

Lyapunov theory demonstrating a fundamental limit on the speed of systems consolidation

Alireza Alemi^{1,*}, Emre R. F. Aksay² and Mark S. Goldman^{1,3,†}

¹*Center for Neuroscience, and Department of Neurobiology, Physiology, and Behavior, University of California, Davis, Davis, California 95616, USA*

²*Institute for Computational Biomedicine and Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York 10021, USA*

³*Department of Ophthalmology and Vision Science, University of California, Davis, Davis, California 95616, USA*



(Received 5 February 2024; accepted 2 May 2025; published 21 May 2025)

This article is part of the Physical Review Research collection titled [Physics of Neuroscience](#).

The nervous system reorganizes memories from an early site to a late site, a commonly observed feature of learning and memory systems known as systems consolidation. Previous work has suggested learning rules by which consolidation may occur. Here, we provide conditions under which such rules are guaranteed to lead to stable convergence of learning and consolidation. We use the theory of Lyapunov functions, which enforces stability by requiring learning rules to decrease an energy-like (Lyapunov) function. We present the theory in the context of a simple circuit architecture motivated by classic models of cerebellum-mediated learning and consolidation. Stability is only guaranteed if the learning rate in the late stage is not faster than the learning rate in the early stage. Further, the slower the learning rate at the late stage, the larger the perturbation the system can tolerate with a guarantee of stability. We provide intuition for this result by mapping a simple example consolidation model to a damped driven oscillator system and showing that the ratio of early- to late-stage learning rates in the consolidation model can be directly identified with the oscillator's damping ratio. We then apply the theory to modeling the tuning by the cerebellum of a well-characterized analog short-term memory system, the oculomotor neural integrator, and find similar stability conditions. This work suggests the power of the Lyapunov approach to provide constraints on nervous system function.

DOI: [10.1103/PhysRevResearch.7.023174](https://doi.org/10.1103/PhysRevResearch.7.023174)

I. INTRODUCTION

Systems consolidation is the process of transferring learned memories from an early-stage site to a late-stage site [1–3] and has been suggested theoretically to enhance the ability of memory systems to simultaneously learn new associations while protecting previously learned memories from being overwritten [2,4]. Various forms of memories undergo consolidation in different brain areas. For example, declarative memories initially learned in the hippocampus get transferred to the neocortex [2,5]. Motor memories initially located in the cerebellar cortex [6] or the basal ganglia [7] get transferred out of the early learning site into direct motor pathways. Furthermore, strong evidence suggests that fear-based memories initially learned in the amygdala later get transferred to a different site [3,8]. Understanding how neural signals and learning rules orchestrate a successful memory transfer requires guiding principles to shed light on the interactions of

brain areas and their plasticity rules. Here we develop a Lyapunov theory that provides a first-principles account for the speed of consolidation and the robustness of the consolidation process.

Neural circuits underlying learning face a fundamental challenge common to many biological and engineered dynamical systems with adaptively tunable parameters: the concurrent presence of time-varying inputs, states, and parameters may cause the dynamics to become unstable, for example, by growing unboundedly or falling into undesirable oscillatory patterns. In addition, the nervous system abounds with various forms of noise and disturbances [9], which may take the system into undesirable regimes. Thus not only should the final desired solution of learning be stable, but also the overall system should remain stable throughout the process of learning.

The theory of adaptive control systems has been successful in providing essential tools, such as the Lyapunov function formalism, for guaranteeing the stability of learning systems [10]. The concept of a Lyapunov function has been used for quantifying the stability of adaptive recurrent neural networks [11–13] as well as for building attractor neural networks to model long-term memory [14–16]. Here, we apply the Lyapunov formalism to the problem of guaranteeing stable systems consolidation. Systems consolidation, like many learning processes, contains feedback loops between the training signals and neural dynamics that drive learning,

*Contact author: alemi@ucdavis.edu

†Contact author: msgoldman@ucdavis.edu

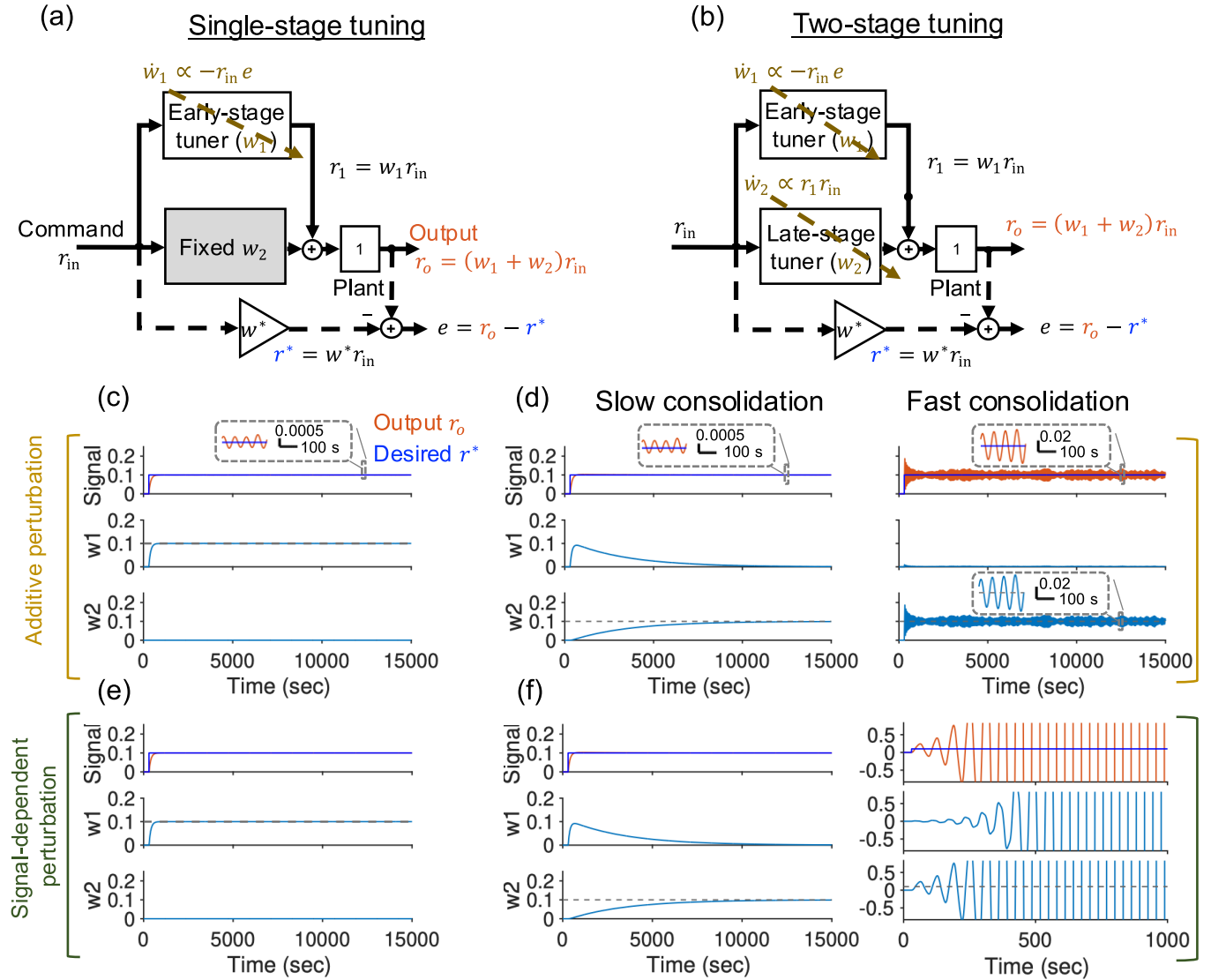


FIG. 1. A toy tracking model demonstrating instability in systems consolidation. (a) Top: single-stage model. The early-stage parameter, w_1 , is directly tuned by the error signal $e(t)$, whereas the late-stage parameter, w_2 , is fixed. (b) Two-stage tuning model. The early stage is as in the single-stage model. The late-stage weight w_2 is tuned using the output of the early stage as a secondary teaching signal. (c) Simulation of a single-stage model showing that the model successfully converges to the desired output r^* and the desired, tuned weight w_1 (dashed line) in the case of additive perturbation. (d) In the two-stage model with additive perturbation, when the consolidation process is slow enough (left panel), the model dynamics successfully converge and tune the weight w_2 to its desired value (dashed line). However, if the consolidation process becomes too fast (right panels), the system can show instability. [(e) and (f)] Similar to (c) and (d) but for a signal-dependent perturbation for which fast consolidation leads to unbounded growth. Note the abbreviated time scale of the right panel in (f), chosen for better visibility of the unbounded growth. See Appendix A for simulation details.

and the weight changes that drive system dynamics. Such feedback can make learning prone to instability. This may be further exacerbated by the fact that, in biological systems, many synapses do not have direct access to the ground-truth performance error and must therefore learn from indirect error signals.

We place our theory in the context of a simple circuit architecture for systems consolidation consisting of an early-stage and a late-stage learner. Critically, the late stage, unlike the early stage, lacks direct access to an error-correcting teaching signal (Fig. 1). In this architecture, the early-stage activity trains the weight(s) of the late stage, such that the memory is eventually transferred from the early to the late stage. This

architecture and set of learning rules correspond to classic theories of learning and systems consolidation in the cerebellum. Learning at the early-stage site of plasticity corresponds to the classic Marr-Albus-Ito theory of cerebellar learning [17,18] in which an error signal conveyed by climbing fibers trains the cerebellar output conveyed by Purkinje cells. These Purkinje cell firing rates then serve as a secondary teaching signal for the late-stage site of plasticity located in the cerebellar output nuclei. Applying the formalism of Lyapunov stability theory, we find that, to guarantee stability, plasticity at the late-stage site must not be updated faster than plasticity at the early-stage site. For a given speed of plasticity at the early-stage site, the slower the tuning in the late stage, the more robust the learning

process is against noise in the primary teaching signal. We then show that the theory can be applied to the learning and tuning of a simple recurrent short-term memory network, the oculomotor neural integrator, that is responsible for stably maintaining the position of the eyes in their orbits [19].

II. RESULTS

A. Toy model for systems consolidation

We first consider a simple toy model of systems consolidation motivated by the basic circuitry thought to be involved in systems consolidation of cerebellum-mediated learning. The task is to track a given time-varying input command $r_{\text{in}}(t)$ and to generate a desired output $r_o^*(t) = w^* r_{\text{in}}(t)$ where the desired input-to-output gain is denoted as w^* . The model, shown in Fig. 1, has an architecture with two pathways: a direct pathway with gain w_2 and an indirect pathway with gain w_1 that provides a learned, online correction to the output of the direct pathway. The output of the model can be written as

$$r_o(t) = (w_1 + w_2)r_{\text{in}}(t) \quad (1)$$

where $r_o(t)$ is the output. The goal of learning is two-fold: (1) to reduce the gain error $\tilde{W} = w_1 + w_2 - w^*$ to zero, and (2) to consolidate the weight changes into the late-stage weight such that asymptotically $w_1 = 0$ and $w_2 = w^*$. The indirect pathway receives the information about the error signal and constitutes the early site of plasticity. The direct pathway constitutes the late site of plasticity and needs to be tuned based on the information received from the early stage, i.e., $r_1 = w_1 r_{\text{in}}$.

The learning rule in the early stage is a supervised delta-like correlational rule proportional to the product of the input and the teaching signal [20,21].

$$\dot{w}_1 = -\eta_1 r_{\text{in}}(t)(e(t) + \xi(t)) \quad (2)$$

where the teaching error signal $e(t)$ is the tracking error $e(t) = r_o(t) - r_o^*(t) = (w_1 + w_2 - w^*)r_{\text{in}}(t)$, η_1 is the learning rate of the early-stage tuner, and $\xi(t)$ is a perturbation to the error signal. We assume that the maximal absolute value of the perturbation at any given time point is signal-dependent [22,23], i.e., the strength of the perturbation depends on the magnitude of error signal $e(t)$, and is given by $|\xi(t)| \leq \mu|e(t)|$ where $\mu > 0$ is a constant parameter of the model that defines the regime of possible perturbations. A bounded constant noise component could be added, as illustrated in Figs. 1(c) and 1(d), but is smaller for large error signals where the biggest issues are expected to arise (such as the unbounded growth in Fig. 1(f)); thus, it is omitted for simplicity in the analytical calculations. Alternatively, such a constant noise term could be offset by adding a decay term in Eq. (2) [24]. The learning rule in the late stage is a heterosynaptic correlational rule between the online corrective signal $r_1(t)$ provided by the early stage, and the direct input to the late stage, $r_{\text{in}}(t)$ [25]:

$$\dot{w}_2 = \eta_2 r_{\text{in}}(t)r_1(t), \quad (3)$$

where η_2 is the learning rate.

We seek a general framework to characterize the conditions for stability and convergence of this two-stage learning system. More specifically, we seek criteria for tuning the system to avoid undesired behaviors of the system, such as

oscillation or unbounded growth. By stability, we intuitively mean that, as time goes to infinity, the system is well-behaved; for example, it should reach its goal and stay close to it. The mathematical definition of Lyapunov stability and its various refined notions are provided in Appendix B.

B. Stability of the single-stage toy model

To illustrate the Lyapunov stability approach and motivate the problem of instability in systems consolidation, we first consider a one-stage model [Fig. 1(a)] in which learning occurs only at the early stage of the circuit (i.e., we set $\frac{dw_2}{dt} = 0$). In this case, so long as the magnitude of the perturbation $\xi(t)$ does not exceed the magnitude of the error signal $e(t)$, the learning is guaranteed to stably converge. This is illustrated for the task of learning to track a step-like input in Figs. 1(a), 1(c), and 1(e) and proven for the more general case in the following paragraphs.

The basic idea behind Lyapunov's direct approach to stability is based on constructing a scalar function and showing that all trajectories of the dynamics of the system decrease this scalar function, guaranteeing that the system safely and stably reaches the fixed point of the dynamics. This scalar function, known as a Lyapunov function, can be considered a generalization of the concept of energy in classical mechanics. Finding one such function is enough to prove the stability of the system. To provide intuition about the Lyapunov function formalism, we first apply it to the simple single-stage tuning model given by Eq. (2).

To prove the uniform global stability of the model, we use the Lyapunov theorem 1 for nonautonomous systems, and to prove asymptotic stability, we use the Lyapunov-like lemma 1 (Appendix B). To prove the uniform global stability, we need to find an appropriate Lyapunov function candidate L such that 1) L is positive definite, 2) the time derivative of L is $\dot{L} \leq 0$, 3) L is decrescent, and 4) L at $t = 0$ is radially unbounded. We propose the following scalar function of the error in the gain $\tilde{W}$:

$$L_1 = \frac{1}{2}\tilde{W}^2, \quad (4)$$

where $\tilde{W} = w_1 + w_2 - w^*$. L_1 is positive definite and radially unbounded by inspection. To show that $\dot{L}_1 \leq 0$, we write the learning dynamics as $\dot{w}_1 = -\eta_1 \tilde{W} r_{\text{in}}^2 - \eta_1 r_{\text{in}} \xi$. Using the learning rule to compute the time derivative yields $\dot{L}_1 = -\eta_1 \tilde{W}^2 r_{\text{in}}^2 - \eta_1 \tilde{W} r_{\text{in}} \xi$. With $|\xi| \leq \mu|e|$, one obtains

$$\begin{aligned} \dot{L}_1 &\leq -\eta_1 \tilde{W}^2 r_{\text{in}}^2 + \eta_1 |\tilde{W} r_{\text{in}}| |\xi| \\ &\leq -\eta_1 \tilde{W}^2 r_{\text{in}}^2 + \eta_1 |\tilde{W} r_{\text{in}}| \mu |\tilde{W} r_{\text{in}}| \\ &= -\eta_1 r_{\text{in}}^2 (1 - \mu) \tilde{W}^2. \end{aligned} \quad (5)$$

$\eta_1 > 0$ by definition, and as long as $\mu \leq 1$, we have $\dot{L}_1 \leq 0$. If $r_{\text{in}} = 0$, we are in a trivial case where $\dot{L}_1 = 0$ and $\dot{w}_1 = 0$ and the system is not changing at all. We assume $r_{\text{in}} \neq 0$ throughout most of the learning period. Therefore, the equilibrium is globally stable. This stability guarantees that w_1 and $\tilde{W}$ are bounded. Since L_1 does not explicitly depend on time and is positive definite, it is decrescent; therefore, the equilibrium is uniformly globally stable. To find conditions for guaranteeing asymptotic stability, i.e., as $t \rightarrow \infty$, $w_1 \rightarrow w^*$, from Lemma 1 what is left to show is that $\dot{L}_1$ is uniformly continuous

in time. A practical way of showing uniform continuity is to show that $\dot{L}_1$ is bounded. Given that the hyperparameters η_1, η_2 and the gain error $\tilde{W}$ are bounded, and r_{in} and ξ are assumed to be smooth functions of time with bounded derivatives, $\dot{L}_1$ is bounded. ■

C. Lyapunov function theory for the two-stage toy consolidation model

We next consider systems consolidation in the two-stage model [Fig. 1(b)]. When the consolidation process is sufficiently slow, the two-stage learning model successfully converges to solving the tracking task [Figs. 1(d) and 1(f), left]. However, when the consolidation process is too fast, the system can exhibit instability in which a small perturbation can cause large perturbations in the output [Figs. 1(d) and 1(f), right]. Below, we analytically find the conditions for guaranteeing stability (Sec. II C 1) and provide intuition for the source of possible instability in regions without stability guarantee by solving a special case of the system (Sec. II C 2).

1. Theory

To investigate the stability and convergence of the two-stage model, we need to find an appropriate Lyapunov function candidate. The two learning rules can be rewritten as $\dot{w}_1 = -\eta_1 \tilde{W} r_{in}^2 - \eta_1 r_{in} \xi$ and $\dot{w}_2 = \eta_2 w_1 r_{in}^2$. We choose the following Lyapunov function candidate for the two-stage model:

$$L = \frac{1}{2}(\tilde{W}^2 + \tilde{w}_2^2), \quad (6)$$

where $\tilde{w}_2 = w_2 - w^*$. We refer to the first term in the above as the (squared) gain error and the second term as the (squared) consolidation error. The nullclines and the fixed points of the learning dynamics are shown in the weight space in Fig. 2(a) for $w^* = 1$ in the limit that the amplitude of the perturbation goes to zero, i.e., $\mu \rightarrow 0$. The first term encourages the gain error to go towards zero in a stable manner and stay close to zero, which is the goal of the learning rule for w_1 . The second term aligns with the goal of consolidating the learned memories into w_2 and is achieved when the desired gain w^* is only due to w_2 [Fig. 2(b)].

We now turn to proving uniform global stability and asymptotic stability of the two-stage model. Using $w_1 = \tilde{W} - \tilde{w}_2$, the time derivative of L can be written as $\dot{L} = \tilde{W}\dot{\tilde{W}} + \tilde{w}_2\dot{\tilde{w}_2} = -\eta_1 r_{in}^2((1-\alpha)\tilde{W}^2 + \alpha\tilde{w}_2^2) - \eta_1 \tilde{W} r_{in} \xi$, where $\alpha = \eta_2/\eta_1$. With $|\xi| \leq \mu|e|$ one obtains:

$$\begin{aligned} \dot{L} &= -\eta_1 r_{in}^2((1-\alpha)\tilde{W}^2 + \alpha\tilde{w}_2^2) - \eta_1 \tilde{W} r_{in} \xi \\ &\leq -\eta_1 r_{in}^2((1-\alpha)\tilde{W}^2 + \alpha\tilde{w}_2^2) + \eta_1 |\tilde{W} r_{in}| \mu |\tilde{W} r_{in}| \\ &= -\eta_1 r_{in}^2((1-\alpha-\mu)\tilde{W}^2 + \alpha\tilde{w}_2^2). \end{aligned} \quad (7)$$

The main requirement for Lyapunov stability is to show $\dot{L} \leq 0$, which is achieved when $\alpha \leq 1 - \mu$. As in the single-stage model, we assume $r_{in} \neq 0$ throughout most of the learning period. For the rest of the proof, we need to verify the other conditions of theorem 1. L is bounded from below, $\min(L) = 0$, and L does not explicitly depend on time. Hence, L is positive definite and decrescent. We therefore conclude that the equilibrium is uniformly stable. Since L is the sum of

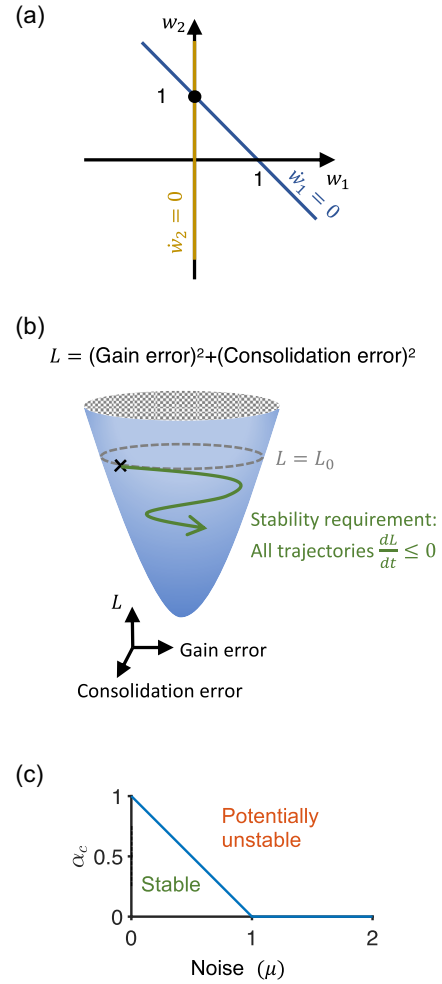


FIG. 2. Lyapunov function theory for stability of the two-stage toy model. (a) In the limit that the perturbation goes to zero, $\mu \rightarrow 0$, the closed-loop learning dynamics has a single fixed point and two nullclines (shown for $w^* = 1$). (b) The Lyapunov function candidate L has two terms: the squared gain error and the squared consolidation error. The most important property in order to have stable convergence in the Lyapunov sense is that the dynamics of the learning rules should avoid going uphill on the Lyapunov function surface. (c) When the ratio of learning rates $\alpha = \eta_2/\eta_1$ is less than a critical value $\alpha_c = 1 - \mu$, the learning is guaranteed to be stable. As the maximum perturbation amplitude reaches $|e|$, i.e., $\mu = 1$, the region of guaranteed stability vanishes.

two quadratic terms, it is radially unbounded by inspection, which guarantees that $\tilde{W}$, $\tilde{w}_2$, w_2 , w_1 are globally bounded. To guarantee asymptotic stability, what is left to show is that $\dot{L}$ is uniformly continuous in time by showing $\dot{L}$ is bounded. Given that the hyperparameters η_1, η_2, α , and the variables $\tilde{W}$, $\tilde{w}_2$, w_1 are bounded, and r_{in} and ξ are assumed to be smooth bounded functions of time with bounded derivatives, $\dot{L}$ is bounded. ■

The key result of the above is that, when the late stage is tuned at a rate not faster than the early stage rate, i.e., $\alpha = \eta_2/\eta_1 \leq \alpha_c = 1 - \mu$, the system provably remains globally stable and is guaranteed to successfully converge [Fig. 2(c)]. Intuitively, in the extreme case where the learning rate of w_1 is much lower than that of w_2 , it is easy to see why

the system may become unstable: w_1 moves infinitesimally slowly towards the goal, but w_2 gets rapidly updated with a secondary teaching signal that is not in the direction of the gradient of the error. This leads to an alteration of the error signal feeding back onto the early (w_1) site of learning, potentially causing the learning process to become unstable. To combat this potential source of stability, the learning rate at w_2 should be slower than that of w_1 to filter out noise and prevent run-away amplification.

When $\alpha > \alpha_c$, the analysis only indicates that the system may become prone to instability, but does not itself say whether the system will become unstable. Such instability can potentially arise if a perturbation brings the system into regions where the derivative of L becomes positive, which for this system occurs when $|\tilde{w}_2| < \sqrt{\frac{\alpha - \alpha_c}{\alpha}} |\tilde{W}|$, showing that the size of this region increases with α .

To check whether we can improve the stability conditions (i.e., find a higher value of α_c) by considering a different relative weighting of the gain and consolidation error terms of L , we consider the Lyapunov function $L_b = \frac{1}{2}(\tilde{W}^2 + b\tilde{w}_2^2)$, where $b > 0$. For simplicity, we work in the regime $\mu \rightarrow 0$, for which $\alpha_c = 1$. Calculating the time derivative $\dot{L}_b = -\eta_1 r_{in}^2 ((1 - \alpha)\tilde{W}^2 + \alpha(1 - b)\tilde{W}\tilde{w}_2 + \alpha b\tilde{w}_2^2)$, we note that $\dot{L}_b$ is again guaranteed to be less than or equal to zero in the whole weight space as long as $\alpha \leq 1$, but not for $\alpha > 1$ (in particular, this is easily seen when $\tilde{w}_2 = 0$). Thus the same fundamental criterion for guaranteeing stability emerges even for different weightings of the two error terms of the Lyapunov function L .

2. Intuition for instability

In the $\alpha > \alpha_c$ regime, our Lyapunov stability analysis only shows that stability is not guaranteed and thus only indicates the potential for instability. Therefore, it is instructive to investigate this regime more closely. Consider the case where a sinusoidal probe perturbation $\xi_p = \epsilon \sin(\omega t)$ with an infinitesimal amplitude ϵ is present in Eq. (2). We examine its effect on the system in the regime that $\mu \rightarrow 0$ while, for simplicity, we set $r_{in} = 1$. To gain intuition about this amplification, we solve the system in the presence of the probe. By eliminating w_1 in the two learning rules, we obtain

$$\ddot{\tilde{w}}_2 + \eta_1 \dot{\tilde{w}}_2 + \eta_1 \eta_2 \tilde{w}_2 = -\eta_1 \eta_2 \epsilon \sin(\omega t). \quad (8)$$

This second-order differential equation is equivalent to forced mass-spring-damper dynamics $m\ddot{x} + c\dot{x} + kx = F$, where x is the object displacement, $m = 1$ is the mass, $c = \eta_1$ is the damping coefficient, $k = \eta_1 \eta_2$ is the spring constant, and $F = -\eta_1 \eta_2 \epsilon \sin(\omega t)$ is the external force. When the damping ratio $\zeta = \frac{c}{2\sqrt{km}} = \frac{1}{2\sqrt{\alpha}} < 1$, we are in the underdamped regime. The steady-state response, which is dominated by $\tilde{w}_{2,ss}(t)$ since w_1 approaches zero in the steady state, then has the following form:

$$\tilde{w}_{2,ss}(t) = -\frac{\epsilon \sin(\omega t + \phi)}{\sqrt{\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\frac{\omega}{\omega_n}\zeta\right)^2}}, \quad (9)$$

where $\phi = \arctan \frac{2\zeta(\frac{\omega}{\omega_n})}{1 - (\frac{\omega}{\omega_n})^2}$ and $\omega_n = \sqrt{\eta_1 \eta_2}$ is the undamped natural frequency. As α increases, the damping ratio ζ decreases, and the amplitude of the $\tilde{w}_{2,ss}(t)$ resonance increases.

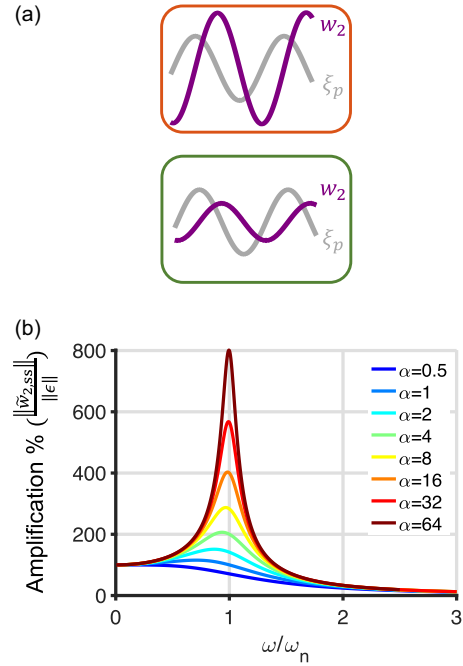


FIG. 3. Amplification of perturbation in the region without stability guarantee in the two-stage toy model. (a) The steady-state of w_2 exhibits an amplification of an infinitesimal sinusoidal perturbation probe $\xi_p = \epsilon \sin(\omega_n t)$ when $\alpha > 1$, shown for $\omega = \omega_n$ in the limit that $\mu \rightarrow 0$. Top: $\alpha = 3$ (red box); bottom: $\alpha = 0.33$ (green box). (b) The steady-state percent amplification of the sinusoidal probe perturbation as a function of the ratio of the normalized frequency (normalized by the undamped natural frequency ω_n) of the probe ξ_p in the limit that $\mu \rightarrow 0$.

At the natural frequency, the amplitude of $\tilde{w}_{2,ss}(t)$ equals $\sqrt{\alpha}\epsilon$, showing that a small error perturbation ξ_p leads to an amplified output when $\alpha > 1$ [Fig. 3(a)]. The mathematical correspondence between the ratio of learning rates α in systems consolidation and the (inverse square of the) damping ratio in a physical oscillator, and the resultant resonant amplification, is the intuition behind the potential instability in the presence of perturbations. This resonant behavior is shown for a sweep of relative frequencies for a larger range of α values in Fig. 3(b).

For the additive perturbation of Fig. 1(d), the resonance caused amplification but this amplification was kept finite by the damping. In the presence of signal-dependent perturbation, the natural frequency of the unperturbed system can interact with the frequency of vibration of the perturbation. When this interaction gives rise to a frequency close to the natural frequency of the unperturbed system, especially when the amplitude of the perturbation is sufficiently large, there can be increasing amplification of the system in each period, leading to unbounded growth. This resonance phenomenon, often referred to as parametric resonance [26], can happen when two oscillators get coupled in such a way that one causes oscillations in the parameters of the other oscillator, and does not necessarily need an external force to exhibit instability.

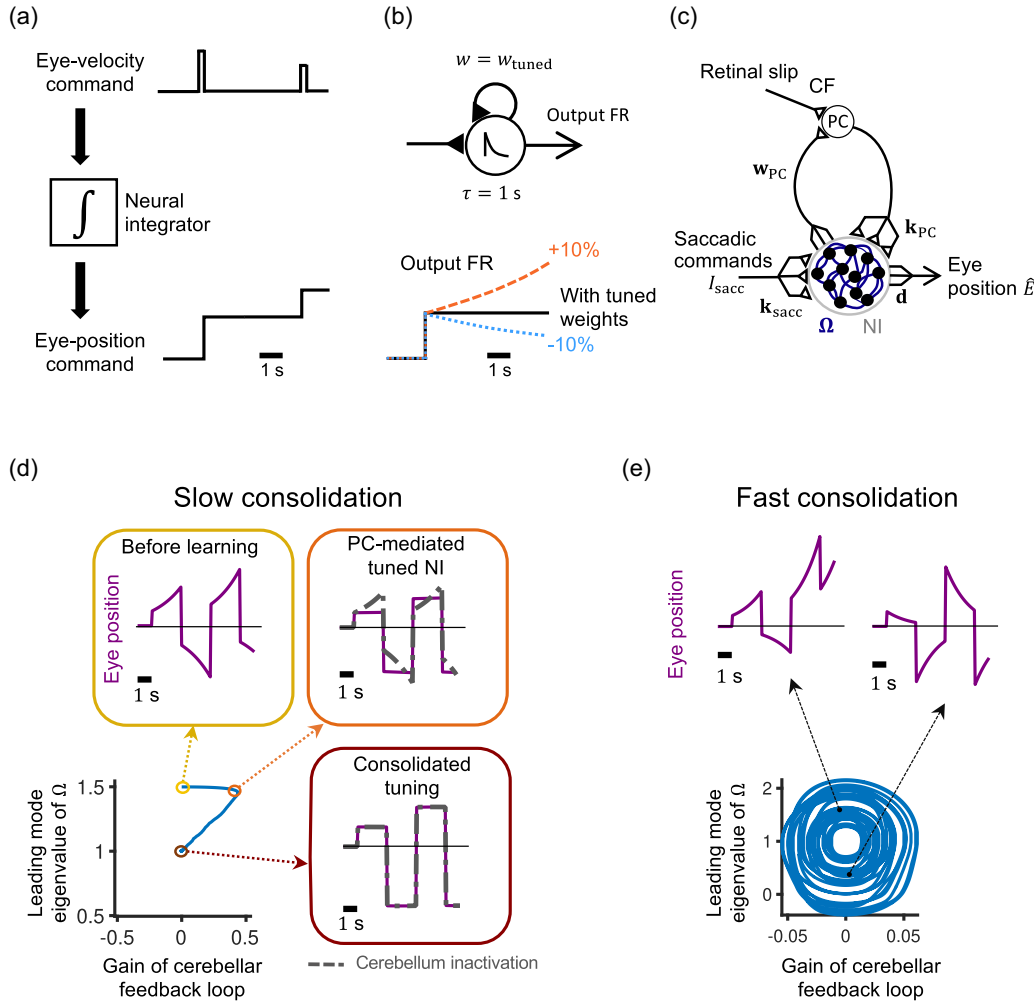


FIG. 4. Cerebellar tuning and subsequent consolidation of the time constant of the oculomotor neural integrator. (a) The computation of a neural integrator. (b) The fine-tuning problem of neural integrators. Increasing or decreasing the strength of recurrent feedback by a small amount causes exponential growth or decay, respectively, of integrator activity (the shown neuronal time constant $\tau = 1$ s is taken from Ref. [46], but the results apply more broadly). FR: firing rate. (c) Two-stage model for tuning the time constant of the neural integrator (NI). PC: cerebellar Purkinje cell. CF: climbing fiber carrying the retinal slip error signal. (d) Tuning the network from an initially unstable integrator condition (yellow), for a model with a fast learning rate in the early site of plasticity (cerebellum) and a slow learning rate in the late site (integrator). In the early stage of learning (orange), behavior is tuned primarily through the plasticity of the early site weights w_{PC} , so that the integrator function depends on the cerebellum (dashed grey line: effect of cerebellar inactivation). The learned memory then gets consolidated into the recurrent connectivity Ω of the neural integrator circuit, making the function independent of the cerebellum (dark magenta). Lower left panel shows strength of feedback in the cerebellar-NI feedback loop (x axis), and within the NI recurrent network (y axis, as characterized by the largest eigenvalue of Ω). (e) Tuning the model from the same initial condition as in (d) but with the learning rates of the early and late sites interchanged, so the early site has the slower learning rate. This can lead to instability in the system, as demonstrated by this example of oscillatory behavior in the weight space, resulting in alternating unstable (left) and leaky (right) eye positions, with a period ranging from 200 to 280 seconds.

D. Lyapunov function theory for tuning the recurrent dynamics of the oculomotor neural integrator

We next apply the theory to analyze the two-stage tuning of the dynamics of a recurrently connected neural circuit that is, furthermore, recurrently connected to the population that represents the early site of learning. We do this within the context of a well-characterized short-term memory circuit of the vertebrate brainstem, the oculomotor neural integrator. This circuit mathematically integrates (in the sense of calculus) motor command signals conveying the desired velocity of the eyes into motor command signals conveying the desired

position of the eyes [Fig. 4(a)]. In the absence of eye velocity command inputs, network activity is maintained persistently at an analog value [27,28]. Dynamically, such persistent activity corresponds to a line of fixed points, or line attractor, that is thought to be created by positive feedback within the network offsetting the inherent tendency of neuronal activity to decay back to a baseline firing level [19]. Models of line attractors require precise tuning of their recurrent weights [19,29] for this offset to occur [Fig. 4(b)]. Within the oculomotor system, this precise tuning might, like other visuomotor control tasks, be driven by visual error-feedback signals conveyed to

the cerebellum [30,31]. However, to our knowledge, no prior work has shown how this tuning can be learned initially by the cerebellum and then consolidated into the oculomotor neural integrator.

The cerebellum in this model is recurrently connected to the neural integrator so that its dynamics are tuned via a cerebello-integrator loop [Fig. 4(c)]. Fast plasticity instructed by retinal slip signals conveyed by climbing fibers (CF) occurs in the synapses onto a single lumped cerebellar Purkinje cell (PC) population. On a slower time scale, Purkinje cell activity then tunes the leading eigenmode of the neural integrator network dynamics. We model the neural integrator as a recurrently connected linear network of N neurons [19] that is reciprocally connected to the Purkinje cell as follows:

$$\tau \dot{r}_i = -r_i + \sum_{j=1}^N \Omega_{ij} r_j - k_{PC,i} r_{PC} + k_{sacc,i} I_{sacc} \quad (10)$$

$$r_{PC} = \sum_{j=1}^N w_{PC,j} r_j \quad (11)$$

$$\dot{E} = \sum_{j=1}^N d_j r_j, \quad (12)$$

where τ is the time constant of an integrator neuron, r_i is the firing rate of neuron i , Ω_{ij} denotes the connection weight from neuron j to neuron i , $k_{PC,i}$ is the strength of the input from the cerebellum to neuron i , r_{PC} is the Purkinje cell firing rate, $k_{sacc,i}$ is the strength of the saccadic input to neuron i , I_{sacc} is the velocity-encoding eye movement (saccadic) commands, $w_{PC,j}$ is the strength of the plastic synapse onto the Purkinje cell, $\dot{E}$ is the internal representation of the eye position, and d_j is the j^{th} component of the eye position decoder, which is assumed to be fixed during learning.

We non-dimensionalize time and rates by measuring them in units of the time constant τ as follows: $(t/\tau) \rightarrow t$, $r_i \tau \rightarrow r_i$, and $I_{sacc} \tau \rightarrow I_{sacc}$. The resulting dimensionless equations, in vector-matrix notation, are

$$\dot{\mathbf{r}} = -\mathbf{r} + \mathbf{\Omega} \mathbf{r} - \mathbf{k}_{PC} \mathbf{w}_{PC}^T \mathbf{r} + \mathbf{k}_{sacc} I_{sacc}, \quad (13)$$

$$\dot{E} = \mathbf{d}^T \mathbf{r}, \quad (14)$$

where $\mathbf{k}_{PC}$, $\mathbf{k}_{sacc}$, $\mathbf{r}$, $\mathbf{w}_{PC}$, and $\mathbf{d}$ are column vectors with N components, $\mathbf{\Omega}$ is an $N \times N$ matrix with zero diagonal elements, T denotes the matrix transpose operation, and $r_{PC} = \mathbf{w}_{PC}^T \mathbf{r}$. The desired behavior is perfect integration so that the eye position during eye fixations (in the absence of eye velocity commands) does not change, i.e., $\dot{E} = I_{sacc}$ or $E = \int I_{sacc} dt$. The retinal slip errors can be written as

$$\dot{e} = \dot{E} - \dot{E} = -\mathbf{d}^T (\mathbf{k}_{PC} \mathbf{w}_{PC}^T + \mathbb{I}_N - \mathbf{\Omega}) \mathbf{r} + (\mathbf{d}^T \mathbf{k}_{sacc} - 1) I_{sacc}, \quad (15)$$

where $\mathbb{I}_N$ is the $N \times N$ identity matrix. To achieve the desired behavior, $\dot{e}$ must be zero. From the second term above, this yields $\mathbf{d}^T \mathbf{k}_{sacc} = 1$. From the first term, denoting the term in parentheses as $\tilde{\mathbf{W}} \equiv \mathbf{k}_{PC} \mathbf{w}_{PC}^T + \mathbb{I}_N - \mathbf{\Omega}$, this yields $\mathbf{d}^T \tilde{\mathbf{W}} = 0$. Further, denoting a desired solution of $\mathbf{\Omega}$ after consolidation (when $\mathbf{w}_{PC} = 0$) as $\mathbf{\Omega}^*$, yields $\mathbf{d}^T \mathbf{\Omega}^* = \mathbf{d}^T$, which gives the expected condition that, after consolidation, the neural integrator weight matrix should have a unity eigenvalue. If

we define the deviation of $\mathbf{\Omega}$ from its desired solution as $\tilde{\mathbf{\Omega}} \equiv \mathbf{\Omega} - \mathbf{\Omega}^*$, then the component of $\tilde{\mathbf{W}}$ along the readout vector $\mathbf{d}$ can be decomposed as $\mathbf{d}^T \tilde{\mathbf{W}} = \mathbf{d}^T (\mathbf{k}_{PC} \mathbf{w}_{PC}^T - \tilde{\mathbf{\Omega}})$, i.e., as a sum of terms due to the deviation of the neural integrator weight matrix from its desired post-consolidation value $\mathbf{d}^T \tilde{\mathbf{\Omega}}$ and due to the (nonzero before consolidation) feedback loop through the cerebellar Purkinje cells.

The correlational learning rules in this nondimensionalized notation are

$$\dot{\mathbf{w}}_{PC} = \eta_1 (\dot{e} + \xi(t)) \mathbf{r}, \quad (16)$$

$$\dot{\tilde{\mathbf{\Omega}}} = -\eta_2 (\mathbf{k}_{PC} r_{PC}) \mathbf{r}^T, \quad (17)$$

where η_1 is the nondimensionalized learning rate for the early site in the cerebellum, η_2 is the corresponding learning rate for the late site in the neural integrator, and $\xi(t)$ is a perturbation to the error signal with maximal value bounded by a fraction of the error signal, $|\xi| \leq \mu |\dot{e}|$. We consider a Lyapunov function candidate for the two-stage tuning of the neural integrator consisting of the sum of squares of the components of $\mathbf{d}^T \tilde{\mathbf{\Omega}}$ and $\mathbf{d}^T \tilde{\mathbf{W}}$:

$$L = \frac{1}{2} (\mathbf{d}^T \tilde{\mathbf{W}}) (\mathbf{d}^T \tilde{\mathbf{W}})^T + \frac{1}{2} (\mathbf{d}^T \tilde{\mathbf{\Omega}}) (\mathbf{d}^T \tilde{\mathbf{\Omega}})^T. \quad (18)$$

As in the Lyapunov function for the two-stage model in the previous section, the first term in L encourages the retinal slip error to vanish, and the second term aligns with the goal of consolidating the learned memories from the early stage to the late stage synapses.

We now prove uniform global stability and asymptotic stability of the two-stage neural integrator tuning model. The time derivative of L is

$$\begin{aligned} \dot{L} &= (\mathbf{d}^T \tilde{\mathbf{W}}) (\mathbf{d}^T \tilde{\mathbf{W}})^T + (\mathbf{d}^T \tilde{\mathbf{\Omega}}) (\mathbf{d}^T \tilde{\mathbf{\Omega}})^T \\ &= \mathbf{d}^T (\mathbf{k}_{PC} \dot{\mathbf{w}}_{PC}^T - \dot{\tilde{\mathbf{\Omega}}}) (\mathbf{d}^T \tilde{\mathbf{W}})^T + (\mathbf{d}^T \dot{\tilde{\mathbf{\Omega}}}) (\mathbf{d}^T \tilde{\mathbf{\Omega}})^T \\ &= -\eta_1 \left(c \left(1 - \frac{\alpha}{c} \right) \dot{e}^2 + c \xi \dot{e} + \alpha (\mathbf{d}^T \tilde{\mathbf{\Omega}} \mathbf{r})^2 \right), \end{aligned} \quad (19)$$

where $\alpha = \eta_2 / \eta_1$ and $c = \mathbf{d}^T \mathbf{k}_{PC}$. Similar to the previous case, we can use the defined noise regime to prove L is nondecreasing:

$$\begin{aligned} \dot{L} &= -\eta_1 \left(c \left(1 - \frac{\alpha}{c} \right) \dot{e}^2 + \alpha (\mathbf{d}^T \tilde{\mathbf{\Omega}} \mathbf{r})^2 \right) - \eta_1 c \xi \dot{e} \\ &\leq -\eta_1 \left(c \left(1 - \frac{\alpha}{c} \right) \dot{e}^2 + \alpha (\mathbf{d}^T \tilde{\mathbf{\Omega}} \mathbf{r})^2 \right) + \eta_1 c \mu |\dot{e}| |\dot{e}| \\ &= -\eta_1 \left(c \left(1 - \frac{\alpha}{c} - \mu \right) \dot{e}^2 + \alpha (\mathbf{d}^T \tilde{\mathbf{\Omega}} \mathbf{r})^2 \right), \end{aligned} \quad (20)$$

when $\alpha \leq (1 - \mu)c$, for $c > 0$, $\dot{L} \leq 0$. If we consider $\mathbf{d}$ and $\mathbf{k}_{PC}$ to be normalized to unit length, then c is the cosine of the angle between $\mathbf{d}$ and $\mathbf{k}_{PC}$. If $c = 1$, i.e., $\mathbf{d}$ and $\mathbf{k}_{PC}$ are parallel, then we recover the same condition as in the toy model of Sec. II C. As the angle between $\mathbf{d}$ and $\mathbf{k}_{PC}$ increases, c gets smaller, which causes the condition for stability to become tighter. We verify the other conditions of theorem 1 as follows: L is bounded from below, $\min(L) = 0$, and L does not explicitly depend on time. Hence, L is positive definite and decrescent. Thus the equilibrium is uniformly stable. Since L is the sum of two quadratic terms, it is radially unbounded by

inspection, which guarantees that $\tilde{\mathbf{W}}$, $\tilde{\mathbf{\Omega}}$, and $\mathbf{w}_{\text{PC}}$ are globally bounded. To guarantee asymptotic stability, what is left to show is that $\tilde{L}$ is uniformly continuous in time by showing $\dot{\tilde{L}}$ is bounded. Given that the hyperparameters $\eta_1, \eta_2, \alpha, \tau$, the vectors $\mathbf{d}, \mathbf{k}_{\text{PC}}$, and the variables $\tilde{\mathbf{W}}, \tilde{\mathbf{\Omega}}, \mathbf{w}_{\text{PC}}$ are bounded, and $\mathbf{r}$ and ξ are assumed to be smooth bounded functions of time with bounded derivatives, $\dot{\tilde{L}}$ is bounded. ■

Simulations confirm that two-stage tuning of the neural integrator in the presence of signal-dependent noise is stable when learning in the early site is faster than learning in the late site [Fig. 4(d), Methods]. Starting from an initial unstable condition, the simulations successfully converged, and the learned memory was consolidated into the neural integrator. In the early phases of learning, the circuit function is dependent on the cerebellum [Fig. 4(d), dashed lines]. Following consolidation, the neural integrator function becomes independent of the cerebellum. Thus experiments that inactivate the cerebellum at different times during training of the oculomotor neural integrator should provide testable predictions of this two-stage model. Notably, if the early, cerebellar site of plasticity is slow and the late site of plasticity in the neural integrator is fast, this can potentially cause instability according to the theory. This potential instability is illustrated in Fig. 4(e), for which the learning rates at the early and late sites have been interchanged relative to that of Fig. 4(d).

III. DISCUSSION

We have provided a framework for studying the stability of systems consolidation and applied it to simple circuit architectures characterized by an early learning area that is directly trained by performance errors, which then trains a late learning area that provides the final site of memory storage. Using a Lyapunov function theory that enforces the stability of the learning and consolidation process, we obtained a fundamental result on the speed of learning: the late stage must not be tuned faster than the early stage, and when the teaching signal is corrupted by perturbation, the late stage should be tuned more slowly. We mapped the consolidation process of a simple example circuit to the dynamics of a driven damped oscillator, providing the intuition that increasing the ratio of late- to early-stage learning rates α is like decreasing the oscillator damping, leading to potential resonant instability. We then applied the theory to a biologically relevant case of tuning a neural integrator circuit.

Previous work on memory consolidation has focused primarily on a fundamental robustness-speed tradeoff in learning, known as the stability-plasticity dilemma, in which having fast plasticity leads to “instability” in the sense that new memories overwrite old ones. This previous work has shown that the tradeoff can be lessened in multistage models [32,33]. Here, within a two-stage model, we show a complementary, dynamical form of instability that occurs for the systems consolidation of graded memories, where having an excessively fast consolidation speed can lead to amplification of a perturbation or even exponential, unbounded growth of activity.

Our toy two-stage model in Fig. 1 maps onto the architecture of classic models of motor learning mediated by the

cerebellum [17,18]. In such models, early learning is thought to occur through plasticity of the weight w_1 between presynaptic parallel fiber inputs and postsynaptic Purkinje cells. This plasticity is thought to be driven by correlations between the activity of parallel fiber inputs and behavioral error signals that are conveyed by separate, climbing fiber inputs to the Purkinje cells. The learning process is particularly well-characterized in the cerebellum-mediated adaptation of eye movement reflexes. For example, in the vestibulo-ocular reflex (VOR), rapid corrective eye movements are generated to offset movements of the head, functioning like a motion-correcting camera. This reflex requires tuning because, for example, the introduction of eyeglasses can alter the relation between eye movement and resulting image motion across the retina. Connecting to the present work, one can map the input r_{in} of Fig. 1(b) to the head velocity, the output r_o to the eye velocity, w^* to the desired VOR gain (i.e., the ratio of eye to head velocity), and the teaching error to the “retinal slip” motion of the visual image on the retina [24]. Learning is then transferred from an initial site in the cerebellum (weight w_1) to a late site of final storage in the vestibular nucleus (weight w_2). Interestingly, to properly model the biological circuit, one should make the climbing-fiber-driven error signals come through discrete spikes rather than the smooth firing rate assumed here. This provides an effective form of perturbation $\xi(t)$ that can decrease the stability of the system if not compensated for by decreasing the learning rate at the late site. Finally, we note that a similar consolidation of learning has been shown to occur in the striato-neocortical reinforcement learning system of the brain [34], suggesting similar fundamental constraints on the speed of learning may be applicable more broadly.

We have shown that the theory also can be applied to tuning the dynamics of a simple network model of the oculomotor neural integrator. Temporal integration is a fundamental operation in the nervous system, where it is thought to mediate not only motor control, as focused upon here, but also the accumulation of evidence for decision making [35]. Within the eye movement system, loss of function of the oculomotor neural integrator is associated with eye movement disorders of gaze-holding, many of which may be due to improper tuning of the neural integrator by the cerebellum [36].

Our work differs from previous research on modeling the tuning of the oculomotor neural integrator in several ways. Early models used direct retinal slip input to the neural integrator to train it [21,37,38]. A later model used corrective saccades to provide negative derivative information as a teaching signal [39], while another used a derivative signal generated in an unsupervised manner to tune the integrator [40]. All of these models were single-stage learning models, in contrast to the present work. These differences highlight the novelty that our model brings, which integrates an adaptive cerebellar controller and a two-stage learning process to tune the neural integrator.

To illustrate the Lyapunov framework, we focused on tuning the dynamics of the leading eigenvector of a linear integrator network. More generally, within linear network theory, additional non-integrating modes of the network could be tuned by including additional Purkinje cells. Further, we

expect the core principles—fast learning at the cerebellum and slower consolidation in the cerebellar nuclei, as demonstrated in our model—should carry over to more complex, nonlinear networks. For such nonlinear networks, we speculate that the nonlinear tuning could be accomplished through having a set of Purkinje cells that serve as learned, nonlinear basis functions, with learning occurring due to changes in the weights of the many granule cell inputs to the Purkinje cells that, for simplicity, were not explicitly modeled here. Recent advances in adaptive control theory [41] provide new techniques that may help with analyzing the tuning of nonlinear target circuits, not only in linearly parameterized nonlinear circuits where adaptive parameters affect the network state linearly but also in nonlinearly parameterized networks.

Although systems consolidation is often considered to be a very slow process in several reported neural systems, this is not always the case. For declarative memories, the late stage can consolidate quickly if the new memories have features that are consistent with the existing structure of knowledge in the late stage [42,43]. Recent evidence from songbird motor consolidation suggests that the consolidation process may happen faster than originally thought, occurring online in the daytime, and not necessarily requiring offline nighttime processes [34]. Enforcing the stability and convergence of consolidation in these scenarios may reveal constraints on the speed of learning and consolidation, similar to what we have found in the current work.

Besides the important implications for neuroscience experiments, the framework we have provided here may have engineering applications. Classically in adaptive control theory, the tracking error or prediction error is directly used to tune the parameters of single-stage adaptive controllers [10]. Our work gives the insight that using a two-stage adaptive controller can give flexibility in terms of having robust storage of memory at the final site, as well as an extra knob to tune the speed of learning in a stable manner. In systems with delayed negative feedback that are subject to inappropriate oscillations, such two-stage learning could be used to avoid deleterious resonance effects. Recent machine learning approaches have focused on using machine learning to generate and control complex dynamical systems [44,45]. Given that artificial neural networks with online, adaptive learning are increasingly in demand throughout society, including in safety-critical tasks, this suggests a compelling need to develop frameworks that guarantee the stability of such algorithms. We hope that the principles of systems consolidation and Lyapunov theory introduced here could help current progress in this area by highlighting the need for real-time, continuously adaptive systems that are safe and stable.

ACKNOWLEDGMENTS

We thank Jay Bhasin for helpful discussions. We acknowledge financial support from Simons Foundation Collaboration on the Global Brain Grants No. 542989SPI and No. NC-GB-CULM-00002734 (M.S.G. and E.R.F.A.), NIH R01

NS104926 (M.S.G. and E.R.F.A.), and NIH R01 EY031972 (M.S.G.).

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: METHODS

1. Toy model simulation

For the simulation of the toy model, we simulated Eqs. (1)–(3) using the MATLAB solver ode45, which is a fifth-order Runge-Kutta method. The learning rate of the first stage was set to a fixed value $\eta_1 = 0.01 \text{ s}^{-1}$ in all simulations shown in Fig. 1. η_2 was zero in the single-stage models [in Figs. 1(c) and 1(e)], 0.0003 s^{-1} in the two-stage model with slow consolidation [in Figs. 1(d) and 1(f), left], and 1 s^{-1} in the two-stage models with fast consolidation [in Figs. 1(d) and 1(f), right]. $r_{in}(t)$ was generated by a step function with amplitude 0.1, smoothed using the filter $100/(100s + 1)$, where s represents the Laplace transform variable. In Figs. 1(c) and 1(d), a sinusoidal perturbation with a very small time-varying amplitude $a(t) \sin(\omega_a t)$ where $\omega_a = 0.1 \frac{\text{rad}}{\text{s}}$ was applied. $a(t)$ was sampled at random from a uniform distribution in the range $[0, 0.004]$ with a 100 s sampling period—the time variation of $a(t)$ was not necessary for our core results, but was included to illustrate the effect of a slow non-stationary amplitude $a(t)$. In Figs. 1(e) and 1(f), a signal-dependent perturbation $\xi(t) = 0.7 \sin(\omega_{sd} t) e(t)$ where $\omega_{sd} = 0.2 \frac{\text{rad}}{\text{s}}$ was applied for all three panels. Parameters for the simulations shown in Fig. 3 are provided in the main text and figure captions.

2. Simulation of the oculomotor neural integrator

We simulated the dimensionless neural integrator Eqs. (13)–(17) using the forward Euler method in MATLAB with 30 neurons, and with $dt = 0.05$, for 6800 time units, then plotted the results by converting to dimensionalized units by setting $\tau = 1 \text{ s}$ [46]. The entries of $\mathbf{d}$, $\mathbf{k}_{PC}$, and $\mathbf{k}_{sacc}$ were chosen from a uniform distribution from $[0, 1]$, and then normalized such that $\mathbf{d}^T \mathbf{d} = 1$, $\mathbf{d}^T \mathbf{k}_{PC} = 1$, and $\mathbf{d}^T \mathbf{k}_{sacc} = 1$. The weight matrix of the integrator was initialized to $\mathbf{\Omega} = 1.5 \mathbf{k}_{sacc} \mathbf{d}^T$ so that it has an unstable leading eigenvalue of 1.5, and then small random Gaussian noise with standard deviation 0.001 was added to the off-diagonal elements such that the other modes remain leaky but with some variation of decay timescales. $\mathbf{w}_{PC}$ was initialized as a vector of zeros. The learning rates for Fig. 4(d) were set to $\eta_1 = 0.014$, $\eta_2 = 1.5217 \times 10^{-4}$, and $\mu = 0.5$. The signal-dependent sinusoidal perturbation $\xi(t) = \mu e(t) \sin(\omega_\xi t)$, with $\omega_\xi = 0.0029$, was chosen to illustrate the potential instability in the system. The saccadic commands were generated as random-amplitude pulses, with amplitudes sampled from a Gaussian distribution with a standard deviation of 48. Pulses were generated from a Poisson process with a mean of 1 pulse per time unit. On top of this Poisson rate, whenever the eye position went out of the range $[-2.625, 2.625]$, an additional corrective saccade with an absolute amplitude of 61.2 in the opposite direction was

generated to bring the eye position back within this range, and the Poisson process for generating regular saccades was reset. In Fig. 4(e), the learning rates η_1 and η_2 were interchanged to illustrate how a network with slow early site and fast late site can potentially become unstable.

The code to generate the results can be found at Ref. [47]

APPENDIX B: STABILITY DEFINITIONS AND THEOREMS

For ease of notation, we present the following definitions, lemmas, and theorems from [10] in the context of a general nonautonomous dynamical system $\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t)$ for the state vector $\mathbf{x} \in \mathbb{R}^N$ with equilibrium point $\mathbf{x}_{eq} = \mathbf{0}$.

1. Definitions

The formal definition of the most basic notion of stability in the Lyapunov sense for a non-autonomous system is

Definition B1. The equilibrium point $\mathbf{0}$ is stable at t_0 if for any $R \geq 0$, there exists a positive scalar $r(R, t_0)$ such that

$$\|\mathbf{x}(t_0)\| < r \Rightarrow \|\mathbf{x}(t)\| < R \quad \forall t \geq t_0.$$

Otherwise, the equilibrium point $\mathbf{0}$ is unstable. If the scalar r in the above can be chosen independently of t_0 , i.e., if $r = r(R)$, then the equilibrium point $\mathbf{0}$ is uniformly unstable.

If the above conditions are true for the whole state space, then the stability is global; otherwise, the stability is local.

A more refined and desirable concept is asymptotic stability, which not only ensures that the state stays in a ball of arbitrarily small radius around the equilibrium but also provides a statement about convergence to the equilibrium:

Definition B2. The equilibrium point $\mathbf{0}$ is asymptotically stable at t_0 if

- (1) it is stable
- (2) $\exists r(t_0) > 0$ such that $\|\mathbf{x}(t_0)\| < r(t_0) \Rightarrow \|\mathbf{x}(t_0)\| \rightarrow 0$ as $t \rightarrow \infty$.

In addition, if there exists a ball of attraction $\mathbf{B}_{\mathbf{R}_0}$, whose radius is independent of t_0 , such that any system trajectory with initial states in $\mathbf{B}_{\mathbf{R}_0}$ converges to $\mathbf{0}$ uniformly in t_0 , then the equilibrium point $\mathbf{0}$ is uniformly asymptotically stable.

Definition B3. A scalar continuous function $L(\mathbf{x})$ is said to be locally positive definite if $L(\mathbf{0}) = 0$ and, in a ball around the origin

$$\mathbf{x} \neq \mathbf{0} \Rightarrow L(\mathbf{x}) > 0.$$

If the inequality in the above is replaced with $L(\mathbf{x}) \geq 0$, then $L(\mathbf{x})$ is (locally) positive semidefinite. If $L(\mathbf{0}) = 0$ and the above property holds over the whole state space, then $L(\mathbf{x})$ is said to be globally positive definite. If $L(\mathbf{x})$ is positive (semi)definite, then $-L(\mathbf{x})$ is negative (semi)definite.

The time-varying function $L(\mathbf{x}, t)$ is said to be positive definite if $L(\mathbf{0}, t) = 0$ and there is a time-invariant positive definite function $L_0(\mathbf{x})$ such that

$$\forall t \geq t_0, \quad L(\mathbf{x}, t) \geq L_0(\mathbf{x}).$$

Definition B4. A scalar function $L(\mathbf{x}, t)$ is said to be decescent if $L(\mathbf{0}, t) = 0$, and if there exists a time-invariant positive definite function $L_l(\mathbf{x})$ such that

$$\forall t \geq 0, \quad L(\mathbf{x}, t) \leq L_l(\mathbf{x}).$$

Definition B5. A function g is said to be uniformly continuous on $[0, \infty)$ if

$$\forall R > 0, \exists \eta(R), \forall t_1 \geq 0, \forall t \geq 0, \text{ such that } |t - t_1| < \eta \Rightarrow |g(t) - g(t_1)| < R.$$

This uniformity in time means that one can always find an η which does not depend on the point t_1 .

2. Lemmas and theorems

Below, we provide the theorems and lemmas needed to prove the stability of the systems in the main text. For their proofs, consult Ref. [10].

Theorem B1 (Lyapunov theorem for non-autonomous systems).

Stability. If, in a ball $\mathbf{B}_{\mathbf{R}_0}$ around the equilibrium point $\mathbf{0}$, there exists a scalar function $L(\mathbf{x}, t)$ with continuous partial derivatives such that

- (1) L is positive definite.
- (2) $\dot{L} = \frac{dL}{dt}$ is negative semi-definite

then the equilibrium point $\mathbf{0}$ is stable in the sense of Lyapunov.

Uniform stability and uniform asymptotic stability. If, furthermore,

- (3) L is decrescent,

then the origin is uniformly stable. If condition 2 is strengthened by requiring that L be negative definite, then the equilibrium point is uniformly asymptotically stable.

Global uniform asymptotic stability. If the ball $\mathbf{B}_{\mathbf{R}_0}$ is replaced by the whole state space, and condition 1, the strengthened condition 2, condition 3, and the condition

- (4) $L(\mathbf{x}, 0)$ is radially unbounded, i.e., $L(\mathbf{x}, 0) \rightarrow \infty$ as $\|\mathbf{x}\| \rightarrow \infty$

are all satisfied, then the equilibrium point at $\mathbf{0}$ is globally uniformly asymptotically stable.

To prove asymptotic stability in cases where it is not easy to prove negative definiteness of $\dot{L}$, we use the following Lyapunov-like lemma, which is a variant of Barbalat's lemma [10], that requires the derivative of L to have some additional smoothness property to ensure $\dot{L}$ converges to zero:

Lemma B1. If a scalar function $L(x, t)$ satisfies the following conditions

- (1) $L(x, t)$ is lower bounded,
- (2) $\dot{L}(x, t)$ is negative semidefinite,
- (3) $\dot{L}(x, t)$ is uniformly continuous in time,

then $\dot{L}(x, t) \rightarrow 0$ as $t \rightarrow \infty$.

The first two conditions in the above Lemma imply that L has a finite limiting value L_∞ , but they do not guarantee that L will remain stationary at L_∞ [10]. The addition of the third condition gives us the ability to conclude that, in the limit $t \rightarrow \infty$, L remains stationary at L_∞ and the convergence will be achieved.

[1] R. Shadmehr and H. H. Holcomb, Neural correlates of motor memory consolidation, *Science* **277**, 821 (1997).

[2] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, Why there are complementary learning systems in the hippocam-

- pus and neocortex: Insights from the successes and failures of connectionist models of learning and memory, *Psychol. Rev.* **102**, 419 (1995).
- [3] J. F. Medina, J. C. Repa, M. D. Mauk, and J. E. LeDoux, Parallels between cerebellum-and amygdala-dependent conditioning, *Nat. Rev. Neurosci.* **3**, 122 (2002).
 - [4] S. Grossberg, Competitive learning: From interactive activation to adaptive resonance, *Cognit. Sci.* **11**, 23 (1987).
 - [5] H. R. Joo and L. M. Frank, The hippocampal sharp wave-ripple in memory retrieval for immediate use and consolidation, *Nat. Rev. Neurosci.* **19**, 744 (2018).
 - [6] J. W. Krakauer and R. Shadmehr, Consolidation of motor memory, *Trends Neurosci.* **29**, 58 (2006).
 - [7] A. S. Andalman and M. S. Fee, A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors, *Proc. Natl. Acad. Sci. USA* **106**, 12518 (2009).
 - [8] J. L. McGaugh, The amygdala modulates the consolidation of memories of emotionally arousing experiences, *Annu. Rev. Neurosci.* **27**, 1 (2004).
 - [9] A. A. Faisal, L. P. J. Selen, and D. M. Wolpert, Noise in the nervous system, *Nat. Rev. Neurosci.* **9**, 292 (2008).
 - [10] J.-J. E. Slotine and W. Li, *Applied Nonlinear Control*, Vol. 199, (Prentice hall Englewood Cliffs, NJ, 1991).
 - [11] R. M. Sanner and J.-J. E. Slotine, Stable adaptive control of robot manipulators using “neural” networks, *Neural Comput.* **7**, 753 (1995).
 - [12] A. Alemi, C. Machens, S. Deneve, and J.-J. Slotine, Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 (AAAI Press, Palo Alto, CA, 2018), pp. 588–595.
 - [13] A. Gilra and W. Gerstner, Predicting non-linear dynamics by stable local learning in a recurrent spiking neural network, *eLife* **6**, e28295 (2017).
 - [14] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982).
 - [15] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks*, (Cambridge University Press, Cambridge, England, 1989).
 - [16] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, *Nature (London)* **405**, 947 (2000).
 - [17] F. A. Miles and S. G. Lisberger, Plasticity in the vestibulo-ocular reflex: a new hypothesis, *Annu. Rev. Neurosci.* **4**, 273 (1981).
 - [18] C. D. Kassardjian, Y.-F. Tan, J.-Y. J. Chung, R. Heskin, M. J. Peterson, and D. M. Broussard, The site of a motor memory shifts with consolidation, *J. Neurosci.* **25**, 7979 (2005).
 - [19] H. S. Seung, How the brain keeps the eyes still, *Proc. Natl. Acad. Sci. USA* **93**, 13339 (1996).
 - [20] B. Widrow and M. A. Lehr, 30 years of adaptive neural networks: perceptron, madaline, and backpropagation, *Proc. IEEE* **78**, 1415 (1990).
 - [21] D. B. Arnold and D. A. Robinson, The oculomotor integrator: testing of a neural network model, *Exp. Brain Res.* **113**, 57 (1997).
 - [22] C. M. Harris and D. M. Wolpert, Signal-dependent noise determines motor planning, *Nature (London)* **394**, 780 (1998).
 - [23] K. E. Jones, A. F. de C. Hamilton, and D. M. Wolpert, Sources of signal-dependent noise during isometric force production, *J. Neurophysiol.* **88**, 1533 (2002).
 - [24] B. J. Bhasin, J. L. Raymond, and M. S. Goldman, Synaptic weight dynamics underlying memory consolidation: Implications for learning rules, circuit organization, and circuit function, *Proc. Natl. Acad. Sci. USA* **121**, e2406010121 (2024).
 - [25] C. Clopath, A. Badura, C. I. De Zeeuw, and N. Brunel, A cerebellar learning model of vestibulo-ocular reflex adaptation in wild-type and mutant mice, *J. Neurosci.* **34**, 7203 (2014).
 - [26] F. Verhulst, Perturbation Analysis of Parametric Resonance, in *Encyclopedia of Complexity and Systems Science*, edited by R. Meyers (Springer, New York, NY, 2009).
 - [27] D. A. Robinson, Integrating with neurons, *Annu. Rev. Neurosci.* **12**, 33 (1989).
 - [28] E. Aksay, R. Baker, H. S. Seung, and D. W. Tank, Anatomy and discharge properties of pre-motor neurons in the goldfish medulla that have eye-position signals during fixations, *J. Neurophysiol.* **84**, 1035 (2000).
 - [29] H. S. Seung, D. D. Lee, B. Y. Reis, and D. W. Tank, Stability of the memory of eye position in a recurrent network of conductance-based model neurons, *Neuron* **26**, 259 (2000).
 - [30] G. Major, R. Baker, E. Aksay, B. Mensh, H. S. Seung, and D. W. Tank, Plasticity and tuning by visual feedback of the stability of a neural integrator, *Proc. Natl. Acad. Sci. USA* **101**, 7739 (2004).
 - [31] J. L. Raymond and J. F. Medina, Computational principles of supervised learning in the cerebellum, *Annu. Rev. Neurosci.* **41**, 233 (2018).
 - [32] A. Roxin and S. Fusi, Efficient partitioning of memory systems and its importance for memory consolidation, *PLoS Comput. Biol.* **9**, e1003146 (2013).
 - [33] M. K. Benna and S. Fusi, Computational principles of synaptic memory consolidation, *Nat. Neurosci.* **19**, 1697 (2016).
 - [34] R. O. Tachibana, D. Lee, K. Kai, and S. Kojima, Performance-dependent consolidation of learned vocal changes in adult songbirds, *J. Neurosci.* **42**, 1974 (2022).
 - [35] M. N. Shadlen and R. Kiani, Decision making as a window on cognition, *Neuron* **80**, 791 (2013).
 - [36] R. J. Leigh and D. S. Zee, *The Neurology of Eye Movements*, (Contemporary Neurology, 2015).
 - [37] D. B. Arnold and D. A. Robinson, A neural network model of the vestibulo-ocular reflex using a local synaptic learning rule, *Philos. Trans. R. Soc. London B* **337**, 327 (1992).
 - [38] X. Xie and H. S. Seung, Spike-based learning rules and stabilization of persistent neural activity, *Advances in Neural Information Processing Systems*, Vol. 12 (MIT Press, Cambridge, MA, 1999).
 - [39] D. MacNeil and C. Eliasmith, Fine-tuning and the stability of recurrent neural networks, *PLoS One* **6**, e22885 (2011).
 - [40] C. Federer and J. Zylberberg, A self-organizing short-term dynamical memory network, *Neural Networks* **106**, 30 (2018).
 - [41] N. M. Boffi and J.-J. E. Slotine, Implicit regularization and momentum algorithms in nonlinearly parameterized adaptive control and prediction, *Neural Comput.* **33**, 590 (2021).

- [42] D. Tse, R. F. Langston, M. Kakeyama, I. Bethus, P. A. Spooner, E. R. Wood, M. P. Witter, and R. G. M. Morris, Schemas and memory consolidation, *Science* **316**, 76 (2007).
- [43] D. Kumaran, D. Hassabis, and J. L. McClelland, What learning systems do intelligent agents need? complementary learning systems theory updated, *Trends Cognit. Sci.* **20**, 512 (2016).
- [44] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, *Advances in Neural Information Processing Systems*, Vol. 31 (Curran Associates, Inc., Red Hook, NY, 2018).
- [45] R. Hasani, M. Lechner, A. Amini, L. Liebenwein, A. Ray, M. Tschaikowski, G. Teschl, and D. Rus, Closed-form continuous-time neural networks, *Nat. Mach. Intell.* **4**, 992 (2022).
- [46] D. Fisher, I. Olasagasti, D. W. Tank, E. R. F. Aksay, and M. S. Goldman, A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit, *Neuron* **79**, 987 (2013).
- [47] <https://github.com/alemi/LyapunovSystemConsolidation>.