence between the true input and the estimated distributions $D_{KL}\left[p^*(r_X^t) \,\|\, p(r_X^t \mid C, W)\right]$ would be minimized (Dayan et al., 1995) (Nessler et al., 2013).

Therefore, synaptic weights learning can be performed by $\mathrm{argmin}_W D_{KL}\left[p^*(r_X^t)\,\|\,p(r_X^t \mid C,W)\right]$.

Likelihood $p(r_X^t \mid C,W)$ is approximated as

$$\begin{aligned} p(r_X^t \mid C,W) &\propto \sum_\mu p(r_X^t \mid s^t=\mu,C,W)p(s^t=\mu \mid C,W) \\ &= \sum_\mu p(s^t=\mu \mid C,W)\exp\left[\sum_j \left(h(\theta_{j,\mu}^{C,W})g(r_{X,j}^t) - A(\theta_{j,\mu}^{C,W}) + B(r_{X,j}^t)\right)\right] \\ &\simeq \sum_\mu p(s^t=\mu)\exp\left[\sum_j \left(q_{j\mu}^{C,W} g(r_{X,j}^t) - \alpha(q_{j\mu}^{C,W}) + B(r_{X,j}^t)\right)\right]. \end{aligned} \tag{10}$$

$\theta_{j,\mu}^{C,W}$ in the second line is the average response estimated from connectivity matrix $C$, and weight matrix $W$. In the last equation, $q_{j\mu}^{C,W}$ is substituted for $h\left(\theta_{j,\mu}^{C,W}\right)$. If we approximate the estimated parameter $q_{j\mu}^{C,W}$ with $q_{j\mu}^{C,W} \simeq \rho_o w_{ij}$ by using the average connectivity $\rho_o$, a synaptic weight plasticity rule is given by stochastic gradient descending as

$$\begin{aligned} \Delta w_{ij} &\propto \frac{\partial \log p(r_X^t \mid C,W)}{\partial w_{ij}} \\ &= p(s^t=\mu \mid r_X^t,C,W)\rho_o\left(g(r_{X,j}^t) - \alpha'(\rho_o w_{ij})\right) \\ &\simeq r_{Y,i}^t \rho_o\left(g(r_{X,j}^t) - \alpha'(\rho_o w_{ij})\right). \end{aligned} \tag{11}$$

Especially, in a Gaussian model, the synaptic weight converges to the weight coding as $w_{ij} = \left\langle r_{Y,i}^t r_{X,j}^t \Big/ \left(\sigma_X^2 \rho_o r_{Y,i}^t\right)\right\rangle = q_{j\mu}/\rho_o$, where $\mu$ is the external state that output neuron $i$ learned to represent (i.e. $i \in \Omega_\mu$).

As we were considering population representation, in which the total number of output neuron is larger than the total number of external states (i.e. $p < N$), there is a redundancy in representation. Thus, to make use of most of population, homeostatic constraint is necessary. For homeostatic plasticity, we set a constraint on the output firing rate. By combining two terms, synaptic weight plasticity rule is given as

$$\Delta w_{ij} = \frac{\eta_X}{\gamma}\left(r_{Y,i}^t\left[g(r_{X,j}^t) - \alpha'\left(\rho_o w_{ij}\right)\right] + b_h\left[r_Y^o / N - r_{Y,i}^t\right]\right). \tag{12}$$

By changing the strength of homeostatic plasticity $b_h$, the network changes its behavior. The learning rate is divided by $\gamma$, because the mean of $w$ is proportional to $1/\gamma$. Although, this learning rule is unsupervised, each output neuron naturally selects an external state in self-organisation manner.

*1.5 Synaptic connection learning*

Wiring plasticity of synaptic connection can be given in a similar manner. As shown in **Figure 3**, if the synaptic connection structure of network is correlated with the external model, the learning performance typically gets better. Therefore, by considering $\mathrm{argmin}_{\rho} D_{KL}\left[p^*(r_X^t) \| p(r_X^t \mid \rho, W)\right]$, the update rule of connection probability is given as

$$\Delta\rho_{ij} \propto r_{Y,i}^t w_o \left[g(r_{X,j}^t) - \alpha'(\rho_{ij} w_o)\right]. \quad (13)$$

Here, we approximated $w_{ij}$ with its average value $w_o$. In this implementation, if synaptic weight is also plastic, convergence of $D_{KL}$ is no longer guaranteed, yet as shown in **Figure 3**, redundant representation robustly provides a good heuristic solution.

Let us next consider the implementation of the rewiring process with local spine elimination and creation based on the connection probability $\rho_{ij}$. To keep the detailed balance of connection probability, creation probability $c_p(\rho)$ and elimination probability $e_p(\rho)$ need to satisfy

$$(1-\rho)c_p(\rho) = \rho e_p(\rho)$$

The simplest functions that satisfy above equation is $c_p(\rho) \equiv \rho/\tau_c$, $e_p(\rho) \equiv (1-\rho)/\tau_c$. In the simulation, we implemented this rule by changing $c_{ij}$ from 1 to 0 with probability $(1-\rho)/\tau_c$ for every connection with $c_{ij}$=1, and shift $c_{ij}$ from 0 to 1 with probability $\rho/\tau_c$ for non-existing connection ($c_{ij}$=0) at every time step.

*1.6 Dual Hebbian rule and estimated transfer entropy*

The results in the main texts suggest that non-random synaptic connection structure can be beneficial either when that increases estimated transfer entropy or is correlated with the structure of the external model. To derive dual Hebbian rule, we used the latter property, yet in the simulation, estimated transfer entropy also increased by the dual Hebbian rule. Here, we consider relationship of two objective functions. Estimation of the external state from the sampled inputs is approximated as

$$\left\langle p(s^t = \mu) \mid \{c_{ij} r_{X,j}^t\} \right\rangle_{i \in \Omega_\mu} \simeq \frac{1}{|\Omega_\mu|} \sum_{i \in \Omega_\mu} \frac{p(s^t = \mu) \exp\left(\sum_j \rho_{ij} \left[q_{\mu j} g(r_{X,j}^t) - \alpha(q_{\mu j}) + B(r_{X,j}^t)\right]\right)}{\sum_\nu p(s^t = \nu) \exp\left(\sum_j c_{ij} \left[q_{\nu j} g(r_{X,j}^t) - \alpha(q_{\nu j}) + B(r_{X,j}^t)\right]\right)} \quad (14)$$

Therefore, by considering stochastic gradient descending, an update rule of $\rho_{ij}$ is given as

$$\Delta\rho_{ij} \propto \left(1 + \log r_{Y,i}^t / r_Y^o\right) r_{Y,i}^t \left[g(r_{X,j}^t) - \alpha(q_{\mu j})/q_{\mu j} + B(r_{X,j}^t)/q_{\mu j}\right] \quad (15)$$

If we compare this equation with the equation for dual Hebbian rule (equation (13)), both of them are monotonically increasing function of $r_{Y,i}^t$ and have the same dependence on $g(r_{X,j}^t)$ although normalization terms are different. Thus, under an adequate normalization, the inner product of

change direction is on average positive. Therefore, although dual Hebbian learning rule does not maximize the estimated maximum transfer entropy, the rule rarely diminishes it.

*1.7 Gaussian model*

We constructed mean response probabilities $\{\theta_{j\mu}\}_{j=1,\ldots,M}^{\mu=1,\ldots,p}$ by following 2 steps. First, non-normalized response probabilities $\{\tilde{\theta}_{j\mu}\}_{j=1,\ldots,M}^{\mu=1,\ldots,p}$ were chosen from a truncated normal distribution $N(\mu_M,\sigma_M)$ defined on $[0,\infty)$. Second, we defined $\{\theta_{j\mu}\}_{j=1,\ldots,M}^{\mu=1,\ldots,p}$ by $\theta_{j\mu}=\tilde{\theta}_{j\mu}/Z_\mu$, where $Z_\mu = r_X^o \Big/ \sqrt{\sum_{j=1}^M \tilde{\theta}_{j\mu}/M}$.

When the noise follows a Gaussian distribution, the response functions in equation (5) are given as

$$h(\theta)=\frac{\theta}{\sigma_x^2},\ g(r)=r,\ A(\theta)=\frac{\theta^2}{2\sigma_x^2}+\log(\sqrt{2\pi}\sigma_x),\ B(r)=-\frac{r^2}{2\sigma_x^2}. \qquad (16)$$

Because $h^{-1}(q)=\sigma_x^2 q$, $\alpha(q)$ is given as $\alpha(q)\equiv A\left(h^{-1}(q)\right)=\sigma_x^2 q^2/2+\log(\sqrt{2\pi}\sigma_x)$. By substituting above values into the original equations, the neural dynamics is given as

$$r_{Y,i}^t = r_Y^o \exp\left[\sum_{j=1}^M c_{ij}\left(w_{ij}r_{X,j}^t - w_o\right) - I_{inh}^t\right]. \qquad (17)$$

Similarly, dual Hebbian rule becomes

$$\Delta w_{ij} = \frac{\eta_X}{\gamma}\left(r_{Y,i}^t\left[r_{X,j}^t - \sigma_X^2\rho_o w_{ij}\right] + b_h\left[r_Y^o/N - r_{Y,i}^t\right]\right) \qquad (18)$$

$$\Delta\rho_{ij} = \eta_\rho r_{Y,i}^t\left(r_{X,j}^t - \sigma_x^2\rho_{ij}w_o\right). \qquad (19)$$

*1.8 Poisson model*

For Poisson model, we defined mean response probabilities $\{\theta_{j\mu}\}_{j=1,\ldots,M}^{\mu=1,\ldots,p}$ from a log-normal distribution instead of a normal distribution. Non-normalized values were sampled from a truncated log-normal distribution $\log N(\mu_M^p,\sigma_M^p)$ defined on $(I_{min}^p, I_{max}^p)$. Normalization was performed as $\theta_{j\mu}=\tilde{\theta}_{j\mu}/Z_\mu$ for $\{\tilde{\theta}_{j\mu}\}_{j=1,\ldots,M}^{\mu=1,\ldots,p}$, where $Z_\mu = r_X^o M/\sum_j \theta_{j\mu}$. Because the noise follows a Poisson distribution $p(r\mid\theta)=\exp[-q+r\log q-\log r!]$, the response functions are given as

$$h(\theta)=\log\theta,\ g(r)=r,\ A(\theta)=\theta,\ B(r)=-\log r!. \qquad (20)$$

As a result, $\alpha(q)$ is defined as $\alpha(q)\equiv A\left(h^{-1}(q)\right)=e^q$. By substituting them to the original equations, the neural dynamics also follows equation (17). If connection is all-to-all, by setting $w_{ij}=\log\theta_{j\mu}/\theta_o$ for $i\in\Omega_\mu$, optimal inference is achievable. Here, we normalized $\theta_j$ by $\theta_o$, which

is defined as $\theta_o = \frac{1}{2}\min_{j,\mu}\theta_{\mu j}$, in order to keep synaptic weights in non-negative values.

Learning rules for synaptic weight and connection are given as

$$\Delta w_{ij} = \frac{\eta_x}{\gamma}\left(r_{Y,i}^t\left[r_{X,j}^t - \theta_{min}\exp[\rho_o w_{ij}]\right] + b_h\left[r_Y^o / N - r_{Y,i}^t\right]\right), \quad (21)$$

$$\Delta\rho_{ij} = \eta_\rho r_{Y,i}^t\left(r_{X,j}^t - \theta_{min}\exp(\rho_{ij}w_o)\right). \quad (22)$$

Note that the first term of the synaptic weight learning rule coincides with a previously proposed optimal learning rule for spiking neurons (Nessler et al., 2013)(Habenschuss et al., 2013). In calculation of model error, error was calculated as $d = \sqrt{\frac{1}{pM}\sum_\mu\sum_j\left(\tilde{q}_{j\mu} - q^*_{j\mu}\right)^2}$, where estimated parameter $\{\tilde{q}_{j\mu}\}$ was given by $\tilde{q}_{j\mu} = \frac{\langle q^*_{j\mu}\rangle\bar{q}_{j\mu}}{\sum_q\sum_j\bar{q}_{j\mu}/pM}$. Here, $\langle q^*_{j\mu}\rangle$ represents the mean of true $\{q_{j\mu}\}$, and non-normalized estimator $\bar{q}_{j\mu}$ was calculated as $\bar{q}_{j\mu} = \frac{1}{\langle c_{ij}\rangle|\Omega_\mu|}\sum_{i\in\Omega_\mu}c_{ij}w_{ij}$. In **Figure S1D**, estimation from connectivity was calculated from $\bar{q}^C_{j\mu} = \frac{1}{\langle c_{ij}\rangle|\Omega_\mu|}\sum_{i\in\Omega_\mu}c_{ij}$, and similarly, estimation from weights was calculated by $\bar{q}^W_{j\mu} = \frac{1}{|\Omega_\mu|\sum_{i\in\Omega_\mu}c_{ij}}\sum_{i\in\Omega_\mu}c_{ij}w_{ij}$. For parameters, we used $\mu^p_M = 0.0$, $\sigma^p_M = 1.0$, $I^p_{min} = 0.2$, $I^p_{max} = 20.0$, $w_o = 1/\gamma$, $r^o_X = 0.3$, and for other parameters, we used same values with the Gaussian model.

## 2 Analytical evaluations

### *2.1 Evaluation of performances in weight coding and connectivity coding*

In Gaussian model, we can analytically evaluate the performance in two coding schemes. As the dynamics of output neurons follows $r_{Y,i} = r_Y^o\exp\left[\sum_j c_{ij}(w_{ij}r_{X,j}^t - w_o) - I_{inh}^t\right]$, membrane potential variable $u_i$, which is defined as

$$u_i \equiv \sum_j c_{ij}(w_{ij}r_{X,j}^t - w_o), \quad (23)$$

determines firing rates of each neuron. Because $\{\theta_{j\mu}\}$ is normalized with $\sum_{j=1}^M\theta_{j\mu}^2/M = (r_X^o)^2$,

584 mean and variance of $\{\theta_{j\mu}\}$ are given as

585 $$\mu_\theta = \frac{\mu_M r_X^o}{\sqrt{\mu_M^2+\sigma_x^2}},\quad \sigma_\theta^2 = \frac{\left(\sigma_M r_X^o\right)^2}{\mu_M^2+\sigma_M^2}, \tag{24}$$

586 where $\mu_M$ and $\sigma_M$ are the mean and variance of the original non-normalized truncated Gaussian
587 distribution $\{\tilde{\theta}_{j\mu}\}$. Because both $r_{X,j}$ and $\{\theta_{j\mu}\}$ approximately follow Gaussian distribution, $u_i$ is
588 expected to follow Gaussian. Therefore, by evaluating its mean and variance, we can characterize the
589 distribution of $u_i$ for a given external state (Babadi and Sompolinsky, 2014).

590 Let us first consider the distribution of $u_i$ in the weight coding. In weight coding scheme,
591 $w_{ij}$ and $c_{ij}$ are defined as

592 $$w_{ij} = \theta_{j\mu}/\rho\sigma_x^2,\quad \Pr\left[c_{ij}=1\right]=\rho \tag{25}$$

593 where $\rho = \gamma\mu_\theta/\sigma_x^2$. By setting $w_o = \mu_\theta^2/\left(\rho\sigma_X^2\right)$, the mean membrane potential of output neuron $i$
594 selective for given signal (i.e. $i\in\Omega_\mu$ for $s^t=\mu$) is calculated as,

595 $$\langle u_i\rangle = \left\langle \sum_j \left(\theta_{j\mu}^2 - \langle\theta_{j\mu}\rangle^2\right)/\sigma_x^2 \right\rangle = M\sigma_\theta^2/\sigma_x^2.$$

596 Similarly, the variance of $u_i$ is given as

597 $$\begin{aligned}\left\langle\left(u_i-\langle u_i\rangle\right)^2\right\rangle &= \left\langle\left(\frac{1}{\rho\sigma_X}\sum_j c_{ij}\theta_{j\mu}\zeta_j + \frac{1}{\rho\sigma_X^2}\sum_j\left(c_{ij}-\rho\right)\left(\theta_{j\mu}^2-\mu_\theta^2\right)+\frac{1}{\sigma_X^2}\sum_j\left(\theta_{j\mu}^2-\left[\mu_\theta^2+\sigma_\theta^2\right]\right)\right)^2\right\rangle \\ &= \frac{M}{\rho\sigma_X^2}\left(\mu_\theta^2+\sigma_\theta^2\right)+\frac{M\sigma_\theta^2}{\rho\sigma_X^4}\left[2\left(2\mu_\theta^2+\sigma_\theta^2\right)+\left(1-\rho\right)\sigma_\theta^2\right]\end{aligned} \tag{26}$$

598 where $\zeta_j$ is a Gaussian random variable. On the other hand, if output neuron $i$ is not selective for the
599 presented stimuli (if $s^t\neq\mu$ and $i\in\Omega_\mu$), $w_{ij}$ and $r_{X,j}$ are independent. Thus, the mean and the
600 variance of $u_i$ are given as,

601 $$\langle u_i\rangle = 0,\quad \left\langle\left(u_i-\langle u_i\rangle\right)^2\right\rangle = \frac{M}{\rho\sigma_x^2}\left(\mu_\theta^2+\sigma_\theta^2\right)+\frac{M\sigma_\theta^2}{\rho\sigma_x^4}\left(2\mu_\theta^2+\sigma_\theta^2\right)$$

602 In addition to that, due to feedforward connection, output neurons show noise correlation. For two
603 output neurons $i$ and $l$ selective for different states (i.e. $i\in\Omega_\mu$ and $l\notin\Omega_\mu$), the covariance
604 between $u_i$ and $u_l$ satisfies

605 $$\left\langle\left(u_i-\langle u_i\rangle\right)\left(u_l-\langle u_l\rangle\right)\right\rangle = \left\langle\rho^2\sum_j w_{ij}w_{lj}\left(r_{X,j}-\theta_{j\mu}\right)^2\right\rangle = M\mu_\theta^2/\sigma_x^2$$

606 Therefore, approximately ($u_i$, $u_l$) follows a multivariable Gaussian distributions

$$\begin{pmatrix} u_i \\ u_l \end{pmatrix} = N\left( \begin{pmatrix} \frac{M\sigma_\theta^2}{\sigma_x^2} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{M\left(\mu_\theta^2+\sigma_\theta^2\right)}{\rho\sigma_X^2} + \frac{M\sigma_\theta^2\left[2\left(2\mu_\theta^2+\sigma_\theta^2\right)+(1-\rho)\sigma_\theta^2\right]}{\rho\sigma_X^4} & \frac{M\mu_\theta^2}{\sigma_x^2} \\ \frac{M\mu_\theta^2}{\sigma_x^2} & \frac{M(\mu_\theta^2+\sigma_\theta^2)}{\rho\sigma_x^2} + \frac{M\sigma_\theta^2\left(2\mu_\theta^2+\sigma_\theta^2\right)}{\rho\sigma_x^4} \end{pmatrix} \right). \quad (27)$$

In maximum likelihood estimation, the estimation fails if a non-selective output neuron shows higher firing rate than the selective neuron. When there are two output neurons, probability for such an event is calculated as

$$\epsilon_w = \Pr\left[ \sum_j c_{lj}(w_{lj} r_{X,j}^t - w_o) > \sum_j c_{ij}(w_{ij} r_{X,j}^t - w_o) \mid s^t = \mu, i \in \Omega_\mu, l \notin \Omega_\mu \right].$$

In the simulation, there are *p-1* distractors per one selective output neuron. Thus, approximately, accuracy of estimation was evaluated by $(1-\epsilon_w)^{p-1}$. In **Figure 2B**, we numerically calculated this value for the analytical estimation.

Similarly, in connectivity coding, $w_{ij}$ and $c_{ij}$ are given as

$w_{ij} = 1/\gamma$, $\Pr[c_{ij} = 1] = \rho_{ij}$, $\rho_{ij} = \gamma\theta_{j\mu}/\sigma_x^2$.

By setting $w_o = \mu_\theta/\gamma$, from a similar calculation done above, the mean and the variance of ($u_i$, $u_l$) are derived as

$$\begin{pmatrix} u_i \\ u_l \end{pmatrix} = N\left( \begin{pmatrix} \frac{M\sigma_\theta^2}{\sigma_x^2} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{M\mu_\theta}{\gamma} + \frac{M\sigma_\theta^2\left[\mu_\theta\sigma_x^2-\gamma\sigma_\theta^2\right]}{\gamma\sigma_x^4} & \frac{M\mu_\theta^2}{\sigma_x^2} + \frac{M\mu_\theta^2\sigma_\theta^2}{\sigma_x^4} \\ \frac{M\mu_\theta^2}{\sigma_x^2} + \frac{M\mu_\theta^2\sigma_\theta^2}{\sigma_x^4} & \frac{M\mu_\theta}{\gamma} + \frac{M\mu_\theta\sigma_\theta^2}{\gamma\sigma_x^2} \end{pmatrix} \right). \quad (28)$$

If we compare the two coding schemes, means are the same for two coding schemes, and as $\gamma$ satisfies $\gamma = \sigma_x^2\rho/\mu_\theta$, variance of non-selective output neuron are similar. The main difference is the second term of signal variance. In the weight coding, signal variance is proportional to *1/γ*, on the other hands, in the connectivity coding, the second term of signal variance is negative, and does not depend on the connectivity. As a result, in the adequately sparse regime, firing rate variability of selective output neuron becomes smaller in connectivity coding, and the estimation accuracy is better. In the sparse limit, the first term of variance becomes dominant and both schemes do not work well, consequently, the advantage for connectivity coding disappears. Coefficient of variation calculated for signal terms is indeed smaller in connectivity coding scheme (blue and red lines in **Fig 2C**), and the same tendency is observed in simulation (cyan and orange lines in **Fig 2C**).

*2.2 Optimality of connectivity*

To evaluate optimality of a given connection matrix *C*, we calculated the posterior probability of the external states estimated from *C* and $r_X$, and compared then to that from the fully connected network

$C_{all}$. Below, we denote the mean KL-divergence $\left\langle D_{KL}\left[p\left(s^t \mid r_X, C_{all}\right) \| p\left(s^t \mid r_X, C\right)\right]\right\rangle_{r_X}$ as $I(C_{all},C)$ for readability. When the true external state is $s^t=v$, firing rates of input neurons are given by $r_{X,j}{}^t \sim N(\theta_{jv}, \sigma_X)$, hence this $I(C_{all},C)$ is approximately evaluated as

$$I\left(C_{all},C\right) \approx \frac{1}{p}\sum_v \left\langle D_{KL}\left[p\left(s^t \mid r_{X\mathrm{I}v}, C_{all}\right) \| p\left(s^t \mid r_{X\mathrm{I}v}, C\right)\right]\right\rangle_{r_X}$$

$$\approx \frac{1}{p}\sum_v D_{KL}\left[\left\langle p\left(s^t \mid \left\{\theta_{jv}+\sigma_X\zeta_j\right\}, C_{all}\right)\right\rangle_{\{\zeta_j\}} \| \left\langle p\left(s^t \mid \left\{\theta_{jv}+\sigma_X\zeta_j\right\}, C\right)\right\rangle_{\{\zeta_j\}}\right]$$

where $\{\zeta_j\}$ are Gaussian random variables, and $C_{all}$ represents the all-to-all connection matrix. By taking integral over Gaussian variables, the posterior probability is evaluated as

$$\left\langle p\left(s^t=\mu \mid \left\{\theta_{jv}+\sigma_X\zeta_j\right\}, C\right)\right\rangle_{\{\zeta_j\}} \cong \frac{1}{\left|\Omega_\mu\right|}\sum_{i\in\Omega_\mu}\frac{\exp\left(\phi_{\mu v}^{i,C}+\frac{1}{2}\psi_\mu^{i,C}\right)}{\sum_{\mu'}\exp\left(\phi_{\mu' v}^{i,C}+\frac{1}{2}\psi_{\mu'}^{i,C}\right)} \equiv p_v\left(s^t=\mu \mid C\right),$$

where

$$\phi_{\mu v}^{i,C} \equiv \sum_j c_{ij}\left(2\theta_{\mu j}\theta_{vj}-\theta_{\mu j}^2\right)/\left(2\sigma_X^2\right), \quad \psi_\mu^{i,C} \equiv \sum_j c_{ij}\left(\theta_{\mu j}/\sigma_X\right)^2.$$

Thus, the KL-divergence between estimations by two connection structures $C_{all}$ and $C$ is approximated as:

$$I\left(C_{all},C\right) \approx \frac{1}{p}\sum_v\sum_\mu p_v\left(s^t=\mu \mid C_{all}\right)\log\frac{p_v\left(s^t=\mu \mid C_{all}\right)}{p_v\left(s^t=\mu \mid C\right)} \qquad (29)$$

In the black lines in **Figures 3C-E**, we maximized the approximated KL-divergence $I(C_{all},C)$ with a hill-climbing method from various initial conditions, thus the lines may not be the exact optimal, but rather lower bounds of the optimal performance. Information gain by a connection $c_{ij}$ was evaluated by

$$\Delta I_{ij} \equiv \left\langle I\left(C_{all},C\right)-I\left(C_{all},C+\eta_{ij}\right)\right\rangle_C \quad , \qquad (30)$$

where $\eta_{ij}$ is a $N\times M$ matrix in which only $(i,j)$ element takes 1, and all other elements are 0. In **Figure 3B**, we took average over 1000 random connection structures with connection probability $\rho$=0.1.

## 3 Model settings

### *3.1 Details of simulation*

In the simulation, the external variable $s^t$ was chosen from 10 discrete variables ($p$ = 10) with equal probability (Pr[$s^t$ = $q$] = 1/$p$, for all $q$). The mean response probability $\theta_{j\mu}$ was given first

by randomly chosen parameters $\left\{\tilde{\theta}_{j\mu}\right\}_{j=1,..,M}^{\mu=0,..,p-1}$ from the truncated normal distribution $N(\mu_M,\sigma_M)$ in $[0,\infty)$, and then normalized using $\theta_{j\mu}=\tilde{\theta}_{j\mu}/Z_\mu$, where $Z_\mu=r_X^o\Big/\sqrt{\sum_{j=1}^{M}\tilde{\theta}_{j\mu}/M}$. Mean weight $w_o$ was defined as $w_o=r_X^o/\gamma$. The normalization factor $h_w$ was defined as $h_w=\bar{q}/\gamma$ in **Figures 1–2** and **4-5**, where $\bar{q}=\frac{1}{Mp}\sum_j\sum_\mu\theta_{j\mu}/\sigma_X^2$, and as $h_w=r_X^o/\gamma$ in **Figures 6–7,** as the mean of $\theta$ depends on $\kappa_m$. In **Figure 3**, we used $h_w=\bar{q}/\gamma$ for the dual coding, and $h_w=\bar{q}/\rho_o$ for the rest. Average connectivity $\bar{\rho}$ was calculated from the initial connection matrix of each simulation. In the calculation of the dynamics, for the membrane parameter $v_i\equiv\sum_j c_{ij}\left(w_{ij}r_{X,j}^t-h_w\right)$, a boundary condition $v_i>\max_\ell\{v_\ell-v_d\}$ was introduced for numerical convenience, where $v_d$ = -60. In addition, synaptic weight $w_{ij}$ was bounded to a non-negative value ($w_{ij}$ > 0), and the connection probability was defined as $\rho\in[0,1]$. For simulations with synaptic weight learning, initial weights were defined as $w_{ij}=\left(1+\sigma_w^{init}\zeta\right)/\gamma$, where $\sigma_w^{init}$ = 0.1, and $\zeta$ is a Gaussian random variable. Similarly, in the simulation with structural plasticity, the initial condition for the synaptic connection matrix was defined as $\Pr\left[c_{ij}=1\right]=\gamma\left\langle\theta_{j\mu}\right\rangle/\sigma_x^2$. In both the dual Hebbian rule and the approximated dual Hebbian rule, the synaptic weight of a newly created spine was given as $w_{ij}=\left(1+\sigma_w^{init}\zeta\right)w_o$, for a random Gaussian variable $\zeta\leftarrow N(0,1)$. In **Figure 8**, simulations were initiated at -20 days (i.e., $2\times10^6$ steps before stimulus onset) to ensure convergence for the control condition. For model parameters, $\mu_M$= 1.0, $\sigma_M$= 1.0, $\sigma_X$= 1.0, $M$ = 200, $N$ = 100 $r_X{}^o$= 1.0, and $r_Y{}^o$= 1.0 were used, and for learning-related parameters, $\eta_X$= 0.01, $b_h$= 0.1, $\eta_\rho$= 0.001, $\tau_c$= $10^6$, $T_2$= $10^5$, and $\kappa_m$= 0.5 were used. In **Figures 7 and 8**, $\eta_\rho$= 0.0001, $\tau_c$= $3\times10^5$, and $\gamma$ = 0.6 were used, unless otherwise stated.

*3.2 Accuracy of estimation*

The accuracy was measured with the bootstrap method. By using data from $t$-$T_o$ <= $t'$ < $t$, the selectivity of output neurons was first decided. $\Omega_\mu$ was defined as a set of output neurons that represents external state $\mu$. Neuron $i$ belongs to set $\Omega_\mu$ if $i$ satisfies

$$\mu=\underset{\mu'}{\operatorname{argmax}}\frac{\sum_{t'=t-T_o}^{t}\left[s^t=\mu'\right]_+r_{Y,i}^t}{\sum_{t'=t-T_o}^{t}\left[s^t=\mu'\right]_+},$$

where operator $[X]_+$ returns 1 if X is true; otherwise, it returns 0. By using this selectivity, based on data from $t <= t' < t+T_o$, the accuracy was estimated as

$$\frac{1}{T_o}\sum_{t'=t}^{t+T_o-1}\left[\frac{1}{\left|\Omega_{s^{t'}}\right|}\sum_{i\in\Omega_{s^{t'}}} r_{Y,i}^{t'} > \max_{\mu\neq s^{t'}}\frac{1}{\left|\Omega_{\mu}\right|}\sum_{i\in\Omega_{\mu}} r_{Y,i}^{t'}\right]_{tof}.$$

In the simulation, $T_o = 10^3$ was used because this value is sufficiently slow compared with weight change but sufficiently long to suppress variability.

### *3.3. Model error*

Using the same procedure, model error was estimated as

$$d = \sqrt{\frac{1}{pM}\sum_{\mu=1}^{p}\sum_{j=1}^{M}\left(\tilde{\theta}_{j\mu} - \theta_{j\mu}\right)^2},$$

where $\tilde{\theta}_{j\mu}$ represents the estimated parameter. $\tilde{\theta}_{j\mu}$ was estimated by

$$\bar{\theta}_{j\mu} = \frac{1}{\left\langle c_{ij}\right\rangle\left|\Omega_{\mu}\right|}\sum_{i\in\Omega_{\mu}} c_{ij}w_{ij}, \quad \tilde{\theta}_{j\mu} = r_o^X\bar{\theta}_{j\mu}\Big/\sqrt{\tfrac{1}{M}\textstyle\sum_{j=1}^{M}\bar{\theta}_{j\mu}^2}.$$

In **Figure 6E**, the estimation of the internal model from connectivity was calculated by

$$\bar{\theta}_{j\mu}^{C} = \frac{1}{\left\langle c_{ij}\right\rangle\left|\Omega_{\mu}\right|}\sum_{i\in\Omega_{\mu}} c_{ij}.$$

Similarly, the estimation from the synaptic weight in **Figure 6F** was performed with

$$\bar{\theta}_{j\mu}^{W} = \frac{1}{\left|\Omega_{\mu}\right|}\sum_{i\in\Omega_{\mu}} c_{ij}w_{ij}\Big/\sum_{i\in\Omega_{\mu}} c_{ij}.$$

### *3.4 Transfer entropy*

Entropy reduction caused by partial information on input firing rates was evaluated by transfer entropy:

$$T_E = \left\langle H\left(s^t\right) - H\left(s^t \mid r_X^t, C\right)\right\rangle_t,$$

where

$$\begin{aligned} H\left(s^t \mid r_X^t, C\right) &= -\textstyle\sum_{\mu=1}^{p} p\left(s^t = s_\mu \mid r_X^t, C\right)\log p\left(s^t = s_\mu \mid r_X^t, C\right) \\ &\cong -\textstyle\sum_{\mu=1}^{p}\left\langle p\left(s^t = s_\mu \mid \left\{c_{ij}r_{X,j}^t\right\}\right)\right\rangle_{i\in\Omega_\mu}\log\left\langle p\left(s^t = s_\mu \mid \left\{c_{ij}r_{X,j}^t\right\}\right)\right\rangle_{i\in\Omega_\mu}, \end{aligned}$$

$$\left\langle p\left(s^t = s_\mu \mid \left\{c_{ij} r^t_{X,j}\right\}\right)\right\rangle_{i\in\Omega_\mu} \cong \frac{1}{\left|\Omega_\mu\right|}\sum_{i\in\Omega_\mu} p\left(s^t = s_\mu\right)\prod_{c_{ij}=1} p\left(r^t_{X,j} \mid s^t = s_\mu\right)$$

$$= \frac{1}{\left|\Omega_\mu\right|}\sum_{i\in\Omega_\mu} \frac{p\left(s^t = s_\mu\right)\exp\left(\sum_{j=1}^{M} c_{ij}\left[q_{\mu j} g\left(r^t_{X,j}\right) - \alpha\left(q_{\mu j}\right) + B\left(r^t_{X,j}\right)\right]\right)}{\sum_\nu p\left(s^t = s_\nu\right)\exp\left(\sum_{j=1}^{M} c_{ij}\left[q_{\nu j} g\left(r^t_{X,j}\right) - \alpha\left(q_{\nu j}\right) + B\left(r^t_{X,j}\right)\right]\right)}.$$

Output group $\Omega_\mu$ was determined as described above. Here, the true model was used instead of the estimated model to evaluate the maximum transfer entropy achieved by the network.

### *Code availability*

C++ codes of the simulation program will be available at ModelDB.

### Acknowledgment

The authors thank Drs. Haruo Kasai and Taro Toyoizumi for their comments on the early version of the manuscript.

### References

Babadi, B., and Sompolinsky, H. (2014). Sparseness and Expansion in Sensory Representations. *Neuron* 83, 1213–1226. doi:10.1016/j.neuron.2014.07.035.

Bartol, T. M., Bromer, C., Kinney, J. P., Chirillo, M. A., Bourne, J. N., Harris, K. M., et al. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife*, e10778. doi:10.7554/eLife.10778.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., et al. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron* 60, 1142–1152. doi:10.1016/j.neuron.2008.09.021.

Bliss, T. V., and Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361, 31–39. doi:10.1038/361031a0.

Brunel, N., Hakim, V., Isope, P., Nadal, J.-P., and Barbour, B. (2004). Optimal Information Storage and the Distribution of Synaptic Weights: Perceptron versus Purkinje Cell. *Neuron* 43, 745–757. doi:10.1016/j.neuron.2004.08.023.

Buzsáki, G., and Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nat. Rev. Neurosci.* 15, 264–278. doi:10.1038/nrn3687.

Caporale, N., and Dan, Y. (2008). Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annu. Rev. Neurosci.* 31, 25–46. doi:10.1146/annurev.neuro.31.060407.125639.

Caroni, P., Donato, F., and Muller, D. (2012). Structural plasticity upon learning: regulation and

functions. *Nat. Rev. Neurosci.* 13, 478–490. doi:10.1038/nrn3258.

Chechik, G., Meilijson, I., and Ruppin, E. (1998). Synaptic Pruning in Development: A Computational Account. *Neural Comput.* 10, 1759–1777. doi:10.1162/089976698300017124.

Chen, B. L., Hall, D. H., and Chklovskii, D. B. (2006). Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci. U. S. A.* 103, 4723–4728. doi:10.1073/pnas.0506806103.

Chklovskii, D. B., Mel, B. W., and Svoboda, K. (2004). Cortical rewiring and information storage. *Nature* 431, 782–788. doi:10.1038/nature03012.

Dayan, P., and Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. 1 edition. Cambridge, Mass.: The MIT Press.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904.

Deger, M., Helias, M., Rotter, S., and Diesmann, M. (2012). Spike-timing dependence of structural plasticity explains cooperative synapse formation in the neocortex. *PLoS Comput. Biol.* 8, e1002689. doi:10.1371/journal.pcbi.1002689.

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory* 52, 1289–1306. doi:10.1109/TIT.2006.871582.

Faisal, A. A., Selen, L. P. J., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303. doi:10.1038/nrn2258.

Fauth, M., Wörgötter, F., and Tetzlaff, C. (2015). The Formation of Multi-synaptic Connections by the Interaction of Synaptic and Structural Plasticity and Their Functional Consequences. *PLoS Comput. Biol.* 11. doi:10.1371/journal.pcbi.1004031.

Feldman, D. E. (2009). Synaptic Mechanisms for Plasticity in Neocortex. *Annu. Rev. Neurosci.* 32, 33–55. doi:10.1146/annurev.neuro.051508.135516.

Fusi, S., Asaad, W. F., Miller, E. K., and Wang, X.-J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron* 54, 319–333. doi:10.1016/j.neuron.2007.03.017.

Ganguli, S., and Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* 35, 485–508. doi:10.1146/annurev-neuro-062111-150410.

Habenschuss, S., Puhr, H., and Maass, W. (2013). Emergence of Optimal Decoding of Population Codes Through STDP. *Neural Comput.* 25, 1371–1407. doi:10.1162/NECO_a_00446.

Hiratani, N., Teramae, J.-N., and Fukai, T. (2013). Associative memory model with long-tail-distributed Hebbian synaptic connections. *Front. Comput. Neurosci.* 6. doi:10.3389/fncom.2012.00102.

Holtmaat, A. J. G. D., Trachtenberg, J. T., Wilbrecht, L., Shepherd, G. M., Zhang, X., Knott, G. W., et al. (2005). Transient and persistent dendritic spines in the neocortex in vivo. *Neuron* 45, 279–291. doi:10.1016/j.neuron.2005.01.003.

Holtmaat, A., and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nat. Rev. Neurosci.* 10, 647–658. doi:10.1038/nrn2699.

Ikegaya, Y., Sasaki, T., Ishikawa, D., Honma, N., Tao, K., Takahashi, N., et al. (2013). Interpyramid Spike Transmission Stabilizes the Sparseness of Recurrent Network Activity. *Cereb. Cortex* 23, 293–304. doi:10.1093/cercor/bhs006.

Kasai, H., Hayama, T., Ishikawa, M., Watanabe, S., Yagishita, S., and Noguchi, J. (2010). Learning rules and persistence of dendritic spines. *Eur. J. Neurosci.* 32, 241–249. doi:10.1111/j.1460-9568.2010.07344.x.

Knoblauch, A., Palm, G., and Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Comput.* 22, 289–341. doi:10.1162/neco.2009.08-07-588.

Knott, G. W., Holtmaat, A., Wilbrecht, L., Welker, E., and Svoboda, K. (2006). Spine growth precedes synapse formation in the adult neocortex in vivo. *Nat. Neurosci.* 9, 1117–1124. doi:10.1038/nn1747.

Ko, H., Cossell, L., Baragli, C., Antolik, J., Clopath, C., Hofer, S. B., et al. (2013). The emergence of functional microcircuits in visual cortex. *Nature* 496, 96–100. doi:10.1038/nature12015.

Lefort, S., Tomm, C., Floyd Sarria, J.-C., and Petersen, C. C. H. (2009). The Excitatory Neuronal Network of the C2 Barrel Column in Mouse Primary Somatosensory Cortex. *Neuron* 61, 301–316. doi:10.1016/j.neuron.2008.12.020.

Lochmann, T., and Deneve, S. (2011). Neural processing as causal inference. *Curr. Opin. Neurobiol.* 21, 774–781. doi:10.1016/j.conb.2011.05.018.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Comput.* 14, 2531–2560. doi:10.1162/089976602760407955.

Masamizu, Y., Tanaka, Y. R., Tanaka, Y. H., Hira, R., Ohkubo, F., Kitamura, K., et al. (2014). Two distinct layer-specific dynamics of cortical ensembles during learning of a motor task. *Nat. Neurosci.* 17, 987–994. doi:10.1038/nn.3739.

Matsui, A., Tran, M., Yoshida, A. C., Kikuchi, S. S., U, M., Ogawa, M., et al. (2013). BTBD3 controls dendrite orientation toward active axons in mammalian neocortex. *Science* 342, 1114–1118. doi:10.1126/science.1244505.

Matsuzaki, M., Honkura, N., Ellis-Davies, G. C. R., and Kasai, H. (2004). Structural basis of long-term potentiation in single dendritic spines. *Nature* 429, 761–766. doi:10.1038/nature02617.

Munz, M., Gobert, D., Schohl, A., Poquérusse, J., Podgorski, K., Spratt, P., et al. (2014). Rapid

Hebbian axonal remodeling mediated by visual stimulation. *Science* 344, 904–909. doi:10.1126/science.1251593.

Navlakha, S., Barth, A. L., and Bar-Joseph, Z. (2015). Decreasing-Rate Pruning Optimizes the Construction of Efficient and Robust Distributed Networks. *PLoS Comput. Biol.* 11. doi:10.1371/journal.pcbi.1004347.

Nessler, B., Pfeiffer, M., Buesing, L., and Maass, W. (2013). Bayesian Computation Emerges in Generic Cortical Microcircuits through Spike-Timing-Dependent Plasticity. *PLoS Comput Biol* 9, e1003037. doi:10.1371/journal.pcbi.1003037.

Perin, R., Berger, T. K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci.* 108, 5419–5424. doi:10.1073/pnas.1016051108.

Poirazi, P., and Mel, B. W. (2001). Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron* 29, 779–796.

Potjans, T. C., and Diesmann, M. (2014). The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. *Cereb. Cortex* 24, 785–806. doi:10.1093/cercor/bhs358.

Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A., and Tonegawa, S. (2015). Engram cells retain memory under retrograde amnesia. *Science* 348, 1007–1013. doi:10.1126/science.aaa5542.

Sacramento, J., Wichert, A., and van Rossum, M. C. W. (2015). Energy Efficient Sparse Connectivity from Imbalanced Synaptic Plasticity Rules. *PLoS Comput Biol* 11, e1004265. doi:10.1371/journal.pcbi.1004265.

Salinas, E., and Romo, R. (1998). Conversion of Sensory Signals into Motor Commands in Primary Motor Cortex. *J. Neurosci.* 18, 499–511.

Sayer, R. J., Friedlander, M. J., and Redman, S. J. (1990). The time course and amplitude of EPSPs evoked at synapses between pairs of CA3/CA1 neurons in the hippocampal slice. *J. Neurosci.* 10, 826–836.

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and Efficiency in the Nervous System—A Synthesis. *PLoS Comput Biol* 9, e1003157. doi:10.1371/journal.pcbi.1003157.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits. *PLoS Biol* 3, e68. doi:10.1371/journal.pbio.0030068.

Stepanyants, A., Hof, P. R., and Chklovskii, D. B. (2002). Geometry and Structural Plasticity of Synaptic Connectivity. *Neuron* 34, 275–288. doi:10.1016/S0896-6273(02)00652-9.

Sul, J. H., Jo, S., Lee, D., and Jung, M. W. (2011). Role of rodent secondary motor cortex in value-based action selection. *Nat. Neurosci.* 14, 1202–1208. doi:10.1038/nn.2881.

Turrigiano, G. G., and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system.

*Nat. Rev. Neurosci.* 5, 97–107. doi:10.1038/nrn1327.

Varshney, L. R., Sjöström, P. J., and Chklovskii, D. B. (2006). Optimal Information Storage in Noisy Synapses under Resource Constraints. *Neuron* 52, 409–423. doi:10.1016/j.neuron.2006.10.017.

Wiegert, J. S., and Oertner, T. G. (2013). Long-term depression triggers the selective elimination of weakly integrated synapses. *Proc. Natl. Acad. Sci. U. S. A.* 110, E4510–4519. doi:10.1073/pnas.1315926110.

Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., et al. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* 462, 915–919. doi:10.1038/nature08389.

Yang, G., Lai, C. S. W., Cichon, J., Ma, L., Li, W., and Gan, W.-B. (2014). Sleep promotes branch-specific formation of dendritic spines after learning. *Science* 344, 1173–1178. doi:10.1126/science.1249098.

Yang, G., Pan, F., and Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. *Nature* 462, 920–924. doi:10.1038/nature08577.

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008). Principles of long-term dynamics of dendritic spines. *J. Neurosci. Off. J. Soc. Neurosci.* 28, 13592–13608. doi:10.1523/JNEUROSCI.0603-08.2008.

Yoshimura, Y., Dantzker, J. L. M., and Callaway, E. M. (2005). Excitatory cortical neurons form fine-scale functional networks. *Nature* 433, 868–873. doi:10.1038/nature03252.

Zheng, P., Dimitrakakis, C., and Triesch, J. (2013). Network self-organization explains the statistics and dynamics of synaptic connection strengths in cortex. *PLoS Comput. Biol.* 9, e1002848. doi:10.1371/journal.pcbi.1002848.

Zuo, Y., Lin, A., Chang, P., and Gan, W.-B. (2005). Development of Long-Term Dendritic Spine Stability in Diverse Regions of Cerebral Cortex. *Neuron* 46, 181–189. doi:10.1016/j.neuron.2005.04.001.

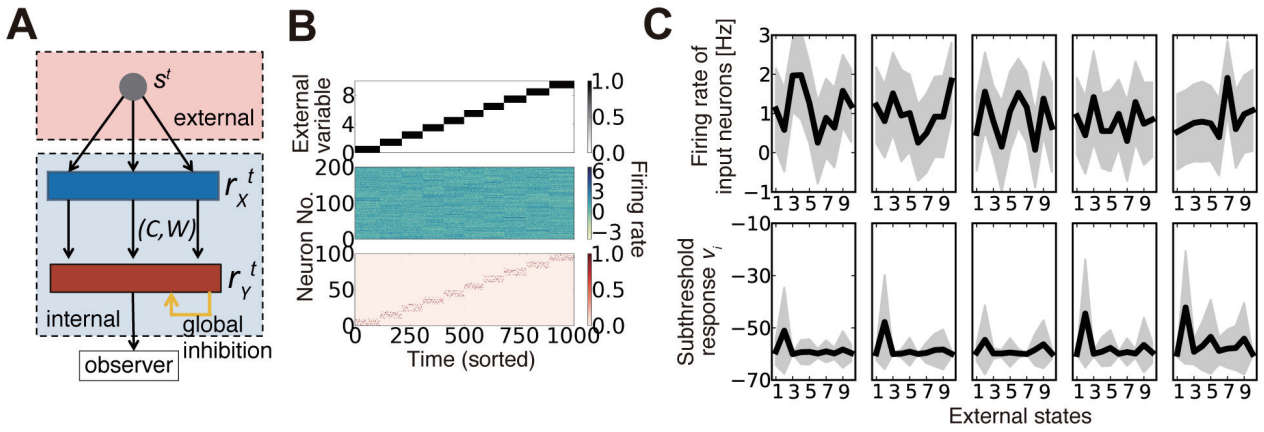



Figure 1: Description of the model. (**A**) Schematic diagram of the model. (**B**) An example of model behavior calculated at $\rho = 0.16$, when the synaptic connection is organized using the weight-coding scheme. The top panel represents the external variable, which takes an integer 0 to 9 in the simulation. The middle panel is the response of input neurons, and the bottom panel shows the activity of output neurons. In the simulation, each external state was randomly presented, but here the trials are sorted in ascending order. (**C**) Examples of neural activity in a simulation. Graphs on the top row represent the average firing rates of five randomly sampled input neurons for given external states (black lines) and their standard deviation (gray shadows). The bottom graphs are subthreshold responses of output neurons that represent the external state $s = 1$. Because the boundary condition for the membrane parameter $v_i \equiv \sum_j c_{ij}(w_{ij} r^t_{X,j} - h_w)$ was introduced as $v_i > \max_{i'}\{v_{i'} - v_d\}$, $v_i$ is typically bounded at $-v_d$. Note that $v_i$ is the unnormalized log-likelihood, and the units on the y-axis are arbitrary.

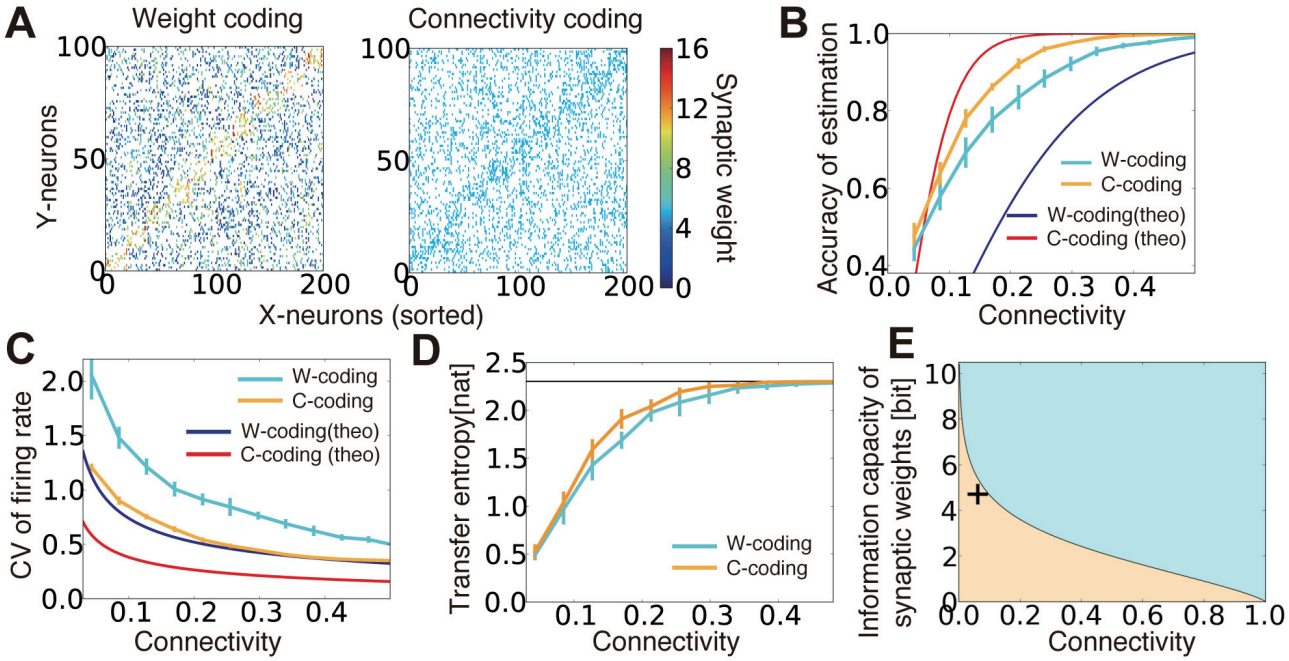



Figure 2: Performance comparison between connectivity coding and weight coding. (**A**) Examples of synaptic weight matrices in weight coding (W-coding) and connectivity coding (C-coding) schemes. X-neurons were sorted by their selectivity for external states. (**B**) Comparison of the performance between connectivity coding and weight coding schemes at various sparseness of connectivity. Orange and cyan lines are simulation results. The error bars represent standard deviation over 10 independent simulations. In the following panels, error bars are trial variability over 10 simulations. Red and blue lines are analytical results. (**C**) Analytically evaluated coefficient of variation (CV) of output firing rate and corresponding simulation results. For simulation results, the variance was evaluated over whole output neurons from their firing rates for their selective external states. (**D**) Estimated maximum transfer entropy for two coding strategies. Black horizontal line is the maximal information $log_e p$. (**E**) Relative information capacity of connection structure versus synaptic weight is shown at various values of synaptic connectivity. In the orange (cyan) area, the synaptic connectivity has higher (lower) information capacity than the synaptic weights. Plus symbol represents the data point obtained from CA3-to-CA1 connections.

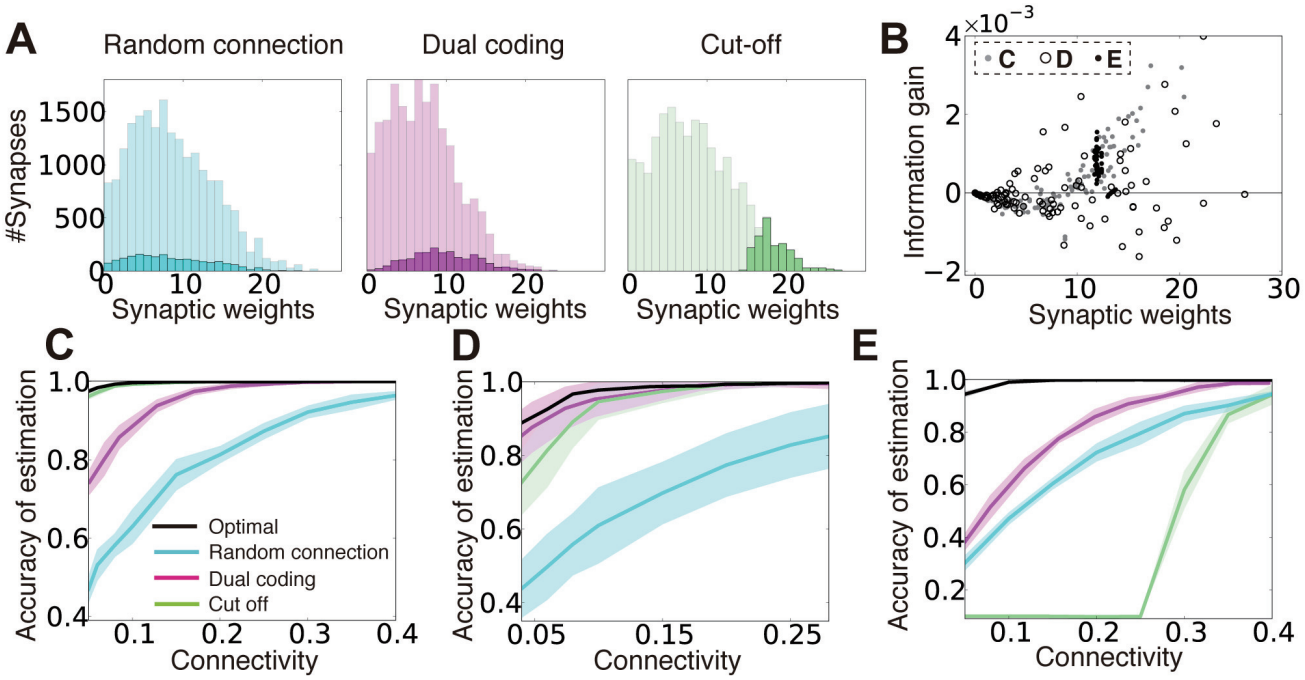



Figure 3: Dual coding yields robust information representation compared to fixed random connections and cut-off strategy. (**A**) Synaptic weight distributions in random connection (left), dual coding (middle), and cut-off (right) strategies. Light colors represent possible connections (i.e. distributions of synaptic weights under all-to-all connections), while dark colors show the actual connections. Connection probability was set at $\rho = 0.1$. (**B**) Relationships between the synaptic weight and the information gain per connection for three input configurations described in panels **C-E**. The open black circles were calculated with $\sigma_r = 2.0$ instead of $\sigma_r = 4.0$ for illustration purpose. (**C-E**) Comparisons of performance among different connection structure organizations. Note that black lines represent lower bounds for the optimal performance, but not the exact optimal solutions. In panel **D**, the means and standard deviations were calculated over 100 simulation trials instead of 10 due to intrinsic variability.

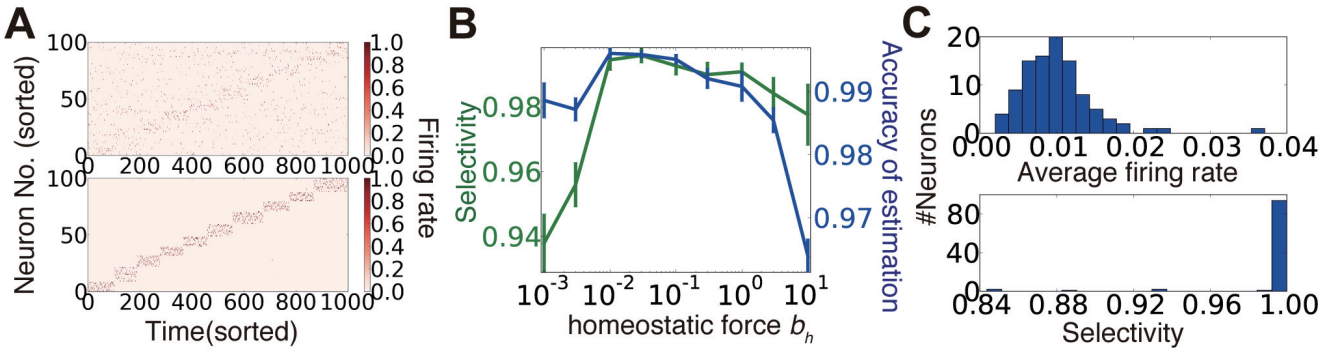



Figure 4: Synaptic weight learning on random connection structures. (**A**) An example of output neuron activity before (top) and after (bottom) synaptic weight learning calculated at connectivity $\rho = 0.4$. (**B**) Selectivity of output neurons and accuracy of estimation at various strengths of homeostatic plasticity at $\rho = 0.4$. Selectivity was defined as $\sum_{s^t=\mu} r^t_{Y,i} / \sum_t r^t_{Y,i}$ for $i \in \Omega_\mu$. (**C**) Histogram of average firing rates of output neurons (top), and selectivity of each neuron calculated for the simulation depicted in panel **A**.

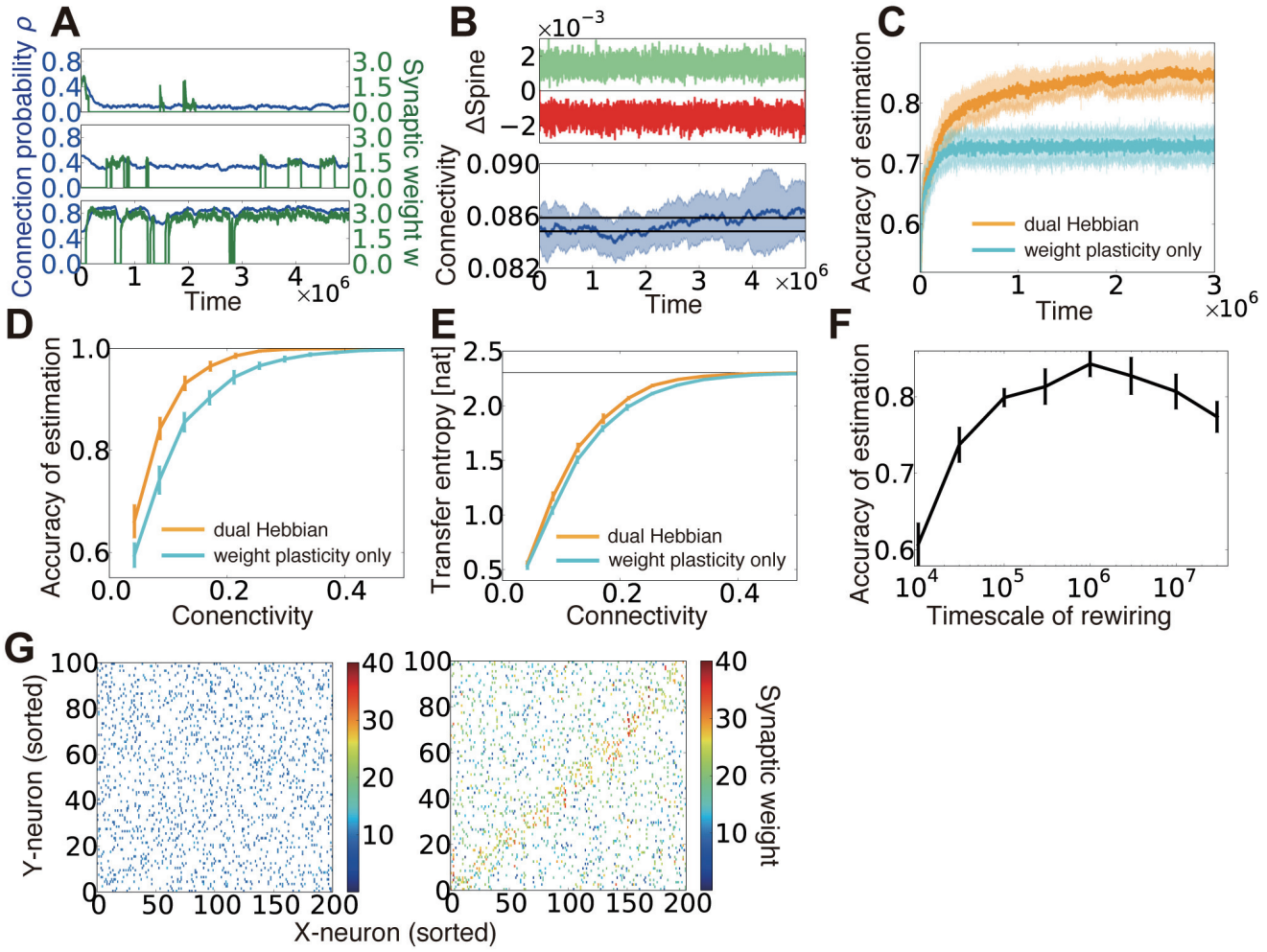



Figure 5: Dual Hebbian learning for synaptic weights and connections. (**A**) Examples of spine creation and elimination. In all three panels, green lines show synaptic weights, and blue lines are connection probability. When there is not a synaptic connection between two neurons, the synaptic weight becomes zero, but the connection probability can take a non-zero value. Simulation was calculated at $\rho = 0.48$, $\eta_\rho = 0.001$, and $\tau_c = 10^5$. (**B**) Change in connectivity due to synaptic elimination and creation. Number of spines eliminated (red) and created (green) per unit time was balanced (top). As a result, connectivity did not appreciably change due to rewiring (bottom). Black lines in the bottom graph are the mean connectivity at $\gamma = 0.1$ and $\gamma = 0.101$ in the model without rewiring. (**C**) Accuracy of estimation for the model with/without wiring plasticity. For the dual Hebbian model, the sparseness parameter was set as $\gamma = 0.1$, whereas $\gamma = 0.101$ was used for the weight plasticity model to perform comparisons at the same connectivity (see panel **B**). (**D, E**) Comparison of the performance (**D**) and the maximum estimated transfer entropy (**E**) after learning between the dual Hebbian model and the model implemented with synaptic plasticity only at various degrees of connectivity. Horizontal line in panel **E** represents the total information $log_e p$. (**F**) Accuracy of estimation with various timescales for rewiring $\tau_c$. Note that the simulation was performed only for $5 \times 10^6$ time steps, and the performance did not converge for the model with a longer timescale. (**G**) Synaptic weight matrices before (left) and after (right) learning. Both X-neurons (input neuron) and Y-neurons (output neurons) were sorted based on their preferred external states.

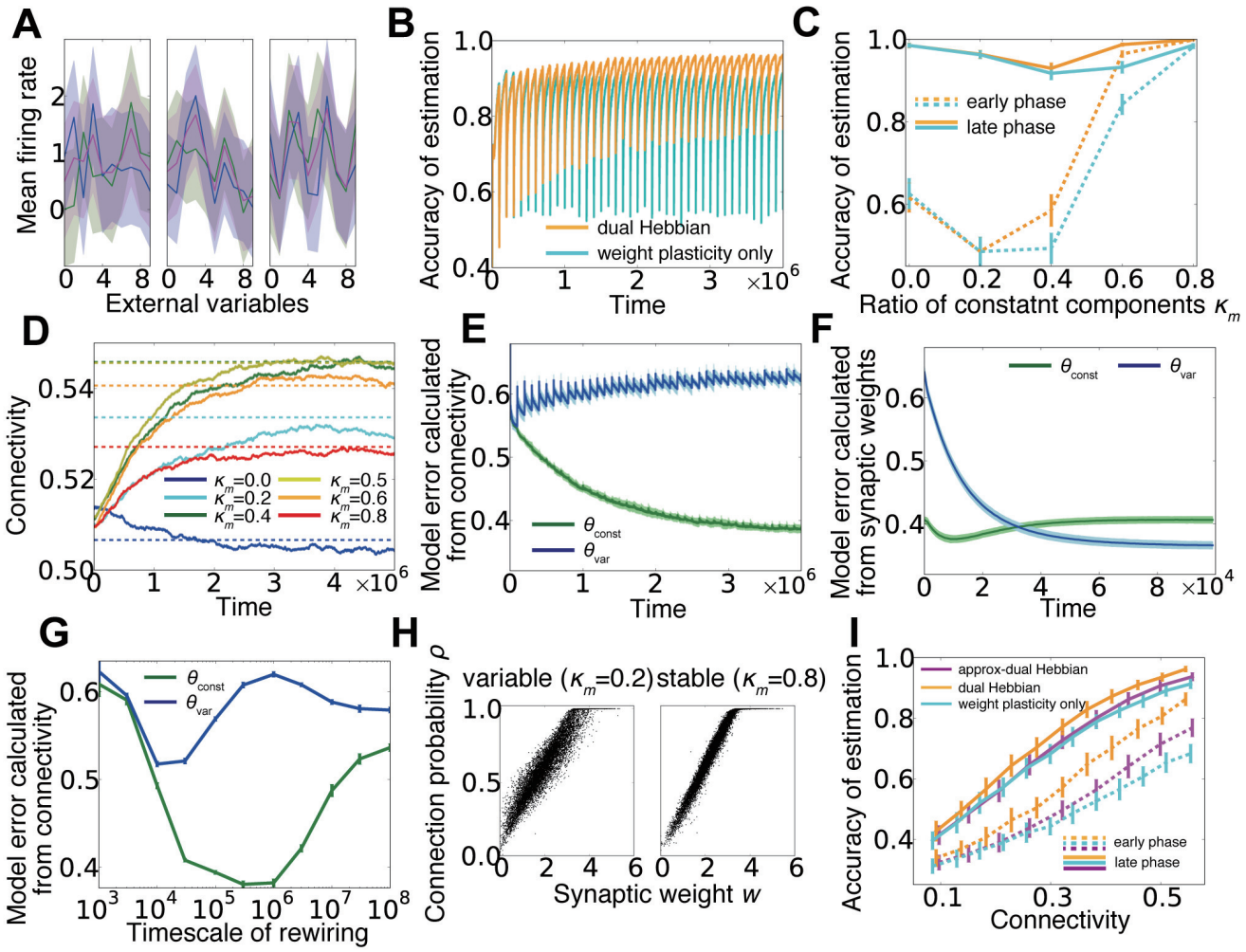



Figure 6: Dual learning under a dynamic environment. (**A**) Examples of input neuron responses. Blue lines represent the constant components $\theta_{const}$, green lines show the variable components $\theta_{var}$, and magenta lines are the total external models $\theta$ calculated from the normalized sum. (**B**) Learning curves for the model with or without wiring plasticity, when the variable components change every $10^5$ time steps. (**C**) Accuracy of estimation for various ratios of constant components. Early phase performance was calculated from the activity within 10,000 steps after the variable component shift, and the late phase performance was calculated from the activity within 10,000 steps before the shift. As in panel **B**, orange lines represent the dual Hebbian model, and cyan lines are for the model with weight plasticity only. (**D**) Trajectories of connectivity change. Connectivity tends to increase slightly during learning. Dotted lines are mean connectivity at $(\kappa_m, \gamma) = (0.0, 0.595)$, $(0.2, 0.625)$, $(0.4, 0.64)$, $(0.5, 0.64)$, $(0.6, 0.635)$, and $(0.8, 0.620)$. In **C**, these parameters were used for the synaptic plasticity only model, whereas $\gamma$ was fixed at $\gamma = 0.6$ for the dual Hebbian model. (**E,F**) Model error calculated from connectivity (**E**) and synaptic weights (**F**). Note that the timescale of **F** is the duration in which the variable component is constant, not the entire simulation (i.e. the scale of x-axis is $10^4$ not $10^6$). (**G**) Model error calculated from connectivity for various rewiring timescales $\tau_c$. For a large $\tau_c$, the learning process does not converge during the simulation. (**H**) Relationship between synaptic weight w and connection probability $\rho$ at the end of learning. When the external model is stable, $w$ and $\rho$ have a more linear relationship than that for the variable case. (**I**) Comparison of performances among the model without wiring plasticity (cyan), the dual Hebbian model (orange), the approximated model (magenta).

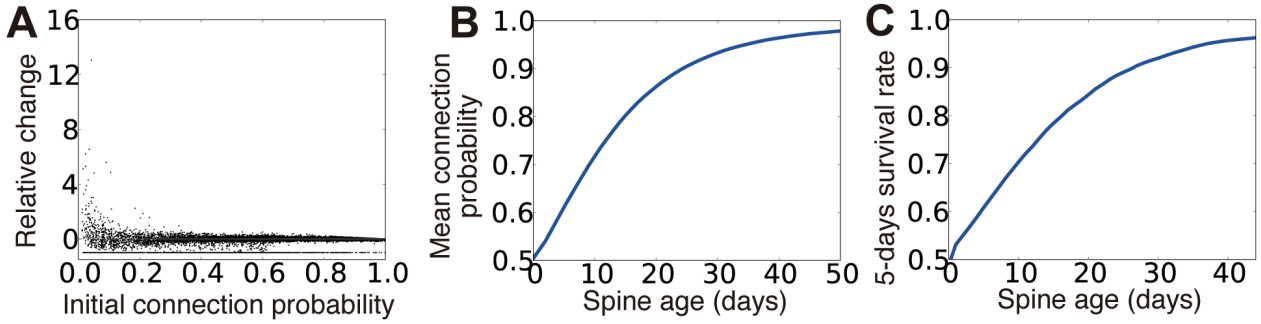



Figure 7: Spine dynamics of the approximated dual Hebbian model. (**A**) Relative change of connection probability in $10^5$ time steps. If the initial connection probability is low, the relative change after $10^5$ time steps has a tendency to be positive, whereas spines with a high connection probability are more likely to show negative changes. The line at the bottom represents eliminated spines (i.e., relative change = -1). (**B,C**) Relationships between spine age and the mean connection probability (**B**) and the 5-days survival rate (**C**). Consistent with the experimental results, survival rate is positively correlated with spine age. 5-days survival rate was calculated by regarding $10^5$ time steps as one day.

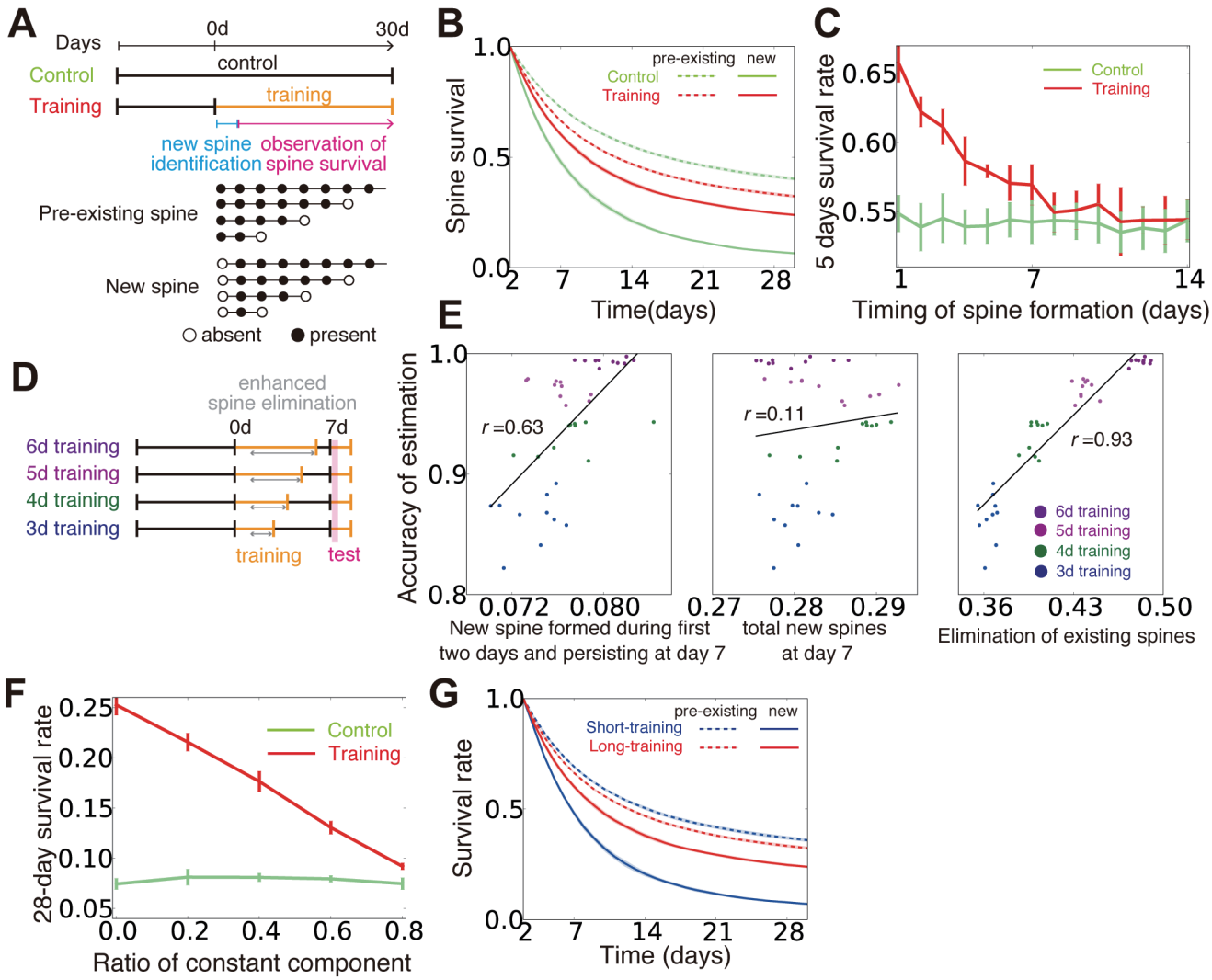
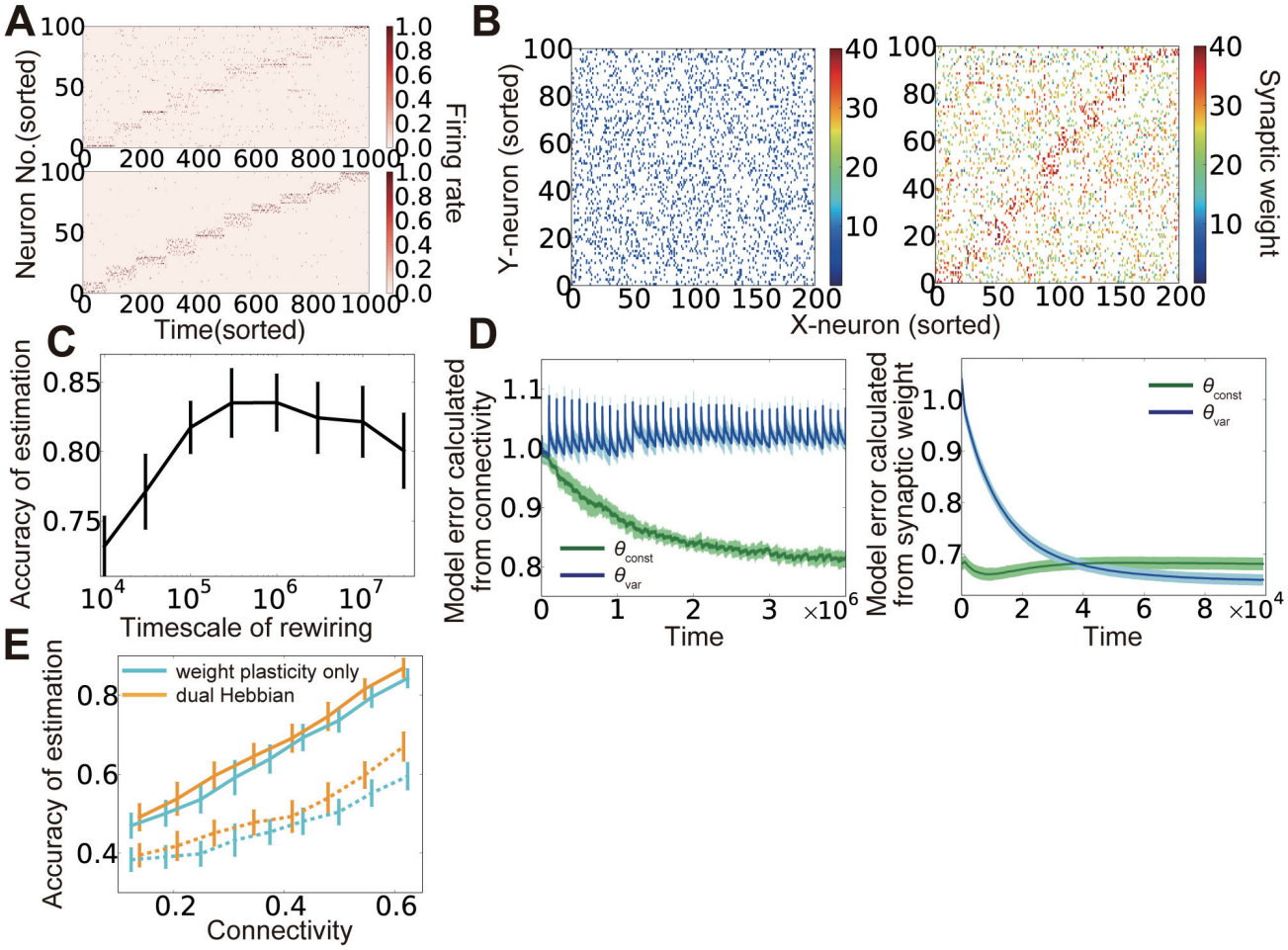



Figure 8: Influence of training on spine dynamics. (**A**) Schematic diagrams of the simulation protocols for panels **B,C**, and **F,G**, and examples of spine dynamics for pre-existing spines and new spines. (**B**) Spine survival rates for control and training simulations. Dotted lines represent survival rates of pre-existing spines (spines created before day 0 and existing on day 2), and solid lines are new spines created between day 0 and day 2. (**C**) The 5-day survival rate of spines created at different stages of learning. (**D,E**) Relationships between creation and elimination of spines and task performance. Performance was calculated from the activity within 2,000-7,000 time steps after the beginning of the test phase. In the simulation, the synaptic elimination was increased fivefold from day 1 to the end of training. (**F**) Effect of similarity between the control condition and training on the new spine survival rate. The value of $\kappa_m$ was changed as in **Figure 6C** to alter the similarity between the two conditions. Note that $\kappa_m = 0$ in panels **A-E**, and **G**. (**G**) Spine survival rates for short-training (2 d) and long-training (30 d) simulations. Pre-existing and new spines were defined as in panels **A,B**.




Supplementary Figure 1: Results in Poisson model. (**A**) An example of output neuron activity before (top) and after (bottom) synaptic weight learning at connectivity $\rho = 0.25$. (**B**) Synaptic weight matrices before (left) and after (right) learning. Both X-neurons and Y-neurons were sorted based on their preferred external states. (**C**) Accuracy of estimation at various timescale of rewiring $\tau_c$. (**D**) Model error calculated from connectivity (left) and synaptic weights (right). (**E**) Comparison of performance among the model without wiring plasticity (cyan), and dual Hebbian model(orange). Corresponding results in the Gaussian model are described in **Fig. 4A**, **Fig. 5F**, **Fig. 5G**, **Fig. 6EF**, **Fig. 6I** respectively.