

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

Alireza Alemi,^{1,*} Emre R. F. Aksay,² and Mark S. Goldman^{1,3,†}

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

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

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

Abstract

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 learning in systems consolidation mediated by the cerebellum. 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 the 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 (square of the) oscillator’s damping ratio. This work suggests the power of the Lyapunov approach to provide constraints on nervous system function.

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 discrete attractor neural

* alemi@ucdavis.edu

† msgoldman@ucdavis.edu

networks to model long-term memory [14, 15]. 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, 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 of systems consolidation in which the late stage, unlike the early stage, lacks direct access to an error-correcting teaching signal (Fig. 1). In this architecture, the activity at this early site then trains the weight of the late site such that the memory is eventually transferred from the early to the late site. This architecture and set of learning rules correspond to classic theories of learning and systems consolidation in the cerebellum. Learning at the early site corresponds to the classic Marr-Albus-Ito theory of cerebellar learning [16, 17] 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 site 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. The slower the tuning in the late stage, the more robust the learning process is against noise in the primary teaching signal.

II. RESULTS

A. Model for systems consolidation

We 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 $\widetilde{W} = w_1 + w_2 - w^*$ 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., $r_1 = w_1r_{\text{in}}$.

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 [18, 19].

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

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 the perturbation is signal-dependent [20, 21], i.e., the strength of the perturbation depends on the amount of error signal $e(t)$. We formalize the regime of the perturbation in the teaching signal during learning by a parameter $\mu = \max \frac{|\xi(t)|}{|e(t)|}$ (where $e(t) \neq 0$) that defines the maximal amount of perturbation during learning.

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, $r_{in}(t)$ [22]:

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

where η_2 is the learning rate.

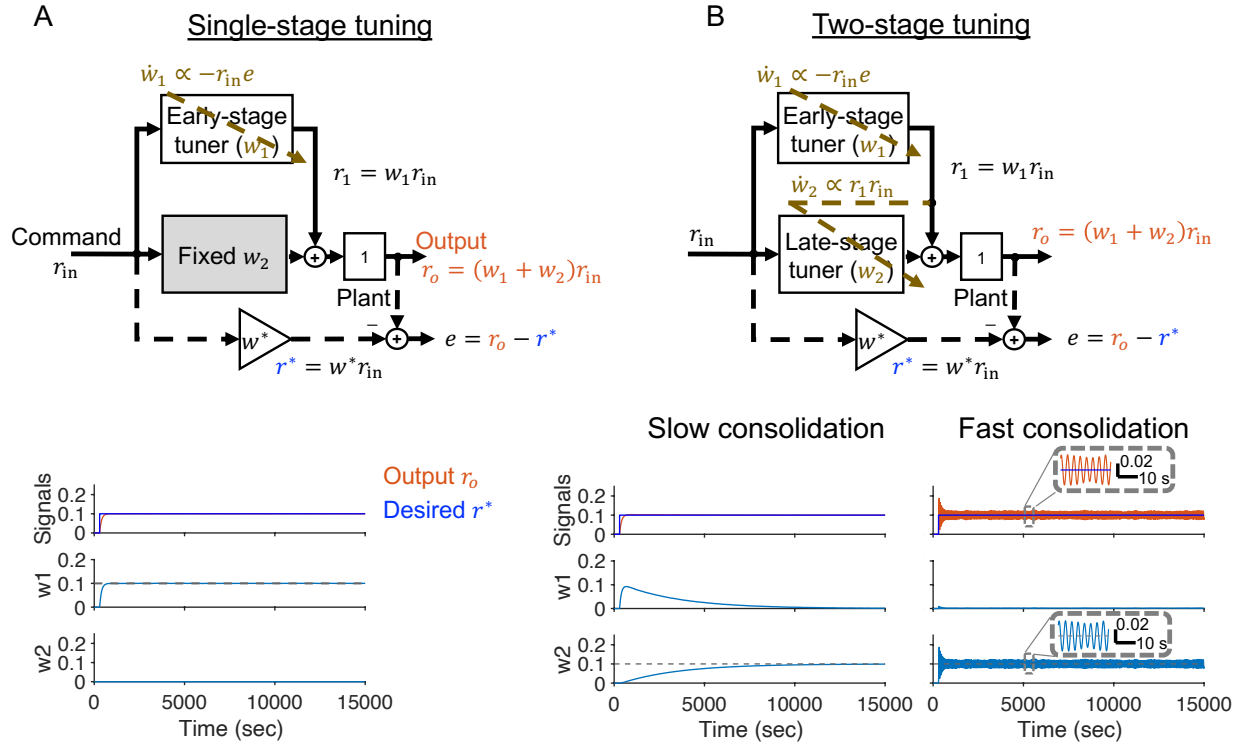


FIG. 1. A toy tracking model demonstrating instability in systems consolidation. (A) Top: Single-stage model. The early-stage parameter, w_1 , is directly tuned by the error signal $e(t)$, whereas the late-stage parameter, w_2 , is fixed. Bottom: Simulation showing that the model successfully converges to the desired output r^* and the desired, tuned weight w_1 (dashed line). (B) Two-stage tuning model. Top: 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. Bottom: When the consolidation process is slow enough (left panels), the model dynamics successfully converges and tunes 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. See Appendix A for simulation details.

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 model

To illustrate the Lyapunov stability approach and motivate the problem of instability in systems consolidation, we first consider a one-stage model (Fig. 1A, top) 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 Fig. 1A (bottom) 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 non-autonomous 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 \widetilde{W} :

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

where $\widetilde{W} = w_1 + w_2 - w^*$. L_1 is positive definite and radially unbounded by inspection. To show that $\dot{L}_1 \leq 0$, we write the learning dynamics as $\dot{w}_1 = -\eta_1\widetilde{W}r_{\text{in}}^2 - \eta_1r_{\text{in}}\xi$. Using the learning rule to compute the time derivative yields $\dot{L}_1 = -\eta_1\widetilde{W}^2r_{\text{in}}^2 - \eta_1\widetilde{W}r_{\text{in}}\xi$. With $|\xi| \leq \mu|e|$, one obtains:

$$\begin{aligned} \dot{L}_1 &\leq -\eta_1\widetilde{W}^2r_{\text{in}}^2 + \eta_1|\widetilde{W}r_{\text{in}}||\xi| \\ &\leq -\eta_1\widetilde{W}^2r_{\text{in}}^2 + \eta_1|\widetilde{W}r_{\text{in}}|\mu|\widetilde{W}r_{\text{in}}| \\ &= -\eta_1r_{\text{in}}^2(1 - \mu)\widetilde{W}^2. \end{aligned} \quad (5)$$

$\eta_1 > 0$ 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 \widetilde{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 uniformity is to show that \dot{L}_1 is bounded. Given that the hyperparameters η_1, η_2 and the gain error \widetilde{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 consolidation model

We next consider systems consolidation in the two-stage model. As proven below, when the consolidation process is sufficiently slow, the two-stage learning model successfully converges to solving the tracking task (Fig. 1B, bottom 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 (Fig. 1B, bottom right). Below, we analytically find the conditions for guaranteeing stability (Section II C 1), provide intuition for the source of possible instability in regions without stability guarantee by solving a special case of the system (Section II C 2), and demonstrate with simulations a case in which loss of stability leads to unbounded growth of activity (Section II C 3).

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 $\dot{w}_1 = -\eta_1 \widetilde{W} r_{\text{in}}^2 - \eta_1 r_{\text{in}} \xi$ and $\dot{w}_2 = \eta_2 w_1 r_{\text{in}}^2$. We choose the following Lyapunov function candidate for the two-stage model:

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

where $\widetilde{w}_2 = w_2 - w^*$. 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. 2A 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. 2B).

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

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

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 $r_{\text{in}} \neq 0$ 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 two quadratic terms, it is radially unbounded by inspection, which guarantees that \widetilde{W} , \widetilde{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 \ddot{L} is bounded.

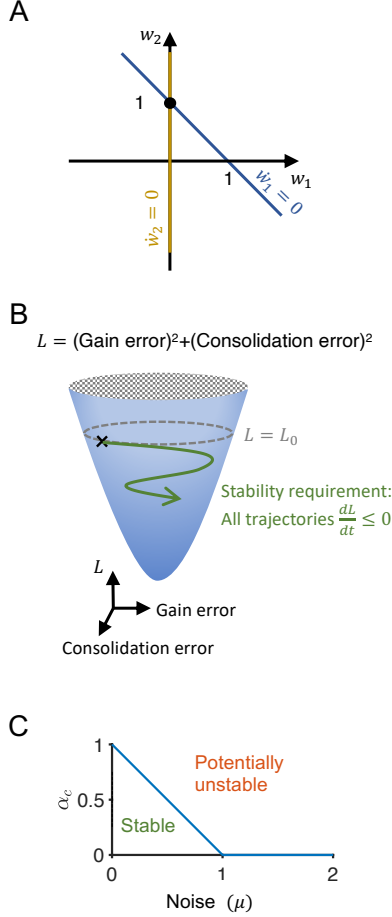


FIG. 2. Lyapunov function theory for stability of the two-stage model. (A) In the limit that the perturbation goes to zero, $\mu \rightarrow 0$, the closed-loop learning dynamics has a single fixed point and two nullclines (shown for $w^* = 1$). (B) The Lyapunov function candidate L has two terms: the 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.

Given that the hyperparameters η_1, η_2, α , and the variables $\widetilde{W}, \widetilde{w}_2, w_1$ are bounded, and r_{in} and ξ are assumed to be smooth bounded functions of time with bounded derivatives, \ddot{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. 2C). 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

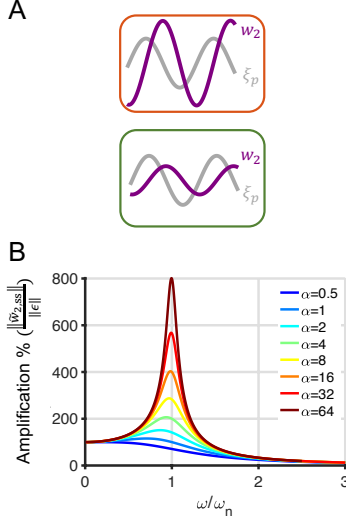


FIG. 3. Amplification of perturbation in the region without stability guarantee. (A) The state-state of w_2 exhibits an amplification of an infinitesimal sinusoidal perturbation probe $\xi_p = \epsilon \sin(\omega_n t)$, shown in the limit that $\mu \rightarrow 0$. Top, $\alpha = 3$ (red box); bottom, $\alpha = 5.33$ (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 $\mu \rightarrow 0$.

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 $|\tilde{w}_2| < \sqrt{\frac{\alpha-1}{\alpha}}|\tilde{W}|$, 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:

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

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. This clearly shows that a small error perturbation ξ_p leads to an amplified output whose amplitude at the natural frequency equals $\sqrt{\alpha}\epsilon$. The simulations for two values of α , i.e., $\alpha = 0.33$ and $\alpha = 3$, show an amplification of the probe in w_2 in the potentially unstable region (Fig. 3A) at the natural frequency. 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. 3B.

3. Simulation of an unbounded growth instability

To further investigate the instability, we simulate the effect of a sinusoidal signal-dependent perturbation, which can alternatively be interpreted as a time-varying learning rate of the early stage. Solving the equations directly in this case is cumbersome, but analyzing their stability is trivial with the Lyapunov theory. The theory we have developed can be used directly: when the maximum $\alpha(t)$ is less than one, the system is guaranteed to be stable; otherwise, stability is not guaranteed. Consider the case $\eta_1(t) = 0.1(1 + 0.7 \sin(0.2\pi t))$. For a slow learning rate of the late site, $\eta_2 = 0.02$, $\max(\alpha(t)) < 1$ so that stability is guaranteed (Fig. 4A). By contrast, for $\eta_2 = 1$, yielding values of $\alpha(t) > 1$ so that stability is not guaranteed, we find that the system not only exhibits amplification of the sinusoidal perturbation but grows unboundedly (Fig. 4B).

We can interpret the instability using the resonance intuition we developed in the previous subsection. For the simple case of Section II C 2, 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

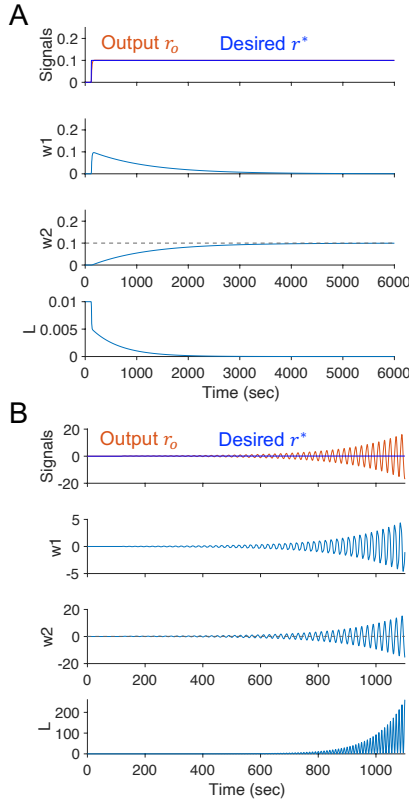


FIG. 4. The presence of a sinusoidal signal-dependent perturbation in the form of a time-varying learning rate of the early stage can cause unbounded growth instability. We consider $\eta_1(t) = 0.1(1 + 0.7 \sin(0.2\pi t))$. (A) A stable case with $\eta_2 = 0.02$ so that $\max(\alpha) < 1$. The first three subplots are in the same format as Figure 1, with the last subplot showing the Lyapunov function L . (B) An unstable case with $\eta_2 = 1$ so that $\alpha > 1$.

[23], 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.

III. DISCUSSION

We have provided a framework for studying the stability of systems consolidation and applied it to a simple circuit architecture characterized by an early learning area that is directly trained by performance errors, which 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 have 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 much more slowly. We mapped the consolidation process 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.

Previous work on memory consolidation has focused primarily on a fundamental robustness-

speed tradeoff in learning with single-stage models, known as the stability-plasticity dilemma. This dilemma states that, in single-stage models, having fast plasticity leads to ‘instability’ in the sense that new memories overwrite old ones, whereas this tradeoff can be lessened in multi-stage models [24, 25]. Here, by contrast, we show a complementary, dynamical form of instability that occurs for the systems consolidation of graded memories, in which having too fast a speed of consolidation can lead to amplification of a perturbation or even exponential unbounded growth of activity. Despite these differences, both our theory and previous ones obey a similar principle: multi-stage learning offers robustness by having slower learning in the later stages.

Our toy two-stage model maps onto the architecture of classic models of motor learning mediated by the cerebellar brain region [16, 17]. 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 the parallel fiber inputs with 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} 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. 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 [26], suggesting similar fundamental constraints on the speed of learning may be applicable more broadly.

Although systems consolidation in the late stage tends to be considered a 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 [27, 28]. 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 [26]. Enforcing the stability and convergence of consolidations 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 a robust storage 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 [29, 30]. 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.

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 $\eta_1 = 0.01$ in all simulations shown in Fig. 1. η_2 was zero in the single-stage model, 0.0003 in the two-stage model with slow consolidation (Fig. 1B, left), and 1 in the two-stage model with fast consolidation (Fig. 1B, right). $r_{\text{in}}(t)$ was generated by a step function with amplitude 0.1 which was smoothed by the filter $100/(100s + 1)$ (with s being the Laplace variable). In Fig. 1, a sinusoidal perturbation with time-varying amplitude $\xi = a(t) \sin(\omega_n t)$ where $\omega_n = 0.1$ (rad/s) was considered. In addition, $a(t)$ was sampled at random from a uniform distribution in the range $[0, 0.002]$ with a 10 s sampling period – this was not necessary for our core results, but was included to illustrate the effect of a slow non-stationary amplitude $a(t)$.

Appendix B: STABILITY DEFINITIONS AND THEOREMS

For ease of notation, we present the following definitions, lemmas, and theorems in the context of a general non-autonomous 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 1

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

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

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 2

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

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

In addition, if there exists a ball of attraction \mathbf{B}_{R_0} , whose radius is independent of t_0 , such that any system trajectory with initial states in \mathbf{B}_{R_0} converges to $\mathbf{0}$ uniformly in t_0 , then the equilibrium point $\mathbf{0}$ is uniformly asymptotically stable.

Definition 3

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

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

If the inequality in the above is replaced with $L(\mathbf{x}) \geq 0$, then $L(\mathbf{x})$ is (locally) positive semi-definite. 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 4

A scalar function $L(\mathbf{x}, t)$ is said to be decreascent 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 5

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

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

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 [10].

Theorem 1 (Lyapunov theorem for non-autonomous systems)

Stability: *If, in a ball \mathbf{B}_{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}_{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 L converges to zero:

Lemma 1

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

- $L(x, t)$ is lower bounded
- $\dot{L}(x, t)$ is negative semi-definite
- $\dot{L}(x, t)$ 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.

ACKNOWLEDGMENTS

We thank Jay Bhasin for helpful discussions. We acknowledge financial support from Simons Foundation Collaboration on the Global Brain grants 542989SPI and NC-GB-CULM-00002734 (MG, EA), NIH R01 NS104926 (MG, EA), and NIH R01 EY031972 (MG).

- [1] Reza Shadmehr and Henry H Holcomb. Neural correlates of motor memory consolidation. *Science*, 277(5327):821–825, 1997.
- [2] James L McClelland, Bruce L McNaughton, and Randall C O’Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419, 1995.
- [3] Javier F Medina, J Christopher Repa, Michael D Mauk, and Joseph E LeDoux. Parallels between cerebellum-and amygdala-dependent conditioning. *Nature Reviews Neuroscience*, 3(2):122–131, 2002.
- [4] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, 11(1):23–63, 1987.
- [5] Hannah R Joo and Loren M Frank. The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience*, 19(12):744–757, 2018.
- [6] John W Krakauer and Reza Shadmehr. Consolidation of motor memory. *Trends in Neurosciences*, 29(1):58–64, 2006.
- [7] Aaron S Andalman and Michale S Fee. A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. *Proceedings of the National Academy of Sciences*, 106(30):12518–12523, 2009.
- [8] James L McGaugh. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review Neuroscience*, 27:1–28, 2004.
- [9] A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, 2008.
- [10] Jean-Jacques E Slotine, Weiping Li, et al. *Applied Nonlinear Control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.
- [11] Robert M Sanner and Jean-Jacques E Slotine. Stable adaptive control of robot manipulators using “neural” networks. *Neural Computation*, 7(4):753–790, 1995.
- [12] Alireza Alemi, Christian Machens, Sophie Deneve, and Jean-Jacques Slotine. Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [13] Aditya Gilra and Wulfram Gerstner. Predicting non-linear dynamics by stable local learning in a recurrent spiking neural network. *Elife*, 6:e28295, 2017.
- [14] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.
- [15] Daniel J Amit and Daniel J Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, 1989.
- [16] FA Miles and SG Lisberger. Plasticity in the vestibulo-ocular reflex: a new hypothesis. *Annual Review of Neuroscience*, 4(1):273–299, 1981.
- [17] Charles D Kassardjian, Yao-Fang Tan, Ji-Yeon J Chung, Raquel Heskin, Michael J Peterson,

- and Dianne M Broussard. The site of a motor memory shifts with consolidation. *Journal of Neuroscience*, 25(35):7979–7985, 2005.
- [18] Bernard Widrow and Michael A Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.
- [19] DB Arnold and DA Robinson. The oculomotor integrator: testing of a neural network model. *Experimental Brain Research*, 113:57–74, 1997.
- [20] Christopher M Harris and Daniel M Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784, 1998.
- [21] Kelvin E Jones, Antonia F de C Hamilton, and Daniel M Wolpert. Sources of signal-dependent noise during isometric force production. *Journal of Neurophysiology*, 88(3):1533–1544, 2002.
- [22] Claudia Clopath, Aleksandra Badura, Chris I De Zeeuw, and Nicolas Brunel. A cerebellar learning model of vestibulo-ocular reflex adaptation in wild-type and mutant mice. *Journal of Neuroscience*, 34(21):7203–7215, 2014.
- [23] Ferdinand Verhulst. *Perturbation Analysis of Parametric Resonance*, pages 6625–6639. Springer New York, New York, NY, 2009.
- [24] Alex Roxin and Stefano Fusi. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Computational Biology*, 9(7):e1003146, 2013.
- [25] Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, 2016.
- [26] Ryosuke O Tachibana, Dahyun Lee, Kazuki Kai, and Satoshi Kojima. Performance-dependent consolidation of learned vocal changes in adult songbirds. *Journal of Neuroscience*, 42(10):1974–1986, 2022.
- [27] Dorothy Tse, Rosamund F Langston, Masaki Takeyama, Ingrid Bethus, Patrick A Spooner, Emma R Wood, Menno P Witter, and Richard GM Morris. Schemas and memory consolidation. *Science*, 316(5821):76–82, 2007.
- [28] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.
- [29] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.
- [30] Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, pages 1–12, 2022.