Reward-based stochastic self-configuration of neural circuits

David Kappel[1,*], Robert Legenstein[*], Stefan Habenschuss, Michael Hsieh and Wolfgang Maass
Institute for Theoretical Computer Science
Graz University of Technology
8010 Graz, Austria
[1]corresponding author: kappel@igi.tugraz.at

[*]equal contribution

April 7, 2017

Abstract

Experimental data suggest that neural circuits configure their synaptic connectivity for a given computational task. They also point to dopamine-gated stochastic spine dynamics as an important underlying mechanism, and they show that the stochastic component of synaptic plasticity is surprisingly strong. We propose a model that elucidates how task-dependent self-configuration of neural circuits can emerge through these mechanisms. The Fokker-Planck equation allows us to relate local stochastic processes at synapses to the stationary distribution of network configurations, and thereby to computational properties of the network. This framework suggests a new model for reward-gated network plasticity, where one replaces the common policy gradient paradigm by continuously ongoing stochastic policy search (sampling) from a posterior distribution of network configurations. This posterior integrates priors that encode for example previously attained knowledge and structural constraints. This model can explain the experimentally found capability of neural circuits to configure themselves for a given task, and to compensate automatically for changes in the network or task. We also show that experimental data on dopamine-modulated spine dynamics can be modeled within this theoretical framework, and that a strong stochastic component of synaptic plasticity is essential for its performance.

keywords

Spine dynamics, rewiring, stochastic synaptic plasticity, reward-modulated STDP, reinforcement learning, policy gradient, sampling.

Introduction

Networks of neurons in the brain are known to rewire themselves on a time scale of hours to days [Holtmaat et al., 2005, Stettler et al., 2006, Yang et al., 2009, Holtmaat and Svoboda, 2009, Ziv and Ahissar, 2009, Minerbi et al., 2009, Kasai et al., 2010, Loewenstein et al., 2011, Loewenstein et al., 2015]. Experimental data suggest that task-dependent self-configuration of neural circuits results from an interplay of

stochastic processes and reward signals that stabilize spines [Yagishita et al., 2014]. Commonly considered deterministic rules for reward-gated synaptic plasticity are not suitable for modelling this process. Hence we propose a stochastic modelling framework that elucidates how neural circuits can employ local stochastic processes in order to install and maintain a concrete computational function. In Fig. 1 we show that the model is consistent with the data of [Yagishita et al., 2014] on reward-gated spine-stabilization. We show in Fig. 2 that our model reproduces data on reward driven emergence of a specific motor response, and the accompanying reorganization of network connectivity and dynamics [Peters et al., 2014].

Other recent data show that the stochastic component of synaptic plasticity is surprisingly strong, at least as strong as the impact of neural activity on synaptic plasticity (see the analysis of [Dvorkin and Ziv, 2016], which includes in Fig. 8 a reanalysis of data from [Kasthuri et al., 2015]). Our theoretical model and network simulations (see Fig. 3) suggest that this strong stochastic component is essential for network self-configuration.

Previous models for reward-gated network plasticity were based on policy-gradient approaches, where policies are defined implicitly through the parameters $\boldsymbol{\theta}$ of the network. They modelled the learning process as gradient ascent in the parameter space in order to maximize rewards. The experimentally found strong contribution of stochastic processes suggests to replace policy gradient by policy sampling models, where stochastic components of synaptic plasticity enable the network to sample continuously from a posterior distribution $p^*(\boldsymbol{\theta})$ of network configurations $\boldsymbol{\theta}$. This posterior distribution $p^*(\boldsymbol{\theta})$ favors those configurations $\boldsymbol{\theta}$ that provide good compromises between recently rewarded network outputs and priors that can encode for example previously attained knowledge and structural constraints.

The resulting model for reward-gated network rewiring and synaptic plasticity provides a new method for deriving rules for reward-gated synaptic plasticity from first principles, and can be applied to a wide range of neuron and network models. In particular, it significantly expands the stochastic approach from [Kappel et al., 2015] for unsupervised learning for a specific neuron model and a specific STDP-rule. It paves the way for moving data-based modelling of neural circuits and systems to the next stage, where one not only models the network dynamics, but also how these networks can attain and maintain a computational function.

Results

Our first goal is to lay out a mathematical framework for understanding stochastic reward-based rewiring of neural networks. We first introduce a general framework for stochastic synaptic rewiring and then extend the model to include reward-based plasticity processes. Consider a network scaffold $\mathcal{N}$ that contains a set of neurons and a set of potential synaptic connections between them. At each time point t this scaffold gives rise to a network configuration $\mathcal{N}(t)$ where some of the potential synaptic connections are realized. The experimentally found presence of multiple synaptic connections between two neurons can easily be accommodated within this framework, and is included in our simulations. We characterize the current configuration $\mathcal{N}(t)$ of a network of neurons at time t through a parameter vector $\boldsymbol{\theta}(t)$. Although $\boldsymbol{\theta}(t)$ can in general contain any network parameter (including for example also neuron excitabilities), we focus on the case where each real-valued value $\theta_i(t)$ encodes the state of a potential synaptic connection i. In order to

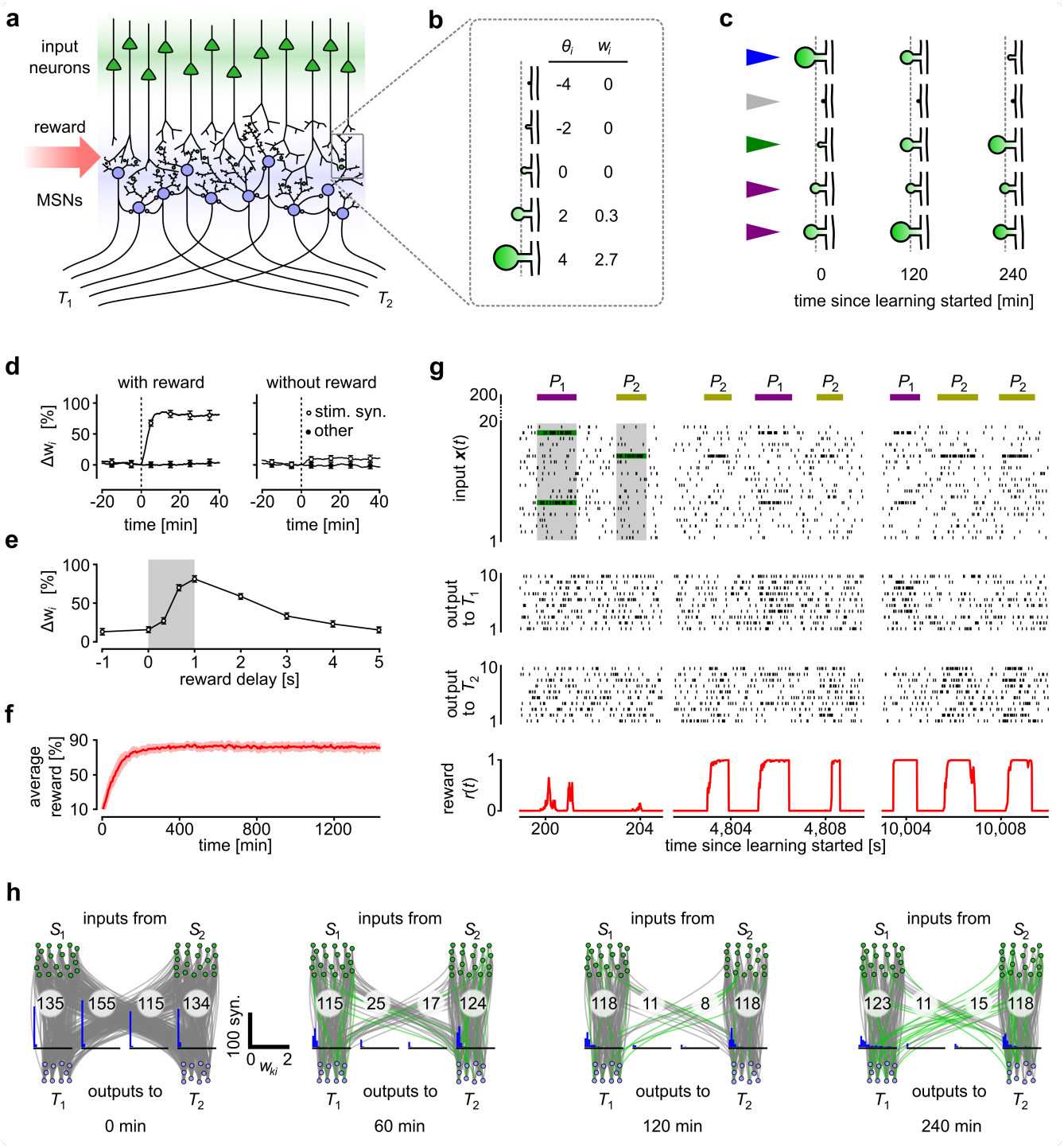

Figure 1: **Reward-based routing of input patterns.** **(a)** Illustration of the network scaffold architecture. A population of model MSNs (blue) receives input from a population of excitatory input neurons (green) that model cortical neurons. Potential synaptic connections between these 2 populations of neurons were subject to reward-based synaptic sampling. In addition, fixed lateral connections provide recurrent inhibitory input to the MSNs. *Caption continued on next page...*

Caption of Fig. 1 continued: The MSNs are divided into two assemblies, each projecting exclusively to one of two target areas T_1 and T_2. Reward is delivered whenever the network manages to route an input pattern P_i primarily to that assembly of MSNs that projects to target area T_i. (**b)** Illustration of the model for spine dynamics. Five potential synaptic connections at different states are shown. Synaptic spines are represented by circular volumes with diameters proportional to $\sqrt[3]{w_i(t)}$, assuming a linear correlation between spine-head volume and synaptic efficacy [Matsuzaki et al., 2001]. Spine neck dimensions are scaled accordingly to facilitate the illustration. (**c)** Snapshots of five potential synaptic connection of the network shown at three different time points throughout learning. Both transient and stable behavior can be found. The color of the arrow heads indicates different behaviors (blue: transiently decaying, gray: stably inactive, green: transiently emerging, purple: stably active). (**d)** Time course of synaptic efficacies under a reward-modulated STDP pairing protocol according to Eq. (3). Reward delivery after STDP pairings results in strong synaptic weight increase (left). This effect is reduced without reward (right), and prevented completely if no presynaptic stimulus is applied (dashed lines: pairing onset time). Compare with Fig. 1F,G in [Yagishita et al., 2014]. (**e)** Dependence of resulting changes in synaptic weights as a function of the delay of reward delivery. Gray time window indicates application of the STDP pairing protocol. Values represent percentage of weight change relative to the pairing onset time (means and s.e.m. over 50 synapses). Compare to Figure 1O in [Yagishita et al., 2014]. (**f)** The average reward throughout learning (mean over 5 independent trial runs; shaded area indicates s.e.m.). (**g)** The spiking activity of the network during learning. Activities of 20 randomly selected input neurons and all MSNs are shown. 3 salient input neurons (belonging to pools S_1 or S_2 in **h**) are highlighted in green. The neurons that project to target areas T_1 and T_2 have learned to respond to their associated input pattern with increased firing activity. The reward delivered to the synapses is shown at the bottom. (**h)** Dynamics of network rewiring throughout learning. Several network configurations $\mathcal{N}(t)$ for times t indicated below the plots. Gray lines indicate active connections between neurons; connections that were not present at previous time points are highlighted in green. All output neurons and two subsets of input neurons that fire strongly in pattern P_1 or P_2 are shown (pools S_1 and S_2, 20 neurons each). The current numbers of connections between pools are printed on top. Histograms of synaptic weights are shown for each pair of pools below (blue). The connectivity was initially dense and then rapidly restructured and became sparser. Rewiring took place all the time throughout learning.

allow our model to describe both rewiring and synaptic plasticity, the value of $\theta_i(t)$ encodes both whether connection i is functional and, in case it is functional, its synaptic efficacy (weight). Concretely, negative values of $\theta_i(t)$ denote a nonfunctional synaptic connection, and positive values of $\theta_i(t)$ encode the efficacy of the corresponding functional synaptic connection. The relation between $\theta_i(t)$ and the corresponding synaptic weight w_i is in our model given by an exponential mapping

$$w_i(t) = \exp(\theta_i(t) - \theta_0) \;, \tag{1}$$

where θ_0 is an offset parameter. This model provides a simple mechanism for synaptic rewiring, since for a sufficiently large offset θ_0 (we used $\theta_0 = 3$ in our simulations) this function maps all negative values of $\theta_i(t)$ (i.e., non-functional synapses) onto approximately vanishing synaptic weights (in simulations, we set the weights of all synapses with $\theta_i(t) < 0$ explicitly to zero). In addition, we will see below that this exponential mapping leads to parameter dynamics that is consistent with experimental findings.

Various experimental data indicate that synaptic parameters follow stochastic dynamics [Loewenstein et al., 2011,Statman et al., 2014,Dvorkin and Ziv, 2016]. In [Loewenstein et al., 2011] it was shown that the

experimentally found random increases and decreases of the logarithm of spine sizes is well described by an Ornstein-Uhlenbeck process. Since spine sizes are strongly (linearly) correlated with synaptic efficacies, and since the logarithm of $w_i(t)$ in our model is given by $\theta_i(t)$ (see Eq. (1)), we characterize the dynamics of the synaptic parameter $\theta_i(t)$ by the following stochastic differential equation that contains Ornstein-Uhlenbeck dynamics as a special case, see below and [Loewenstein et al., 2011, Kappel et al., 2015]:

$$d\theta_i \;=\; \beta\,\frac{\partial}{\partial\theta_i}\,\log p^*(\boldsymbol{\theta})\,dt \;+\; \sqrt{2T\beta}\,d\mathcal{W}_i\;. \tag{2}$$

For the sake of brevity we have suppressed the time dependencies of parameters in our notation (see *Methods* for a more rigorous notation). Using the Fokker-Planck equation, we show (see Theorem 1 in *Methods*) that if the stochastic dynamics of each parameter θ_i is given by the stochastic differential equation Eq. (2), then the network samples after some burn-in period from the distribution $p^*(\boldsymbol{\theta})$ over network configurations that appears in Eq. (2). In other words, $p^*(\boldsymbol{\theta})$ is the stationary distribution of all the parameters θ_i that results from the stochastic processes Eq. (2), i.e. the global probability distribution over all synaptic parameters θ_i that emerges when the process is observed over long time. One can insert in principle any target distribution $p^*(\boldsymbol{\theta})$ into Eq. (2). We will focus here on the case where $p^*(\boldsymbol{\theta})$ assigns the highest probability to network configurations that support a desirable computational capability of the network. The constant $\beta > 0$ in Eq. (2) controls the speed of the synaptic dynamics. The last term $d\mathcal{W}_i$ of Eq. (2) describes infinitesimal stochastic increments and decrements of a Wiener process $\mathcal{W}_i$ – a standard model for Brownian motion in one dimension (see [Gardiner, 2004]). If we set $p^*(\boldsymbol{\theta})$ to be a Gaussian distribution, we recover the Ornstein-Uhlenbeck process that was observed in [Loewenstein et al., 2011]. The stationary distribution over the synaptic weights w_i is then given by a log-normal distribution in accordance with experimental data [Loewenstein et al., 2011, Buzsáki and Mizuseki, 2014]. The amplitude of the noise term is scaled by the temperature parameter $T > 0$, which can be used to increase or decrease random exploration of the parameter space. For small temperatures the stationary distribution becomes narrower, for large temperatures the variance increases (see *Methods*).

The resulting dynamics of synaptic connections (e.g. for connections from cortex to striatum, see Fig. 1a) is illustrated in Fig. 1b-c. Five example spines of different sizes and corresponding values of synaptic parameters $\theta_i(t)$ and synaptic efficacies $w_i(t)$ are shown in Fig. 1b. An example for the temporal evolution of these five spines is shown in Fig. 1c for three different time points of the subsequently described learning process. Different patterns of temporal evolution emerged for the spines in our model, analogous to those found in experimental data [Holtmaat et al., 2005, Yasumatsu et al., 2008]. Many synapses were transient, while some synapses stabilized (persistently functional or non-functional).

Reward-based synaptic plasticity and rewiring as Bayesian policy sampling

To integrate structural constraints with reward-based learning in the stochastic rewiring framework, we assume that $p^*(\boldsymbol{\theta})$ in Eq. (2) is proportional to the product of a prior $p_{\mathcal{S}}(\boldsymbol{\theta})$ with the expected discounted reward $\mathcal{V}(\boldsymbol{\theta})$. Both of these terms will be described in detail below. In this case, $p^*(\boldsymbol{\theta})$ describes the posterior distribution over network configurations that combines structural constraints and previous learnt knowledge with the goal of maximizing $\mathcal{V}(\boldsymbol{\theta})$ in a Bayes optimal manner, see *A Bayesian framework for*

reward-modulated learning in *Methods*. Using this definition of $p^*(\boldsymbol{\theta})$ in Eq. (2), we obtain the following stochastic parameter dynamics:

$$d\theta_i \;=\; \beta \left(\frac{\partial}{\partial \theta_i} \log p_{\mathcal{S}}\left(\boldsymbol{\theta}\right) \;+\; \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta}) \right) dt \;+\; \sqrt{2T\beta}\, d\mathcal{W}_i \;. \tag{3}$$

Due to the temperature-dependent noise term that acts on each network parameter, the synapses continuously probe different states that balance the constraints induced by the prior $p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)$ and the expected reward $\mathcal{V}(\boldsymbol{\theta})$.

We now discuss the two terms $p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)$ and $\mathcal{V}(\boldsymbol{\theta})$ in more detail. The prior $p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)$ can take for example into account that each configuration $\mathcal{N}(t)$ of a network needs to satisfy structural constraints, such as sparse connectivity. We use a Gaussian distribution that prefers small but nonzero weights throughout all simulations (see *Methods*). When the contribution of the second term $\mathcal{V}(\boldsymbol{\theta})$ can be neglected, such a Gaussian prior $p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)$ leads to the experimentally observed Ornstein-Uhlenbeck spine-size dynamics as discussed above. The second term $\mathcal{V}(\boldsymbol{\theta})$ is the expected reward associated with a given set of parameters $\boldsymbol{\theta}$. We assume that the network scaffold $\mathcal{N}$ receives reward signals $r(t)$ at certain times t, e.g., in the form of dopamine. Formally, the objective function $\mathcal{V}(\boldsymbol{\theta})$ is defined at time $t = 0$ by the expected discounted reward for a given parameter vector $\boldsymbol{\theta}$, where rewards in the immediate future are strongly preferred

$$\mathcal{V}(\boldsymbol{\theta}) \;=\; \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, r(\tau)\, d\tau \right\rangle_{p(\boldsymbol{r}|\boldsymbol{\theta})} \;. \tag{4}$$

In Eq. (4) we integrate over all future rewards $r(\tau)$, while discounting more remote rewards exponentially with a discount rate τ_e, which for simplicity was set equal to 1 s in this paper. We find (see Eq. (15) in Methods) that this time constant τ_e is immediately related to the experimentally studied time window or eligibility trace for the influence of dopamine on synaptic plasticity [Yagishita et al., 2014]. The expectation in Eq. (4) is taken over multiple learning episodes, where in each episode one realization of the reward sequence $\boldsymbol{r} = \{r(\tau), \tau \geq 0\}$ is encountered. More precisely, this expectation $\langle \cdot \rangle_{p(\boldsymbol{r}|\boldsymbol{\theta})}$ is taken with respect to the distribution $p(\boldsymbol{r}|\boldsymbol{\theta})$ over sequences $\boldsymbol{r}$ of future rewards that result from the given set of synaptic parameters $\boldsymbol{\theta}$. The probabilities $p(\boldsymbol{r}|\boldsymbol{\theta})$ represent averages over the influences of initial network activity, the stochastic effects of network inputs, network responses, and stochastic reward delivery, see *Methods* for details. Furthermore we develop in *Methods* an online learning theory that does not require us to explicitly compute the expectation over episodes in Eq. (4). The input-output behavior of the network parametrized by $\boldsymbol{\theta}$ is referred to as policy in the context of reinforcement learning.

When the parameter dynamics are given solely by the second term in the parentheses of Eq. (3), $\frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$, we recover the standard policy gradient method [Williams, 1992, Baxter and Bartlett, 2000, Peters and Schaal, 2006] for reinforcement learning. In this method, the parameters are gradually changed such that the expected discounted reward $\mathcal{V}(\boldsymbol{\theta})$ is increased locally. This is achieved by parameter dynamics that follows the gradient of $\mathcal{V}(\boldsymbol{\theta})$, i.e., $\frac{d\theta_i}{dt} = \beta \frac{\partial}{\partial \theta_i} \mathcal{V}(\boldsymbol{\theta})$ or equivalently $\frac{d\theta_i}{dt} = \beta \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$, where $\beta > 0$ is a small learning rate.

In contrast to policy gradient, the reinforcement learning model proposed here does not converge to a locally optimal network configuration, but produces permanently changing configurations, with a preference

for configurations that both satisfy structural constraints and provide a large expected reward $\mathcal{V}(\boldsymbol{\theta})$. We therefore refer to this learning model as *Bayesian policy sampling*, and to the family of reward-based plasticity rules that is defined by Eq. (3) as *reward-based synaptic sampling*.

Relationship to previous models for reward-based learning using eligibility traces

For the simulations described below, we considered standard models for networks of spiking neurons (see *Network model* in *Methods*). In this case, the derivative $\frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$ defines synaptic updates at a synapse i that are essentially given by the product of the current reward signal $r(t)$ and an eligibility trace $e_i(t)$, see *Reward-modulated synaptic plasticity* in *Methods*. The dynamics of the eligibility trace is given by

$$\frac{de_i(t)}{dt} \;=\; -\frac{1}{\tau_e} e_i(t) \;+\; w_i(t)\, y_{\mathrm{PRE}_i}(t)\, (z_{\mathrm{POST}_i}(t) - f_{\mathrm{POST}_i}(t)) \;, \tag{5}$$

where $w_i(t)\, y_{\mathrm{PRE}_i}(t)$ is the value at time t of the postsynaptic potential evoked by synapse i , $z_{\mathrm{POST}_i}(t)$ denotes the postsynaptic spike train (formally, a sum of Dirac delta pulses at the times of spikes), and $f_{\mathrm{POST}_i}(t)$ is the instantaneous firing rate of the postsynaptic neuron at time t (see Eq. (14) in *Methods*). The eligibility trace accumulates spike-timing dependent eligibilites of the synapse, where postsynaptic events that follow a presynaptic spike tend to increase the eligibility while presynaptic spikes alone decrease the eligibility in proportion to the instantaneous firing rate. In the absence of pre- or postsynaptic spikes, the eligibility trace decays with a time constant τ_e (we used 1 s in our simulations). Hence, a reward induces a significant synaptic change only if there have been such events on the time-scale of τ_e. Such plasticity rules for policy gradient learning in spiking neural networks have previously been proposed by [Pfister et al., 2006, Florian, 2007, Legenstein et al., 2008, Urbanczik and Senn, 2009]. For non-spiking neural networks, a similar update rule was first introduced by Williams and termed the REINFORCE rule [Williams, 1992]. In fact, when discretizing time and under the assumption that rewards and parameter updates are only induced at the end of each episode, we recover the REINFORCE rule in the limit $\tau_e \to \infty$.

Eq. (5) induces multiplicative synaptic dynamics, such that the amount of changes in a given time window is proportional to the current efficacy of the synapse (due to the multiplication with $w_i(t)$). This implies (see *Methods*) that for disconnected synapses the influence of the eligibility trace and therefore of the policy-gradient term in Eq. (3) vanishes. Hence, the dynamics of disconnected synapses does not depend on neural activity or reward, which is important since non-functional synapses do not have access to such information. In our model, synapses reappear randomly according to the stochastic process of the form Eq. (2). However, an explicit update of $\theta_i(t)$ for non-functional synapses is not necessary, since this process results in a distribution over reappearance times that was found in our simulations to be similar to the distribution of inter-event times of a Poisson point process (see [Ding and Rangarajan, 2004] for a detailed analysis).

Our approach places previously proposed rules for reward-gated synaptic plasticity based on eligibility traces in a new context. A key difference to previous models for reward-gated synaptic plasticity and policy gradient learning is that Eq. (3) also contains a first term that arises from a prior for network configurations, e.g., sparse connectivity; and a term $d\mathcal{W}_i$ (last term in Eq. (3)) that models experimentally observed synapse-autonomous stochastic processes such as spine dynamics. This stochastic term is not

compatible with policy gradient learning, since it creates a permanently ongoing search, i.e., Bayesian policy sampling. We propose that the temperature of this Bayesian policy sampling is regulated by a biological mechanism that optimizes the trade-off between exploration of new network configurations and exploitation of the currently found configuration. In Fig. 1 we show that this stochastic exploration allows us to model rewiring in spiking neural networks and that only synapses that are functionally relevant are maintained with high probability. We further investigate the role of the temperature parameter T to enhance parameter exploration in a model for motor cortex plasticity in Fig. 3.

A model for task-dependent rewiring of synaptic connections from cortex to medium spiny neurons (MSNs) in the basal ganglia

Here, we ask whether our model is sufficient for explaining task-dependent rewiring of synaptic connections from cortex to MSNs as reported in [Yagishita et al., 2014]. We examine this question in the context of a simple functional goal: that two different distributed activity patterns P_1, P_2 of upstream neurons in the cortex are routed to two different ensembles of MSNs, and thereby to two different downstream targets T_1 and T_2 (see Fig. 1a,g). In this way we can address the question whether the mechanisms reported in [Yagishita et al., 2014] enable the brain to activate (or rather, disinhibit) specific behaviors for different activation patterns P_1, P_2 in the cortex.

The experimental data of [Yagishita et al., 2014] elucidated the interaction between reward signals (dopamine) and spine dynamics. In particular, they reported in Fig. 1 that the volumes of excitatory synaptic spines show significant changes only when pre- and postsynaptic activity is paired with precisely timed delivery of dopamine (see [Yagishita et al., 2014] Fig. 1 E-G, O). More precisely, an STDP pairing protocol followed by dopamine uncaging induces strong LTP in corticostriatal synapses, whereas the same protocol without dopamine uncaging leads only to a minor increase of synaptic efficacies. We applied the same STDP pairing protocol to our synapse model and found that these experimental data can be reproduced (Fig. 1d). The parameters of the model that reproduced (according to Fig. 1d) the results from Figures 1F,G of [Yagishita et al., 2014] were used throughout subsequent analyses. Another important result of [Yagishita et al., 2014] is the existence of a rather narrow time window after synapse activation during which dopamine promotes spine enlargement, see their Fig. 1O. This result was also reproduced by our model (Fig. 1e).

To answer the question whether the Bayesian policy sampling mechanism in Eq. (3) is sufficient to explain the creation of different striatal pathways for different activity patterns P_1, P_2 of upstream neurons in the cortex, we analyzed the network scaffold illustrated in Fig. 1a. It consisted of 20 inhibitory model MSNs with lateral recurrent connections. These received feedforward excitatory input from 200 input neurons that model neurons distributed throughout the cortex. The synapses from input neurons to model MSNs were subject to plasticity and rewiring. Multiple connections were allowed between each pair of input neuron and MSN (see *Methods*). The MSNs were randomly divided into two assemblies, each projecting exclusively to one of two downstream target areas T_1 and T_2. Cortical input $\boldsymbol{x}(t)$ was modeled as Poisson spike trains from the 200 input neurons with instantaneous rates defined by two prototype rate patterns P_1 and P_2, see Fig. 1g. The task was to learn to activate T_1-projecting neurons and to silence T_2-projecting

neurons whenever pattern P_1 was presented as cortical input. For pattern P_2, the activation should be reversed (activate T_2-projecting neurons and silence those projecting to T_1). This desired function was defined through a reward signal $r(t)$ that was proportional to the ratio between the mean firing rate of the assembly projecting to the associated target and that of the non-target projecting assembly (see *Methods*).

Fig. 1g shows the firing activity and reward signal of the network during segments of one simulation run. After about 80 minutes of simulated biological time, each assembly of MSNs neurons had learned to increase its firing rate when the activity pattern P_i associated with its projection target T_i was presented to the network. Fig. 1f shows the average reward throughout learning. After 3 hours of learning about 82% of the maximum reward was acquired on average, and this level was maintained during prolonged learning.

Fig. 1h depicts several network configurations $\mathcal{N}(t)$ for times t indicated below the plots. Snapshots of the network connectivity are shown at five different times throughout learning. We considered two subsets S_1 and S_2 of the input neurons, that fired strongly for activity patterns P_1 and P_2, respectively (see *Methods* for details). Learning started from a densely connected network with small synaptic strengths. During learning the connectivity became sparser and the output neurons received significantly more and stronger connections from the pool S_i of input neurons that fired strongly during the corresponding pattern P_i. However, small fractions of synapses remained present between the other pairs of neurons. We propose that these (at this stage) functionally useless connections will support the exploration of new network configurations when the task or input patterns change. Although the network performance in the task did not change significantly after about 3 hours (see Fig. 1f), permanent network rewiring was observed throughout prolonged learning (new synaptic connections are marked in green). Hence in our model the network configuration does not remain fixed after good task performance has been achieved, but keeps alternating between different but functionally equivalent connectivity patterns.

A model for task-dependent self-configuration of a recurrent network of excitatory and inhibitory spiking neurons

Changes of network activity and spine turnover in motor cortex were monitored in [Peters et al., 2014] through calcium imaging over 2 weeks, while mice acquired a forelimb lever-press task through reward-based learning. A reward was given when a lever press crossed two thresholds within a given time window marked by an auditory cue. We examined to what extent a simple model based on the previously described framework for network plasticity would be able to reproduce the observed changes in neural activity, the observed transient turnover in spine dynamics, and the learning of the task.

We adapted the learning task of [Peters et al., 2014] in the following way for our model (see Fig. 2a). The beginning of a trial was indicated through the presentation of a cue input pattern $\boldsymbol{x}(t)$ (a fixed, randomly generated rate pattern for all 200 input neurons that lasted until the task was completed, but at most 10 s). When the lever position crossed the threshold +5 after first crossing a lower threshold -5 (black horizontal lines in Fig. 2a,b) within 10 s after cue onset a 400 ms reward window was initiated during which $r(t)$ was set to 1 (red vertical bars in Fig. 2b). Unsuccessful trials were aborted after 10 seconds and no reward was delivered. After each trial a brief holding phase of random length was inserted, during which input neurons were set to a background input rate of 2 Hz.

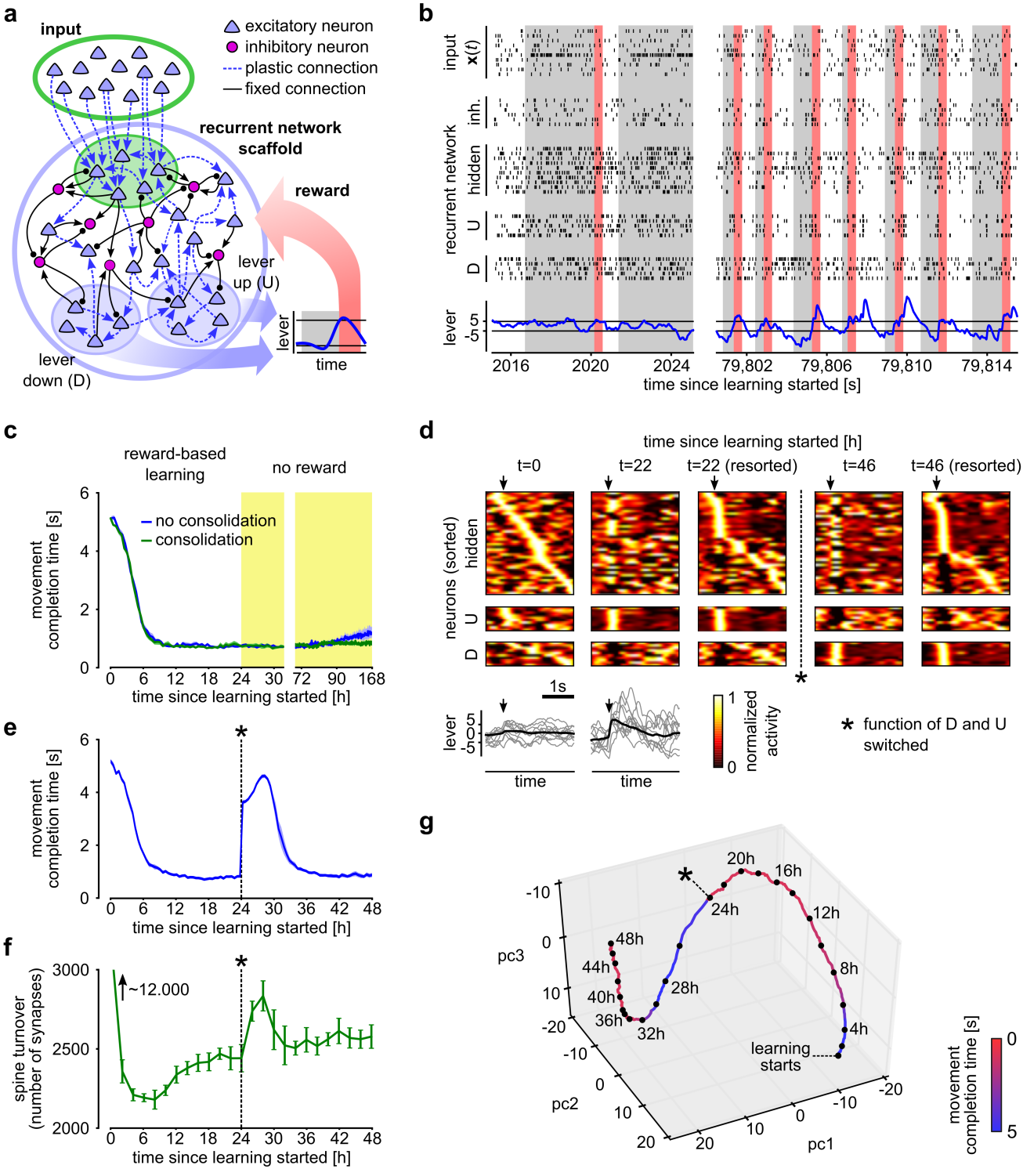

Figure 2: **Reward-based self-configuration of a recurrent neural network in a model for the task of [Peters et al., 2014].** **(a)** Network scaffold and task schematic. A recurrent network scaffold of excitatory and inhibitory neurons (large blue circle); a subset of excitatory neurons received input from afferent excitatory neurons (indicated by green shading). *Caption continued on next page...*

Caption of Fig. 2 continued: From the remaining excitatory neurons, two pools D and U were randomly selected to control lever movement (blue shaded areas). Multiple plastic synaptic connections were allowed between each pair of excitatory neurons. Dashed blue lines with arrows indicate potential excitatory synapses. Synapses to and from inhibitory neurons were kept fixed (full black lines with arrows and dots respectively). The inset at the bottom shows the stereotypical movement that had to be generated to receive a reward. (**b)** Spiking activity of the network at learning onset and after 22 hours of learning. Activities of random subsets of neurons from all populations are shown (hidden: excitatory neurons of the recurrent network, which are not in pool D or U). Lever position inferred from the neural activity in pools D and U is shown at the bottom. Rewards are indicated by red bars. Gray shaded areas indicate cue presentation. (**c)** Task performance quantified by the average time from cue presentation onset to completion of the task. The network was able to solve this task in less than 1 seconds on average after about 8 hours of learning. After 24 hours the reward delivery was stopped. Despite continued stochastic synaptic dynamics, the network maintained stable performance for three subsequently simulated days. When a simple consolidation mechanism was included in the synaptic dynamics, the performance remained stable without rewards for more than a week (note the change in time scale). (**d)** Trial-averaged network activity (top) and lever movements (bottom). Activity traces are aligned to movement onsets (arrows). Y-axis of trial-averaged activity plots are sorted by the time of highest firing rate. Sorting of the first and second plot is based on the activity at $t = 0$, third and fourth by that at $t = 22$, fifth is resorted by the activity at $t = 46$. The functional roles of D and U were switched after 24 hours. Average lever movement (black) and 10 individual movements (gray) are shown at the bottom. (**e)** Task performance as in **c** for the experiment in **d**. The task switch at 24 hours is accompanied by a transient increase of the temperature (see also **f**). (**f)** Analysis of the spine turnover measured in the total number of spines that appeared or disappeared during a time window of 2 hours, for the experiment shown in **e**. Y-axis is clipped at 3000. During the first two hours around 12.000 synapses turned over ($\sim 25\%$ of total number of spine). Another increased in spine turnover rate can be observed during learning the new task after 24 hours. (**g)** PCA of a random subset of the parameters θ_i. The plot suggests continuing dynamics in task-irrelevant dimensions after the learning goal has been reached (indicated by red color). When the function of the neuron pools U and D was switched after 24 h, the synaptic parameters migrated to a new region. Mean and s.e.m. were computed over 5 independent simulation runs in all plots.

We asked whether a generic recurrent network scaffold of excitatory and inhibitory spiking neurons with connectivity parameters taken from layer 2/3 in mouse cortex [Avermann et al., 2012] would learn to accomplish this task. The recurrent network scaffold consisted of 60 excitatory and 20 inhibitory neurons (see Fig. 2a). Half of the excitatory neurons received connections from 200 afferent excitatory input neurons. From the remaining 30 neurons we randomly selected one pool D of 10 excitatory neurons to cause downwards movements of the lever, and another pool U of 10 neurons for upwards movements. We refer to the 40 excitatory neurons that were not members of D or U as hidden neurons. Synaptic connections and weights from and to inhibitory neurons were randomly chosen and fixed (see *Methods*). All excitatory synaptic connections from the external input (cue) and between the 60 excitatory neurons (including those in the pools D and U) in the network were subjected to reward-based synaptic sampling. Thus, the network had to learn without any guidance, except for the reward in response to good performance, to create after the onset of the cue first higher firing in pool D, and then higher firing in pool U. This task was challenging, since no information was provided about which neurons belonged to pools D and U. Moreover, the synapses did not "know" whether they connected to hidden neurons, neurons within a pool, hidden neurons and pool-neurons, or input neurons with other neurons. Furthermore the plasticity of all these

different synapses was gated by the same global reward signal. Since the pools D and U did not receive direct connections from the input neurons, the network also had to learn to communicate the presence of the cue pattern to these pools.

Network responses before and after learning are shown in Fig. 2b. Initially, the rewarded goal was only reached occasionally. After learning for 8 hours the network was able to solve the task in most of the trials, and the average trial duration (movement completion time) decreased to less than 1 seconds (851 ± 46 ms, Fig. 2c). Decreased trial durations were accompanied by more stereotyped network activity and lever movement patterns as in the experimental data of [Peters et al., 2014]: compare our Fig. 2d with Fig. 1b and Fig. 2j of [Peters et al., 2014]. In Fig. 2d we show the trial-averaged activity of the 60 excitatory neurons before and after learning for 22 hours. The neurons are sorted in the first two plots of Fig. 2d by the time of maximum activity after movement onset times in the right plot after 22 hours of learning, i.e. the time point when the lever movement speed first exceeded a certain threshold (see *Methods* and [Peters et al., 2014]). These plots show that reward-based learning led to a restructuring of the network activity. In particular, an assembly of neurons emerged that controlled a sharp upwards movement. Also, less background activity was observed after 22 hours of learning, in particular for neurons with early activity peaks. Lower panels in Fig. 2d show the average lever movement and 10 individual movement traces at the beginning and after 22 hours of learning. The lever movements became more stereotyped during learning featuring a sharp upwards movement at cue onset followed by a slower downwards movement in preparation for the next trial.

Without a consolidation mechanism the synaptic sampling model leads in the absence of reward on a large timescale to slow deterioration due to the continuing noise in synapses. However, we found that the rate of forgetting is rather slow, see blue curve in Fig. 2c. This effect can be further reduced by adding a simple consolidation mechanism to the model: All synapses i for which the synaptic parameter θ_i became larger than 3 and remained above this threshold for more than 24 hours were consolidated by setting the mean of the prior $p_{\mathcal{S}}(\theta_i)$ to the current value of θ_i. In addition the standard deviation of the prior was set to a small value of $\sigma = 0.001$. This simple consolidation mechanism kept the task performance stable (see green curve in Fig. 2c).

Our model shows permanent rewiring of around 2500 synapses ($\sim 5\%$) turning over in a time window of 2 hours (see Fig. 2f). During the first 2 hours the synaptic turnover rate peaked at 12.000 synapses, presumably to drive the network from the random initial configuration to a functioning parameter regime. Experimental data suggests that acquisition of a new behavioral task is accompanied by an increase of spine dynamics [Peters et al., 2014, Xu et al., 2009]. Here, we show that this effect can be reproduced in our model. To mimic a new behavioral task we modified the learning goal by inverting the function of the neuron pools D and U after an initial learning phase of 24 hours. D now caused upwards and U downwards lever movement. Bayesian policy sampling compensates for this perturbation within about 10 hours of ongoing learning (Fig. 2e). We found that during this phase of learning a new task, a significant increase in spine turnover rate could be observed (Fig. 2f). The turnover rate than remained slightly elevated during the subsequent learning time. A new assembly emerged through continuous Bayesian policy sampling, that controls the neural pools D and U to generate the modified behavior (Fig. 2d).

A structural difference between stochastic learning models such as Bayesian policy sampling and learn-

ing models that focus on convergence of parameters to a (locally) optimal setting becomes apparent when one tracks the temporal evolution of the network parameters $\boldsymbol{\theta}$ over larger periods of time during the previously discussed learning process (Fig. 2g). Although performance no longer improved after 5 hours, both network connectivity and parameters kept changing in task-irrelevant dimensions, as often observed in experimental data, see e.g., [Todorov and Jordan, 2002]. For Fig. 2g we randomly selected 5% of the roughly 47000 parameters θ_i and plotted the first 3 principal components of their dynamics. The network change after 24 hours caused the synaptic parameters to migrate to a new region within about 8 hours of continuing learning. Again we observe that Bayesian policy sampling keeps exploring different equally good solutions after the learning process has reached stable performance. This property of our model is compatible with experimental findings on degenerate neural systems that utilize redundancies to enhances robustness [Marder, 2011].

Relative contribution of spontaneous and activity-dependent processes to synaptic plasticity

[Dvorkin and Ziv, 2016] analyzed the correlation of sizes of postsynaptic densities and spine volumes for synapses that shared the same pre- and post-synaptic neuron, called commonly innervated (CI) synapses, and also for synapses that shared in addition the same dendrite ($\mathrm{CI_{SD}}$). Activity-dependent rules for synaptic plasticity, such as Hebbian or STDP rules on which previous models for network plasticity relied, suggest that the strength of CI and especially $\mathrm{CI_{SD}}$ synapses should be highly correlated. But both data from ex-vivo [Kasthuri et al., 2015] and neural circuits in culture [Dvorkin and Ziv, 2016] show that postsynaptic density sizes and spine volumes of $\mathrm{CI_{SD}}$ synapses are only weakly correlated: even in a conservative estimate that corrects for possible influences of their experimental procedure, more than 50% of the observed synaptic strength appears to result from activity-independent stochastic processes (Fig. 8E of [Dvorkin and Ziv, 2016]).

We tested our model by asking whether such a strong contribution of activity-independent stochastic plasticity processes could be consistent with task-dependent network self-organization as in the experiment of Fig. 2. We were able to carry out this test because many synaptic connections between neurons that were formed in that model consisted of more than one synapse (to be precise: 49% of connections consisted of multiple synapses). We classified pairs of synapses that had the same pre- and post-synaptic neuron as CI synapses (one could also call them $\mathrm{CI_{SD}}$ synapses, since the neuron model did not have different dendrites), and pairs with the same post-synaptic but different pre-synaptic neurons as non-CI synapses. Example traces of synaptic weights for CI and non-CI synapse pairs are shown in Fig. 3a,b. CI pairs were found to be more strongly correlated than non-CI pairs (Fig. 3c). However also the correlation of CI pairs was quite low, and varied with the temperature parameter T in Eq. (3). The correlation was measured in terms of the Pearson correlation (covariance of synapse pairs normalized between -1 and 1).

[Dvorkin and Ziv, 2016] reported that a certain degree of uncertainty could be attributed to their experimental setup. The maximum detectable correlation coefficient was limited to $0.76 - 0.78$, due to the variability of light fluorescence intensities which were used to estimate the sizes of postsynaptic densities. To account for this unknown noise source we tested our network over a wide range of temperatures between

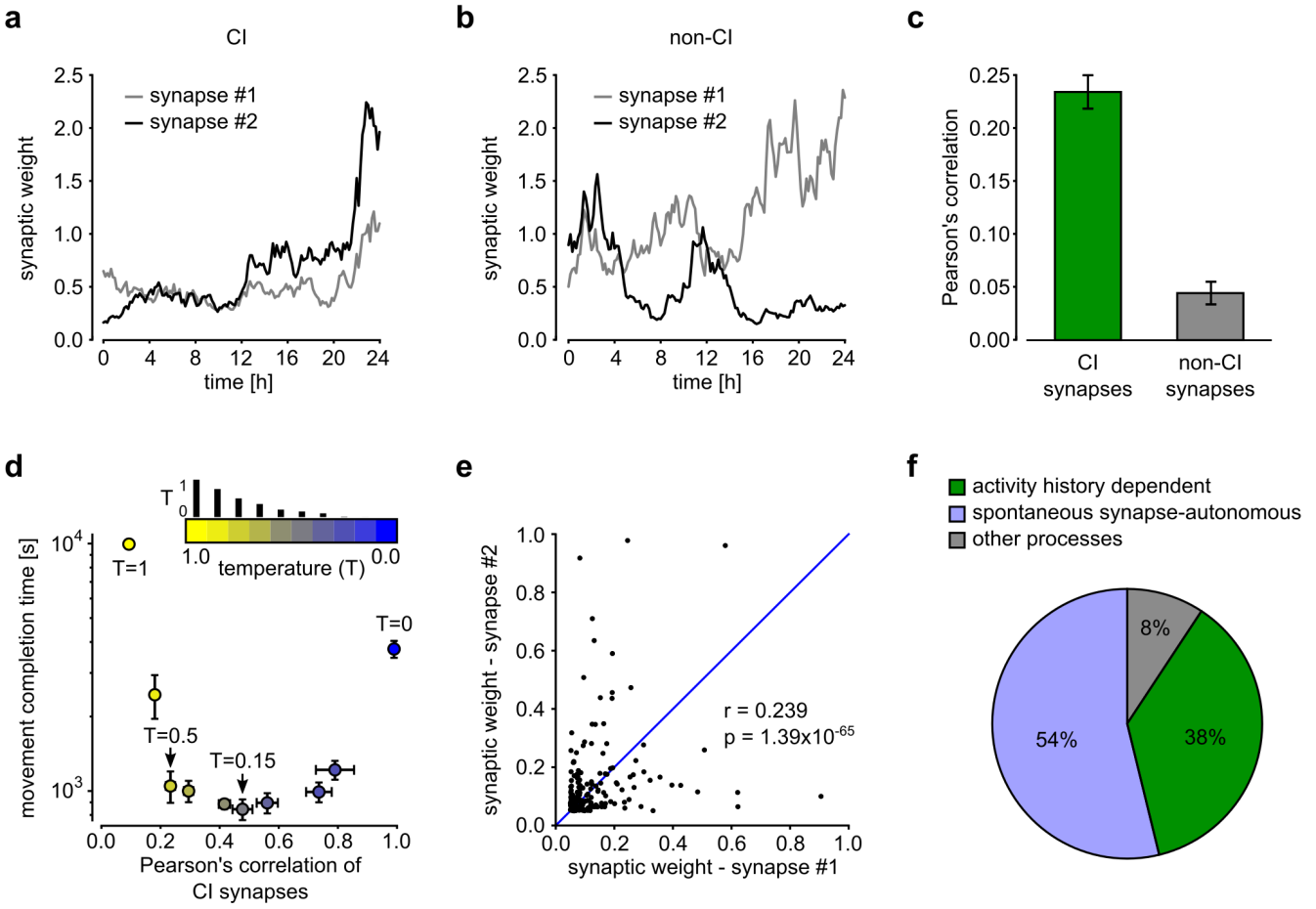

Figure 3: **Contribution of spontaneous stochastic and activity-dependent processes to synaptic plasticity** **(a,b)** Evolution of synaptic weights w_i plotted against time for a pair of CI synapses in **a**, and non-CI synapses in **b**, for $T = 0.5$. **(c)** Pearson's correlation coefficient computed between synaptic weights of CI and non-CI synapses of a network with $T = 0.5$ after 48h of network plasticity as in Fig. 2g. CI synapses were only weakly correlated, but significantly stronger correlated than non-CI synapses. **(d)** Impact of T on correlation of CI synapses (x-axis) and learning performance (y-axis). Each dot represents averaged data for one particular temperature value, indicated by the color. Values for T were 1.0, 0.75, 0.5, 0.35, 0.2, 0.15, 0.1, 0.01, 0.001, 0.0. These values are proportional to the small bars that protrude from the color bar. The performance (measured in movement completion time) is measured after 48 hours for the learning experiment as in Fig. 2g. Good performance was achieved for a range of temperature values between 0.01 and 0.5. Too low (< 0.01) or too high (> 0.5) values impaired learning. Means + s.e.m. over 5 independent trials are shown. **(e)** Synaptic weights of 100 pairs of CI synapses that emerged from a run with $T = 0.5$. Pearson's correlation is 0.239, comparable to the experimental data in Fig. 8A-D of [Dvorkin and Ziv, 2016]. **(f)** Estimated contributions of activity history dependent (green), spontaneous synapse-autonomous (blue) and other (gray) processes to the synaptic plasticity for a run with $T = 0.15$. The resulting fractions are very similar to the experimental data shown in Fig. 8E of [Dvorkin and Ziv, 2016].

$T = 0.0$ and $T = 1.0$. In Fig. 3d we analyzed the impact of the temperature T on correlations of pairs of CI synapses, as well as on task performance. The Pearson correlation coefficient for CI synapses is plotted here together with the average performance achieved on the task of Fig. 2d-g by networks that learn with different temperatures. The best performing temperature region for the task ($0.01 \leq T \leq 0.5$) roughly coincided with the region of experimentally measured values of Pearson's correlation for CI-synapses. Fig. 3e shows the correlation of 100 CI synapse pairs that emerged from a run with $T = 0.5$. We found a value of $r = 0.239$ in this case. This value is in the order of the lowest experimentally found correlation coefficients in [Dvorkin and Ziv, 2016] (both in culture and ex-vivo, see Fig. 8A-D in [Dvorkin and Ziv, 2016]). For $T = 0.15$ we found the best task performance and the closest match to experimentally measured correlations when the results of [Dvorkin and Ziv, 2016] were corrected for measurement limitations : A correlation coefficient of $r = 0.46 \pm 0.034$ for CI synapses and 0.08 ± 0.015 for non-CI synapse pairs (mean $\pm$ s.e.m. over 5 trials, 2-tailed p-value below 0.005 in all trials).

[Dvorkin and Ziv, 2016] further analyzed what the main contributors to the measured synaptic plasticity were. Since in our computer simulations we can directly read out values of the synaptic parameters we were not required to correct our results for noise sources in the experimental procedure (see p. 16ff and equations on p. 18 of [Dvorkin and Ziv, 2016]). This is also reflected in our data by the fact that we got a correlation coefficient that was close to 1.0 in the case $T = 0$ (see Fig. 3d). Following the procedure of [Dvorkin and Ziv, 2016] we estimated in our model the contributions of activity history dependent and spontaneous synapse-autonomous processes as in Fig. 8E of [Dvorkin and Ziv, 2016]. Using the assumption of zero measurement error and thus a theoretically achievable maximum correlation coefficient of $r = 1.0$ we estimated the fraction of contributions of specific activity histories to synaptic changes (for $T = 0.15$) as

$$\frac{0.46 - 0.08}{1.0} = 0.38$$

and of spontaneous synapse-autonomous processes as

$$\frac{1.0 - 0.46}{1.0} = 0.54.$$

The remaining 8% resulted from processes that were not specific to pre-synaptic input, but specific to the activity of the post-synaptic neuron. These results from our model, that are plotted in Fig. 3f, match quite closely the experimentally found values 0.36, 0.56 and 0.08, see Fig. 8E in [Dvorkin and Ziv, 2016]. Altogether we found that the results of [Dvorkin and Ziv, 2016] are best explained by our model for a temperature parameter between $T = 0.5$ (corresponding to their lowest measured correlation coefficient) and $T = 0.15$ (corresponding to their most conservative estimate). Importantly, this range of parameters coincided with well-functioning learning behavior (Fig. 3d). Therefore, the results in Fig. 2 and Fig. 3 show that the underlying scaffold for a generic recurrent network of excitatory and inhibitory neurons was able to configure its connections and synaptic weights to perform the reward-based learning task, while at the same time reproducing the experimentally found quite high level of stochastic synapse-autonomous process. In fact, Fig. 3d shows that noise was necessary for good learning performance.

Discussion

Recent experimental data ([Dvorkin and Ziv, 2016], where in Fig. 8 also ex-vivo data from [Kasthuri et al., 2015] were reanalyzed) suggest that common models for learning in neural networks of the brain need to be revised, since synapses are subject to surprisingly strong activity-independent stochastic processes. In particular, these data are in conflict with the common exclusive reliance on deterministic and activity-dependent rules for synaptic plasticity in neural network models, and also with common models for reward-gated network plasticity based on policy gradient. In addition, experimentally found network rewiring has so far not been integrated into models for reward-gated network plasticity. We have presented a theoretical framework that enables us to investigate and understand reward-based network rewiring and synaptic plasticity in the context of the experimentally found high level of activity-independent stochastic fluctuations of synaptic connectivity and synaptic strength ("synaptic sampling"). We have shown that the Fokker-Planck equation from theoretical physics allows us to understand how local stochastic processes at numerous synapses can orchestrate global goal-directed network learning. This approach provides a new normative model for local plasticity rules based on given functional goals for network plasticity.

We have shown in Fig. 1 that the resulting model can reproduce data on dopamine-dependent spine dynamics reported in [Yagishita et al., 2014], and that it provides an understanding how these local processes can produce function-oriented cortical-striatal connectivity. We have shown in Fig. 2 that the resulting model also elucidates reward-based self-organization of generic recurrent neural networks (consisting of excitatory and spiking neurons, as in [Avermann et al., 2012]) for a given computational task. We chose as benchmark task the production of a specific motor output in response to a cue, like in the experiments of [Peters et al., 2014]. Similarly as reported in the experimental data of [Peters et al., 2014], the network connectivity and dynamics reorganized itself in our model, just driven by stochastic processes and rewards for successful task completion. Analysis of the impact of the amount of stochasticity on network learning performance has shown in Fig. 3 that the network learns best when the stochastic component of synaptic plasticity is as high as reported in the experimental data of [Kasthuri et al., 2015, Dvorkin and Ziv, 2016].

Although our approach is based on experimental data for the biological implementation level of network plasticity, i.e., for the lowest level of the Marr hierarchy of models [Marr and Poggio, 1976], it turns out to have significant implications for modelling network plasticity on the top level ("what is the functional goal?") and the algorithmic level of the Marr hierarchy. It suggests for the top level that the goal of network plasticity is to sample continuously from a posterior distribution of network configurations. This posterior integrates functional demands with priors that represent structural constraints as well as results of preceding learning experiences and innate programs. In other words, our model suggests to view reward-gated network learning as Bayesian inference. On the side, the model also proposes a solution to the general question how neural networks can encode and learn a posterior distribution, which has been highlighted as a major open question in computational neuroscience [Pouget et al., 2013]: It proposes that neural networks of the brain represent a posterior in the form of the stationary distribution of network reconfigurations, from which they sample through synaptic sampling. This Bayesian perspective also creates a link to previous work on Bayesian reinforcement learning [Vlassis et al., 2012, Rawlik et al., 2013]. The essence of the resulting model for reward-gated network learning is illustrated in Fig. 4: The traditional view (panel a) of

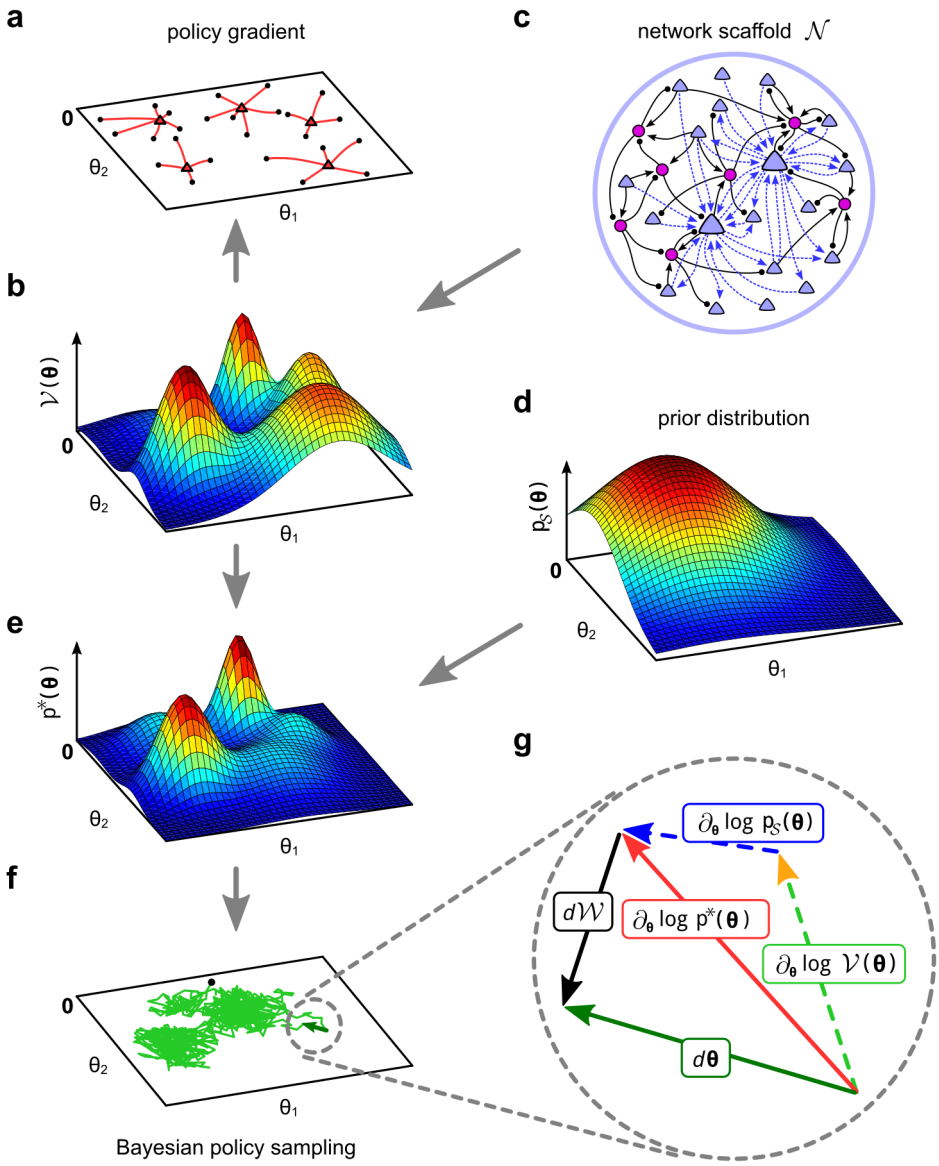

Figure 4: **Illustration of policy gradient and the new Bayesian policy sampling approach.** **(a,b,c)** Illustration of policy gradient learning for two parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ of a neural network scaffold $\mathcal{N}$ shown in **c**, where only synaptic connections from and to inhibitory neurons are fixed. Potential synaptic connections of only two excitatory neurons are shown in **c** to keep the figure uncluttered. **(a)** Illustration of policy gradient on the objective function (reward expectation) shown in **b**. Multiple gradient ascent trajectories from random initial values (black dots) are shown. Red triangles indicate local maxima. **(d)** Example prior that prefers small values for each θ_i. **(e)** The posterior distribution $p^*(\boldsymbol{\theta})$ that results as product of the prior from panel **d** and the objective function of panel **b**. **(f)** A single trajectory of Bayesian policy sampling from the posterior distribution of panel **e**, starting at the black dot. The parameter vector $\boldsymbol{\theta}$ fluctuates between different solutions, and the visited values cluster near local maxima of the posterior. **(g)** Illustration of the dynamic forces (plasticity rule Eq. (3)) that act on $\boldsymbol{\theta}$ in each sampling step $d\boldsymbol{\theta}$ while sampling from the posterior distribution. The deterministic drift term (red), which consists of the first two terms (prior and reward expectation) in Eq. (3), is directed to the next local maximum of the posterior. The stochastic diffusion term $d\mathcal{W}$ (black) of Eq. (3) has a random direction.

gradient ascent (policy gradient) in the landscape (panel b) of reward expectation is first modified through the integration of a prior (panel d), and then through the replacement of gradient ascent by continuously ongoing stochastic sampling (Bayesian policy sampling) from the posterior distribution of panel e, which is illustrated in panels f and g.

This model makes a number of experimentally testable predictions. Continuously ongoing stochastic sampling of network configurations suggests that synaptic connectivity does not converge to a fixed point solution but rather undergoes permanent modifications (Fig. 2g). This prediction is compatible with reports of continuously ongoing spine dynamics and axonal sprouting even in the adult brain [Holtmaat and Svoboda, 2009, Yasumatsu et al., 2008, Stettler et al., 2006, Yamahachi et al., 2009, Loewenstein et al., 2011,Holtmaat et al., 2005,Loewenstein et al., 2015]. These continuously ongoing parameter changes predict continuously ongoing changes in the assembly sequences that accompany and control a motor response (see Fig. 2d). Our model predicts, that these changes do not impair the performance of the network, but rather induce the network to explore different but equally good solutions when exposed for many hours to the same task (see Fig. 2g). Such continuously ongoing drifts of neural codes in functionally less relevant dimensions have already been observed experimentally [Rokni et al., 2007, Ziv et al., 2013, Driscoll and Harvey, 2016]. This effect also explains why the same computational function is found to be realized by the same neural circuit in different individuals with drastically different parameters [Tang et al., 2010, Grashow et al., 2010, Marder, 2011, Prinz et al., 2004]. In fact, this *degeneracy* of neural circuits is thought to be an important property of biological neural networks [Marder, 2011, Prinz et al., 2004, Marder and Goaillard, 2006]. In addition, our model predicts that neural networks automatically compensate for disturbances by moving their continuously ongoing sampling of network configurations to a new region of the parameter space, as illustrated by the response to the disturbance marked by $*$ in Fig. 2g.

In conclusion the mathematical framework presented in this article provides a principled way of understanding the complex interplay of deterministic and stochastic processes that underlie the implementation of goal-directed learning in neural circuits of the brain.

Methods

A Bayesian framework for reward-modulated learning. The classical goal of reinforcement learning is to maximize the expected future discounted reward $\mathcal{V}(\boldsymbol{\theta})$ given by Eq. (4). The expectation in Eq. (4) is taken with respect to the distribution $p(\boldsymbol{r}|\boldsymbol{\theta})$ over sequences $\boldsymbol{r} = \{r(\tau), \tau \geq 0\}$ of future rewards that result from the given set of synaptic parameters $\boldsymbol{\theta}$. The stochasticity of the reward sequence $\boldsymbol{r}$ arises from stochastic network inputs, stochastic network responses, and stochastic reward delivery. The resulting distribution $p(\boldsymbol{r}|\boldsymbol{\theta})$ of reward sequences $\boldsymbol{r}$ for the given parameters $\boldsymbol{\theta}$ can also include influences of network initial conditions by assuming some distribution over these initial conditions. Network initial conditions include for example initial values of neuron membrane voltages and refractory states of neurons. The role of initial conditions on network learning is further discussed below when we consider the online learning scenario in *Reward-modulated synaptic plasticity approximates gradient ascent on the expected discounted reward.*

There exists a close relationship between reinforcement learning and Bayesian inference [Vlassis et al., 2012, Rawlik et al., 2013, Botvinick and Toussaint, 2012]. To make this relationship apparent, we define our model for reward-gated network plasticity by introducing a binary random variable $v_{\mathfrak{B}}$ that represents the currently expected future discounted reward in a probabilistic manner. The likelihood $p_{\mathcal{N}}(v_{\mathfrak{B}} = 1 \,|\, \boldsymbol{\theta})$ is determined in this theoretical framework by the expected future discounted reward Eq. (4) that is achieved by a network with parameter set $\boldsymbol{\theta}$ (see e.g., [Rawlik et al., 2013]):

$$p_{\mathcal{N}}(v_{\mathfrak{B}} = 1 \,|\, \boldsymbol{\theta}) \;\equiv\; \frac{1}{\mathcal{Z}_{\mathcal{V}}} \mathcal{V}(\boldsymbol{\theta}) \;, \tag{6}$$

where $\mathcal{Z}_{\mathcal{V}}$ denotes a constant, that assures that Eq. (6) is a correctly normalized probability distribution. Thus reward-based network optimization can be formalized as maximizing the likelihood $p_{\mathcal{N}}(v_{\mathfrak{B}} = 1 \,|\, \boldsymbol{\theta})$ with respect to the network configuration $\boldsymbol{\theta}$. Structural constraints can be integrated into a stochastic model for network plasticity through a prior $p_S(\boldsymbol{\theta})$ over network configurations. Hence reward-gated network optimization amounts from a theoretical perspective to learning of the posterior distribution $p^*(\boldsymbol{\theta}|v_{\mathfrak{B}} = 1)$, which by Bayes' rule is defined (up to normalization) by $p_{\mathcal{S}}(\boldsymbol{\theta}) \cdot p_{\mathcal{N}}(v_{\mathfrak{B}} = 1 \,|\, \boldsymbol{\theta})$. Therefore, the learning goal can be formalized in a compact form as evaluating the posterior distribution $p^*(\boldsymbol{\theta}|v_{\mathfrak{B}} = 1)$ of network parameters $\boldsymbol{\theta}$ under the constraint that the abstract learning goal $v_{\mathfrak{B}} = 1$ is achieved.

More generally, one is often interested in a tempered version of the posterior

$$p^*(\boldsymbol{\theta}) \equiv \frac{1}{\mathcal{Z}} p^*(\boldsymbol{\theta}|v_{\mathfrak{B}} = 1)^{\frac{1}{T}} \;, \tag{7}$$

where $\mathcal{Z}$ is a suitable normalization constant and $T > 0$ is the temperature parameter that controls the "sharpness" of $p^*(\boldsymbol{\theta})$. For $T = 1$, $p^*(\boldsymbol{\theta})$ is given by the original posterior, $T < 1$ emphasizes parameter values with high probability in the posterior, while $T > 1$ leads to parameter distributions $p^*(\boldsymbol{\theta})$ which are more uniformly distributed than the posterior.

Analysis of Bayesian policy sampling. Here we prove that the stochastic parameter dynamics Eq. (3) samples from the tempered posterior distribution $p^*(\boldsymbol{\theta})$ given in Eq. (7). In *Results* we suppressed time-dependencies in order to simplify notation. We reiterate Eq. (2) with explicit time-dependencies of parameters:

$$d\theta_i(t) \;=\; \beta \;\frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta}|v_{\mathfrak{B}} = 1)\Big|_{\boldsymbol{\theta}(t)} dt \;+\; \sqrt{2T\beta}\, d\mathcal{W}_i \;, \tag{8}$$

where the notation $\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}(t)}$ denotes the derivative of $f(\boldsymbol{\theta})$ with respect to θ_i evaluated at the current parameter values $\boldsymbol{\theta}(t)$. By Bayes' rule, the derivative of the log posterior is the sum of the derivatives of the prior and the likelihood:

$$\frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta}|v_{\mathfrak{B}} = 1) \;=\; \frac{\partial}{\partial \theta_i} \log p_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\partial}{\partial \theta_i} \log p_{\mathcal{N}}(v_{\mathfrak{B}} = 1 \,|\, \boldsymbol{\theta}) \;=\; \frac{\partial}{\partial \theta_i} \log p_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta}) \;,$$

which allows us to rewrite Eq. (8) as

$$d\theta_i(t) \;=\; \beta \left(\frac{\partial}{\partial \theta_i} \log p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}(t)} \;+\; \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}(t)} \right) dt \;+\; \sqrt{2T\beta}\, d\mathcal{W}_i \;, \tag{9}$$

which is identical to the form Eq. (3), where the contributions of $p_{\mathcal{S}}\left(\boldsymbol{\theta}\right)$ and $\mathcal{V}(\boldsymbol{\theta})$ are given explicitly.

We prove the correctness of reward-based synaptic sampling for the more general synaptic dynamics:

$$d\theta_i(t) \;=\; \left(b(\boldsymbol{\theta}(t))\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1)\Big|_{\boldsymbol{\theta}(t)} + T\, b'(\boldsymbol{\theta}(t)) \right) \, dt \;+ \sqrt{2\, T\, b(\boldsymbol{\theta}(t))}\, d\mathcal{W}_i \;, \tag{10}$$

where $b(\boldsymbol{\theta}) > 0$ is a twice differentiable function that may scale the learning rate depending on the synaptic parameters $\boldsymbol{\theta}$ and $b'(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_i} b(\boldsymbol{\theta})$. Note that Eq. (8) is a special case of this form, with $b(\boldsymbol{\theta}) = \beta$. To simplify notation we drop in the following the explicit time dependence of the synaptic parameters $\boldsymbol{\theta}$. The result can be formalized in the following theorem:

Theorem 1. *Let $p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1)$ be a strictly positive, continuous probability distribution over parameters $\boldsymbol{\theta}$, twice continuously differentiable with respect to $\boldsymbol{\theta}$. Let $b(\boldsymbol{\theta})$ be a strictly positive, twice continuously differentiable function. Then the set of stochastic differential equations Eq.* (10) *leaves the distribution $p^*(\boldsymbol{\theta})$ invariant. Furthermore, $p^*(\boldsymbol{\theta})$ is the unique stationary distribution of the sampling dynamics.*

Proof. The proof is analogous to the one provided in [Kappel et al., 2015]. The stochastic differential equation Eq. (10) translates into a Fokker-Planck equation [Gardiner, 2004] that describes the evolution of the distribution over parameters $\boldsymbol{\theta}$

$$\frac{\partial}{\partial t} p_{\mathrm{FP}}(\boldsymbol{\theta}, t) = \sum_i -\frac{\partial}{\partial \theta_i} \left(\left(b(\boldsymbol{\theta})\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1) + T\, b'(\boldsymbol{\theta}) \right) p_{\mathrm{FP}}(\boldsymbol{\theta}, t) \right) + \frac{\partial^2}{\partial \theta_i^2} \left(T\, b(\boldsymbol{\theta}) p_{\mathrm{FP}}(\boldsymbol{\theta}, t) \right), \tag{11}$$

where $p_{\mathrm{FP}}(\boldsymbol{\theta}, t)$ denotes the distribution over network parameters at time t. Plugging in the presumed stationary distribution $p^*(\boldsymbol{\theta})$ for $p_{\mathrm{FP}}(\boldsymbol{\theta}, t)$ on the right hand side of Eq. (11), one obtains

$$\begin{aligned}
\frac{\partial}{\partial t} p_{\mathrm{FP}}(\boldsymbol{\theta}, t) &= \sum_i -\frac{\partial}{\partial \theta_i} \left(\left(b(\boldsymbol{\theta})\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1) + T\, b'(\boldsymbol{\theta}) \right) p^*(\boldsymbol{\theta}) \right) + \frac{\partial^2}{\partial \theta_i^2} \left(T\, b(\boldsymbol{\theta}) p^*(\boldsymbol{\theta}) \right) \\
&= \sum_i -\frac{\partial}{\partial \theta_i} \left(b(\boldsymbol{\theta})\, p^*(\boldsymbol{\theta})\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1) \right) + \frac{\partial}{\partial \theta_i} \left(T\, b(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} p^*(\boldsymbol{\theta}) \right) \\
&= \sum_i -\frac{\partial}{\partial \theta_i} \left(b(\boldsymbol{\theta})\, p^*(\boldsymbol{\theta})\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1) \right) + \frac{\partial}{\partial \theta_i} \left(T\, b(\boldsymbol{\theta})\, p^*(\boldsymbol{\theta})\, \frac{\partial}{\partial \theta_i} \log p^*(\boldsymbol{\theta}) \right) ,
\end{aligned}$$

which by inserting $p^*(\boldsymbol{\theta}) \;=\; \frac{1}{\mathcal{Z}} p^*(\boldsymbol{\theta} \,|\, v_{\mathfrak{B}} = 1)^{\frac{1}{T}}$, with normalizing constant $\mathcal{Z}$, becomes

$$\frac{\partial}{\partial t}p_{\mathrm{FP}}(\boldsymbol{\theta},t) = \frac{1}{\mathcal{Z}}\sum_i -\frac{\partial}{\partial\theta_i}\left(b(\boldsymbol{\theta})\,p^*(\boldsymbol{\theta})\,\frac{\partial}{\partial\theta_i}\log p^*(\boldsymbol{\theta}\,|\,v_{\mathfrak{B}}=1)\right) + \frac{\partial}{\partial\theta_i}\left(T\,b(\boldsymbol{\theta})\,p^*(\boldsymbol{\theta})\,\frac{1}{T}\frac{\partial}{\partial\theta_i}\log p^*(\boldsymbol{\theta}\,|\,v_{\mathfrak{B}}=1)\right)$$
$$= \sum_i 0 = 0 \ .$$

This proves that $p^*(\boldsymbol{\theta})$ is a stationary distribution of the parameter sampling dynamics Eq. (10). Under the assumption that $b(\boldsymbol{\theta})$ is strictly positive, this stationary distribution is also unique (see Section 3.7.2 in [Gardiner, 2004]).

The unique stationary distribution of Eq. (11) is given by $p^*(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}}p^*(\boldsymbol{\theta}|v_{\mathfrak{B}}=1)^{\frac{1}{T}}$, i.e. $p^*(\boldsymbol{\theta})$ is the only solution for which $\frac{\partial}{\partial t}p_{\mathrm{FP}}(\boldsymbol{\theta},t)$ becomes 0, which completes the proof. □

Network model. Plasticity rules for this general framework were derived based on a specific spiking neural network model, which we describe in the following. All reported computer simulations were performed with this network model. We considered a general network scaffold $\mathcal{N}$ of K neurons with potentially asymmetric recurrent connections. We denote by $w_i(t)$ the synaptic efficacy of the i-th synapse in the network at time t and we define SYN_k to be the index set of synapses that project to neuron k. Further we denote by PRE_i and POST_i the index of the pre- and postsynaptic neuron of synapse i, respectively. We denote the output spike train of a neuron k by $z_k(t)$. It is defined as the sum of Dirac delta pulses positioned at the spike times $t_k^{(1)}, t_k^{(2)}, \ldots$, i.e., $z_k(t) = \sum_l \delta(t - t_k^{(l)})$.

Network neurons were modeled by a standard stochastic variant of the spike response model [Gerstner et al., 2014]. In this model, the membrane potential of a neuron k at time t is given by

$$u_k(t) \ = \ \sum_{i\,\in\,\mathrm{SYN}_k} y_{\mathrm{PRE}_i}(t)\,w_i(t) \ + \ \vartheta_k(t)\ , \tag{12}$$

where $\vartheta_k(t)$ denotes the slowly changing bias potential of neuron k, and $y_{\mathrm{PRE}_i}(t)$ denotes the trace of the (unweighted) postsynaptic potentials (PSPs) that neuron PRE_i leaves in its postsynaptic synapses at time t. More precisely, it is defined as $y_{\mathrm{PRE}_i}(t) = z_{\mathrm{PRE}_i}(t) * \epsilon(t)$ given by spike trains filtered with a double-exponential PSP kernel of the form $\epsilon(t) = \Theta(t)\,\frac{\tau_r}{\tau_m-\tau_r}\left(e^{-\frac{t}{\tau_m}} - e^{-\frac{t}{\tau_r}}\right)$, with time constants $\tau_m = 20$ ms and $\tau_r = 2$ ms, if not stated otherwise. Here $*$ denotes convolution and $\Theta(\cdot)$ is the Heaviside step function, i.e. $\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise. In general we allowed multiple synapses between each pair of pre- and postsynaptic neuron.

The synaptic weights $w_i(t)$ in Eq. (12) were determined by the synaptic parameters $\theta_i(t)$ through the mapping Eq. (1) for $\theta_i(t) > 0$. Synaptic connections with $\theta_i(t) \leq 0$ were interpreted as not functional (disconnected) and $w_i(t)$ was therefore set to 0 in that case.

The bias potential $\vartheta_k(t)$ in Eq. (12) implements a slow adaptation mechanism of the intrinsic excitability, which ensures that the output rate of each neuron stays near the firing threshold and the neuron maintains responsiveness [Desai et al., 1999, Fan et al., 2005]. We used a simple adaptation mechanism which was

updated according to

$$\tau_\vartheta \frac{d\vartheta_k(t)}{dt} \;=\; \nu_0 - z_k(t)\;, \tag{13}$$

where $\tau_\vartheta = 50$ s is the time constant of the adaptation mechanism and $\nu_0 = 5$ Hz is the desired output rate of the neuron. In our simulations, the bias potential $\vartheta_k(t)$ was initialized at -3 and then followed the dynamics given in Eq. (13). This regularization is a simplified version of the mechanism proposed in [Remme and Wadman, 2012] to balance activity in networks of excitatory and inhibitory neurons. We found that this regularization significantly increased the performance and learning speed of our network model, presumably due to the substantial change in neural fan-in (due to rewiring as discussed above) that may take place during learning which is counteracted by such a mechanism.

We used a simple refractory mechanism for our neuron model. The firing rate, or intensity, of neuron k at time t is defined by the function $f_k(t) \;=\; f(u_k(t), \rho_k(t))$, where $\rho_k(t)$ denotes a refractory variable that measures the time elapsed since the last spike of neuron k. We used an exponential dependence between membrane potential and firing rate, such that the instantaneous firing rate of the neuron k at time t can be written as

$$f_k(t) \;=\; f(u_k, \rho_k) \;=\; \exp(u_k)\Theta(\rho_k - t_{\text{ref}})\;. \tag{14}$$

Furthermore, we denote by $f_{\text{POST}_i}(t)$ the firing rate of the neuron postsynaptic to synapse i. If not stated otherwise we set the refractory time t_{ref} to 5 ms. In addition, a subset of neurons was clamped to some given firing rates (input neurons), such that $f_k(t)$ of these input neurons was given by an arbitrary function. We denote the spike train from these neurons by $\boldsymbol{x}(t)$, the network input.

Synaptic dynamics for the reward-based synaptic sampling model. The synaptic dynamics was given by Eq. (3). For the neural network model described above, the gradient $\frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$ (the gradient of the expected reward), was estimated through a plasticity mechanism that uses two variables $e_i(t)$ and $g_i(t)$ in each synapse which were updated according to the differential equations

$$\frac{de_i(t)}{dt} \;=\; -\frac{1}{\tau_e} e_i(t) \;+\; w_i(t)\, y_{\text{PRE}_i}(t)\, (z_{\text{POST}_i}(t) - f_{\text{POST}_i}(t))\;, \tag{15}$$

$$\frac{dg_i(t)}{dt} \;=\; -\frac{1}{\tau_g} g_i(t) \;+\; \left(\frac{r(t)}{\hat{r}(t)} + \alpha\right) e_i(t)\;, \tag{16}$$

where $\tau_e = 1$ s and $\tau_g = 50$ s, are time constants of the synaptic dynamics. In Eq. (15) $z_{\text{POST}_i}(t)$ denotes the postsynaptic spike train, $f_{\text{POST}_i}(t)$ denotes the instantaneous firing rate (Eq. (14)) of the postsynaptic neuron and $w_i(t)\, y_{\text{PRE}_i}(t)$ denotes the postsynaptic potential under synapse i. The variable $e_i(t)$ plays the role of an eligibility trace which averages over a brief history of past synaptic changes. The variable $g_i(t)$ combines the eligibility trace and the reward, and averages over the time scale τ_g. α is an arbitrary constant offset on the reward signal. In our simulations, this offset α was chosen slightly above 0 ($\alpha = 0.02$) such that small parameter changes were also present without any reward, as observed in [Yagishita et al., 2014]. In the next section we show that $g_i(t)$ approximates the gradient of the expected future reward with respect to the synaptic parameter, i.e. $g_i(t) \approx \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$ for all $t > \tau_g$. Note, that for retracted synapses ($w_i(t) = 0$), both $e_i(t)$ and $g_i(t)$ decay to zero (within few minutes in our simulations). Therefore, we find that the

dynamics of retracted synapses is only driven by the first (prior) and last (random fluctuations) term of Eq. (3). Thus, retracted synapses spontaneously reappear also in the absence of reward after a random amount of time.

$\hat{r}(t)$ in Eq. (16) is a low-pass filtered version of $r(t)$ that scales the synaptic updates. It was implemented through $\tau_g \frac{d\hat{r}(t)}{dt} = -\hat{r}(t) + r(t)$, with $\tau_g = 50$ s. This scaling of the reward signal has the following effect. If the current reward $r(t)$ exceeds the average reward $\hat{r}(t)$, the effect of the neuromodulatory signal $r(t)$ will be greater than 1. On the other hand, if the current reward is below average synaptic updates will be weighted by a term significantly lower than 1. Therefore, parameter updates are preferred for which the current reward signal exceeds the average.

The first term in Eq. (3) is the gradient of the prior distribution. We used a prior distribution that pulls the synaptic parameters towards $\theta_i(t) = 0$ such that unused synapses tend to disappear and new synapses are permanently formed. Throughout all simulations we used independent Gaussian priors for the synaptic parameters

$$p_{\mathcal{S}}(\boldsymbol{\theta}) \;=\; \prod_i p_{\mathcal{S}}(\theta_i(t)) \;, \qquad \text{with} \qquad p_{\mathcal{S}}(\theta_i(t)) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta_i(t)-\mu)^2}{2\sigma^2}\right) .$$

Using this, we find that the contribution of the prior to the online parameter update equation is given by

$$\frac{\partial}{\partial\theta_i} \log p_{\mathcal{S}}(\boldsymbol{\theta}) \;=\; \frac{1}{\sigma^2}(\mu - \theta_i(t)) \;. \tag{17}$$

Finally by plugging Eq. (17) and (16) into Eq. (3) the synaptic parameter changes at time t are given by

$$d\theta_i(t) \;=\; \beta \left(\frac{1}{\sigma^2}(\mu - \theta_i(t)) \;+\; g_i(t)\right) dt \;+\; \sqrt{2T\beta}\, d\mathcal{W}_i \;, \tag{18}$$

where σ is the standard deviation of the prior. If not stated otherwise we used $\sigma = 2$ and $\mu = 0$, and a learning rate of $\beta = 10^{-5}$.

Reward-modulated synaptic plasticity approximates gradient ascent on the expected discounted reward. We first consider a theoretical setup where the network is operated in arbitrarily long episodes such that in each episode a reward sequence $\boldsymbol{r}$ is encountered. The reward sequence $\boldsymbol{r}$ can be any discrete or real-valued function that is positive and bounded. The episodic scenario is useful to derive exact batch parameter update rules, from which we will then deduce online learning rules. Due to stochastic network inputs, stochastic network responses, and stochastic reward delivery, the reward sequence $\boldsymbol{r}$ is stochastic.

The classical goal of reinforcement learning is to maximize the function $\mathcal{V}(\boldsymbol{\theta})$ of discounted expected rewards Eq. (4), which we reiterate here for convenience:

$$\mathcal{V}(\boldsymbol{\theta}) \;=\; \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, r(\tau)\, d\tau \right\rangle_{p(\boldsymbol{r}|\boldsymbol{\theta})} \;. \tag{19}$$

Policy gradient algorithms perform gradient ascent on $\mathcal{V}(\boldsymbol{\theta})$ by changing each parameter θ_i in the direction

of the gradient $\partial \log \mathcal{V}(\boldsymbol{\theta})/\partial\theta_i$. Here, we show that the parameter dynamics Eq. (15), (16) approximate this gradient, i.e., $g_i(t) \approx \partial \log \mathcal{V}(\boldsymbol{\theta})/\partial\theta_i$ for all $t > \tau_g$.

It is natural to assume that the reward signal $r(\tau)$ only depends indirectly on the parameters $\boldsymbol{\theta}$, through the history of network spikes $z_k(\tau)$ up to time τ, which we write as $\mathbf{z}(\tau) = \{z_k(s) \mid 0 \le s < \tau,\, 1 \le k \le K\}$, i.e., $p_{\mathcal{N}}\left(r(t), \mathbf{z}(t) \,|\, \boldsymbol{\theta}\right) = p\left(r(t) \,|\, \mathbf{z}(t)\right) p_{\mathcal{N}}\left(\mathbf{z}(t) \,|\, \boldsymbol{\theta}\right)$. We can first expand the expectation $\langle \cdot \rangle_{p(\boldsymbol{r}|\boldsymbol{\theta})}$ in Eq. (19) to be taken over the joint distribution $p(\boldsymbol{r}, \mathbf{z}|\boldsymbol{\theta})$ over reward sequences $\boldsymbol{r}$ and network trajectories $\mathbf{z}$. The derivative

$$\frac{\partial}{\partial\theta_i} \log \mathcal{V}(\boldsymbol{\theta}) \;=\; \frac{1}{\mathcal{V}(\boldsymbol{\theta})} \frac{\partial}{\partial\theta_i} \mathcal{V}(\boldsymbol{\theta}) \;=\; \frac{1}{\mathcal{V}(\boldsymbol{\theta})} \frac{\partial}{\partial\theta_i} \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, r(\tau)\, d\tau \right\rangle_{p(\boldsymbol{r},\mathbf{z}|\boldsymbol{\theta})} \tag{20}$$

can be evaluated using the well-known identity $\frac{\partial}{\partial x} \langle\, f(a) \,\rangle_{p(a|x)} = \left\langle\, f(a) \frac{\partial}{\partial x} \log p(a|x) \,\right\rangle_{p(a|x)}$:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} \log \mathcal{V}(\boldsymbol{\theta}) \;&=\; \frac{1}{\mathcal{V}(\boldsymbol{\theta})} \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, r(\tau)\, \frac{\partial}{\partial\theta_i} \log p\left(r(\tau), \mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right)\, d\tau \right\rangle_{p(\boldsymbol{r},\mathbf{z}|\boldsymbol{\theta})} \\
&=\; \frac{1}{\mathcal{V}(\boldsymbol{\theta})} \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, r(\tau)\, \frac{\partial}{\partial\theta_i} \left(\log p\left(r(\tau) \,|\, \mathbf{z}(\tau)\right) + \log p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right) \right)\, d\tau \right\rangle_{p(\boldsymbol{r},\mathbf{z}|\boldsymbol{\theta})} \\
&=\; \left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, \frac{r(\tau)}{\mathcal{V}(\boldsymbol{\theta})}\, \frac{\partial}{\partial\theta_i} \log p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right)\, d\tau \right\rangle_{p(\boldsymbol{r},\mathbf{z}|\boldsymbol{\theta})} .
\end{aligned} \tag{21}$$

Here, $p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right)$ is the probability of observing the spike train $\mathbf{z}(\tau)$ in the time interval 0 to τ. For the definition of the network $\mathcal{N}$ given above, the gradient $\frac{\partial}{\partial\theta_i} \log p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right)$ of this distribution can be directly evaluated. Using Eq. (12) and (1) we get Eq. [Pfister et al., 2006]

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} \log p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right) \;&=\; \frac{\partial w_i}{\partial\theta_i} \frac{\partial}{\partial w_i} \int_0^\tau z_{\mathrm{POST}_i}(s)\, \log\left(f_{\mathrm{POST}_i}(s)\right) \,-\, f_{\mathrm{POST}_i}(s)\, ds \\
&=\; \int_0^\tau w_i\, y_{\mathrm{PRE}_i}(s)\, \left(z_{\mathrm{POST}_i}(s) - f_{\mathrm{POST}_i}(s)\right) ds \,,
\end{aligned} \tag{22}$$

where we have used that by construction only the rate function $f_{\mathrm{POST}_i}(s)$ depends on the parameter θ_i. This learning rule is similar to previous ones which were found in the context of maximum likelihood and reinforcement learning in neural networks [Pfister et al., 2006, Florian, 2007]. The main difference is the factor w_i which induces multiplicative synaptic dynamics and is a consequence of the exponential mapping Eq. (1).

Online learning. Eq. (21) defines a batch learning rule with an average taken over learning episodes where in each episode network responses and rewards are drawn according to the distribution $p(\boldsymbol{r}, \mathbf{z}|\boldsymbol{\theta})$. In a biological setting, there are typically no clear episodes but rather a continuous stream of network inputs and rewards and parameter updates are performed continuously (i.e., learning is online). The analysis of online policy gradient learning is far more complicated than the batch scenario, and typically only approximate results can be obtained that however perform well in practice, see e.g., [Seung, 2003, Xie and Seung, 2004] for discussions.

In order to arrive at an online learning rule for this scenario, we consider an estimator of Eq. (21) that approximates its value at each time $t > \tau_g$ based on the recent network activity and rewards during time $[t - \tau_g, t]$ for some suitable $\tau_g > 0$. We denote the estimator at time t by $G_i(t)$ where we want $G_i(t) \approx \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$ for all $t > \tau_g$. To arrive at such an estimator, we approximate the average over episodes in Eq. (21) by an average over time where each time point is treated as the start of an episode. The average is taken over a long sequence of network activity that starts at time t and ends at time $t + \tau_g$. Here, one systematic difference to the batch setup is that one cannot guarantee a time-invariant distribution over initial network conditions as we did there since those will depend on the current network parameter setting. However, under the assumption that the influence of initial conditions (such as initial membrane potentials and refractory states) decays quickly compared to the time scale of the environmental dynamics, it is reasonable to assume that the induced error is negligible. We thus rewrite Eq. (21) in the form (we use the abbreviation $PSP_i(s) = w_i(s)\, y_{\text{PRE}_i}(s)$).

$$\frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta}) \;\approx\; G_i(t) = \frac{1}{\tau_g} \int_t^{t+\tau_g} \int_\zeta^{t+\tau_g} e^{-\frac{\tau-\zeta}{\tau_e}}\, \frac{r(\tau)}{\mathcal{V}(\boldsymbol{\theta})} \int_\zeta^{\tau} PSP_i(s)\, (z_{\text{POST}_i}(s) \,-\, f_{\text{POST}_i}(s))\, ds\, d\tau\, d\zeta \;,$$

where τ_g is the length of the sequence of network activity over which the empirical expectation is taken. Finally, we can combine the second and third integral into a single one, rearrange terms and substitute s and τ so that integrals run into the past rather than the future, to obtain

$$G_i(t) \;\approx\; \frac{1}{\tau_g} \int_{t-\tau_g}^{t} \frac{r(\tau)}{\mathcal{V}(\boldsymbol{\theta})} \int_0^{\tau} e^{-\frac{s}{\tau_e}}\, PSP_i(\tau - s)\, (z_{\text{POST}_i}(\tau - s) \,-\, f_{\text{POST}_i}(\tau - s))\, ds\, d\tau \;, \tag{23}$$

We now discuss the relationship between $G_i(t)$ and Eq. (15), (16) to show that the latter equations approximate $G_i(t)$. Solving Eq. (15) with zero initial condition $e_i(0) = 0$ yields

$$e_i(t) = \int_0^t e^{-\frac{s}{\tau_e}}\, PSP_i(t - s)\, (z_{\text{POST}_i}(t - s) \,-\, f_{\text{POST}_i}(t - s))\, ds \;. \tag{24}$$

This corresponds to the inner integral in Eq. (23) and we can write

$$G_i(t) \;\approx\; \frac{1}{\tau_g} \int_{t-\tau_g}^{t} \frac{r(\tau)}{\mathcal{V}(\boldsymbol{\theta})}\, e_i(\tau)\, d\tau \;=\; \left\langle \frac{r(t)}{\mathcal{V}(\boldsymbol{\theta})}\, e_i(t) \right\rangle_{\tau_g} \;\approx\; \left\langle \frac{r(t)}{\hat{r}(t)}\, e_i(t) \right\rangle_{\tau_g} \;, \tag{25}$$

where $\langle \cdot \rangle_{\tau_g}$ denotes the temporal average from $t - \tau_g$ to t and $\hat{r}(t)$ estimates the expected discounted reward through a slow temporal average.

Finally, we observe that any constant α can be added to $r(\tau)/\mathcal{V}(\boldsymbol{\theta})$ in Eq. (21) since

$$\left\langle \int_0^\infty e^{-\frac{\tau}{\tau_e}}\, \alpha\, \frac{\partial}{\partial \theta_i} \log p_{\mathcal{N}}\left(\mathbf{z}(\tau) \,|\, \boldsymbol{\theta}\right)\, d\tau \right\rangle_{p(\boldsymbol{r}, \mathbf{z} | \boldsymbol{\theta})} = 0 \tag{26}$$

for any constant α (cf. [Williams, 1992, Urbanczik and Senn, 2009]).

Hence, we have $G_i(t) \;\approx\; \left\langle \left(\frac{r(t)}{\hat{r}(t)} + \alpha \right)\, e_i(t) \right\rangle_{\tau_g}$. Eq. (16) implements this in the form of a running average and hence $g_i(t) \approx G_i(t) \approx \frac{\partial}{\partial \theta_i} \log \mathcal{V}(\boldsymbol{\theta})$ for $t > \tau_g$. Note that this result assumes that the parameters

$\boldsymbol{\theta}$ change slowly on the time-scale of τ_g.

Simulation details. Simulations were preformed with NEST [Gewaltig and Diesmann, 2007] using an in-house implementation of the synaptic sampling model; additional tests were run in Matlab R2011b (Mathworks). The differential equations of the neuron and synapse models were approximated using the Euler method, with fixed time steps $\Delta t = 1$ ms. All network variables were updated based on this time grid, except for the synaptic parameters $\theta_i(t)$ according to Eq. (18) which were updated only every 100 ms to reduce the computation time. Control experiments with $\Delta t = 0.1$ ms, and 1 ms update steps for all synaptic parameters showed no significant differences. If not stated otherwise synaptic parameters were initially drawn from a Gaussian distribution with $\mu = -0.5$ and $\sigma = 0.5$ and the temperature was set to $T = 0.1$. Synaptic delays were 1 ms. Synaptic parameter changes were clipped at $\pm 4 \times 10^{-4}$ and synaptic parameters were not allowed to exceed the interval $[-2, 5]$ for the sake of numerical stability.

Details to: A model for task-dependent rewiring of synaptic connections from cortex to medium spiny neurons (MSNs) in the basal ganglia. The number of potential excitatory synaptic connections between each pair of input and MSN neurons was initially drawn from a Binomial distribution ($p = 0.5$, $n = 10$). The connections then followed the reward-based synaptic sampling dynamics Eq. (3) as described above. Lateral inhibitory connections were fixed and thus not subject to learning. These connections between MSN neurons were drawn from a Bernoulli distribution with $p = 0.5$ and synaptic weights were drawn from a Gaussian distribution with $\mu = -1$ and $\sigma = 0.2$, truncated at zero. Two subsets of ten neurons were connected to either one of the targets T_1 or T_2.

To generate the input patterns we adapted the method from [Kappel et al., 2015]. The inputs were representations of a simple symbolic environment, realized by Poisson spike trains that encoded sensory experiences P_1 or P_2. The 200 input neurons were assigned to Gaussian tuning curves ($\sigma = 0.2$) with centers independently and equally scattered over the unit cube. The sensory experiences P_1 and P_2 were represented by two different, randomly selected points in this 3-dimensional space. The stimulus positions were overlaid with small-amplitude jitter ($\sigma = 0.05$). For each sensory experience the firing rate of an individual input neuron was given by the support of the sensory experience under the input neuron's tuning curve (maximum firing rate was 60 Hz). An additional offset of 2 Hz background noise was added. The lengths of the spike patterns were uniformly drawn from the interval [750 ms, 1500 ms]. The spike patterns were alternated with time windows (durations uniformly drawn from the interval [1000 ms, 2000 ms]) during which only background noise of 2 Hz was presented.

The network was rewarded if the assembly associated to the current sensory experience fired stronger than the other assembly. More precisely, we used a sliding window of 500 ms length to estimate the current output rate of the neural assemblies. Let $\hat{\nu}_1(t)$ and $\hat{\nu}_2(t)$ denote the estimated output rates of assemblies A_1 and A_2, respectively, at time t and let $I(t)$ be a function that indicates the identity of the input pattern at time t, i.e. $I(t) = 1$ if pattern P_1 is present and $I(t) = -1$ if pattern P_2 is present. If $I(t)(\hat{\nu}_1(t) - \hat{\nu}_2(t)) < 0$ the reward was set to $r(t) = 0$. Otherwise the reward signal was given by $r(t) = S\left(\frac{1}{5}(I(t)\hat{\nu}_1(t) - I(t)\hat{\nu}_2(t) - \nu_0)\right)$, where $\nu_0 = 25$ Hz is a soft firing threshold and $S(\cdot)$ denotes the logistic sigmoid function. The reward was recomputed every 5 ms. During the presentation of the

background patterns no reward was delivered.

In Fig. 1d,e we tested our reward-gated synaptic plasticity mechanism with the reward-modulated STDP pairing protocol reported in [Yagishita et al., 2014]. Briefly we presented 15 pre/post pairings; one per 10 seconds. In each pre/post pairing 10 presynaptic spikes were presented at 10 Hz. Each presynaptic spike was followed ($\Delta t = 10$ ms) by a brief postsynaptic burst of 3 spikes at 100 Hz. During the pairings the membrane potential was clamped to $u(t) = -2.4$. Reward was delivered here in the form of a rectangular-shaped continuous wave of constant amplitude and duration of 1 s to mimic puff application of dopamine. Rewards were delivered for each pairing and delays were relative to the onset of pairings.

Details to: A model for task-dependent self-configuration of a recurrent network of excitatory and inhibitory spiking neurons. Neuron and synapse parameters were as reported above, except for the inhibitory neurons for which we used faster dynamics with a refractory time $t_{\text{ref}} = 2$ ms and time constants $\tau_m = 10$ ms and $\tau_r = 1$ ms for the PSP kernel. The network connectivity between excitatory and inhibitory neurons was as suggested in [Avermann et al., 2012]. Excitatory (pools D, U and hidden) and inhibitory neurons were randomly connected with connection probabilities given in Table 2 in [Avermann et al., 2012]. Connections include lateral inhibition between excitatory and inhibitory neurons. The connectivity to and from inhibitory neurons was kept fixed throughout the simulation (not subject to synaptic plasticity or rewiring). The connection probability from excitatory to inhibitory neurons was given by 0.575. The synaptic weights were drawn from a Gaussian distribution (truncated at zero) with $\mu = 0.5$ and $\sigma = 0.1$. Inhibitory neurons were connected to their targets with probability 0.6 (to excitatory neurons) and 0.55 (to inhibitory neurons) and the synaptic weights were drawn from a truncated normal distribution with $\mu = -1$ and $\sigma = 0.2$. The number of potential excitatory synaptic connections between each pair of excitatory neurons was drawn from a Binomial distribution ($p = 0.5$, $n = 10$). These connections were subject to the reward-based synaptic sampling and rewiring described above.

To infer the lever position from the network activity, we weighted spikes from the neuron pool D with -1 and spikes from U with $+1$, summed them and then filtered them with a double-exponential filter kernel with $\tau_r = 50$ ms (rise) and $\tau_m = 500$ ms (decay). The cue input pattern was realized by the same method that was used to generate the patterns P_1 and P_2 outlined above. If a trial was completed successfully the reward signal $r(t)$ was set to 1 for 400 ms and was 0 otherwise. After each trial a short holding phase was inserted during which the input neurons were set to 2 Hz background noise. The lengths of these holding phases were uniformly drawn from the interval [1 s, 2 s]. In Fig. 2d-f the reward policy was changed after 24 hours by switching the decoding functions of the neural pools D and U and by randomly re-generating the input cue pattern.

To identify the movement onset times in Fig. 2d we adapted the method from [Peters et al., 2014]. Lever movements were recorded at a sampling rate of 5 ms. Lever velocities were estimated by taking the difference between subsequent time steps and filtered with a moving average filter of 5 time steps length. A Hilbert transform was applied to compute the envelope of the lever velocities. The movement onset time for each trial was then defined as the time point where the estimated lever velocity exceeded a threshold of 1.5 in the upward movement direction. If this value was never reached throughout the whole trial the time point of maximum velocity was used (most cases at learning onset).

The trial-averaged activity traces in Fig. 2d were generated by filtering the spiking activity of the network with a Gaussian kernel with $\sigma = 75$ ms. The activity traces were aligned with the movement onset times (indicated by black arrows in Fig. 2d) and averaged across 100 trials. The resulting activity traces were then normalized by the neuron's mean activity over all trials and values below the mean were clipped. The resulting activity traces were normalized to the unit interval.

Spine turnover statistics in Fig. 2f were measured as follows. The synaptic parameters were recorded in intervals of 2 hours. The number of synapses that appeared (crossed the threshold of $\theta_i(t) = 0$ from below) or disappeared (crossed $\theta_i(t) = 0$ from above) between two measurements were counted and the total number was reported as turnover rate.

The consolidation mechanism in Fig. 2c was realized as follows. All synapses that were above a threshold of $\theta_i(t) > 3$ for longer than 24 hours were consolidated. For consolidation, the mean of the prior was set to the current value of the synaptic parameter. The standard deviation of the synapse prior was set to a small value of $\sigma = 0.001$. All synapses that became consolidated (about 2%) persisted in this state for the rest of the experiment.

In Fig. 2g we randomly selected 5% of the synaptic parameters θ_i and recorded their traces over a learning experiment of 48 hours (1 sample per minute). The principal component analysis (PCA) was then computed over these traces, treating the parameter vectors at each time point as one data sample. The high-dimensional trace was then projected to the first three principal components in Fig. 2g, and colored according to the average movement completion time that was acquired by the network at the corresponding time points.

Details to: Relative contribution of stochastic and activity-dependent processes to synaptic plasticity. Synaptic weights were recorded Fig. 3a,b in intervals of 10 minutes. We selected all pairs of synapses with common pre- and post-synaptic neurons as CI synapses and synapse pairs with the same post- but not the same pre-synaptic neuron as non-CI synapses. In Fig. 3d-f we took a snapshot of the synaptic weights after 48 hours of learning and computed the Pearson correlation of all CI and non-CI pairs for random subsets of around 5000 pairs. Data for 100 randomly chosen CI synapse pairs are plotted of Fig. 3e.

Acknowledgments

Written under partial support by the Human Brain Project of the European Union #604102 and #720270.

References

Avermann et al., 2012. Avermann, M., Tomm, C., Mateo, C., Gerstner, W., and Petersen, C. (2012). Microcircuits of excitatory and inhibitory neurons in layer 2/3 of mouse barrel cortex. *Journal of Neurophysiology*, 107(11):3116–3134.

Baxter and Bartlett, 2000. Baxter, J. and Bartlett, P. L. (2000). Direct gradient-based reinforcement learning. In *The 2000 IEEE International Symposium on Circuits and Systems*, volume 3, pages 271–274. IEEE.

Botvinick and Toussaint, 2012. Botvinick, M. and Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488.

Buzsáki and Mizuseki, 2014. Buzsáki, G. and Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15:264–278.

Desai et al., 1999. Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience*, 2(6):515–520.

Ding and Rangarajan, 2004. Ding, M. and Rangarajan, G. (2004). First passage time problem: A fokker-planck approach. In Wille, L., editor, *New Directions in Statistical Physics*, pages 31–46. Springer.

Driscoll and Harvey, 2016. Driscoll, L. and Harvey, C. (2016). Dynamic reorganization of neuronal activity patterns in parietal cortex. In *Proc. of Cosyne 2016*, pages 110–111.

Dvorkin and Ziv, 2016. Dvorkin, R. and Ziv, N. E. (2016). Relative contributions of specific activity histories and spontaneous processes to size remodeling of glutamatergic synapses. *PLoS Biology*, 14(10):e1002572.

Fan et al., 2005. Fan, Y., Fricker, D., Brager, D. H., Chen, X., Lu, H.-C., Chitwood, R. A., and Johnston, D. (2005). Activity-dependent decrease of excitability in rat hippocampal neurons through increases in I_h. *Nature Neuroscience*, 8(11):1542–1551.

Florian, 2007. Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19(6):1468–1502.

Fusi et al., 2005. Fusi, S., Drew, P. J., and Abbott, L. (2005). Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611.

Gardiner, 2004. Gardiner, C. (2004). *Handbook of Stochastic Methods*. 3rd ed. Springer.

Gerstner et al., 2014. Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

Gewaltig and Diesmann, 2007. Gewaltig, M.-O. and Diesmann, M. (2007). NEST (NEural Simulation Tool). *Scholarpedia*, 2(4):1430.

Gopnik et al., 2015. Gopnik, A., Griffiths, T. L., and Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, 24(2):87–92.

Grashow et al., 2010. Grashow, R., Brookings, T., and Marder, E. (2010). Compensation for variable intrinsic neuronal excitability by circuit-synaptic interactions. *The Journal of Neuroscience*, 20(27):9145–9156.

Holtmaat and Svoboda, 2009. Holtmaat, A. and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10(9):647–658.

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

Izhikevich, 2007. Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral cortex*, 17(10):2443–2452.

Kappel et al., 2015. Kappel, D., Habenschuss, S., Legenstein, R., and Maass, W. (2015). Network plasticity as Bayesian inference. *PLoS Computational Biology*, 11(11):e1004485.

Kasai et al., 2010. Kasai, H., Fukuda, M., Watanabe, S., Hayashi-Takagi, A., and Noguchi, J. (2010). Structural dynamics of dendritic spines in memory and cognition. *Trends in neurosciences*, 33(3):121–129.

Kasthuri et al., 2015. Kasthuri, N., Hayworth, K. J., Berger, D. R., Schalek, R. L., Conchello, J. A., Knowles-Barley, D. L., Vzquez-Reina, A., Kaynig, V., Jones, T. R., Roberts, M., Morgan, J. L., Tapia, J. C., Seung, H. S., Roncal, W. G., Vogelstein, J. T., Burns, R., Sussman, D. L., Priebe, C. E., Pfister, H., and Lichtman, J. (2015). Saturated reconstruction of a volume of neocortex. *Cell*, 3:648–661.

Kirkpatrick et al., 1983. Kirkpatrick, S., Vecchi, M., et al. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kober et al., 2013. Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1278.

Legenstein et al., 2008. Legenstein, R., Pecevski, D., and Maass, W. (2008). A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS Computational Biology*, 4(10):e1000180.

Loewenstein et al., 2011. Loewenstein, Y., Kuras, A., and Rumpel, S. (2011). Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo. *The Journal of Neuroscience*, 31(26):9481–9488.

Loewenstein et al., 2015. Loewenstein, Y., Yanover, U., and Rumpel, S. (2015). Predicting the dynamics of network connectivity in the neocortex. *The Journal of Neuroscience*, 35(36):12535–12544.

Lucas et al., 2014. Lucas, C. G., Bridgers, S., Griffiths, T. L., and Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2):284–299.

Marder, 2011. Marder, E. (2011). Variability, compensation, and modulation in neurons and circuits. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15542–15548.

Marder and Goaillard, 2006. Marder, E. and Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563–574.

Marr and Poggio, 1976. Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.

Matsuzaki et al., 2001. Matsuzaki, M., Ellis-Davies, G. C., Nemoto, T., Miyashita, Y., Iino, M., and Kasai, H. (2001). Dendritic spine geometry is critical for ampa receptor expression in hippocampal ca1 pyramidal neurons. *Nature Neuroscience*, 4(11):1086–1092.

Minerbi et al., 2009. Minerbi, A., Kahana, R., Goldfeld, L., Kaufman, M., Marom, S., and Ziv, N. E. (2009). Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. *PLoS Biology*, 7(6):e1000136.

Peters et al., 2014. Peters, A. J., Chen, S. X., and Komiyama, T. (2014). Emergence of reproducible spatiotemporal activity during motor learning. *Nature*, 510(7504):263–267.

Peters and Schaal, 2006. Peters, J. and Schaal, S. (2006). Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2219–2225. IEEE.

Pfister et al., 2006. Pfister, J.-P., Toyoizumi, T., Barber, D., and Gerstner, W. (2006). Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural Computation*, 18(6):1318–1348.

Pouget et al., 2013. Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178.

Prinz et al., 2004. Prinz, A. A., Bucher, D., and Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nature Neuroscience*, 7(12):1345–1352.

Rawlik et al., 2013. Rawlik, K., Toussaint, M., and Vijayakumar, S. (2013). On stochastic optimal control and reinforcement learning by approximate inference. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3052–3056. AAAI Press.

Remme and Wadman, 2012. Remme, M. W. and Wadman, W. J. (2012). Homeostatic scaling of excitability in recurrent neural networks. *PLoS Computational Biology*, 8(5):e1002494.

Rokni et al., 2007. Rokni, U., Richardson, A. G., Bizzi, E., and Seung, H. S. (2007). Motor learning with unstable neural representations. *Neuron*, 54(4):653–666.

Seung, 2003. Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6):1063–1073.

Spitzer, 2015. Spitzer, N. C. (2015). Neurotransmitter switching? No surprise. *Neuron*, 86(5):1131–1144.

Statman et al., 2014. Statman, A., Kaufman, M., Minerbi, A., Ziv, N. E., and Brenner, N. (2014). Synaptic size dynamics as an effectively stochastic process. *PLoS Computational Biology*, 10(10):e1003846.

Stettler et al., 2006. Stettler, D. D., Yamahachi, H., Li, W., Denk, W., and Gilbert, C. D. (2006). Axons and synaptic boutons are highly dynamic in adult visual cortex. *Neuron*, 49:877–887.

Tang et al., 2010. Tang, L. S., Goeritz, M. L., Caplan, J. S., Taylor, A. L., Fisek, M., and Marder, E. (2010). Precise temperature compensation of phase in a rhythmic motor pattern. *PLoS Biology*, 8(8):e1000469.

Thompson-Schill et al., 2009. Thompson-Schill, S. L., Ramscar, M., and Chrysikou, E. G. (2009). Cognition without control when a little frontal lobe goes a long way. *Current Directions in Psychological Science*, 18(5):259–263.

Todorov and Jordan, 2002. Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235.

Urbanczik and Senn, 2009. Urbanczik, R. and Senn, W. (2009). Reinforcement learning in populations of spiking neurons. *Nature Neuroscience*, 12(3):250–252.

Vlassis et al., 2012. Vlassis, N., Ghavamzadeh, M., Mannor, S., and Poupart, P. (2012). Bayesian reinforcement learning. In *Reinforcement Learning*, pages 359–386. Springer.

Williams, 1992. Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Xie and Seung, 2004. Xie, X. and Seung, H. S. (2004). Learning in neural networks by reinforcement of irregular spiking. *Physical Review E*, 69(4):041909.

Xu et al., 2009. Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., Jones, T., and Zuo, Y. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature*, 462(7275):915–919.

Yagishita et al., 2014. Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620.

Yamahachi et al., 2009. Yamahachi, H., Marik, S. A., McManus, J. N. J., Denk, W., and Gilbert, C. D. (2009). Rapid axonal sprouting and pruning accompany functional reorganization in primary visual cortex. *Neuron*, 64(5):719–729.

Yang et al., 2009. Yang, G., Pan, F., and Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924.

876 Yasumatsu et al., 2008. Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008).
877 Principles of long-term dynamics of dendritic spines. *The Journal of Neuroscience*, 28(50):13592–13608.

878 Ziv and Ahissar, 2009. Ziv, N. E. and Ahissar, E. (2009). Neuroscience: New tricks and old spines.
879 *Nature*, 462(7275):859–861.

880 Ziv et al., 2013. Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L., Gama,
881 A. E., and Schnitzer, M. J. (2013). Long-term dynamics of CA1 hippocampal place codes. *Nature*
882 *Neuroscience*, 16(3):264–266.