Optimal synaptic dynamics for memory maintenance in the presence of noise

Dhruva V Raman[1] **and Timothy O'Leary**[1]

[1]**Department of Engineering, University of Cambridge, UK**

ABSTRACT

Synaptic connections in many brain areas have been found to fluctuate significantly, with substantial turnover and remodelling occurring over hours to days. Remarkably, this flux in connectivity persists in the absence of overt learning or behavioural change. What proportion of these ongoing fluctuations can be attributed to systematic plasticity processes that maintain memories and neural circuit function? We show under general conditions that the optimal magnitude of systematic plasticity is typically less than the magnitude of perturbations due to internal biological noise. Thus, for any given amount of unavoidable noise, 50% or more of total synaptic turnover should be effectively random for optimal memory maintenance. Our analysis does not depend on specific neural circuit architectures or plasticity mechanisms and predicts previously unexplained experimental measurements of the activity-dependent component of ongoing plasticity.

Keywords: Learning and Memory, Synaptic Plasticity, Optimization

INTRODUCTION

Learning depends systematic changes to the connectivity and strengths of synapses in neural circuits. This has been shown across experimental systems (Moczulska et al., 2013; Lai et al., 2012; Hayashi-Takagi et al., 2015) and is assumed by most theories of learning (Hebb, 1949; Bienenstock et al., 1982; Gerstner et al., 1996).

Neural circuits are required not only to learn, but to retain previously learned information. One might therefore expect synaptic stability in the absence of an explicit learning signal. However, many recent experiments in multiple brain areas have documented substantial ongoing synaptic modification in the absence of any obvious learning or change in behaviour (Attardo et al., 2015; Pfeiffer et al., 2018; Holtmaat et al., 2005; Loewenstein et al., 2015; Yasumatsu et al., 2008; Loewenstein et al., 2011).

It is natural to ask whether this apparently irreducible flux in neural connectivity is due to random biochemical noise, or due to systematic plasticity processes that have not been accounted for. A number of experimental studies have attempted to dissect and quantify both systematic and random synaptic changes at the level of synaptic physiology, either by directly interfering with synaptic plasticity or by correlating changes to circuit-wide measurements of ongoing physiological activity (Nagaoka et al., 2016; Quinn et al., 2019; Yasumatsu et al., 2008; Minerbi et al., 2009; Dvorkin and Ziv, 2016). Consistently, these studies find that the total rate of ongoing synaptic change is reduced by only 50% or less in the absence of neural activity. Similarly, systematic processes that are correlated across synapses only account for less than 50% of ongoing changes. Thus the bulk of ongoing synaptic fluctuations seem to be due to internal biological noise.

Put another way, these experimental findings imply that at steady state, systematic plasticity processes exert a weaker effect on synaptic strength than random fluctuations. This is surprising, because maintenance of neural circuit properties and learned behaviour would intuitively require random fluctuations to be dominated by systematic plasticity. To our knowledge, there is no theoretical account or model prediction that explains these observations.

In this study we consider neural circuits attempting to optimally retain previously learned information through some active, systematic plasticity process, in the presence of unavoidable, learning-independent, synaptic fluctuations. We conduct a first-principles mathematical analysis that is independent of specific plasticity mechanism and circuit architectures. We find that the magnitude of systematic plasticity should not exceed those of the intrinsic fluctuations, in direct agreement with experimental data. Furthermore, these fluctuations should dominate when systematic plasticity mechanisms are relatively precise, suggesting that random fluctuations will often dominate synaptic dynamics in neural circuits that exhibit learning-related plasticity. We validate these theoretical predictions in simulations. Together, our results provide a simple and general theory that explains a number of convergent but puzzling experimental findings, and suggest that synaptic plasticity mechanisms are optimised for the dynamic maintenance of stored information.

RESULTS

We begin with a brief survey of quantitative, experimental measurements of synaptic dynamics. We focused on studies that measured 'baseline' synaptic changes that occur outside of any behavioural learning paradigm, and which controlled for stimuli that induce widespread adaptive changes in synaptic strength.

Reference	Experimental system	Total baseline synaptic change	% synaptic change that is random / learning-independent
Pfeiffer et al. (2018)	Adult mouse hippocampus	40% turnover over 4 days	NA
Loewenstein et al. (2011)	Adult mouse auditory cortex	$>70\%$ of spines changed size by $>50\%$ over 20 days	NA
Zuo et al. (2005)	Adult mouse (barrel, primary motor, frontal) cortex	$3-5\%$ turnover over 2 weeks for all regions. $73.9\pm2.8\%$ of spines stable over 18 months (barrel cortex)	NA
Nagaoka et al. (2016)	Adult mouse visual cortex	8% turnover per 2 days in visually deprived environment. 15% in visually enriched environment. $7-8\%$ in both environments under pharmacological suppression of spiking.	$\approx 50\%$ (turnover)
Quinn et al. (2019)	Glutamatergic synapses, dissociated rat hippocampal culture	$28.2\pm3.7\%$ of synapses formed over 24 hour period. $28.6\pm2.3\%$ eliminated. Activity suppression through tetanus neurotoxin -light chain. Plasticity rate unmeasured.	$\approx 75\%$ (turnover)
Yasumatsu et al. (2008)	CA1 pyramidal neurons, primary culture, rat hippocampus	Measured rates of synaptic turnover and spine-head volume change. Baseline conditions vs activity suppression (NMDAR inhibitors). Turnover rates: $32.8\pm3.7\%$ generation/elimination per day (control) vs $22.0\pm$ 3.6% (NMDAR inhibitor). Rate of spine-head volume change:	$\approx 67\pm17\%$ (turnover). Size-dependent, but consistently $>50\%$ (spine-head volume)
Dvorkin and Ziv (2016)	Glutamatergic synapses in cultured networks of mouse cortical neurons	Partitioned commonly innervated (CI) synapses sharing same axon and dendrite, and non-CI synapses. Quantified covariance in fluorescence change for CI vs non-CI synapses to estimate relative contribution of activity histories to synaptic remodelling	$62-64\%$ (plasticity)
Minerbi et al. (2009)	Rat cortical neurons in primary culture	Created "*relative synaptic remodeling measure*" (RRM) based on frequency of changes in the rank ordering of synapses by fluorescence. Compared baseline RRM to when neural activity was suppressed by tetrodotoxin (TTX). RRM: 0.4 (control) vs 0.3 (TTX) after 30 hours.	$\approx 75\%$ (plasticity)
Ziv and Brenner (2018)	Literature review across multiple systems	"*Collectively these findings suggest that the contributions of spontaneous processes and specific activity histories to synaptic remodeling are of similar magnitudes*"	$\approx 50\%$

Table 1. Synaptic plasticity rates across experimental models, and the effect of activity suppression

Table 1 shows a breakdown of measured baseline synaptic modifications from multiple studies and brain

preparations. We note that there is large heterogeneity in the rates of baseline synaptic turnover across preparations and experimental conditions. For example, the expected lifetime of synapses in adult mouse hippocampus was estimated as $1-2$ weeks Attardo et al. (2015); Pfeiffer et al. (2018), while $>70\%$ of synapses in mouse barrel cortex persisted over 18 months Zuo et al. (2005).

It is reasonable to assume that some component of the total ongoing synaptic changes measured in these experiments arise from unavoidable, noisy fluctuations. Left unchecked, such random perturbations would eventually disrupt circuit function and erase any memories stored in the synaptic weight distribution. This suggests that there should be an additional, systematic component of ongoing plasticity that compensate for the deleterious effect of such fluctuations.

In order to isolate and quantify these components, several experiments in Table 1 blocked known synaptic plasticity pathways, reduced environmental stimuli or statistically factored out the effect of ongoing neural activity. Intriguingly, the rates of ongoing synaptic change remained high. Moreover, the relative reduction in synapse dynamics was remarkably consistent: across multiple brain regions, *in vivo* and *in vitro*, and despite large methodological differences, these studies consistently reported that at least half of ongoing synaptic change persisted.

These surprising observations motivated the central question that we address in this study:

> *how much systematic plasticity is expected in a neural circuit that needs to maintain overall function on a previously learned task while being subjected to unavoidable, task-independent synaptic fluctuations?*

We emphasize that our goal is not to explain the source of the intrinsic synaptic fluctuations, the mechanism of systematic plasticity, nor the *total* magnitude of ongoing synaptic change. Our goal is to derive a general relationship between the sources of ongoing change in synapses, and in doing so explain why random synaptic fluctuations seem to dominate.

For generality, we wanted to make minimal assumptions about the mechanisms of synaptic plasticity, as well as circuit architecture and function. The problem we are considering is outlined in Figure 1a. A neural network can perform some learned task. The level of task performance depends on the the state of various network parameters, such as synaptic connections strengths, and intrinsic neuron properties. Our results are independent of which kinds of parameters are involved, so we name them 'synaptic weights' for convenience. Task performance can be quantified in terms of an error function, $F[\mathbf{w}(t)]$, which depends upon the vector $\mathbf{w}(t)$ of synaptic weights at time t.

We assume that at least some ongoing plasticity processes are 'task-independent', and collectively refer to these processes as 'synaptic fluctuations'. Our definition of a task-independent process is one for which the probability of the process increasing or decreasing a particular synaptic weight, in a small time window, is independent of whether such a change is beneficial to task performance. One such process would be molecular noise affecting synaptic connection strengths. Another might be homeostatic mechanisms internal to each neuron. These perturb the weights in a direction $\boldsymbol{\varepsilon}[\mathbf{w}(t)]$. Our definition of 'task-independent' implies that such changes are on average uncorrelated with the direction of change in $\mathbf{w}(t)$ that would elicit maximal improvement in task performance, namely $-\nabla F[\mathbf{w}(t)]$. For this reason, we will often refer to them as 'random' or 'noise' components, even though they may have a deterministic origin. By definition, we have:

$$\mathbb{E}[\boldsymbol{\varepsilon}[\mathbf{w}(t)]^T \nabla F[\mathbf{w}(t)]] = 0. \tag{1a}$$

Intuitively, one would expect random fluctuations to degrade task performance. To formalise this, we will say that the network is in a **partially trained** state if a small, random change in synaptic weights satisfying equation (1a), degrades memory quality in expectation. Mathematically (see SI section 1.2), this is equivalent to the following condition:

$$Tr(\nabla^2 F[\mathbf{w}(t)]) > 0. \tag{1b}$$

The intuition for (1b) is as follows. We conceptualise F as a landscape, where $\mathbf{w}$ are the coordinates of a point on the landscape, and $F[\mathbf{w}]$ is the height of the landscape at $\mathbf{w}$ (Figure 1b). Improving task

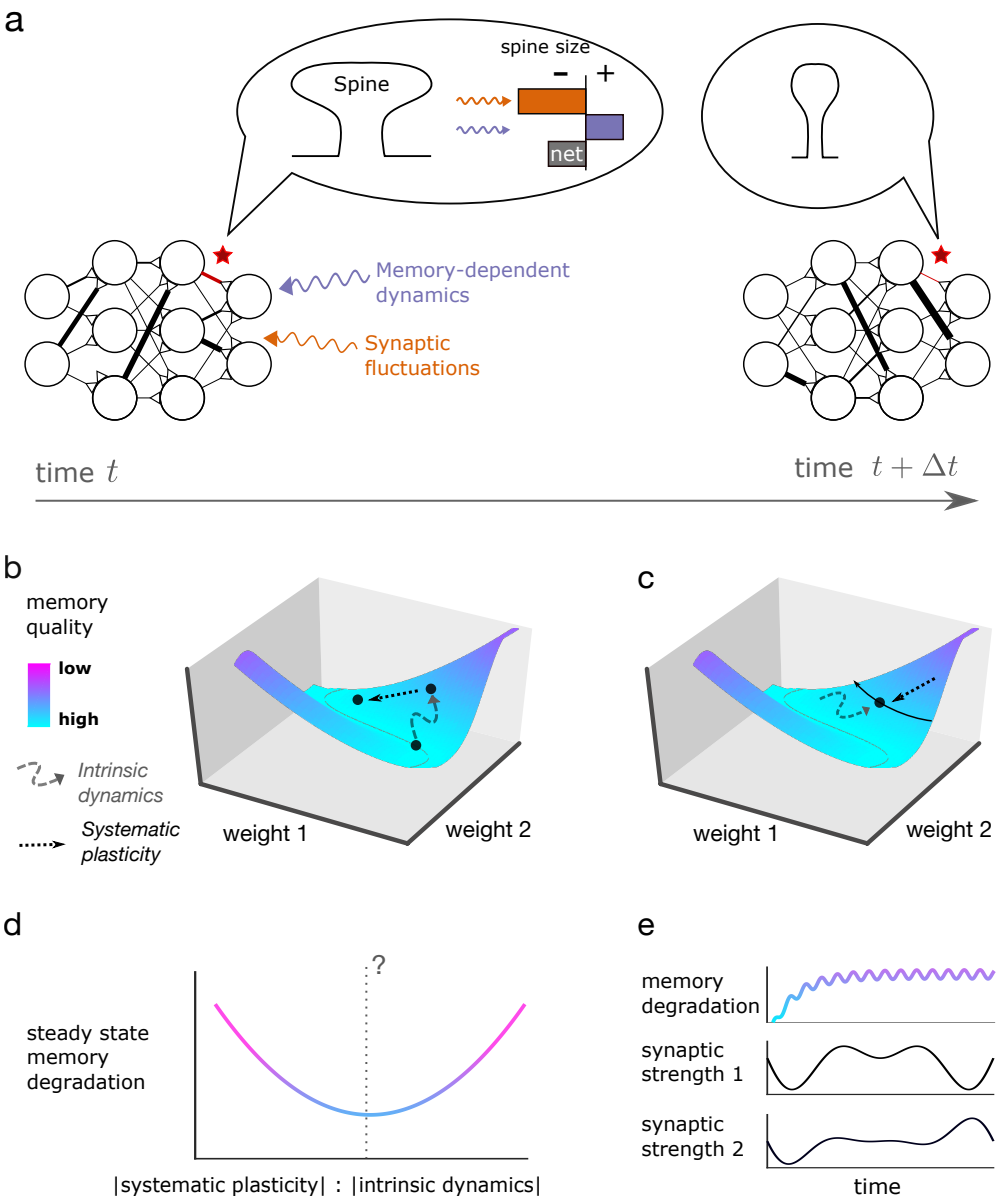

Figure 1. a: A learned task corresponds to an enforced input-output transformation for a neural circuit. This depends on the configuration of network parameters such as synaptic connection strengths, which change over time. Some changes arise from task-independent sources (e.g. molecular noise in the dynamic processes maintaining synapses). Unchecked, these processes will degrade task performance. Therefore some systematic 'relearning' term must nullify their functional effect. The combination of synaptic processes results in stable task performance, even as network parameters change systematically (e.g. see synapse labelled with red star). b: We represent task performance as a 'landscape'. Lateral co-ordinates represent the values of network parameters (only two are visualised). Height represents task error, so lower regions correspond to better performing configurations of network parameters. Ongoing change in the network parameters corresponds to 'wandering' across the landscape. Near the bottom (high task performance), task-independent dynamics will tend to move upwards on the landscape. Systematic plasticity moves doenwards, since it improves task performance. c: Eventually, a steady state is reached at which the effect of the two competing terms on memory quality cancel out. The network parameters wander over a level set of the landscape. d: For a fixed magnitude of task-independent synaptic fluctuations, what magnitude of systematic plasticity maximises memory quality (i.e. gives us the 'best level set' on the landscape)? e: Individual synapses may experience large, systematic changes at this steady state.

performance corresponds to moving downhill on the landscape. Equation (1b) says a randomly chosen direction on the landscape exhibits upward curvature on average. This is because each eigenvector of $\nabla^2 F[\mathbf{w}]$ corresponds to a direction on the landscape with curvature specified by the associated eigenvalue. If the sum of eigenvalues is positive (i.e. equation (1b)), then positive curvature is more prevalent than negative. For an isolated (local) minimum of the loss function, all eigenvalues are positive.

What is the relevance of curvature in determining the effect of random weight changes? A weight perturbation satisfying equation (1a) will be biased neither uphill nor downhill. We can pick a two-sided vector $\mathbf{v}$ along which to perturb $\mathbf{w}$ (Figure 2b). Moving in the downhill direction of $\mathbf{v}$ improves (decreases) $F[\mathbf{w}]$. However, upward curvature diminishes the degree of improvement. On the other hand, moving in the uphill direction increases $F[\mathbf{w}]$, and this increase is magnified by upward curvature. Thus a random fluctuation along $\mathbf{v}$, biased in neither the positive or negative direction of $\mathbf{v}$, will decrease task performance (increase $F[\mathbf{w}]$) in expectation. Equation (1b) follows by averaging over all possible directions $\mathbf{v}$ in a partially trained network.

This reasoning explains why synaptic fluctuations are expected to degrade task performance in a partially trained network. If such fluctuations are unavoidable but task performance is maintained, then a systematic plasticity term must counteract the fluctuations. We will denote this term $\mathbf{c}[\mathbf{w}(t)]$. For notational clarity, we will subsequently omit explicit dependence on $\mathbf{w}(t)$, i.e. $\mathbf{c}[\mathbf{w}(t)]$ and $\varepsilon[\mathbf{w}(t)]$ become $\mathbf{c}(t)$ and $\varepsilon(t)$. Overall synaptic dynamics can now be written as

$$\dot{\mathbf{w}}(t) = \mathbf{c}(t) + \varepsilon(t). \tag{2}$$

Note that many studies attempt to model the distribution of fluctuations $\varepsilon(t)$. For instance, Statman et al. (2014); Yasumatsu et al. (2008); Loewenstein et al. (2011) identify the importance of synapse size, age, and morphology in determining fluctuation magnitude. Our study is agnostic to such considerations, as long as fluctuations are memory independent (i.e. satisfy (1a)).

The systematic plasticity term $\mathbf{c}(t)$ lumps together the contribution of all synaptic plasticity mechanisms that dynamically maintain the learned task. These might be referred to as 'learning rules' but we emphasize that most of our interest lies in the steady-state *maintenance* of a learned task, i.e. when no additional learning is occurring.

Consider the plasticity rates of $\mathbf{c}(t)$ and $\varepsilon(t)$ over a small time interval, Δt. We define:

$$\Delta\mathbf{c} := \int_{t^*}^{t+\Delta t} \mathbf{c}(t')dt' \text{ and } \Delta\varepsilon := \int_{t^*}^{t+\Delta t} \varepsilon(t')dt'.$$

$\Delta\mathbf{c}$ represents the cumulative effect of systematic plasticity and $\Delta\varepsilon$ the cumulative effect of fluctuations over the time interval Δt. Now consider the change in memory quality during this time, $\Delta F := F[\mathbf{w}(t^* + \Delta t)] - F[\mathbf{w}(t^*)]$. A second order Taylor approximation of ΔF gives:

$$\begin{aligned}\Delta F = {} & \Delta\varepsilon^T \nabla F[\mathbf{w}(t^*)] + \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] \\ & + \frac{1}{2}\Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\mathbf{c} + \frac{1}{2}\Delta\varepsilon^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\varepsilon \\ & + \Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\varepsilon + \mathcal{O}(\|\Delta\mathbf{c} + \Delta\varepsilon\|_2^3).\end{aligned} \tag{3}$$

We can choose Δt sufficiently small that the higher order terms in $\mathcal{O}(\|\Delta\mathbf{c} + \Delta\varepsilon\|_2^3)$ can be ignored. We do not make specific assumptions on the mechanism by which the learning rule produces synaptic changes $\Delta\mathbf{c}$. By definition, however, the effect of $\Delta\mathbf{c}$ is to improve task performance. For $\Delta\mathbf{c}$ to have such an effect on ΔF, equation (3) shows that we require

$$\Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] + \frac{1}{2}\Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\mathbf{c} < 0. \tag{4a}$$

Indeed for a sufficiently small Δt, we additionally require that

$$\Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] < 0. \tag{4b}$$

This shows that $\Delta\mathbf{c}$ must be anticorrelated with the gradient $\nabla F[\mathbf{w}]$. Additionally, it might exploit information on $\nabla^2 F[\mathbf{w}]$. Geometrically, $\Delta\mathbf{c}$ points in a descending direction on the loss landscape of $F[\mathbf{w}]$.

Our first task is to find what magnitude of systematic plasticity (i.e. $\|\Delta\mathbf{c}\|_2$) optimises the degree of learning ΔF, assuming any fixed direction $\Delta\hat{\mathbf{c}}$. We use hats to denote normalised variables (i.e. $\hat{x} = \frac{x}{\|x\|_2}$). We will make extensive use of the following operator:

$$\mathscr{Q}_{\mathbf{w}}[\mathbf{v}] = \hat{\mathbf{v}}^T \nabla^2 F[\mathbf{w}] \hat{\mathbf{v}}.$$

Geometrically, $\mathscr{Q}_{\mathbf{w}}[\mathbf{v}]$ represents the relative degree of upward curvature of the loss landscape of F at the point $\mathbf{w}$, in the direction $\mathbf{v}$. This is shown graphically in Figure 2c. Note also that $\mathscr{Q}$ is scale invariant, i.e. $\mathscr{Q}_{\mathbf{w}}[\mathbf{v}] = \mathscr{Q}_{\mathbf{w}}[k\mathbf{v}]$ for any scalar $k \in \mathbb{R}$. It depends on the direction, not the magnitude, of $\mathbf{v}$.

We can rewrite equation (3), using the operator $\mathscr{Q}$ and omitting higher order terms (as previously justified):

$$\begin{aligned}\Delta F = {} & \Delta\boldsymbol{\varepsilon}^T \nabla F[\mathbf{w}(t^*)] + \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] \\ & + \frac{1}{2}\|\Delta\mathbf{c}\|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\mathbf{c}] + \frac{1}{2}\|\Delta\boldsymbol{\varepsilon}\|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\boldsymbol{\varepsilon}] \\ & + \Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\boldsymbol{\varepsilon}.\end{aligned} \tag{5}$$

Since $\Delta\boldsymbol{\varepsilon}$ consists of memory-independent processes, we can consider them as coming from some unknown probability distribution that is uncorrelated, in expectation, with the derivatives of F. Thus, any term in equation (3) that is linear in $\Delta\boldsymbol{\varepsilon}$ disappears in expectation. In particular,

$$\mathbb{E}[\nabla F[\mathbf{w}(t^*)]^T \Delta\boldsymbol{\varepsilon}] = 0. \tag{6a}$$

$$\mathbb{E}[\Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\boldsymbol{\varepsilon}] = 0, \tag{6b}$$

which collectively imply

$$\mathbb{E}[\Delta F] = \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] + \frac{1}{2}\|\Delta\mathbf{c}\|_2^2 \mathscr{Q}[\Delta\mathbf{c}] + \frac{1}{2}\|\Delta\boldsymbol{\varepsilon}\|_2^2 \mathscr{Q}[\Delta\boldsymbol{\varepsilon}]. \tag{7}$$

The requirement for assumption (6b) can be removed (see SI section 1.1). We can differentiate equation (7) in $\|\Delta\mathbf{c}\|_2$, to get:

$$\frac{\mathrm{d}\mathbb{E}[\Delta F]}{\mathrm{d}\|\Delta\mathbf{c}\|_2} = \Delta\hat{\mathbf{c}}^T \nabla F[\mathbf{w}(t^*)] + \|\Delta\mathbf{c}\|_2 \mathscr{Q}[\Delta\mathbf{c}].$$

The root of this derivative gives a global minimum of the equation (7) in $\|\Delta\mathbf{c}\|_2$, as long as $\mathscr{Q}[\Delta\mathbf{c}] \geq 0$ holds (justified in SI section 1.2). We get

$$\|\Delta\mathbf{c}\|_2^* = \frac{-\Delta\hat{\mathbf{c}}^T \nabla\hat{F}}{\mathscr{Q}[\mathbf{c}]}\|\nabla F\|_2, \tag{8}$$

which defines the magnitude of systematic plasticity that minimises ΔF, and thus maximises task performance at time $t^* + \Delta t$.

If the memory improved over the interval Δt, then we would have $\Delta F < 0$. However, we are interested in the special case of memory maintenance, where improvements from $\Delta\mathbf{c}$ are cancelled out by decrements from $\Delta\boldsymbol{\varepsilon}$, and we have $\mathbb{E}[\Delta F] = 0$. Substituting this into equation (7), we get

$$0 = \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] + \frac{1}{2}\|\Delta\mathbf{c}\|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\mathbf{c}] + \frac{1}{2}\|\Delta\boldsymbol{\varepsilon}|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\boldsymbol{\varepsilon}].$$

Next, we substitute in our optimal reconsolidation magnitude (equation (8)). This gives

$$0 = -\frac{1}{2}\|\Delta\mathbf{c}\|_2^2 \mathscr{Q}[\Delta\mathbf{c}] + \frac{1}{2}\|\Delta\boldsymbol{\varepsilon}|_2^2 \mathscr{Q}[\Delta\boldsymbol{\varepsilon}].$$

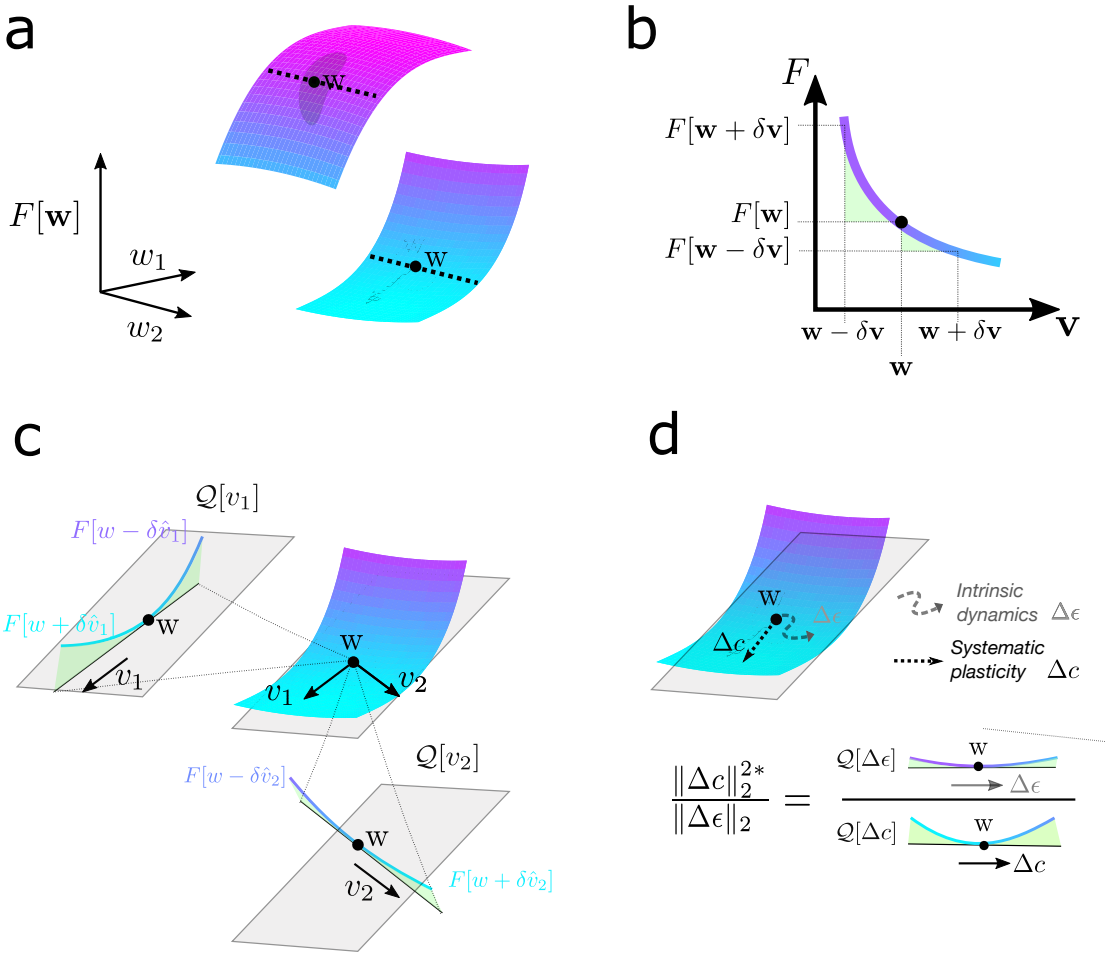

Figure 2. a: Two different error surfaces. Shaded patches are probability distributions of changes to $\mathbf{w}$. The distributions are uncorrelated, in expectation, with the gradient. We see this as half the probability density lies on each side of the dotted line denoting the level set. The concave (convex) surface on top (bottom) curves downwards (upwards) in all directions, which means that a random change in $\mathbf{w}$ following the probability distribution will decrease (increase) error in expectation. If a network state is such that random changes, uncorrelated with $\nabla F[\mathbf{w}]$, increase error in expectation, then the error surface curves upwards along the probability distribution more than it curves downwards. b: Geometrical intuition behind the operator $\mathcal{Q}_{\mathbf{w}}$. The operator charts the degree to which a direction lifts off the tangent plane (grey). In other words, the relative upward curvature of the surface in a given direction. The green, shaded areas are proportional to $\mathcal{Q}_{\mathbf{w}}[v_1]$, and $\mathcal{Q}_{\mathbf{w}}[v_2]$. Note that the operator considers normalised directions, so does not take account the magnitude of either of these vectors. c: The weights of a networks change due to memory-independent 'intrinsic dynamics', as well as systematic plasticity that acts to reconsolidate the memory. What is the optimal magnitude of systematic plasticity, relative to the magnitude of the intrinsic dynamics? As $\mathcal{Q}_{\mathbf{w}}[\Delta\mathbf{c}]$ increases and/or $\mathcal{Q}_{\mathbf{w}}[\Delta\boldsymbol{\varepsilon}]$ increases, the optimal magnitude of systematic plasticity increases.

Therefore,

$$\frac{\|\Delta\mathbf{c}\|_2^2}{\|\Delta\varepsilon\|_2^2} = \frac{\mathscr{Q}[\Delta\varepsilon]}{\mathscr{Q}[\Delta\mathbf{c}]}, \tag{9}$$

and moreover,

$$\mathscr{Q}[\Delta\mathbf{c}] \geq \mathscr{Q}[\Delta\varepsilon] \tag{10}$$

implies that the magnitude of synaptic fluctuation should outcompete that of systematic plasticity for optimal memory maintenance.

Note that an alternative derivation of equation (9) (see SI section 1.1) removes the need for assumption (6b). Equation (9) is valid when the numerator and denominator of the right hand side are both positive. The converse is unlikely in a partially trained network, and impossible in a highly trained network (see SI section 1.2).

Now let us provide more intuition for equation (9). Recall from Figure **??** that $\mathscr{Q}_{\mathbf{w}}[\Delta\mathbf{c}]$ represents the relative upward curvature of $F[\mathbf{w}]$ in the direction $\Delta\mathbf{c}$ (and the same for $\Delta\varepsilon$). Therefore we can interpret (9) as follows:

Draw two unit-length lines on the loss function landscape, both starting at $\mathbf{w}(t^*)$, and in the directions $\Delta\hat{\mathbf{c}}$ and $\Delta\hat{\varepsilon}$ respectively. Measure the upward curvature of the loss function along both of these lines. If the curvature is greater (or equal) for line $\Delta\hat{\mathbf{c}}$, then optimal memory retention over Δt requires $\|\Delta\mathbf{c}\|_2^2 \leq \|\Delta\varepsilon\|_2^2$.

Geometric intuition can give a glimpse into why equation (10) should generically hold, although we also demonstrate this mathematically (SI section 1.3). $\Delta\hat{\mathbf{c}}$ is a descent direction on the loss landscape, since it acts to improve memory quality (see equation (4)). Descent directions will generically have high upwards curvature in highly trained states (i.e. when $F[\mathbf{w}]$ is close to a minimum). In other words, a cross section of the loss landscape along such a descent direction will be U-shaped. For this to be false, F would have to consistently decrease along the cross section. However, the degree to which this decrease can occur is limited by the already low value of $F[\mathbf{w}]$. On the other hand, arbitrary directions (such as $\Delta\varepsilon$) may not act to decrease F, and thus may not have U-shaped cross sections.

We have intuitively justified why (10) should hold. This equation implies that systematic plasticity should generically be outcompeted by synaptic fluctuations, for optimal memory retention. We can justify the same assertion analytically by quantifying $\mathscr{Q}[\Delta\varepsilon]$ and $\mathscr{Q}[\Delta\mathbf{c}]$.

First note that $\mathscr{Q}[\Delta\varepsilon]$ is task-independent. Thus, it should have no systematic relationship with directions of curvature at $F[\mathbf{w}]$. In other words, it should project unbiasedly onto the different eigenvectors of $\nabla^2 F[\mathbf{w}]$. This assumption implies (see SI section 1.2)

$$\mathbb{E}[\mathscr{Q}[\Delta\varepsilon]] = \frac{Tr(\nabla^2 F[\mathbf{w}])}{N}. \tag{11}$$

Note that the RHS of the above equation corresponds to the mean of the eigenvalues of $\nabla^2 F[\mathbf{w}]$.

We now turn to quantifying $\mathscr{Q}[\Delta\mathbf{c}]$. Regardless of the mechanism generating systematic plasticity, it must act to improve task performance. This constraint gave us equation (4): $\Delta\mathbf{c}$ must anticorrelate with the gradient $\nabla F[\mathbf{w}]$, and it can also benefit from information on $\nabla^2 F[\mathbf{w}]$. We can consider two extremal cases:

1. $\Delta\mathbf{c}$ is computed with perfect access to $\nabla F[\mathbf{w}]$, and (possibly) $\nabla^2 F[\mathbf{w}]$.
2. The quantity of information on $\nabla F[\mathbf{w}]$ available with which to compute $\Delta\mathbf{c}$ tends towards zero.

In the SI (section 1.3), we show that synaptic fluctuations outcompete/equal systematic plasticity in both of these extremal cases, and intermediately by interpolation. We therefore suggest that in biological systems maintaining a memory through reconsolidation, the appropriate null hypothesis is that the magnitude of synaptic fluctuations outcompetes the magnitude of reconsolidation plasticity.

Another interesting phenomenon follows from the calculations of SI section 1.3, in the case that only information on $\nabla F[\mathbf{w}]$ is avilable. If $\Delta\mathbf{c}$ can perfectly access $\nabla F[\mathbf{w}]$, then $\Delta\mathbf{c} \propto -\nabla F[\mathbf{w}]$ optimally decreases $F[\mathbf{w}]$. This corresponds to perfect implementation of backpropagation (i.e. gradient descent) by $\Delta\mathbf{c}$. In this case, (9) is smaller than one, and intrinsic fluctuations outcompete systematic plasticity. As we decrease the accuracy of backpropagation (i.e. corrupt access to $\nabla F[\mathbf{w}]$ with task-independent noise), (9) increases towards one: the optimal magnitude of systematic plasticity increases. In other words, a less precise systematic plasticity mechanism has to do 'more work' to optimally counteract intrinsic fluctuations. This is demonstrated numerically in Figures 4 and 3.

We now verify our conclusions in simulations. We consider simple, feedforward, rate-based neural networks. We emphasise that the results do not depend on a particular choice of network model, learning rule, or task. Our aim is to verify the ratios of systematic to intrinsic plasticity that result in optimal steady-state performance on a particular learned task.

The task we provide our neural networks is to maintain their initial input-output behaviour over time as well as possible, even as individual weights fluctuate due to systematic and intrinsic plasticity. This models maintenance of a previously learned task. Initial network behaviour is set by randomly setting the initial neural network weights. We 'save' the initial network behaviour by keeping a fixed copy of the network at time zero, with fixed weights. We call this fixed network the 'teacher', and the adaptive network the 'student'. Our learning problem then recreates the student-teacher framework described in e.g. Levin et al. (1990); Seung et al. (1992).

As before, we model weight change over a time interval T as

$$\Delta\mathbf{w}_t = \Delta\mathbf{c}_t + \Delta\boldsymbol{\varepsilon}_t,$$

where the index t defines the number of previously elapsed intervals. We model $\Delta\boldsymbol{\varepsilon}_t$, which represents integrated intrinsic fluctuations over the time interval, as a scaled, i.i.d, white-noise process at each synapse:

$$\Delta\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \gamma_3\mathbb{I}),$$

for some scaling factor $\gamma_3 > 0$. In order to describe $\Delta\mathbf{c}_t$ we first have to define an error function $F[\mathbf{w}]$ for the student network. We take

$$F[\mathbf{w}] = \frac{1}{|\mathscr{U}|}\sum_{u\in\mathscr{U}} \|y(\mathbf{w},u) - y^*(u)\|_2^2,$$

where $\mathscr{U}$ is a set of inputs with cardinality $|\mathscr{U}|$, $y^*(u)$ is the output of the teacher for input $u \in \mathscr{U}$, and $y(\mathbf{w},u)$ is the output of the student, with weights $\mathbf{w}$. We generated inputs $u \in \mathscr{U}$ as i.i.d Gaussian vectors, and took $\mathscr{U}$ as a set of 1000 such vectors.

We model the systematic plasticity term $\Delta\mathbf{c}_t$ as a noise-corrupted gradient descent term. Any such term must anticorrelate to some degree with the gradient (see equation (4) and surrounding discussion, as well as Raman et al. (2019)). So we take

$$\Delta\mathbf{c}_t = \gamma_1\nabla\hat{F}[\mathbf{w}]_t + \gamma_2\hat{\nu}_t,$$

where $\nu_t \sim \mathcal{N}(0,\mathbb{I})$ models imperfections in the approximation of the gradient, and $\gamma_1, \gamma_2 > 0$ are scaling parameters. By changing the ratio $\frac{\gamma_1}{\gamma_2}$, we can interpolate between high and low quality learning rules. Meanwhile $\sqrt{\gamma_1^2+\gamma_2^2}$ represents the overall magnitude of systematic plasticity in the time interval, which can be compared with γ_3, the magnitude of intrinsic fluctuations.

We ran simulations of a sigmoidally nonlinear multilayer perceptron with a single hidden layer, tuning the vector $\boldsymbol{\gamma}$ of hyperparameters to investigate different qualities of learning rule, and different ratios of systematic to intrinsic fluctuations (see Figure 3). Under all conditions, optimal steady state performance was achieved when the magnitude of systematic plasticity was less or equal to the level of intrinsic fluctuations, corroborating analytical calculations.

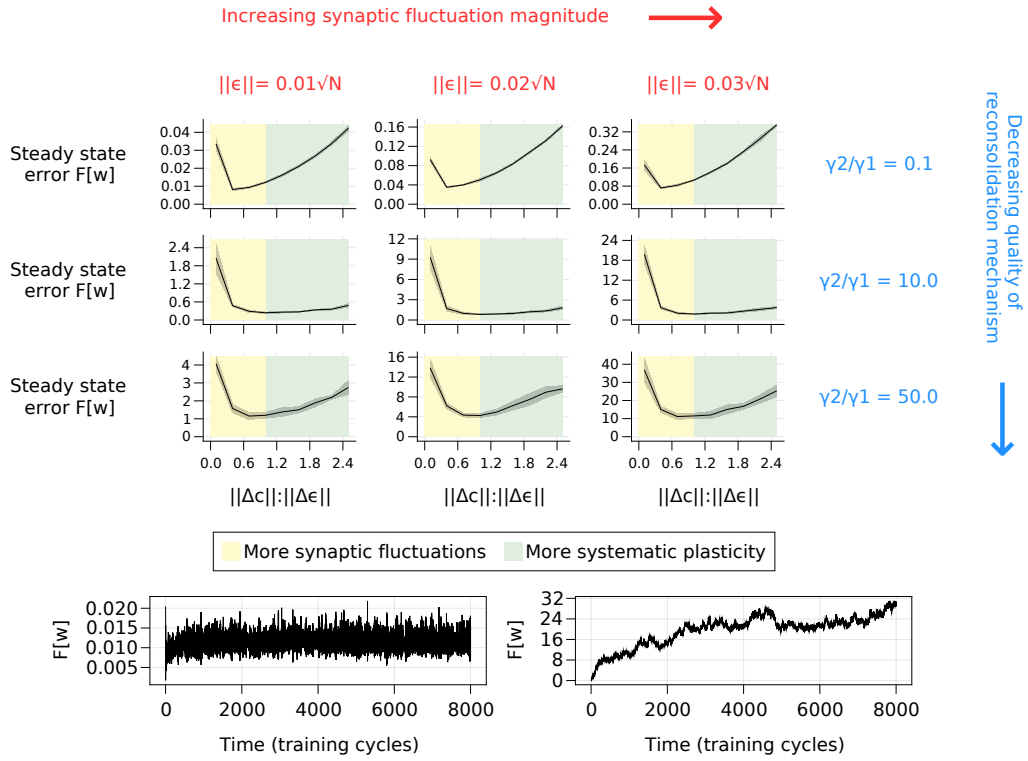

Figure 3. The relationship between steady state task performance and the ratio of systematic plasticity to intrinsic fluctuations in a nonlinear network. We consider multilayer perceptrons with 12 inputs, 10 outputs, and 20 neurons in the hidden layer. Each neuron has a sigmoidal nonlinearity. Weight dynamics are described in the 'simulations' section. Each subgraph in the top pane shows steady state error over 8 repeats, with standard deviation over the repeats shaded. The x-axis of each subgraph is equivalently $\frac{\sqrt{\gamma_1^2+\gamma_2^2}}{\gamma_3}$. The bottom pane depicts sample trajectories of task error over time, for different choices of hyperparameters γ_1, γ_2, and γ_3.

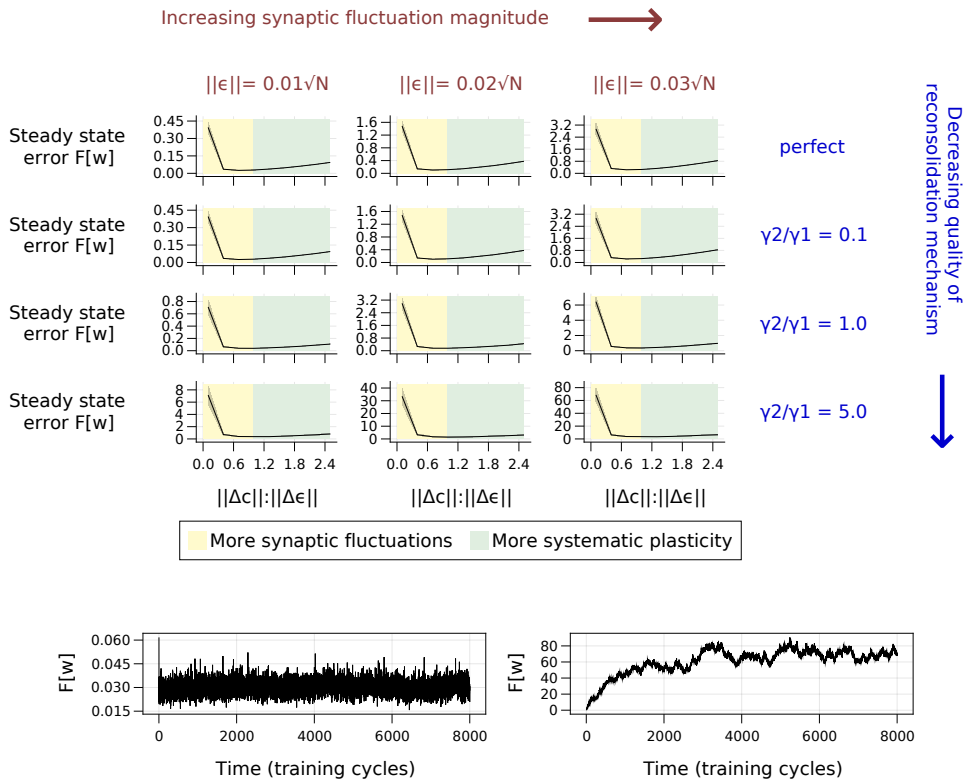

Figure 4. The relationship between steady state task performance and the ratio of systematic plasticity to intrinsic fluctuations in a linear network. We consider linear networks with output $y(\mathbf{w}, u) = Wu$, where $W \in \mathbb{R}^{12\times 10}$ is a matrix representation of the weight vector $\mathbf{w}$. Weight dynamics are described in the 'simulations' section. We ran the dynamics for 8000 training cycles, to allow task performance to settle to a steady state. Each subgraph in the top pane shows steady state error over 8 repeats, with standard deviation over the repeats shaded. **Imperfect learning rules (columns 2-4)**: The x-axis of each subgraph is equivalently $\frac{\sqrt{\gamma_1^2+\gamma_2^2}}{\gamma_3}$. The bottom pane depicts sample trajectories of task error over time, for different choices of hyperparameters γ_1, γ_2, and γ_3. **Perfect learning rule (top row)**: The systematic learning rule updates weights with the Newton step ($\Delta\mathbf{c}_t \propto (\nabla^2 F[\mathbf{w}]_t)^{-1} F[\mathbf{w}]_t$), which becomes (in the linear case): $\Delta\mathbf{c}_t \propto \mathbf{w}^* - \mathbf{w}$. Overall weight dynamics are $\Delta\mathbf{w}_t = \Delta\mathbf{c}_t + \Delta\varepsilon_t$, as before.

DISCUSSION

A long-standing question in neuroscience is how neural circuits optimally maintain memory of a learned task while being buffeted by synaptic fluctuations from noise and other task-independent processes (Fusi et al., 2005). There are several hypotheses that offer potential answers, none of which is mutually exclusive. One possibility is that fluctuations only occur in a subset of volatile connections that are relatively unimportant for learned behaviours Moczulska et al. (2013); Chambers and Rumpel (2017); Kasai et al. (2003). Following this line of thought, circuit models have been proposed that only require stability in a subset of synapses for stable function Clopath et al. (2017); Mongillo et al. (2018); Susman et al. (2018). Another hypothesis is that any memory degradation due to fluctuations is counteracted by restorative, systematic plasticity processes that allow circuits to continually 'relearn' stored associations. The source of information for the systematic plasticity term could come from an external reinforcement signal Kappel et al. (2018), from interactions with other circuits Acker et al. (2018), or spontaneous, network-level reactivation events Fauth and van Rossum (2019). A final possibility is that we rarely observe behavioural states in which an animal is not learning, and that unobserved behavioural changes account for apparent fluctuations in brain connectivity in any given experiment.

Our work does not argue exclusively for or against any of these three broad hypotheses. Rather, we extracted logical consequences from assuming that all hypotheses are viable to some extent. An important caveat to our work is that we do make specific assumptions whose validity depends on the state of current knowledge, and might vary depending on the organism or brain area in question. Most crucially, we assumed that not all fluctuations in synaptic strength are explained by behavioural adaptation and learning. To the extent that this is true, the residual fluctuations must come from ongoing, endogenous processes and irreducible noise that continually perturb synaptic strengths. Our analysis then proceeded by assuming that learned information does not decay appreciably over time, which requires degradation from these task-independent processes to be counteracted by systematic plasticity mechanisms.

Several predictions follow immediately from our analysis. Foremost among these, we predict that for optimal retention of circuit function and learned behaviour, the systematic plasticity contribution should not outcompete task-independent fluctuations. This prediction is somewhat unsettling yet it is borne out across a number of experimental studies (Nagaoka et al., 2016; Quinn et al., 2019; Yasumatsu et al., 2008; Minerbi et al., 2009; Dvorkin and Ziv, 2016). It is intuitively clear that memories should degrade when systematic plasticity is far weaker than noise. It is far less intuitive that maintenance of learned behaviours will also suffer if a learning rule outcompetes task-independent fluctuations at steady state.

By parameterising all possible qualities of systematic plasticity - from precise to highly inaccurate - we also show that a larger task-independent component of ongoing synaptic change predicts a more accurate systematic plasticity mechanism. In other words, sophisticated learning rules need to do less work to overcome the damage done by task-independent synaptic fluctuations. Experimental estimates (see Table 1) suggest task-independent fluctuations often outcompete systematic changes in synaptic strength. Our theory implies that this is consistent with relatively precise learning rules in biological synapses that can approximate the gradient of task error relatively well.

We must qualify the meaning of 'relatively precise' in this conclusion. We conceptualised memory quality as a landscape whose height denotes error. We noted that any systematic plasticity mechanism should act to increase memory quality, and therefore change the weights in a downhill direction on the landscape. Therefore it must have at least some local information on the slope (gradient) and possibly curvature (hessian) of the landscape. We parameterised precision by the quality of access to these quantities. The assertion of the previous paragraph assumed no access to the curvature, as is consistent with biologically plausible learning rules we have seen in the literature (e.g. Williams (1992); Mazzoni et al. (1991); Seung (2003); Lillicrap et al. (2016); Sussillo and Abbott (2009)).

If it were the case that biological learning rules could perfectly access both the gradient and second derivative of the landscape, then the optimal contributions of systematic plasticity and intrinsic fluctuations would in fact be equal (SI section 1.3.2). This would correspond to the systematic plasticity mechanism directly undoing any synaptic changes induced by task-independent fluctuations. We suggest that this is not biologically realistic. First, it is widely believed that even accurate gradient-based learning rules are biologically implausible. Second, there is widespread evidence that neural circuit reconfiguration occurs

in the absence of learning and that many synapses have finite lifetimes, indicating that synapses do not continually revert to some learned state.

There are important caveats to how our results should be interpreted in light of existing experimental data. It is technically difficult to experimentally isolate systematic and random components of synaptic change. Approaches often rely on reduced preparations where 'learning' and 'behaviour' have no direct relationship to neural circuit dynamics. On the other hand, in *in vivo* studies it is extraordinarily difficult to accurately measure synaptic changes and to control for confounding changes in behaviour or physiology. We may therefore simply take these measurements as the best available data and assume that *at least some* ongoing synaptic dynamics are noise-driven.

Thus, while our results offer a surprising agreement with a number of experimental observations, we believe it is important to further replicate measurements of synaptic turnover and synaptic modification in a variety of settings, both *in vivo* and *in vitro*. To this end, we hope our results provide an impetus for this difficult experimental work, because it offers a principled framework for understanding the surprising volatility of connections in neural circuits.

ACKNOWLEDGMENTS

This work was supported by ERC grant StG 2016 716643 FLEXNEURO.

1 SUPPLEMENTAL METHODS

1.1 Alternative derivation of equation (9)

We provide an alternative derivation of equation (9) that removes the need for assumption (6b). We did not put this main derivation in the main text as we perceive it to have less clarity.

The derivation proceeds identically to that given in the main text until equation (5). We can then use (6a) to simplify equation (5). We get

$$\begin{aligned}\mathbb{E}[\Delta F] = \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] \\ + \frac{1}{2}\|\Delta\mathbf{c}\|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\mathbf{c}] + \frac{1}{2}\|\Delta\boldsymbol{\varepsilon}\|_2^2 \mathscr{Q}_{\mathbf{w}(t^*)}[\Delta\boldsymbol{\varepsilon}] \\ + \Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\boldsymbol{\varepsilon}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\Delta F] = \Delta\mathbf{c}^T \nabla F[\mathbf{w}(t^*)] \\ + \frac{1}{2}\Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\mathbf{c} + \frac{1}{2}\Delta\boldsymbol{\varepsilon}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\boldsymbol{\varepsilon} \\ + \Delta\mathbf{c}^T \left(\nabla^2 F[\mathbf{w}(t^*)]\right)\Delta\boldsymbol{\varepsilon}.\end{aligned}$$

Recall that expectation is taken over an unknown probability distribution from which $\Delta\boldsymbol{\varepsilon}$ is drawn, which satisfies equation (6a).

We then assume that we are in a phase of stable memory retention, so that $\mathbb{E}[\Delta F] = 0$. Now if the magnitude of systematic plasticity $\|\Delta\mathbf{c}\|_2$ is tuned to minimise steady state error F, then any change to $\|\Delta\mathbf{c}\|_2$ will result in an increase in $\mathbb{E}[\gtrless\mathbb{F}]$. So $\mathbb{E}[\Delta F]$ is locally minimal in $\|\Delta\mathbf{c}\|_2$. This implies

$$\frac{\mathrm{d}\mathbb{E}[\Delta F]}{\mathrm{d}\|\Delta\mathbf{c}\|_2} = 0.$$

We also claim that local minimality implies

$$\frac{\mathrm{d}\mathbb{E}[\frac{\Delta F}{\|\Delta\mathbf{c}\|_2}]}{\mathrm{d}\|\Delta\mathbf{c}\|_2} = 0. \tag{12}$$

Why? $\mathbb{E}[\Delta F] = 0$ implies that $\mathbb{E}[\frac{\Delta F}{\|\Delta\mathbf{c}\|_2}] = 0$. If a small change to $\|\Delta\mathbf{c}\|_2$ results in $\mathbb{E}[\Delta F] \geq 0$, then it also results in $\mathbb{E}[\frac{\Delta F}{\|\Delta\mathbf{c}\|_2}] \geq 0$, since $\|dC\|_2$ is non-negative.

Expanding the LHS of equation (12), we get

$$\frac{\mathrm{d}}{\mathrm{d}\|\Delta\mathbf{c}\|_2}\Bigg\{\Delta\hat{\mathbf{c}}^T\nabla F[\mathbf{w}(t^*)]+\frac{1}{2}\|\Delta\mathbf{c}\|_2\Delta\hat{\mathbf{c}}^T\left(\nabla^2F[\mathbf{w}(t^*)]\right)\Delta\hat{\mathbf{c}} + \frac{1}{2}\frac{\Delta\varepsilon^T\left(\nabla^2F[\mathbf{w}(t^*)]\right)\Delta\varepsilon}{\|\Delta\mathbf{c}\|_2}+\Delta\hat{\mathbf{c}}^T\left(\nabla^2F[\mathbf{w}(t^*)]\right)\Delta\varepsilon\Bigg\}=0.$$

Differentiating, we get

$$\frac{1}{2}\Delta\hat{\mathbf{c}}^T\left(\nabla^2F[\mathbf{w}(t^*)]\right)\Delta\hat{\mathbf{c}}=\frac{1}{2}\frac{\Delta\varepsilon^T\left(\nabla^2F[\mathbf{w}(t^*)]\right)\Delta\varepsilon}{\|\Delta\mathbf{c}\|_2^2}.$$
$$\Rightarrow\mathscr{Q}[\Delta\mathbf{c}]=\frac{\|\Delta\varepsilon\|_2^2}{\|\Delta\mathbf{c}\|_2^2}\mathscr{Q}[\Delta\varepsilon],$$

from which (9) follows.

1.2 Positivity of the numerator and denominator in equation (9)

Equation (9) of the main text asserts that

$$\frac{\|\Delta\mathbf{c}\|_2^2}{\|\Delta\varepsilon\|_2^2}=\frac{\mathscr{Q}[\Delta\varepsilon]}{\mathscr{Q}[\Delta\mathbf{c}]}$$

holds as long as both the numerator and denominator of the RHS are positive. Here we describe sufficient conditions for positivity.

The inequality $\nabla^2F[\mathbf{w}]\succeq 0$ must hold in some neighbourhood of any minimum $\mathbf{w}^*$. Recall that we referred to such a neighbourhood as a highly trained state of the network in the main text. In such a state, our assertion follows immediately, as $\mathscr{Q}[\mathbf{v}]:=\frac{1}{\|\mathbf{v}\|_2^2}\mathbf{v}^T(\nabla^2F[\mathbf{w}])\mathbf{v}\geq 0$, for any vector $\mathbf{v}$. Therefore $\mathscr{Q}[\Delta\varepsilon]\geq 0$ and $\mathscr{Q}[\Delta\mathbf{c}]\geq 0$.

We now consider a partially trained network state, which we defined in the main text as any $\mathbf{w}$ satisfying $Tr(\nabla^2F)\geq 0$. Note that

$$F[\mathbf{w}+\Delta\varepsilon]=F[\mathbf{w}]+\nabla F[\mathbf{w}]^T\Delta\varepsilon+\frac{1}{2}\Delta\varepsilon^T\nabla F[\mathbf{w}]\Delta\varepsilon+\mathscr{O}(\|\Delta\varepsilon\|_2^3).$$

We assumed in the main text (equation (1a)), that $\Delta\varepsilon$ is uncorrelated with the gradient $\nabla F[\mathbf{w}]$ in expectation, since $\Delta\varepsilon$ is realised by memory-independent processes. Similarly we can assume that $\Delta\varepsilon$ is unbiased in how it projects onto the eigenvectors of $\nabla^2F[\mathbf{w}]$. In other words,

$$\mathbb{E}[\hat{v}_i^T\Delta\varepsilon]=\mathbb{E}[\hat{v}_j^T\Delta\varepsilon],$$

for any normalised eigenvectors v_i, v_j of $\nabla^2F[\mathbf{w}]$. In expectation, we can therefore simplify to

$$\mathbb{E}[F[\mathbf{w}+\Delta\varepsilon]]=F[\mathbf{w}]+\mathbb{E}\left[\nabla F[\mathbf{w}]^T\Delta\varepsilon+\frac{1}{2}\Delta\varepsilon^T\nabla F[\mathbf{w}]\Delta\varepsilon\right]+\mathscr{O}(\|\Delta\varepsilon\|_2^3).$$
$$=0+\|\Delta\varepsilon\|_2^2\frac{Tr(\nabla^2F[\mathbf{w}])}{N}+\mathscr{O}(\|\Delta\varepsilon\|_2^3),$$

where N is the dimensionality of the vector $\mathbf{w}$. So a partially trained network is one for which small, memory-independent weight fluctuations (such as $\Delta\varepsilon$, or white noise) are expected to decrease memory quality.

Now recall that $\mathscr{Q}[\Delta\varepsilon]=\frac{1}{\|\Delta\varepsilon\|_2^2}\Delta\varepsilon^T\nabla^2F[\mathbf{w}]\Delta\varepsilon$. So we have

$$\mathbb{E}\left[\mathscr{Q}[\Delta\varepsilon]\right]=\frac{Tr(\nabla^2F[\mathbf{w}])}{N}>0,$$

where the positivity constraint comes from being in a partially trained network.

We now consider why $\mathscr{Q}[\Delta\mathbf{c}]$ should be generically positive in a partially trained network. Suppose $\mathscr{Q}[\Delta\mathbf{c}] < 0$ holds. We can rewrite this as $\Delta\mathbf{c}^T\nabla^2 F[\mathbf{w}]\Delta\mathbf{c} \leq 0$. In this case, maintaining the same systematic plasticity direction $\Delta\mathbf{c}$ over the time interval $[t^*+\Delta t, t^*+2\Delta t]$ would result in increased improvement in loss, as

$$\nabla F[\mathbf{w}+\Delta\mathbf{c}]^T\Delta\mathbf{c} = \nabla F[\mathbf{w}]^T\Delta\mathbf{c} + \Delta\mathbf{c}^T\nabla^2 F[\mathbf{w}]\Delta\mathbf{c} + \mathscr{O}(\|\Delta\mathbf{c}\|_2^2).$$

Effectively, memory improvement due to systematic plasticity $\Delta\mathbf{c}$ would be in an 'accelerating' direction, and maintaining the same direction $\Delta\mathbf{c}$ of systematic plasticity would lead to ever faster learning. However, by assumption, we are in a regime of steady state task performance, where

$$\mathbb{E}[F(t^*+2\Delta t) - F(t^*+\Delta t)] = \mathbb{E}[F(t^*+\Delta t) - F(t^*)] = 0.$$

1.3 Optimal plasticity ratios in specific learning rules

1.3.1 Noise-free learning rules (first-order)

Let us first consider the case where $\Delta\mathbf{c}$ can be computed with perfect access to the gradient $\nabla F[\mathbf{w}]$, but without access to $\nabla^2 F[\mathbf{w}]$. Such a $\Delta\mathbf{c}$ is known as a first-order learning rule, as it has access only to the first derivative of F Polyak (1987). Imperfect access is considered subsequently. In this case, the optimal direction of systematic plasticity is

$$\Delta\mathbf{c} \propto -\nabla F[\mathbf{w}].$$

In other words, $\Delta\mathbf{c}$ would implement perfect gradient descent on $F[\mathbf{w}]$. The condition (10) for synaptic fluctuations to outcompete reconsolidation plasticity evaluates to

$$\mathscr{Q}[\nabla F[\mathbf{w}]] \geq \mathscr{Q}[\Delta\boldsymbol{\varepsilon}].$$

To what extent can we quantify $\mathscr{Q}[\nabla F[\mathbf{w}]]$? First let us relate the gradient and Hessian of $F[\mathbf{w}]$. Let $\mathbf{w}^*$ be an optimal state of the network (i.e. one where F is minimised). Let us parameterise the straight line connecting $\mathbf{w}$ with $\mathbf{w}^*$:

$$\gamma(s) = s\mathbf{w}^* + (1-s)\mathbf{w}, \quad s \in [0,1].$$

Then

$$\nabla F[\mathbf{w}] = (\mathbf{w}-\mathbf{w}^*)^T M, \text{ where}$$
$$M = \int_0^1 \nabla^2 F[\gamma(s)]\, ds.$$

This gives

$$\mathscr{Q}[\nabla F[\mathbf{w}]] = \frac{(\mathbf{w}-\mathbf{w}^*)^T M^T \nabla^2 F[\mathbf{w}] M(\mathbf{w}-\mathbf{w}^*)}{(\mathbf{w}-\mathbf{w}^*)^T M^T M(\mathbf{w}-\mathbf{w}^*)}.$$

First let us rewrite

$$(\mathbf{w}-\mathbf{w}^*) := \sum_i^N c_i v_i,$$
$$M(\mathbf{w}-\mathbf{w}^*) := \sum_i^N d_i v_i$$

where (λ_i, v_i) is the i^{th} eigenvalue/eigenvector pair of $\nabla^2 F$ (sorted in ascending order of λ_i), and c_i, d_i are some scalar weights. Now

$$\mathscr{Q}[\nabla F[\mathbf{w}]] = \frac{\sum_{i=1}^N d_i^2 \lambda_i}{\sum_{i=1}^N d_i^2}. \tag{13}$$

The value of $\mathscr{Q}[\nabla F[\mathbf{w}]]$ now depends upon the distribution of mass of the sequence $\{d_i\}$. If later elements of the sequence are larger (i.e. $M(\mathbf{w}=\mathbf{w}^*)$ projects more highly onto eigenvectors of $\nabla^2 F[\mathbf{w}]$) with large eigenvalue), then $\mathscr{Q}[\nabla F[\mathbf{w}]]$ becomes larger, and the optimal magnitude of reconsolidation plasticity decreases, relative to the magnitude of synaptic fluctuations. The opposite is true if earlier elements of the sequence are larger.

Guaranteed bounds on the value of equation (13) are vacuous. If we do not restrict M, then we can tailor the sequence $\{d_i\}$ as we like, and we end up with $\lambda_1 \leq \mathscr{Q}[\nabla F[\mathbf{w}]] \leq \lambda_N$. However, pragmatic bounds are much tighter. Let us now consider two plausibly extremal cases.

First consider the simplest case of a network that linearly transforms its outputs, and which has a quadratic loss function $F[\mathbf{w}]$. In this case $\nabla^2 F$ is a constant, (independent of $\mathbf{w}$), positive-semidefinite matrix, and $M = \nabla^2 F$. This means that

$$d_i = c_i \lambda_i v_i$$
$$\mathscr{Q}[\nabla F[\mathbf{w}]] = \frac{\sum_{i=1}^N c_i^2 \lambda_i^3}{\sum_{i=1}^N c_i^2 \lambda_i^2}.$$

Condition (10) then becomes

$$\mathscr{Q}[\nabla F[\mathbf{w}]] \geq \mathscr{Q}[\Delta\varepsilon] \quad \Leftrightarrow \quad \frac{\sum_{i=1}^N c_i^2 \lambda_i^3}{\sum_{i=1}^N c_i^2 \lambda_i^2} \geq \frac{\sum_{i=1}^N \lambda_i}{N}. \tag{14}$$

A conservative sufficient condition for (14), using Chebyshev's summation inequality, is that

$$c_i^2 \lambda_i^2 \geq c_{i-1}^2 \lambda_{i-1}^2, \text{ for all } i \in \{1,\ldots,N\}. \tag{15}$$

Under what conditions would a plausible reconsolidation mechanism choose to 'outcompete' synaptic fluctuations, in this linear example? For $\mathscr{Q}[\nabla F[\mathbf{w}]] < \mathscr{Q}[\Delta\varepsilon]$ to even hold, (15) would have to be broken, and significantly so due to conservatism in the inequality. In other words, $\mathbf{w}-\mathbf{w}^*$ must project quite biasedly onto the eigenvectors of $\nabla^2 F$ with smaller-than-average eigenvalue. If the discrepancy between $\mathbf{w}$ and $\mathbf{w}^*$ were caused by fluctuations (which are independent of $\nabla^2 F$), then this would not be the case, in expectation. Even if this were the case, the reconsolidation mechanism would have to know about the described bias. This requires knowledge of both $\mathbf{w}^*$ and $\nabla^2 F$, and is thus implausible.

Now let us consider the case of a generic nonlinear network. At one extreme, if $\|\mathbf{w}-\mathbf{w}^*\|_2$ is small, then $M \approx \nabla^2 F[\mathbf{w}]$, and the discussion of the linear case is valid. This corresponds to the case where steady state error is close to the minimum achievable by the network. As $\|\mathbf{w}-\mathbf{w}^*\|_2$ increases (i.e. steady state error gets worse), the correspondence between M and $\nabla^2 F[\mathbf{w}]$ will likely decrease. Thus the optimal magnitude of reconsolidation plasticity, relative to the level of synaptic fluctuations, will rise.

We could consider another 'extreme' case in which M and $\nabla^2 F[\mathbf{w}]$ were completely independent of each other. In this case,

$$d_i^2 \approx \frac{1}{N}\sum_{i=1}^N d_i^2. \tag{16}$$

In other words, the projection of $M(\mathbf{w}-\mathbf{w}^*)$ onto the different eigenvectors of $\nabla^2 F[\mathbf{w}]$ is approximately even. Using (13), this gives

$$\mathscr{Q}[\nabla F[\mathbf{w}]] \approx \frac{\sum_{i=1}^N \lambda_i}{N} = \mathscr{Q}[\Delta\varepsilon].$$

In summary, we have two plausible extremes. One occurs where $M = \nabla^2 F[\mathbf{w}]$, and another occurs where M is completely independent of $\nabla^2 F[\mathbf{w}]$. In either case, $\mathscr{Q}[\nabla F[\mathbf{w}]] \geq \mathscr{Q}[\Delta\varepsilon]$, and so the magnitude of synaptic fluctuations should optimally outcompete/equal the magnitude of reconsolidation plasticity. Of course, there might be particular values of $\mathbf{w}$ where the correspondence between M and $\nabla^2 F[\mathbf{w}]$ is 'worse' than chance. In other words, eigenvectors of M with large eigenvalue preferentially project

onto eigenvectors of $\nabla^2 F[\mathbf{w}]$ with small eigenvalue. In such cases, we would have $\mathscr{Q}[\nabla F[\mathbf{w}]] \leq \mathscr{Q}[\Delta\varepsilon]$. However, we find it implausible that a reconsolidation mechanism would be able to gain sufficient information on M to determine this at particular points in time, and thereby increase its plasticity magnitude.

1.3.2 Noise-free learning rules (second order)

Let us now suppose that $\Delta\mathbf{c}$ can be computed with perfect access to both $\nabla F[\mathbf{w}]$ and $\nabla^2 F[\mathbf{w}]$. In this case the reconsolidation mechanism would optimally apply plasticity in the direction of the Newton step: we would have

$$\nabla^2 F[\mathbf{w}]\Delta\mathbf{c} = -\nabla F[\mathbf{w}].$$

Note that the Newton step is often conceptualised as a weighted form of gradient descent, where movement in weight space is biased towards direction of lower curvature. Thus we would expect $\mathscr{Q}[\Delta\mathbf{c}]$ to be smaller, and the optimal proportion of reconsolidation plasticity to be larger. This is indeed the case. For mathematical tractability, we will restrict our discussion to the case in which $\nabla^2 F[\mathbf{w}] \succ 0$, and $M \succ 0$. This would hold if $F[\mathbf{w}]$ were convex, or if $\mathbf{w}$ were sufficiently close to a local minimum $\mathbf{w}^*$. In this case we can rewrite

$$\Delta\mathbf{c} = -\nabla^2 F[\mathbf{w}]^{-1}\nabla F[\mathbf{w}],$$

which gives

$$\mathscr{Q}[\Delta\mathbf{c}] = \frac{\nabla F[\mathbf{w}]^T \left(\nabla^2 F[\mathbf{w}]\right)^{-1} \nabla F[\mathbf{w}]}{\nabla F[\mathbf{w}]^T \left(\nabla^2 F[\mathbf{w}]\right)^{-2} \nabla F[\mathbf{w}]} \tag{17a}$$

$$= \frac{(\mathbf{w}-\mathbf{w}^*)M\left(\nabla^2 F[\mathbf{w}]\right)^{-1} M(\mathbf{w}-\mathbf{w}^*)}{(\mathbf{w}-\mathbf{w}^*)M\left(\nabla^2 F[\mathbf{w}]\right)^{-2} M(\mathbf{w}-\mathbf{w}^*)} \tag{17b}$$

$$= \frac{\sum_{i=1}^N d_i^2 \lambda_i^{-1}}{\sum_{i=1}^N d_i^2 \lambda_i^{-2}}. \tag{17c}$$

Once again, we first consider the case of a linear network with quadratic loss function, and hence with constant Hessian $\nabla^2 F$. This gives $M = \nabla^2 F$, and

$$\mathscr{Q}[\Delta\mathbf{c}] = \frac{(\mathbf{w}-\mathbf{w}^*)\nabla^2 F[\mathbf{w}](\mathbf{w}-\mathbf{w}^*)}{\|\mathbf{w}-\mathbf{w}^*\|_2^2}$$

$$= \frac{\sum_{i=1}^N c_i^2 \lambda_i}{\sum_{i=1}^N c_i^2}.$$

We again assume that the reconsolidation mechanism does not have knowledge of the relative projections of $\mathbf{w}-\mathbf{w}^*$ onto the different eigenvectors of $\nabla^2 F$, which requires knowledge of $\mathbf{w}^*$. Without such information, we can use an analogous argument to that preceding (16) to argue that the approximation $c_i^2 \approx \frac{1}{N}\sum_{i=1}^N c_i^2$ is reasonable. This gives $\mathscr{Q}[\Delta\mathbf{c}] \approx \mathscr{Q}[\Delta\varepsilon]$.

Note that the Newton step, in the linear-quadratic case just considered, corresponds to a direction $\mathbf{w}^* - \mathbf{w}$, i.e. a direct path to a local minimum. So we could consider a systematic plasticity mechanism implementing the Newton step as one directly undoing synaptic changes caused by the intrinsic flucatuations $\Delta\varepsilon$.

We now consider the case of a nonlinear network. As before, if $\|\mathbf{w}-\mathbf{w}^*\|_2$ is small, then we have $M \approx \nabla^2 F[\mathbf{w}]$, and the arguments of the linear network hold. As $\|\mathbf{w}-\mathbf{w}^*\|_2$ increases, the correspondence between M and $\nabla^2 F$ will decrease. We again consider the plausible extreme where M is completely uncorrelated with $\nabla^2 F[\mathbf{w}]$, and so the approximation (16) holds. In this case, equation (17c) can be simplified to give

$$\mathscr{Q}[\Delta\mathbf{c}] \approx \frac{\sum_{i=1}^N \lambda_i^{-1}}{\sum_{i=1}^N \lambda_i^{-2}}.$$

We assumed that $\nabla^2 F[\mathbf{w}] \succ 0$. Therefore, all eigenvalues are positive. This allows us to use Chebyshev's summation inequality to arrive at

$$\frac{\sum_{i=1}^{N} \lambda_i^{-1}}{\sum_{i=1}^{N} \lambda_i^{-2}} \leq \frac{\sum_{i=1}^{N} \lambda_i}{N} = \mathscr{Q}[\Delta\varepsilon].$$

So as $\|\mathbf{w} - \mathbf{w}^*\|_2$ increases, the magnitude of reconsolidation plasticity will optimally outcompete that of synaptic fluctuations.

1.3.3 Imperfect learning rules

The previous section applied in the implausible case where a reconsolidation mechanism had perfect access to $\nabla F[\mathbf{w}]$ and/or $\nabla^2 F[\mathbf{w}]$. Recall, from the discussion surrounding equation (4), that at least some information on $\nabla F[\mathbf{w}]$ is required. What if $\Delta\mathbf{c}$ contains a mean-zero noise term, corresponding to imperfect access to these quantities? We will now show how such noise pushes $\mathscr{Q}[\Delta\mathbf{c}]$ towards equality with $\mathscr{Q}[\Delta\varepsilon]$, and thus pushes the optimal magnitude of reconsolidation plasticity towards the magnitude of synaptic fluctuations. Let us use the model

$$\Delta\mathbf{c} = \tilde{\Delta\mathbf{c}} + \nu, \tag{18}$$

where ν is some mean-zero term, and $\tilde{\Delta\mathbf{c}}$ is the ideal output of the reconsolidation mechanism, assuming perfect access to the derivatives of $F[\mathbf{w}]$. Here ν represents the portion of systematic plasticity attributable to systematic error in the algorithm, due to imperfect information on $F[\mathbf{w}]$. This could arise due to imperfect sensory information or limited communication between synapses. We can therefore assume, as for $\Delta\varepsilon$, that it does not contain information on $\nabla^2 F[\mathbf{w}]$. We therefore get

$$\mathscr{Q}[\nu] \approx \frac{Tr(\nabla^2 F[\mathbf{w}])}{N},$$

analogously to (11). Now the operator $\mathscr{Q}$ satisfies

$$\mathscr{Q}[\Delta\mathbf{c}] = \mathscr{Q}[\tilde{\Delta\mathbf{c}}]\left(1 + \frac{\|\nu\|_2^2}{\|\tilde{\Delta\mathbf{c}}\|_2^2}\right)^{-1} + \mathscr{Q}[\nu]\left(1 + \frac{\|\tilde{\Delta\mathbf{c}}\|_2^2}{\|\nu\|_2^2}\right)^{-1}.$$

So depending upon the relative magnitudes of $\tilde{\Delta\mathbf{c}}$ and ν, $\mathscr{Q}[\Delta\mathbf{c}]$ interpolates between $\mathscr{Q}[\tilde{\Delta\mathbf{c}}]$ and $\mathscr{Q}[\nu]$. In particular, as the crudeness of the learning rule (i.e. the ratio $\frac{\|\nu\|}{\|\tilde{\Delta\mathbf{c}}\|}$) grows, $\mathscr{Q}[\Delta\mathbf{c}]$ approaches equality (from below) with $\mathscr{Q}[\nu]$, and thus $\mathscr{Q}[\Delta\varepsilon]$, completing our argument.

REFERENCES

Acker, D., Paradis, S., and Miller, P. (2018). Stable memory and computation in randomly rewiring neural networks. *bioRxiv*, page 367011.

Attardo, A., Fitzgerald, J. E., and Schnitzer, M. J. (2015). Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature*, 523(7562):592–596.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48.

Chambers, A. R. and Rumpel, S. (2017). A stable brain from unstable components: Emerging concepts and implications for neural computation. *Neuroscience*, 357:172–184.

Clopath, C., Bonhoeffer, T., Hübener, M., and Rose, T. (2017). Variance and invariance of neuronal long-term representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1715):20160161.

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.

Fauth, M. J. and van Rossum, M. C. (2019). Self-organized reactivation maintains and reinforces memories despite synaptic turnover. *eLife*, 8:e43717.

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

Gerstner, W., Kempter, R., Van Hemmen, J. L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76–78.

Hayashi-Takagi, A., Yagishita, S., Nakamura, M., Shirai, F., Wu, Y. I., Loshbaugh, A. L., Kuhlman, B., Hahn, K. M., and Kasai, H. (2015). Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, 525(7569):333–338.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. J. Wiley; Chapman & Hall.

Holtmaat, A. J. G. D., 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(2):279–291.

Kappel, D., Legenstein, R., Habenschuss, S., Hsieh, M., and Maass, W. (2018). A Dynamic Connectome Supports the Emergence of Stable Computational Function of Neural Circuits through Reward-Based Learning. *eNeuro*, 5(2).

Kasai, H., Matsuzaki, M., Noguchi, J., Yasumatsu, N., and Nakahara, H. (2003). Structure–stability–function relationships of dendritic spines. *Trends in Neurosciences*, 26(7):360–368.

Lai, C. S. W., Franke, T. F., and Gan, W.-B. (2012). Opposite effects of fear conditioning and extinction on dendritic spine remodelling. *Nature*, 483(7387):87–91.

Levin, E., Tishby, N., and Solla, S. A. (1990). A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10):1568–1574.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):13276.

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. *Journal of Neuroscience*, 31(26):9481–9488.

Loewenstein, Y., Yanover, U., and Rumpel, S. (2015). Predicting the Dynamics of Network Connectivity in the Neocortex. *Journal of Neuroscience*, 35(36):12535–12544.

Mazzoni, P., Andersen, R. A., and Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, 88(10):4433–4437.

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.

Moczulska, K. E., Tinter-Thiede, J., Peter, M., Ushakova, L., Wernle, T., Bathellier, B., and Rumpel, S. (2013). Dynamics of dendritic spines in the mouse auditory cortex during memory formation and memory recall. *Proceedings of the National Academy of Sciences*, 110(45):18315–18320.

Mongillo, G., Rumpel, S., and Loewenstein, Y. (2018). Inhibitory connectivity defines the realm of excitatory plasticity. *Nature Neuroscience*, 21(10):1463.

Nagaoka, A., Takehara, H., Hayashi-Takagi, A., Noguchi, J., Ishii, K., Shirai, F., Yagishita, S., Akagi, T., Ichiki, T., and Kasai, H. (2016). Abnormal intrinsic dynamics of dendritic spines in a fragile X syndrome mouse model *in vivo*. *Scientific Reports*, 6:26651.

Pfeiffer, T., Poll, S., Bancelin, S., Angibaud, J., Inavalli, V. K., Keppler, K., Mittag, M., Fuhrmann, M., and Nägerl, U. V. (2018). Chronic 2P-STED imaging reveals high turnover of dendritic spines in the hippocampus in vivo. *eLife*, 7:e34700.

Polyak, B. T. (1987). Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1.

Quinn, D. P., Kolar, A., Harris, S. A., Wigerius, M., Fawcett, J. P., and Krueger, S. R. (2019). The Stability of Glutamatergic Synapses Is Independent of Activity Level, but Predicted by Synapse Size. *Frontiers in Cellular Neuroscience*, 13.

Raman, D. V., Rotondo, A. P., and O'Leary, T. (2019). Fundamental bounds on learning performance in neural circuits. *Proceedings of the National Academy of Sciences*, 116(21):10537–10546.

Seung, H. S. (2003). Learning in Spiking Neural Networks by Reinforcement of Stochastic Synaptic Transmission. *Neuron*, 40(6):1063–1073.

Seung, H. S., Sompolinsky, H., and Tishby, N. (1992). Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056.

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.

Susman, L., Brenner, N., and Barak, O. (2018). Stable memory with unstable synapses. *arXiv:1808.00756 [q-bio]*.

Sussillo, D. and Abbott, L. F. (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron*, 63(4):544–557.

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

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008). Principles of Long-Term Dynamics of Dendritic Spines. *Journal of Neuroscience*, 28(50):13592–13608.

Ziv, N. E. and Brenner, N. (2018). Synaptic Tenacity or Lack Thereof: Spontaneous Remodeling of Synapses. *Trends in Neurosciences*, 41(2):89–99.

Zuo, Y., Lin, A., Chang, P., and Gan, W.-B. (2005). Development of Long-Term Dendritic Spine Stability in Diverse Regions of Cerebral Cortex. *Neuron*, 46(2):181–189.