Organizing memories for generalization in complementary learning systems

Weinan Sun[1], Madhu Advani[2], Nelson Spruston[1], Andrew Saxe[2,3,4,5*], and James E. Fitzgerald[1*]

[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA
[2]Center for Brain Science, Harvard University, Cambridge, MA, USA
[3]Department of Experimental Psychology, University of Oxford, Oxford, UK
[4]Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, UCL, London, UK
[5]CIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada

*Equal contribution, order determined by coin flip
Correspondence: a.saxe@ucl.ac.uk, fitzgeraldj@janelia.hhmi.org

ABSTRACT

Our ability to remember the past is essential for guiding our future behavior. Psychological and neurobiological features of declarative memories are known to transform over time in a process known as systems consolidation. While many theories have sought to explain the time-varying role of hippocampal and neocortical brain areas, the computational principles that govern these transformations remain unclear. Here we propose a theory of systems consolidation in which hippocampal-cortical interactions serve to optimize generalizations that guide future adaptive behavior. We use mathematical analysis of neural network models to characterize fundamental performance tradeoffs in systems consolidation, revealing that memory components should be organized according to their predictability. The theory shows that multiple interacting memory systems can outperform just one, normatively unifying diverse experimental observations and making novel experimental predictions. Our results suggest that the psychological taxonomy and neurobiological organization of declarative memories reflect a system optimized for behaving well in an uncertain future.

INTRODUCTION

Memory, the process by which experience is stored and transformed in neural circuits, lies at the heart of our ability to make adaptive decisions[1]. It is threaded through cognition, from perception through spatial navigation to decision-making and explicit conscious recall. Befitting the central importance of memory, brain regions including the hippocampus appear specifically dedicated to this challenge[2–4].

The concept of memory has refracted through psychology and neurobiology into diverse subtypes and forms that have been difficult to reconcile. Taxonomies of memory have been drawn on the basis of psychological content, for instance differences between memories for detailed episodes and semantic facts[5]; on the basis of anatomy, for instance differences between memories that are strikingly dependent on hippocampus versus those that are not[6]; and on the basis of computational properties, for instance differences between memories reliant on pattern-separated or distributed neural representation[7]. Many previous theories have tried to align and unify psychological, neurobiological, and computational memory taxonomies[8–12]. However, none yet resolve long-standing debates on where different kinds of memories are stored in the brain, and, fundamentally, why different kinds of explicit memories exist.

Classical views of systems consolidation, such as the standard theory of systems consolidation[8,13,14], have held that memories reside in hippocampus before transferring completely to neocortex (Fig. 1a). Related neural network models, such as the complementary learning systems theory, have further offered a computational rationale for systems consolidation based on the benefits of coupling complementary fast and slow learning systems[9,15]. However, these theories lack explanations for why some memories remain forever hippocampal dependent, as shown in a growing number of experiments[16,17]. On the other hand, more recent theories, like multiple trace theory[11,18] and trace transformation theory[19,20], hold that the amount of consolidation can depend on memory content, but they do not provide a quantitatively clear criterion for what content will consolidate, nor why this might be beneficial to behavior.

One possible way forward is to see that memories serve not only as veridical records of experience, but also to support generalization in novel circumstances[21–26]. Here we introduce a mathematical neural network theory of systems consolidation founded on the principle that memory systems and their interactions collectively optimize generalization. The resulting theory unifies diverse experimental phenomena that have vexed prior theories, explains why multiple interacting brain areas are beneficial, and reveals that the predictability of experiences should determine when and where memories reside. Our results provide a quantitative and unified picture of the organization of explicit memories based on their utility for future adaptive behavior.

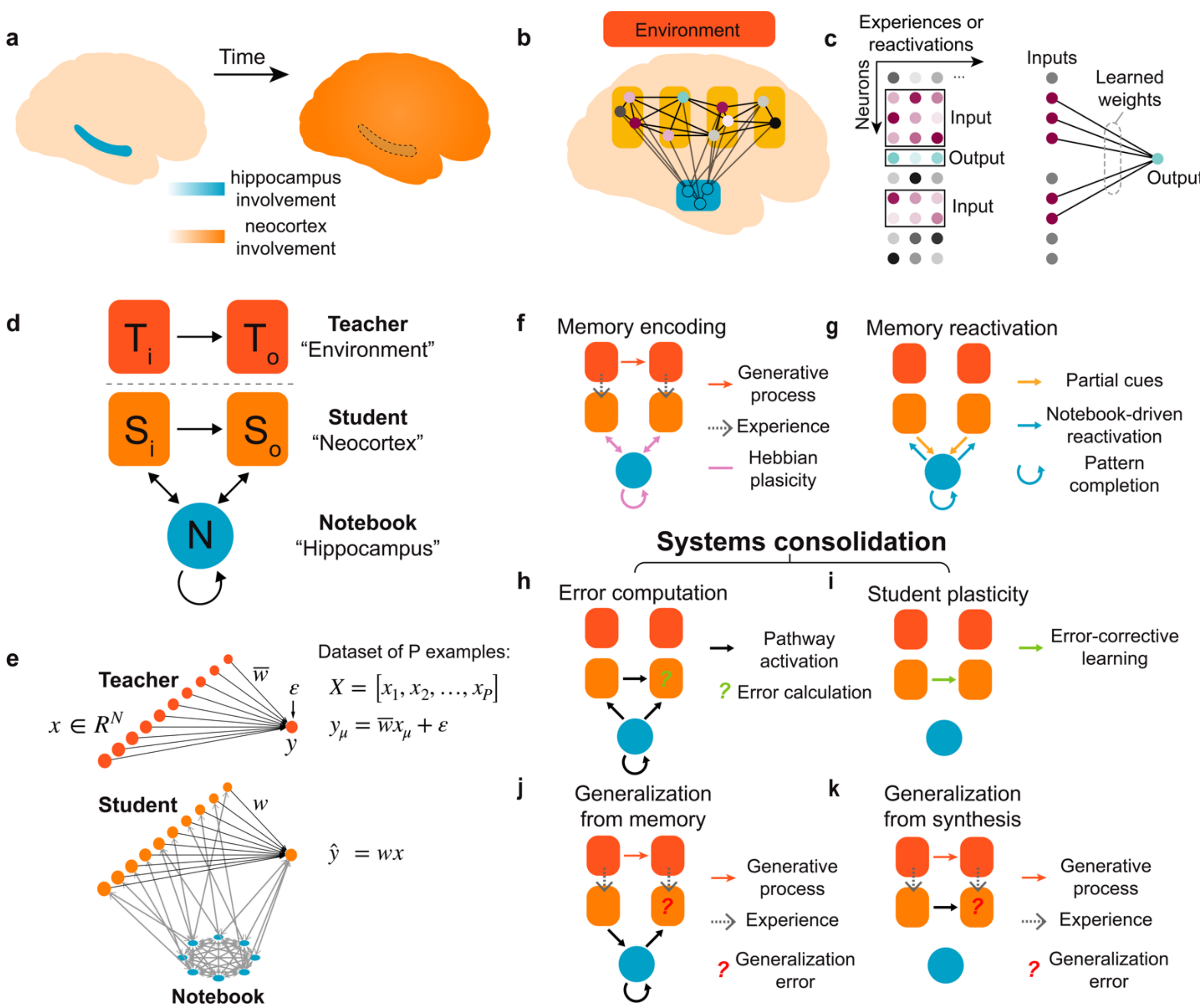

Fig. 1. Neural network model of systems consolidation. (a) The standard theory of systems consolidation. **(b, c)** Our theoretical framework assumes that neocortex extracts and encodes environmental relationships within the weights between distributed neocortical neurons in a process mediated by hippocampal reactivations. **(d)** Cartoon of the teacher-student-notebook formalism; subscripts "i" and "o" refer to input and output layers. **(e)** Core neural network model architecture (see Methods for details). **(f-k)** Stages of learning and inferences in the model.

RESULTS

Formalizing systems consolidation

We conceptualize an animal's experiences in the environment as structured neuronal activity patterns that the hippocampus rapidly encodes and the neocortex gradually learns to produce internally[9,15,27,28] (Fig. 1b). We hypothesize that systems consolidation allows neocortical circuits to learn many structured relationships between different subsets of these active neurons. Focusing on one of these relationships at a time, neocortical circuitry might learn to produce the responses of a particular *output* neuron from the responses of other *input* neurons (Fig. 1c). For example, in a human, an output neuron contributing to a representation of the word "bird" might receive strong inputs from neurons associated with wings and flight. In a mouse, an output neuron associated with freezing might receive strong inputs from neurons associated with the sound of an owl, the smell of a snake, or the features of a laboratory cage where it had been shocked[29].

We first sought to develop a theoretically rigorous mathematical framework to formalize this view of how systems consolidation contributes to learning. Our framework builds on the complementary learning systems hypothesis[9,15], which posits that fast learning in hippocampus guides slow learning in neocortex to provide an integrated learning system that outperforms either subsystem on its own. Here we formalize this notion as a neocortical "student" that learns to predict an environmental "teacher," aided by past experiences recorded in a hippocampal "notebook" (Fig. 1d). We note that other anatomical mappings may be possible[30–32].

We modeled each of these theoretical elements with a simple neural network that permitted analytical analysis (Fig. 1e, Methods). Specifically, we modeled the teacher as a linear feedforward network that generates input-output pairs through fixed weights with additive output noise, the student as a size-matched linear feedforward network with learnable weights[33–35], and the notebook as a sparse Hopfield network[36–38]. The student learns its weights from a finite set of examples (experiences) that contain both signal and noise. We modeled the standard theory of systems consolidation by optimizing weights for memory. This means that the squared difference between the teacher's output and the student's prediction should be as small as possible, averaged across the set of past experiences. Alternatively, we hypothesize that a major goal of the neocortex is to optimize generalization. This means that the squared difference between the teacher's output and the student's prediction should be as small as possible, averaged across possible future experiences that could be generated by the teacher.

Learning starts when the teacher activates student neurons (Fig. 1f, gray arrows). The notebook encodes this student activity by associating it with a random pattern of sparse notebook activity using Hebbian plasticity (Methods; Fig. 1f, pink arrows). This effectively models hippocampal activity as a pattern-separated code for indexing memories[39,40]. The dynamics of the notebook's recurrent notebook network implement pattern completion[36,41], whereby full notebook indices can be reactivated randomly from spontaneous activity or purposefully from partial cues[42] (Methods; Fig. 1g). Student-to-notebook connections allow the student to provide the partial cues that drive pattern completion (Fig. 1g, orange arrows). Notebook-to-student connections then allow the completed notebook index to reactivate whatever student representations were active during encoding (Fig. 1g, blue arrows). Taken together, these three processes permit the student to use the notebook to recall memories from related experiences in the environment. Thus, our theory concretely models how the neocortex could use the hippocampus for memory recall.

We model systems consolidation as plasticity of the student's internal synapses (Figs. 1h, 1i). The student's plasticity mechanism is guided by notebook reactivations (Fig. 1h), similar to how hippocampal replay is hypothesized to contribute to systems consolidation[43]. Slow, error-corrective learning aids generalization[44], and here we adjust internal student weights with gradient-descent learning (Fig. 1i). Specifically, we assume that offline notebook reactivations provide targets for student learning (Methods), where the notebook-reactivated student output is compared with student's internal prediction to calculate an error signal for learning. We consider models that set the number of notebook reactivations to optimize either memory transfer or generalization. Furthermore, the system can use the notebook (Fig. 1j) or only the learned internal student weights (Fig. 1k) to make output predictions from any input generated by the teacher. Our model uses whichever pathway makes statistically better predictions.

Generalization-optimized complementary learning systems (Go-CLS)

We next simulated the dynamics of memorization and generalization in the teacher-student-notebook framework to investigate the impact of systems consolidation. We first modeled the standard theory of systems consolidation as limitless notebook reactivations that optimized student memory recall (Fig. 2a, c, e, Methods). Learning begins when the notebook stores a small batch of examples, which are then repetitively reactivated by the notebook in each epoch to drive student learning (Methods). In separate simulations, examples were generated by one of three teachers that differed in their degree of predictability, here controlled by the signal-to-noise ratio (SNR) of the teacher network's output (Fig. 1e, Methods). The notebook was able to accurately recall the examples provided by each teacher from the beginning (Fig. 2a, c, e, *dashed blue lines*), and we showed mathematically that recall accuracy scaled with the size of the notebook (Supplementary Material 5.2). Notebook-mediated generalization (Student In → Notebook → Student Out) was poor for all three teachers (Fig. 2a, c, e, *dashed red lines*), as rote memorization poorly predicts high-dimensional stimuli that were not previously presented or memorized (Supplementary Material 5.3). The student gradually reproduced past examples accurately (Fig. 2a, c, e, *solid blue lines*), but the signal in each example was contaminated by whatever noise was present during encoding and repetitively replayed throughout learning. Therefore, although the generalization error decreased monotonically for the noiseless teacher (Fig. 2a, *solid red line*), noisy teachers resulted in the student eventually generalizing poorly (Fig. 2c, e, *solid red lines*). From a mathematical point of view, this is expected, as the phenomenon of overfitting to noisy data are well appreciated in statistics and machine learning[45,46].

The implications of these findings for psychology and neuroscience are far reaching, as the standard theory of systems consolidation assumes that generalization follows naturally from hippocampal memorization and replay; it does not consider when systems consolidation is detrimental to generalization. For example, previous neural network models of complementary learning systems focused on learning scenarios where the mapping from input to output was fully reliable[8,9]. A goose is always a bird, and a rose is always a flower. Within our teacher-student-notebook framework, this means that the teacher is noiseless and perfectly predictable by the student architecture. In such scenarios, standard systems consolidation continually improved both memorization and generalization in our model (Fig. 2a, *solid red line*). However, for less predictable environments, our theory suggests that too much systems consolidation can severely degrade generalization performance by leading the neocortex to overfit to unpredictable elements of the environment (Fig. 2c, *solid red*). In highly unpredictable environments, any systems consolidation at all can be detrimental to generalization (Fig. 2e, *solid red*). If the goal of systems consolidation is full memory transfer, then our theory illustrates that the system pays a price in reduced ability to generalize in uncertain environments.

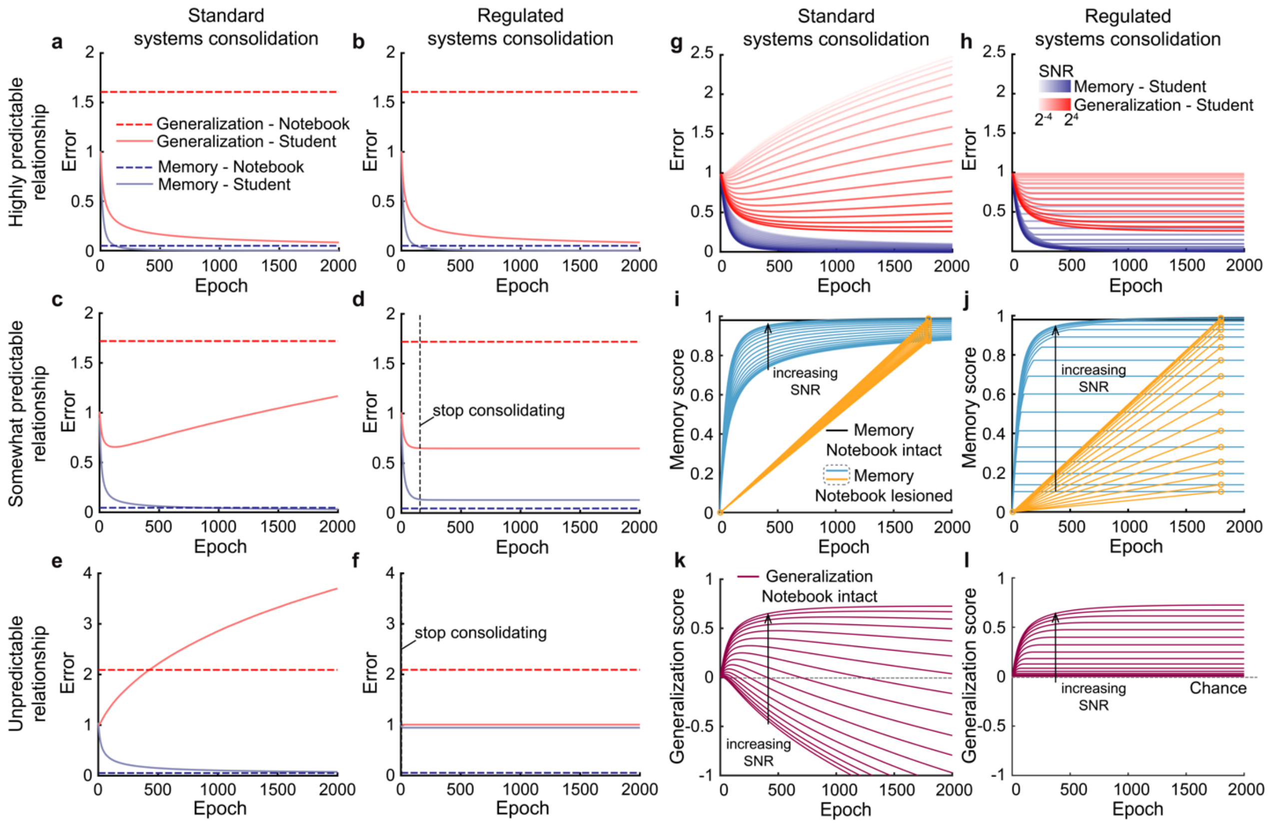

Fig. 2. The predictability of experience controls the dynamics of systems consolidation. (a-h) Dynamics of student generalization error, student memorization error, notebook generalization error, and notebook memorization error when optimizing for student memorization (a, c, e, g) or generalization (b, d, f, h) performance. The student's input dimension $N = 100$, and the number of patterns stored in the notebook $P = 100$ (all encoded at epoch = 1; epochs in the x-axis correspond to the time passage during systems consolidation). Notebook contains $M = 2000$ units, with a sparsity $a = 0.05$. During each epoch, 100 patterns are randomly sampled from the P stored patterns for reactivation and training the student. The student's learning rate is 0.015. Teachers differed in their levels of predictability (a, b: SNR = infinity; c, d: SNR = 4; e, f: SNR = 0.05; g, h: SNR ranges from 2^{-4} to 2^{4}). **(i-l)** Memorization and generalization scores for the integrated student-notebook system as a function of time and SNR, when optimized for student memorization (i, k) or generalization (j, l). Memory and generalization scores are translated from respective error values by Score = $(E_0 - E_t)/E_0$, E_0 and E_t are the generalization or memory errors before training starts and at epoch = t, respectively. Effect of notebook lesion on memory performance (cyan and orange lines, orange lines are simply the two-point version of the cyan lines) depended on optimization objective and time (i, j).

Given the intricate effect of training example replay on generalization, what systems consolidation strategy would optimize generalization? Here we propose generalization-optimized complementary learning systems (Go-CLS) theory, which considers the normative hypothesis that the amount of systems consolidation is adaptively regulated to optimize the student's generalization accuracy based on the predictability of the input-output mapping (Fig. 2b, d, f). For the teacher with high degree of predictability, the student's generalization error always decreased with more systems consolidation (Fig. 2b, *solid red line*), and the student could eventually recall all stored memories (Fig. 2b, *solid blue line*). Memory transfer therefore arises as a property of a student that learns to generalize well from this teacher's examples. In contrast, a finite amount of consolidation (here modeled by a fixed number of notebook reactivations) was necessary to minimize the generalization error when the teacher had limited predictability (Fig. 2d, f), and our normative hypothesis is that systems consolidation halts at the point where further consolidation harms generalization (Fig. 2d, f, *vertical black dashed line*). The resulting student could generalize near optimally from each of the teachers' examples (Fig. 2d, f, *solid red lines,* Supplementary Material 7.2), but its memory performance was hurt by incomplete memorization of the training data (Fig. 2d, f, *solid blue lines*). Nevertheless, the notebook could still recall the memorized examples (Fig. 2b, d, f, *dashed blue lines*). Go-CLS thus results in an integrated system that can both generalize and memorize by using two systems with complementary properties. We note that implementing this strategy for regulated systems consolidation requires a supervisory process capable of estimating the predictability of experience and the optimal amount of consolidation, a topic we address in the Discussion.

These examples show that the dynamics of systems consolidation models interestingly depend on the degree of predictability of the teacher. We therefore leveraged our analytical results to comprehensively compare the standard theory of systems consolidation to the Go-CLS theory for regulated systems consolidation for all degrees of predictability (Supplementary Material 6, 7). Standard systems consolidation eventually consolidated all memories for any teacher (Fig. 2g, blue). As anticipated by Fig. 2a-f, the generalization performance varied dramatically with the teacher's degree of predictability (Fig. 2g, red). Generalization errors were higher for less predictable teachers, and optimal consolidation amounts were lower. Therefore, regulated systems consolidation removed the detrimental effects of overfitting (Fig. 2h, red) but ended before the student could achieve perfect memorization (Fig. 2h, blue, non-zero error). Both the generalization performance and the memory performance improved as the teacher's degree of predictability increased (Fig. 2h).

The experimental literature on the time course of systems consolidation and time-dependent generalization provides important constraints on our theory. We thus sought to model these effects by translating mean square errors (Fig. 2 g, h) into memory retrieval scores, where 0 indicates random performance and 1 indicates perfect performance (Fig. 2i-l, Methods). Our framework can use either the student or the notebook to recall memories or generalize (Fig. 1j, k), and we model the combined system by making predictions with whichever subsystem is more accurate (Methods). We simulated hippocampus lesions by disallowing the combined system from using notebook outputs and ending systems consolidation at the time of the lesion (Fig. 2i, j, cyan). Note that the combined memory (Figs. 2i, j) and generalization scores (Figs. 2k, l) often map onto the notebook and student performances, respectively, but it is also possible for the better subsystem to switch over time (Supplemental Material 6.1). As it takes time for the student to learn accurate generalizations, our systems consolidation models exhibit time-dependent generalization (Fig. 2k, l, purple). In contrast, the notebook permitted accurate memory retrieval from the start (Fig. 2i, j, black).

Standard systems consolidation and regulated systems consolidation make strikingly different predictions for how retrograde amnesia and time-dependent generalization curves depend on the teacher's degree of predictability (Fig. 2i-l). As expected, notebook lesions always produced temporally graded retrograde amnesia curves in the standard theory[13] (Fig. 2i, orange). When systems consolidation was instead optimized for generalization, the effects of notebook lesions depended strongly on the predictability of the teacher (Fig. 2j, orange). In particular, the model could produce both graded and flat retrograde amnesia curves, with the slope of the amnesia curve increasing with the degree of predictability. Diverse generalization curves resulted from either model of systems consolidation (Fig. 2k, l), with maximal generalization performance increasing with the predictability of the teacher. However, student overfitting meant that only regulated systems consolidation maintained this performance over time. Standard systems consolidation could even result in a student that generalized maladaptively, resulting in worse-than-chance performance where the trained student interpolates noise in past examples to produce wildly inaccurate outputs (Fig. 2k).

Go-CLS explains diverse experimental results

These results allow Go-CLS theory to explain diverse experimental findings. Since real world experiences are composed of many elements that differ in their degree of predictability, our theory predicts that different components of human memory will consolidate to different degrees. In human memory research, patients with selective hippocampal damage indeed show retrograde amnesia reflecting diverse dynamics of systems consolidation[17,47]. Researchers usually classify hippocampal amnesia dynamics according to whether memory deficits are similar for recent and remote memories (flat retrograde amnesia), more pronounced for recent memories (graded retrograde amnesia), or absent for both recent and remote memories (no retrograde amnesia) (Fig. 3a). Some patients show graded retrograde amnesia consistent with the standard theory, while others either have flat retrograde amnesia or no retrograde amnesia[17] (Fig. 3b). Regulated systems consolidation can recapitulate this diversity of retrograde amnesia curves (Fig. 3c). High and low predictability experiences lead to graded and flat retrograde amnesia, respectively (Figs. 2j, 3c, *solid lines*). A period of prior consolidation of highly predictable experiences decreases the slope of graded retrograde amnesia (Fig. 3c, *dashed light-blue lines*), and it's possible to see no retrograde amnesia at all when the prior consolidation was extensive (Fig. 3c, *dashed orange line*, Methods). This conceptually resembles schema-consistent learning[48]. Similarly diverse retrograde amnesia curves have been seen in rodent memory tasks (Figs. 3d-g). For example, hippocampal lesions can result in either graded or flat retrograde amnesia in different individuals performing the same task[49–51] (Figs. 3d-f), and individual animals can exhibit different types of amnesia on different tasks[51]. (Figs. 3f, g). In summary, our theory accounts parsimoniously for this wide range of experimental observations through the tuning of two parameters: the predictability of experience and the amount of prior consolidation.

These empirical patterns are interpretable in light of Go-CLS theory. Generally reliable facts about public events and famous faces contain content of high degrees of predictability, and many patients can recall remote facts and faces without a functioning hippocampus[17] (Fig. 3b). In contrast, the idiosyncratic content of an autobiographical memory, such as remembering specific events that happened during a birthday party is much less predictable[52], because many incidental influences shape how complex real-life events unfold. Most patients cannot recall these memories without a hippocampus[17] (Fig. 3b). Similarly, the Morris water maze requires a mouse to remember the detailed arrangement of environmental cues and platform positions[53,54], both chosen arbitrarily and unpredictably by the experimenter, and this task consistently requires the hippocampus[51,55,56] (Figs. 3g).

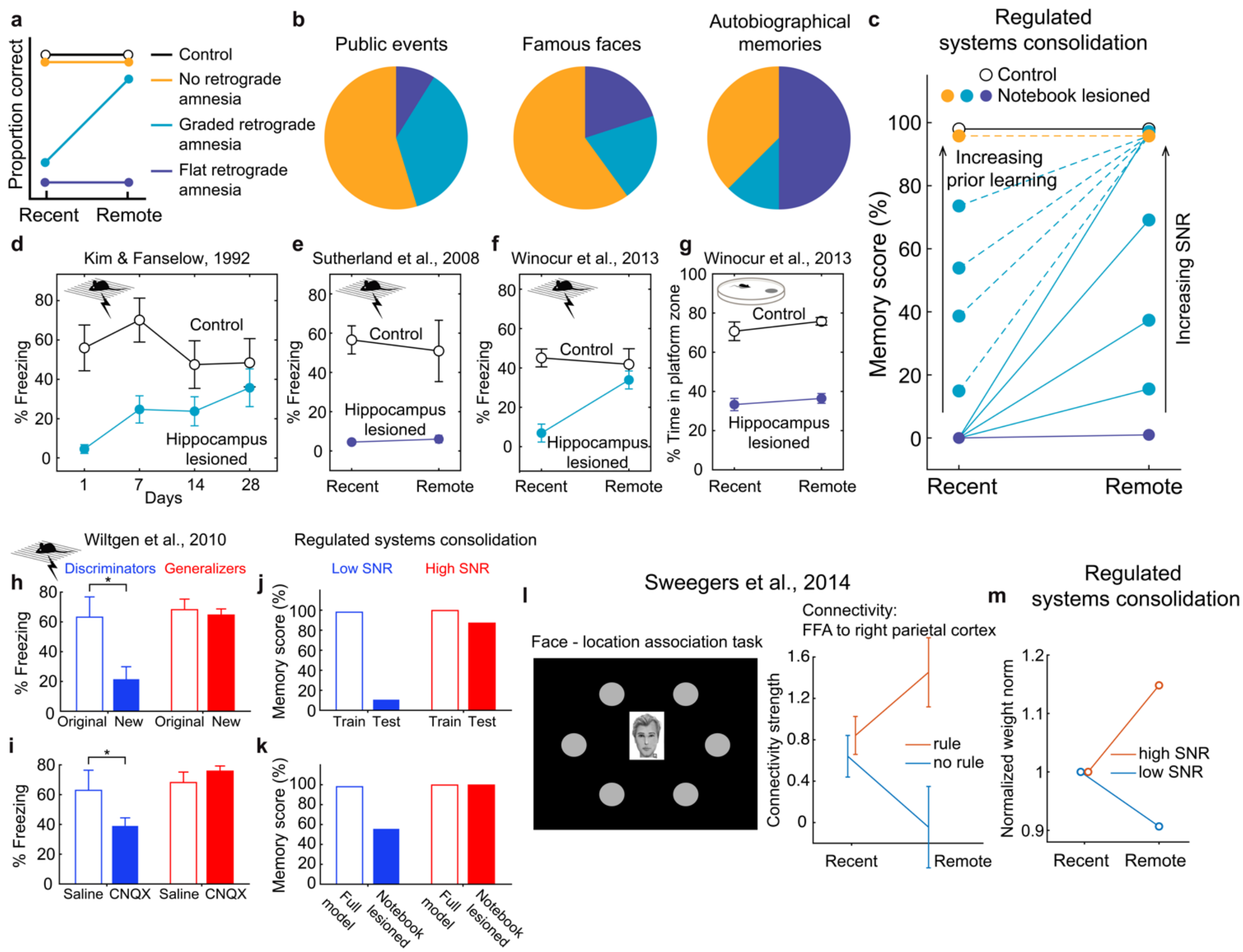

Fig. 3. Regulated systems consolidation mirrors memory research findings in both humans and rodents. (a) Schematic of retrograde amnesia curves. **(b)** Reports of retrograde amnesia in human patients with selective hippocampal damage show diverse dynamics. Figure adapted from Yonelinas et al., 2019[17]. **(c)** Regulated systems consolidation can reproduce the diversity of retrograde amnesia curves (see Methods for model details). **(d, e)** Lesioning hippocampus in rodents can produce both graded and flat retrograde amnesia. Figure adapted from Kim & Fanselow, 1992[57], Sutherland et al., 2008[50]. Lesioning the hippocampus can result in graded **(f)** or flat **(g)** retrograde amnesia in the same animal performing different tasks (contextual fear conditioning and Morris water maze, respectively). Figure adapted from Winocur et al., 2013[51]. **(h)** Discriminators can differentiate the original fear-conditioning context with another similar but novel context, whereas generalizers show similar amount of fear response to both contexts. **(i)** Silencing the hippocampus in mice 15 days after contextual fear conditioning differentially impact fear memory of the original context, depending on whether the animal show time-dependent fear generalization. panels h and i are adapted from Wiltgen et al., 2010[58]. **(j, k)** Regulated systems consolidation can reproduce similar correlation between time-dependent generalization and reduced hippocampal dependence of memories. High SNR (1000) and low SNR (0.6) simulations based on analytical solutions are used to model the “generalizers” and “discriminator”. 2000 total epochs are simulated with $N = P = 100$, notebook size $M = 5000$, and *learnrate* = 0.005. **(l)** Face-location association task with rules vs no rules show different time-dependent change in functional connectivity between cortical areas. Figure adapted from Sweegers et al., 2014[59]. **(m)** Regulated systems consolidation shows similar connectivity changes over time, as reflected in

the norm of the student's weights. Student weight w is drawn *i.i.d.* from $N(0, 0.5)$, where the weights' non-zero initial condition reflect the brain's preexisting connectivity between these two regions. The student then learns from a high SNR teacher (SNR = 2) or a low SNR teacher (SNR = 0.05), while the weight norm is monitored through time (normalized to the initial norm). Note that a decrease in weight norm is expected on the low-SNR learning task, as a large weight norm generates substantial output variance that is uncorrelated with the teacher's noisy output. 2000 total epochs are simulated with $N = P = 100$, notebook size $M = 2000$, and $learnrate = 0.015$.

Go-CLS theory also explains diverse experimental results observed for time-dependent generalization[49,60–62]. For example, some mice showed increased fear responses to similar but not identical contexts in fear-conditioning experiments ("generalizers", Fig. 3h, red bars), while others maintained distinct behavioral responses over time ("discriminators", Fig. 3h, blue bars)[49]. Strikingly, only the discriminators required their hippocampus for memory recall of the original context (Fig. 3i). Our theory predicts that memory transfer and generalization improvement should be similarly correlated (Figs. 3j, k), as the same systems consolidation process leads to both memory transfer and time-dependent generalization. Unpredictable experiences should be susceptible to strong retrograde amnesia and avoid transfer leading to maladaptive generalization (Fig. 3j, k, blue bars). In contrast, predictable experiences should be associated with weak retrograde amnesia and useful learned generalizations (Fig. 3j, k, red bars). As with the amnesia results, our theory explains the diversity of these patterns through variability in the degree of predictability. Later, we further explore the diversity of fear conditioning dynamics (Figs. 3d-f, h, i), and human amnesia curves (Fig. 3b), by explaining why the predictability of experience varies across individuals.

Direct tests of Go-CLS theory require experimental task designs that vary the degree of predictability and assess the effect on systems consolidation. One such experiment has been carried out by Sweegers et al[59] (Fig. 3l, m). In their task design, healthy human participants had to associate specific faces with positions on a computer screen (Fig. 3l, left). Half of the locations were assigned faces through an unpredictable random process, whereas the other locations were assigned faces according to a hidden but fully reliable rule. The authors then used functional magnetic resonance imaging (fMRI) to assess how systems consolidation changed the functional connectivity between the fusiform face area (FFA) and right parietal cortex during memory recall. Remarkably, with time the functional connectivity increased for the rule-based locations and decreased for the no-rule locations (Fig. 3l, right). This result is expected from regulated systems consolidation (Fig. 3m). In particular, the right parietal cortex is involved in spatial processing, and we interpret its functional connectivity from FFA in the teacher-student-notebook model as student weights used to predict neural activity coding location from neural activity coding faces. Our theory predicts that the predictability of the face-location relationship determines whether systems consolidation drives neocortical learning that links FFA to right parietal cortex. Indeed, these connections strengthened only when the face-location relationship was predictable. This empirical difference can be quantitatively captured by regulated systems consolidation (Fig. 3m).

Normative benefits of complementary learning systems for generalization

In addition to reproducing diverse experimental observations, our framework also provides theoretical insights into the complementary learning systems hypothesis, which posits that hippocampal and neocortical systems exploit fundamental advantages provided by coupled fast and slow learning modules[9,15,63–65]. We first investigated its basic premise by comparing generalization in the optimally regulated student-notebook network (Fig. 4a) to what is achievable with isolated student (Fig. 4b) and notebook networks (Fig. 4c). These simpler networks model learning with only neocortex or only hippocampus, respectively.

Both the degree of predictability and the amount of available data impact the time course of systems consolidation in the student-notebook network (Supplementary Material 6, 7), so we used our analytical solutions to systematically examine how late-time memory and generalization jointly depend on the amount of training data and degree of predictability (Supplementary Fig. 2). With just a student (Fig. 4b), the system must learn online from each example with no ability to revisit it. This limitation prevented the optimal student-only network, which modulated its learning rate online to achieve best-case generalization performance (Supplementary Material 4.2), from generalizing as efficiently from predictable teacher-generated data as the optimal student-notebook network (Fig. 4d, *blue vs red curves*). We also confirmed that both networks generalized better than the notebook-only network (Fig. 4d). This is expected, because in high dimensions any new random pattern is almost always far from the nearest memorized pattern (Supplementary Material 5.3); this is the so-called "curse of dimensionality"[66].

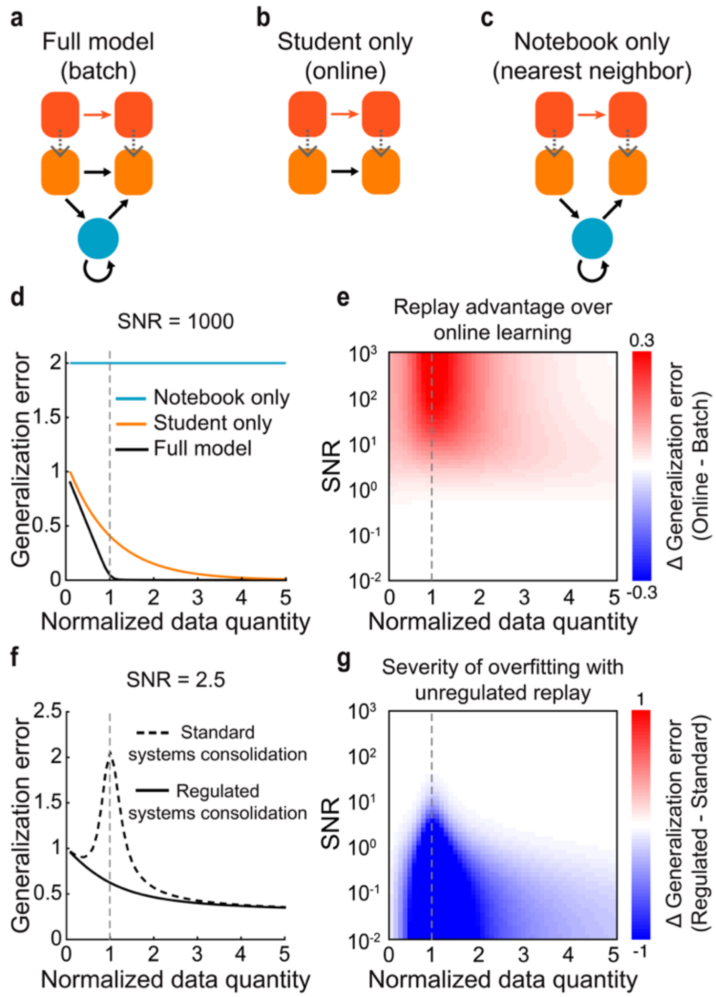

Fig. 4. Normative benefits of complementary learning systems for generalization. (a-c) Schematics illustrating learning systems that can use both the student and the notebook (a), only the student's weights (b), and only the notebook weights (c) for inference. In machine learning terminology, these systems implement batch learning, online learning, and nearest neighbor regression. **(d)** Generalization error as a function of normalized data quantity (or *alpha* (α), defined as $\alpha = P/N$) for each learning system (SNR = 1000), dashed gray line indicates $\alpha = 1$. **(e)** Advantage of regulated systems consolidation over optimal online learning as a function of SNR and normalized data quantity, measured by the difference in generalization error. **(f)** Generalization error as a function of normalized data quantity for the combined system, learning either through regulated systems consolidation or standard systems consolidation (SNR = 2.5). **(g)** Severity of overfitting, measured by the difference in generalization error between standard systems consolidation and regulated systems consolidation.

The generalization gain provided by the student-notebook network over the student-only network was most substantial when the teacher provided a moderate amount of predictable data (Figs. 4d-e, *dashed gray*). This result follows because the student-notebook network was unable to learn much when the data were too few or too noisy, and notebook-driven encoding and reactivation of data was unnecessary when the student had direct access to a large amount of teacher-generated data (Supplementary Material 4, 7). Hence an integrated dual memory system was normatively superior when experience was available, but limited, and the environment was at least somewhat predictable. This ethologically relevant regime is frequently experienced by animals, such as when limited past experiences with predators provide high-fidelity sensory cues for identifying them in the future.

Regulated systems consolidation was most advantageous when the number of memorized examples equaled the number of learnable weights in the student (Fig. 4e, *dashed gray*). Remarkably, this amount of data was also the worst-case scenario for overfitting to noise in standard systems consolidation (Figs. 4f-g, *dashed gray*, Supplementary Fig. 2c), similar to the "double descent" phenomenon in machine learning[67,68], where the worst overfitting also happens at a finite amount of data related to the network size. Intuitively, neural networks must fine-tune their weights to minimize their training error when the number of memorized patterns is close to the maximal achievable number (capacity). This often requires drastic changes in weights to reduce a small training error residual, producing noise-corrupted weights that generalize poorly. The optimal student-notebook network avoided this issue by regulating the amount of systems consolidation according to the predictability of the teacher. We propose that the brain might similarly regulate the amount of systems consolidation according to the predictability of experiences (see Discussion).

Many facets of unpredictability

Our simulations and analytical results show that the degree of predictability controls the consolidation dynamics that optimize generalization. We emphasized the example of a linear student (Fig. 5a) that learns from a noisy linear teacher (Fig. 5b). However, inherent noise is only one of several forms of unpredictability that can cause poor generalization without regulated systems consolidation. For example, when the teacher implements a deterministic transformation

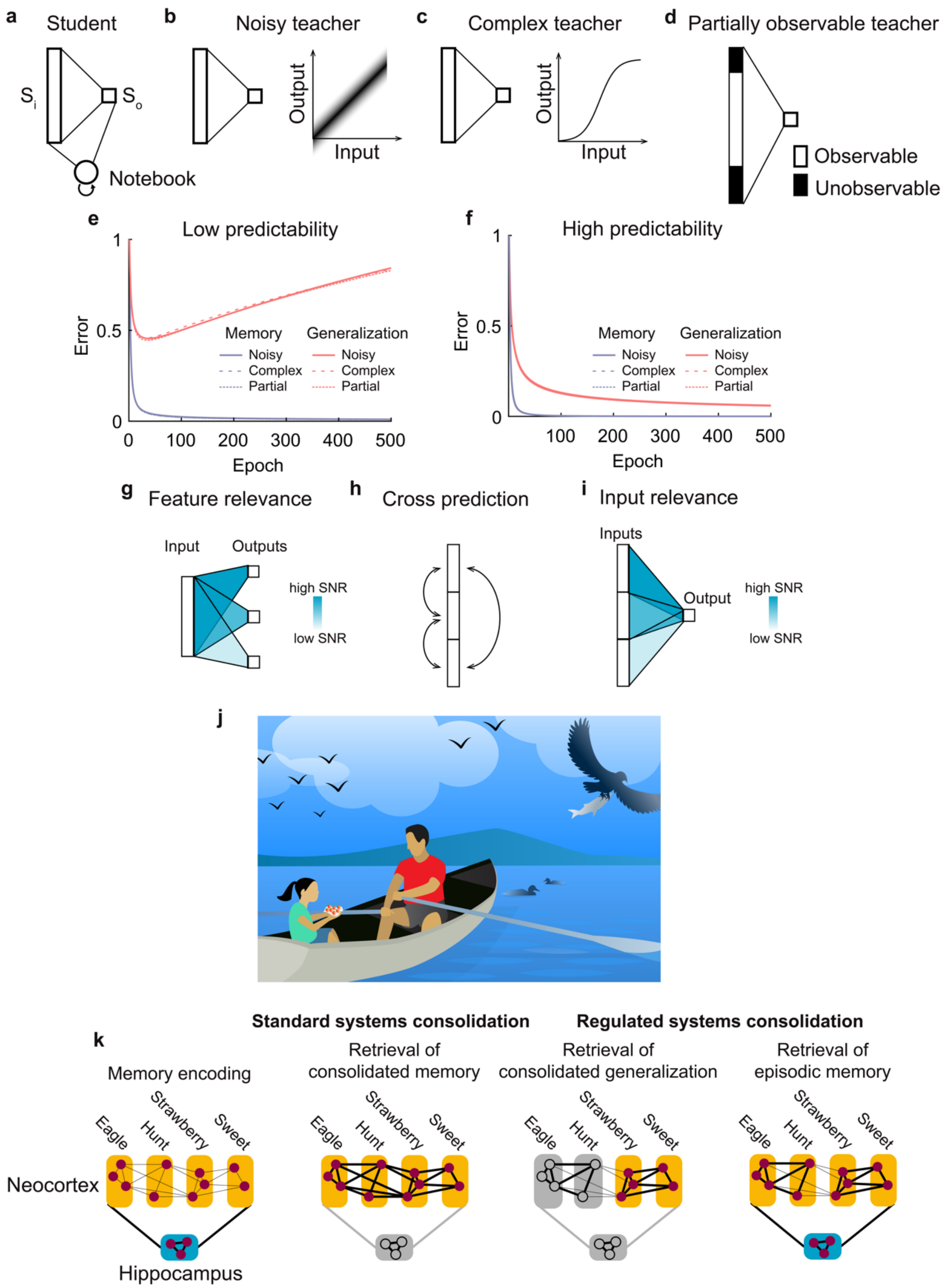

Fig. 5. Many forms of unpredictability demand regulated systems consolidation. (a) The student-notebook learning system. **(b-d)** Example teachers with unpredictable elements. (b) A teacher that linearly transforms inputs into noisy outputs. (c) A teacher that applies a nonlinear activation function at the output unit. (d) A teacher that only partially reveals the relevant inputs to the learning system. **(e, f)** Varying predictability within the three different teachers all lead to quantitatively similar learning dynamics

(complex teacher implements a sine function at the output unit, see Methods for simulation details). **(g-i)** The degree of predictability can vary in many ways. For example, the same inputs can differentially predict various outputs (g), features can cross predict each other with varying levels of predictability (h), and different learning systems could attend to different teacher features to predict the same output (i). **(j)** Cartoon illustrating a child's experience at a lake with her father. **(k)** A cartoon illustrating conceptual differences between what is consolidated in standard systems consolidation and regulated systems consolidation.

that is impossible for the student architecture to implement, the unmodellable parts of the teacher mapping are unpredictable and act like noise (Supplementary Material 9). For instance, a linear student cannot perfectly model a deterministic teacher with nonlinearities (Fig. 5c). Similarly, when the teacher's mapping involves relevant input features that the student cannot observe, the contribution of the unobserved inputs to the output are generally impossible to model (Fig. 5d). This again results in unpredictability from the student's perspective. These sources of unpredictability all consist of a modellable signal and an unmodellable residual (noise) (Supplementary Material 9), and they yield similar training and generalization dynamics (Fig. 5e, f). The real world is noisy and complicated, and the brain's perceptual access to relevant information is limited. Realistic experiences thus frequently combine these sources of unpredictability.

All the above-mentioned cases can be generally understood within the framework of approximation theory[69]. The unmodellable part represents a nonzero optimal approximation error for the student-teacher pair. For this unmodellable part to be generalization limiting, the student must also be expressive enough that the student weights overfit when attempting to fit limited data perfectly. Overfitting is also seen in more complex model architectures, such as modern deep learning models[68] and we expect that the essential concepts presented here will also apply to broader model classes. For all of these types of generalization-limiting unpredictability, generalization is optimized when systems consolidation is limited for unpredictable experiences. Importantly, not all unpredictability limits generalization (Supplementary Material 8).

Previously we have focused on the scenario of learning a single mapping. All real-life experiences are composed of many components, with relationships that can differ in predictability. Many relationships therefore must be learned simultaneously, and these representations are widely distributed across the brain. For instance, the same input features may have different utility in predicting several outputs (Fig. 5g). Furthermore, neocortical circuits may cross-predict between different sets of inputs and outputs (Figs. 1e, 5h); for example, perhaps predicting auditory representations from visual representations and vice versa. In this setting, each cross prediction has its own predictability determined by the noise, the complexity of the mapping, and the features it is based upon. Predictability may also depend on overt and/or covert attention processes in the student. For example, a student may selectively attend to a subset of the inputs it receives (Fig. 5i), making the predictability of the same external experience dependent on internal states that can differ across individuals. A similar attention process could account for the diversity in fear generalization shown in Fig. 3h, if "generalizers" attend to generalizable features in the conditioning chamber while receiving the shock, whereas "discriminators" attend to more unique features. For all the above-mentioned scenarios, Go-CLS theory requires the student to optimize generalization by regulating systems consolidation according to the specific degree of predictability of each modelled relationship contained in an experience. The theory therefore provides a novel predictive framework for quantitatively understanding how diverse relationships within memorized experiences should differentially consolidate to produce optimal general-purpose neocortical representations.

DISCUSSION

The theory presented here — Go-CLS — provides a normative and quantitative framework for assessing the conditions under which systems consolidation is advantageous or deleterious. As such, it differs from previous theories that sought to explain experimental results without explicitly considering when systems consolidation could be counterproductive[8,9,13,17–19]. The central claim of this work is that systems consolidation from the hippocampus to neocortex is most adaptive if it is regulated such that it improves generalization, an essential ability enabling animals to make predictions that guide behaviors promoting survival in an uncertain world. Crucially, we show that unregulated systems consolidation results in inaccurate predictions by neural networks when limited data contain a mixture of predictable and unpredictable components. These errors result directly from the well-known overfitting problem that occurs in artificial neural networks when weights are fine-tuned to account for data containing noise and/or unlearnable structure[34,35,45,46,68,70–72].

For example, consider the experience of a girl spending a day in a boat with her father (Fig. 5j, k). It may contain predictable relationships about birds flying, swimming, and perhaps even catching fish, as well as predictable relationships about fresh-picked strawberries tasting sweet. Our theory posits that these relationships should be extracted from the experience and integrated with memories of related experiences, through regulated systems consolidation, to produce, reinforce, and revise predictions (generalizations). On the other hand, unpredictable correlations, such as the color of her father's shirt matching the color of the strawberries, should not be consolidated in the neocortex. They could nevertheless remain part of an episodic memory of the day, which would reside permanently in the hippocampus.

Go-CLS reconciles many previous experimental results and highlights the normative benefits of complementary learning systems. It explains the diversity of retrograde amnesia dynamics in both humans and animal studies[17,47] (Fig. 3a-g), the intriguing correlation between fear generalization and memory preservation after hippocampal lesioning[49] (Fig. 3h-k), and different functional connectivity changes for learning rule-based or rule-lacking tasks[59] (Fig. 3l, m). It also makes testable predictions that could affirm or refute the theory (Supplementary Material 10). Moreover, Go-CLS provides novel insights into the normative benefits of a dual-memory system[9,15]. Specifically, gradual consolidation of past experiences benefits generalization performance the most when experience is limited and relationships are partially predictable (Fig. 4), mirroring ethologically realistic regimes experienced by animals living in an uncertain world[73]. In addition, this benefit occurs in a regime where the danger overfitting is the highest[34,35,68,70–72], highlighting the need for a regulated systems consolidation process.

Previous theories have also sought to reconcile these and other experimental observations. Several early models—standard systems consolidation[8,74] and complementary learning systems (CLS)[9,15]—posited gradual consolidation of memories from the hippocampus to neocortex. These theories have been highly influential, including motivating other theories that have attempted to address experimental demonstrations of the permanent role of the hippocampus in episodic memory. For example, multiple trace theory[18] and trace transformation theory[19,20] posit that episodic memories are consolidated as multiple memory traces, with the most detailed components permanently residing in the hippocampus. Contextual binding theory[17] posits that items and their context remain permanently bound together in the hippocampus. These theories emphasize the role of the hippocampus in permanent storage of episodic details[17–19,52,75], with the neocortex storing less detailed semantic components of memories. In contrast, Go-CLS posits that predictability, rather than detail, determines consolidation. Similarly, Go-CLS favors predictability over frequency (or feature overlap) or salience as the central determinant of systems consolidation[76–78]. For example, frequent misinformation should not be consolidated, whereas rare gems from a wise source should be. Similarly, emotionally salient events might be prioritized for memory retention[78,79], but only the predictable relationships contained in the experience should be consolidated in the neocortex.

The Go-CLS theory does not specify the mechanisms by which memory consolidation should be regulated. Given the prominent role of sequence replay in existing mechanistic hypotheses about systems consolidation, this would be a natural target for regulation[80–89]. One possibility would be that memory elements reflecting predictable relationships could be replayed together, while unrelated elements are left out or replayed separately. Another would be that entire experiences are replayed, while other processes (e.g., attention mechanisms enabled by the prefrontal cortex[90,91]) regulate how replayed events are incorporated into neocortical circuits that store generalizations[52]. Regardless of the role of replay, a central question is how collections of elements are selected for systems consolidation. We posit that numerous bottom-up and top-down processes could participate in this process. For example, innate attention to facial features or other biologically salient cues could be prioritized[92], and lifelong meta-learning[93,94] could shape regulatory processes that label groups of elements as likely (or not) to contain predictable relationships.

The proposed principle that the degree of predictability regulates systems consolidation reveals complexities about the traditional distinctions between empirically defined episodic and semantic memories[5]. Most episodic memories contain both predictable and unpredictable elements. Unpredictable coincidences in place, time, and content are fundamentally caused by the complexity of the world, which animals cannot fully discern or model. Memorizing such unpredictable events in the hippocampus is consistent with previous proposals suggesting that the hippocampus is essential for incidental conjunctive learning[64], associating discontiguous items[95], storing flexible associations of arbitrary elements[10], relational or configural information[96], and high-resolution binding[97]. However, our theory holds that predictable components of these episodic memories would consolidate separately to inform generalizations. In addition, some key aspects of an experience might themselves be semanticized[98,99] — e.g., by consolidating the *fact* that x, y, and z happened together at time t, or in sequence at times t_1, t_2, and t_3, rather than consolidating the *event* as a remembered experience. In our model, in the extreme case of fully predictable events, the consolidated version fully captures the memorized event. However, real-world events may never be 100% predictable. Furthermore, whether semanticized information about an experience should be considered an episodic memory is a matter of debate. We anticipate that psychologists and neurobiologists will be motivated by the Go-CLS theory to test and challenge it, with the long-range goal of providing new conceptual insight into the organizational principles and biological implementation of memory.

ACKNOWLEDGMENTS

The authors thank Tim Behrens, Brad Hulse, David Kastner, Jay McClelland, and Sandro Romani for helpful comments on the manuscript. We thank Julia Kuhl for illustrations and helpful discussions of the figures. This work was supported by the Howard Hughes Medical Institute (WS, NS, JEF) and the Janelia Visiting Scientist Program (AS). MA and AS were additionally supported by the Swartz Foundation, and AS by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 216386/Z/19/Z), the Gatsby Charitable Foundation, and the CIFAR Azrieli Global Scholars program.

METHODS

Teacher-Student-Notebook framework

Please refer to the Supplementary Material for a detailed description of the Teacher-Student-Notebook framework. The following sections provide a brief description of the framework and simulation details.

Architecture. The teacher network is usually a linear shallow neural network generating input-output pairs $\{x^\mu, y^\mu\}$, $\mu = 1,\cdots, P$, through $y^\mu = \bar{w}x^\mu + \varepsilon$, as training examples. Components of the teacher's weight vector, $\bar{w}$, are drawn *i.i.d.* from $N(0, \sigma^2_w)$. ε is a Gaussian additive noise drawn *i.i.d.* from $N(0, \sigma^2_\varepsilon)$. The signal-to-noise ratio (SNR) of the teacher's mapping is defined as SNR $= \sigma^2_w/\sigma^2_\varepsilon$., and we set $\sigma^2_w + \sigma^2_\varepsilon = 1$ to generate output examples of unit variance. For the simulations in Figs. 2, 3, and 4, the student is a linear shallow neural network whose architecture matches the teacher (both with input dimension = 100 and output dimension = 1). We relaxed this requirement in Fig. 5 to allow mismatch between the teacher and student architectures (see **Generative models** section below). Student's weight vector, w, is initialized as zeros (*i.e. tabula rasa*), unless otherwise noted. The notebook is a sparse Hopfield network containing M binary units (states can be 0 or 1, M = 2000 to 5000 unless otherwise noted). The input and output layers of the student network are bidirectionally connected to the notebook with all-to-all connections.

Training procedure. Training starts with the teacher network generating P input-output pairs, with certain predictability (SNR), as described above. For each of these P examples, the teacher activates the student's input and output layers via identity mapping; at the same time, the notebook randomly generates a binary activity pattern, ξ^μ, $\mu = 1,\cdots, P$, with sparsity a, such that exactly aM units are in the 1 state for each memory. At each of the example presentations, all of the notebook-to-notebook recurrent weights and the student-to-notebook and notebook-to-student interconnection weights undergo Hebbian learning (Supplementary Material 1). This Hebbian learning essentially encodes ξ^μ as an attractor state and associates it with the student's activation $\{x^\mu, y^\mu\}$, for $\mu = 1,\cdots, P$.

After all P examples are encoded through this one-shot Hebbian learning, at each of the following training epochs, P notebook-encoded attractors are randomly retrieved by initializing the notebook with random patterns and letting the network settle into an attractor state through its recurrent dynamics. Notebook activations are updated synchronously for 9 recurrent activation cycles, and we found that each memory was activated with near uniform probability. Once an attractor is retrieved, it activates the student's input and output layers through notebook-to-student weights. Since the number of patterns is far smaller than the number of notebook units ($P << M$) in our model, the Hopfield network is well below capacity, and most of the retrieved attractors were perfect recalls of the original encoded indices. However, real hippocampal networks exhibit active forgetting that may enhance generalization or memory capacity[21,100], and it would be interesting to consider alternate notebook models that incorporate forgetting effects[101]. Reactivation of the student's output through the notebook, $\tilde{y}^{\mu}$, is then compared to the original output, y^{μ}, activated by the teacher to calculate how well the reactivation resembles the original experience, quantified as the mean squared error. For error corrective learning, the student uses the notebook reactivated $\tilde{x}^{\mu}$ and $\tilde{y}^{\mu}$. By comparing the student output that is generated by the reactivated $\hat{\tilde{y}}^{\mu} = w\tilde{x}^{\mu}$ and the reactivated student output $\tilde{y}^{\mu}$ for all P examples, the student updates w using gradient descent with $\frac{1}{P}\sum_{\mu=1}^{P} (\tilde{y}^{\mu} - \hat{\tilde{y}}^{\mu})^2$. The weight update follows:

$$\Delta w = learnrate \times \left(\tilde{Y}\tilde{X}^T - w\tilde{X}\tilde{X}^T\right),$$

where $\tilde{X}$ and $\tilde{Y}$ are the column wise stacked matrix form of the 100 reactivated input and output data points, respectively. Training continues for 500-5000 epochs, and *learnrate* ranges from 0.005 to 0.1. P_{test} number (typically 1000) of additional teacher-generated examples are used to numerically estimate the generalization error at each time step by $\frac{1}{P_{test}}\sum_{\mu=1}^{P_{test}} ({y_{test}}^{\mu} - w{x_{test}}^{\mu})^2$. For some simulations we have applied optimal early-stopping regularization, where we stop the training when the generalization error reaches minimum.

Retrograde amnesia curves. We draw the following connections from network performance in terms of mean squared error to memory and generalization scores, which is typically measured by behavior responses in a task designed to test memorization or generalization performances. At the time when no training occurs, the network error corresponds to a random performance in a task, which is typically set as the zero of a memory retrieval metric. As the error decreases with training, the error is related to the memory retrieval score as follows: score = $(E_0 - E_t)/E_0$, where E stands for memorization error or generalization error, and the subscript indicates the timestep during training. This is stating that the memory retrieval score at each time point is negatively correlated to the error at that time and normalized into a range where 0 indicates chance performance and 1 indicates perfect performance. During memory retrieval, the system chooses whichever available module with lower memorization error. Due to the poor generalization performance of the notebook, we assume the system only uses the student for predicting novel examples. To simulate notebook lesioning at time t, the system starts to use only the student for memory recall, in addition, student's memory score will remain unchanged with time due to the lack of notebook-mediated systems consolidation. In Fig. 3c, both the SNR and amount of prior learning were varied to produce the diverse shapes of retrograde amnesia curves. For the control simulation, SNR is set to infinity. For the solid lines of retrograde amnesia curves, SNR values are 0.01, 0.1, 0.3, 1, and 8. SNR is set to 50 for the dotted lines simulating the effect of prior consolidation. Each line is a different simulation with the amount of prior consolidation ranging from 8 epochs to 2000 epochs ($learnrate = 0.005$). $N = 100$ and notebook size $M = 5000$. $P = 100$ for the varying SNR simulations and $P = 300$ for varying prior consolidation simulations.

Generative models for diverse teachers. To explore different ways unpredictability can exist in the environment, we generalize the teacher-student-notebook framework by relaxing the linear and size matched settings to allow for more complex teachers as generative models for producing training data. For the nonlinear teacher setting, a nonlinear activation function is applied to the linear transformation generate the teacher's output. A sine function was chosen for the simulation in Fig. 5e. The corresponding noisy teacher's SNR is numerically determined from the complex teacher's nonlinearity detailed in Supplementary Material, Section 9. For the partially observable teacher, the input layer is larger than the student's, and the student can only perceive a fixed subregion of the teacher input layer. The exact size of the partially observable teacher is set to match the calculated equivalent SNR of the complex teacher.

Code availability:

Code reproducing the results is available at https://github.com/neuroai/Go-CLS

REFERENCES

1. Shohamy, D. & Daw, N. D. Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences* **5**, 85–90 (2015).
2. *The Hippocampus Book*. (Oxford University Press, 2006). doi:10.1093/acprof:oso/9780195100273.001.0001.
3. Lisman, J. *et al.* Viewpoints: how the hippocampus contributes to memory, navigation and cognition. *Nat Neurosci* **20**, 1434–1447 (2017).
4. Scoville, W. B. & Milner, B. LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *J Neurol Neurosurg Psychiatry* **20**, 11–21 (1957).
5. Tulving, E. Episodic and semantic memory. in *Organization of memory* xiii, 423–xiii, 423 (Academic Press, 1972).
6. Rubin, R. D. & Cohen, N. J. Insights into Hippocampal-Dependent Declarative Memory: Recent Findings and Clinical Implications. *ASHA Wire* https://pubs.asha.org/doi/pdf/10.1044/nnsld24.2.34 (2018) doi:10.1044/nnsld24.2.34.
7. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends in Neurosciences* **34**, 515–525 (2011).
8. Alvarez, P. & Squire, L. R. Memory consolidation and the medial temporal lobe: a simple network model. *PNAS* **91**, 7041–7045 (1994).
9. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. 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**, 419–457 (1995).
10. Moscovitch, M., Cabeza, R., Winocur, G. & Nadel, L. Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annu Rev Psychol* **67**, 105–134 (2016).

11. Nadel, L., Samsonovich, A., Ryan, L. & Moscovitch, M. Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**, 352–368 (2000).

12. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience* **7**, 286–294 (2004).

13. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).

14. Squire, L. R., Genzel, L., Wixted, J. T. & Morris, R. G. Memory Consolidation. *Cold Spring Harb Perspect Biol* **7**, a021766 (2015).

15. Kumaran, D., Hassabis, D. & McClelland, J. L. What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences* **20**, 512–534 (2016).

16. Sutherland, R. J., Sparks, F. T. & Lehmann, H. Hippocampus and retrograde amnesia in the rat model: a modest proposal for the situation of systems consolidation. *Neuropsychologia* **48**, 2357–2369 (2010).

17. Yonelinas, A. P., Ranganath, C., Ekstrom, A. D. & Wiltgen, B. J. A contextual binding theory of episodic memory: systems consolidation reconsidered. *Nat Rev Neurosci* **20**, 364–375 (2019).

18. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).

19. Winocur, G., Moscovitch, M. & Bontempi, B. Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia* **48**, 2339–2356 (2010).

20. Winocur, G. & Moscovitch, M. Memory Transformation and Systems Consolidation. *J Int Neuropsychol Soc* **17**, 766–780 (2011).

21. Richards, B. A. & Frankland, P. W. The Persistence and Transience of Memory. *Neuron* **94**, 1071–1084 (2017).

22. Cowan, E. T., Schapiro, A. C., Dunsmoor, J. E. & Murty, V. P. Memory consolidation as an adaptive process. *Psychon Bull Rev* (2021) doi:10.3758/s13423-021-01978-x.

23. Momennejad, I. Learning Structures: Predictive Representations, Replay, and Generalization. *Current Opinion in Behavioral Sciences* **32**, 155–166 (2020).

24. Mack, M. L., Love, B. C. & Preston, A. R. Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters* **680**, 31–38 (2018).

25. Dudai, Y. & Carruthers, M. The Janus face of Mnemosyne. *Nature* **434**, 567–567 (2005).

26. Schacter, D. L., Addis, D. R. & Buckner, R. L. Remembering the past to imagine the future: the prospective brain. *Nat Rev Neurosci* **8**, 657–661 (2007).

27. Rudy, J. W. & O'Reilly, R. C. Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behav Neurosci* **113**, 867–880 (1999).

28. Takehara-Nishiuchi, K. & McNaughton, B. L. Spontaneous changes of neocortical code for associative memory during consolidation. *Science* **322**, 960–963 (2008).

29. Kitamura, T. *et al.* Engrams and Circuits Crucial for Systems Consolidation of a Memory. *Science* **356**, 73–78 (2017).

30. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160049 (2017).

31. Kumaran, D. & McClelland, J. L. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol Rev* **119**, 573–616 (2012).

32. McNaughton, B. L. Cortical hierarchies, sleep, and the extraction of knowledge from memory. *Artificial Intelligence* **174**, 205–214 (2010).

33. Seung, H. S., Sompolinsky, H. & Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A* **45**, 6056–6091 (1992).

34. Advani, M. S., Saxe, A. M. & Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks* **132**, 428–446 (2020).

35. Krogh, A. & Hertz, J. A. Generalization in a linear perceptron in the presence of noise. *J. Phys. A: Math. Gen.* **25**, 1135–1147 (1992).

36. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A* **79**, 2554–2558 (1982).

37. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Information storage in neural networks with low levels of activity. *Phys. Rev. A* **35**, 2293–2303 (1987).

38. Buhmann, J., Divko, R. & Schulten, K. Associative memory with high information content. *Phys Rev A Gen Phys* **39**, 2689–2692 (1989).

39. Teyler, T. J. & DiScenna, P. The hippocampal memory indexing theory. *Behav. Neurosci.* **100**, 147–154 (1986).

40. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus* **17**, 1158–1169 (2007).

41. Rolls, E. The mechanisms for pattern completion and pattern separation in the hippocampus. *Front. Syst. Neurosci.* **7**, (2013).

42. Rothschild, G., Eban, E. & Frank, L. M. A cortical-hippocampal-cortical loop of information processing during memory consolidation. *Nat Neurosci* **20**, 251–259 (2017).

43. Ólafsdóttir, H. F., Bush, D. & Barry, C. The Role of Hippocampal Replay in Memory and Planning. *Curr Biol* **28**, R37–R50 (2018).

44. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

45. MacKay, D. J. C. *Information Theory, Inference and Learning Algorithms*. (Cambridge University Press, 2003).

46. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, 2009).

47. Frankland, P. W., Teixeira, C. M. & Wang, S.-H. Grading the gradient: Evidence for time-dependent memory reorganization in experimental animals. *Debates in Neuroscience* **1**, 67–78 (2007).

48. Tse, D. *et al.* Schemas and memory consolidation. *Science* **316**, 76–82 (2007).

49. Wiltgen, B. J. *et al.* The hippocampus plays a selective role in the retrieval of detailed context memories. *Curr Biol* **20**, 1336–1344 (2010).

50. Sutherland, R. J., O'Brien, J. & Lehmann, H. Absence of systems consolidation of fear memories after dorsal, ventral, or complete hippocampal damage. *Hippocampus* **18**, 710–718 (2008).

51. Winocur, G., Sekeres, M. J., Binns, M. A. & Moscovitch, M. Hippocampal lesions produce both nongraded and temporally graded retrograde amnesia in the same rat. *Hippocampus* **23**, 330–341 (2013).

52. Gilboa, A. & Moscovitch, M. No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron* **109**, 2239–2255 (2021).

53. Morris, R. Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods* **11**, 47–60 (1984).

54. Richards, B. A. *et al.* Patterns across multiple memories are identified over time. *Nature Neuroscience* **17**, 981–986 (2014).

55. Bolhuis, J. J., Stewart, C. A. & Forrest, E. M. Retrograde Amnesia and Memory Reactivation in Rats with Ibotenate Lesions to the Hippocampus or Subiculum. *The Quarterly Journal of Experimental Psychology Section B* **47**, 129–150 (1994).

56. Ocampo, A. C., Squire, L. R. & Clark, R. E. Hippocampal area CA1 and remote memory in rats. *Learn Mem* **24**, 563–568 (2017).

57. Kim, J. J. & Fanselow, M. S. Modality-specific retrograde amnesia of fear. *Science* **256**, 675–677 (1992).

58. Wiltgen, B. J. *et al.* The hippocampus plays a selective role in the retrieval of detailed context memories. *Curr Biol* **20**, 1336–1344 (2010).

59. Sweegers, C. C. G., Takashima, A., Fernández, G. & Talamini, L. M. Neural mechanisms supporting the extraction of general knowledge across episodic memories. *Neuroimage* **87**, 138–146 (2014).

60. Wiltgen, B. J. & Silva, A. J. Memory for context becomes less specific with time. *Learn. Mem.* **14**, 313–317 (2007).

61. de Voogd, L. D. *et al.* The role of hippocampal spatial representations in contextualization and generalization of fear. *NeuroImage* **206**, 116308 (2020).

62. Biedenkapp, J. C. & Rudy, J. W. Context preexposure prevents forgetting of a contextual fear memory: implication for regional changes in brain activation patterns associated with recent and remote memory tests. *Learning & Memory (Cold Spring Harbor, N.Y.)* **14**, 200–203 (2007).

63. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

64. O'Reilly, R. C., Bhattacharyya, R., Howard, M. D. & Ketz, N. Complementary learning systems. *Cogn Sci* **38**, 1229–1248 (2014).

65. Norman, K. A. & O'Reilly, R. C. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review* **110**, 611–646 (2003).

66. Bellman, R. Dynamic programming. *Science* **153**, 34–37 (1966).

67. Advani, M. S. & Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667 [physics, q-bio, stat]* (2017).

68. Nakkiran, P. *et al.* Deep Double Descent: Where Bigger Models and More Data Hurt. *arXiv:1912.02292 [cs, stat]* (2019).

69. Engel, A. & Broeck, C. V. den. *Statistical Mechanics of Learning*. (Cambridge University Press, 2001).

70. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS* **116**, 15849–15854 (2019).

71. Spigler, S. *et al.* A jamming transition from under- to over-parametrization affects loss landscape and generalization. *J. Phys. A: Math. Theor.* **52**, 474001 (2019).

72. Cun, Y. L., Kanter, I. & Solla, S. A. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters* **66**, 2396–2399 (1991).

73. Trimmer, P. C. *et al.* Decision-making under uncertainty: biases and Bayesians. *Anim Cogn* **14**, 465–476 (2011).

74. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).

75. Moscovitch, M. & Gilboa, A. Systems consolidation, transformation and reorganization: Multiple Trace Theory, Trace Transformation Theory and their Competitors. (2021) doi:10.31234/osf.io/yxbrs.

76. Singer, A. C. & Frank, L. M. Rewarded outcomes enhance reactivation of experience in the hippocampus. *Neuron* **64**, 910–921 (2009).

77. Tompary, A. & Davachi, L. Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. *Neuron* **96**, 228-241.e5 (2017).

78. Kuriyama, K., Soshi, T., Fujii, T. & Kim, Y. Emotional memory persists longer than event memory. *Learn. Mem.* **17**, 130–133 (2010).

79. Conway, A. R. A., Skitka, L. J., Hemmerich, J. A. & Kershaw, T. C. Flashbulb memory for 11 September 2001. *Applied Cognitive Psychology* **23**, 605–623 (2009).

80. Carr, M. F., Jadhav, S. P. & Frank, L. M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat Neurosci* **14**, 147–153 (2011).

81. Ólafsdóttir, H. F., Bush, D. & Barry, C. The Role of Hippocampal Replay in Memory and Planning. *Curr Biol* **28**, R37–R50 (2018).

82. Buzsáki, G. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073–1188 (2015).

83. Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C. & Norman, K. A. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat Commun* **9**, 3920 (2018).

84. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Hippocampal Replay is Not a Simple Function of Experience. *Neuron* **65**, 695–705 (2010).

85. Denis, D. *et al.* The roles of item exposure and visualization success in the consolidation of memories across wake and sleep. *Learn. Mem.* **27**, 451–456 (2020).

86. Noack, H., Doeller, C. F. & Born, J. Sleep strengthens integration of spatial memory systems. *Learn Mem* **28**, 162–170 (2021).

87. Lewis, P. A. & Durrant, S. J. Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences* **15**, 343–351 (2011).

88. Durrant, S. J., Cairney, S. A., McDermott, C. & Lewis, P. A. Schema-conformant memories are preferentially consolidated during REM sleep. *Neurobiology of Learning and Memory* **122**, 41–50 (2015).

89. Durrant, S. J., Cairney, S. A. & Lewis, P. A. Cross-modal transfer of statistical information benefits from sleep. *Cortex* **78**, 85–99 (2016).

90. Aly, M. & Turk-Browne, N. B. Attention promotes episodic encoding by stabilizing hippocampal representations. *Proc Natl Acad Sci U S A* **113**, E420-429 (2016).

91. Aly, M. & Turk-Browne, N. B. How Hippocampal Memory Shapes, and Is Shaped by, Attention. in *The Hippocampus from Cells to Systems* (eds. Hannula, D. E. & Duff, M. C.) 369–403 (Springer International Publishing, 2017). doi:10.1007/978-3-319-50406-3_12.

92. Silva, B. A., Gross, C. T. & Gräff, J. The neural circuits of innate fear: detection, integration, action, and memorization. *Learn Mem* **23**, 544–555 (2016).

93. Wang, J. X. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences* **38**, 90–95 (2021).

94. Lu, Q., Hasson, U. & Norman, K. A. *When to retrieve and encode episodic memories: a neural network model of hippocampal-cortical interaction*. 2020.12.15.422882 https://www.biorxiv.org/content/10.1101/2020.12.15.422882v2 (2021) doi:10.1101/2020.12.15.422882.

95. Wallenstein, G. V., Hasselmo, M. E. & Eichenbaum, H. The hippocampus as an associator of discontiguous events. *Trends in Neurosciences* **21**, 317–323 (1998).

96. Whittington, J. C. R. *et al.* The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell* **183**, 1249-1263.e23 (2020).

97. Yonelinas, A. P. The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behav. Brain Res.* **254**, 34–44 (2013).

98. Lehmann, H. *et al.* Making context memories independent of the hippocampus. *Learning & Memory* **16**, 417–420 (2009).

99. Cermak, L. S. The episodic semantic distinction in amnesia. *Squire, L R And N Butters (Ed) Neuropsychology Of Memory Xv+655p The Guilford Press* 55–62 (1984).

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

101. Mézard, M., Nadal, J. P. & Toulouse, G. Solvable models of working memories. *J. Phys. France* **47**, 1457–1462 (1986).

Supplementary Material: Organizing memories for generalization in complementary learning systems

Weinan Sun[1], Madhu Advani[2], Nelson Spruston[1], Andrew Saxe[2,3,4,5*], and James E. Fitzgerald[1*]

[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA
[2]Center for Brain Science, Harvard University, Cambridge, MA, USA
[3]Department of Experimental Psychology, University of Oxford, Oxford, UK
[4]Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, UCL, London, UK
[5]CIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada

[*]Equal contribution, order determined by coin flip
Correspondence: a.saxe@ucl.ac.uk, fitzgeraldj@janelia.hhmi.org

1 The Teacher-Student-Notebook framework

We consider a setting in which an agent receives experience about a relationship in the environment in the form of P pairs of activity patterns $\{x^\mu, y^\mu\}, \mu = 1, \cdots, P$. Given the *input* activity vector $x^\mu \in R^N$ of dimension N, the agent desires to both memorize the associated scalar *output* activity y^μ, and develop the ability to predict outputs for new, unseen inputs. For example, x^μ might represent activity in visual cortex in response to an event like seeing a bird, and y^μ might represent activity in a higher association cortex derived from a caregiver's speech: "Look, a bird!" The agent wishes to memorize what was said in this specific instance, but also learn more broadly what birds look like.

For any given event, there will be many such relationships to learn, which collectively encode many diverse features and relations in the environment. For instance, while viewing the bird, other neural circuits may encode the spatial location of the event, the time of day, other objects in the scene, and so on. We emphasize that our teacher-student-notebook model first considers just one of these relationships, and we return to having multiple relationships in the final sections of this document. We now describe the three principal components of the teacher-student-notebook framework.

Teacher Network. The ground truth relationship between inputs and output is represented by a "teacher" network, which generates an input-output pair by first drawing an input vector x in which each element is *i.i.d.* and normally distributed with variance $1/N$, i.e. $x_i \sim \mathcal{N}(0, 1/N),\ i = 1, \cdots, N$, such that the overall norm of the input is one in expectation. Next, the teacher labels this input according to the rule

$$y = \bar{w}x + \epsilon, \tag{1}$$

where $\bar{w} \in R^{1\times N}$ are the teacher weights, and ϵ is the teacher output noise. That is, the teacher takes the form of a simple shallow linear network with output noise. We take the teacher weights to be *i.i.d.* Gaussian with variance $\sigma^2_{\bar{w}}$, $\bar{w}_i \sim \mathcal{N}(0, \sigma^2_{\bar{w}}), i = 1, \cdots, N$, and fixed for all examples. The output noise is *i.i.d.* Gaussian with variance σ^2_ϵ on each sample.

A key parameter of this setting is the *signal-to-noise* ratio (SNR),

$$\mathcal{S} = \frac{\sigma^2_{\bar{w}}}{\sigma^2_\epsilon}. \tag{2}$$

This ratio measures the extent to which the teacher's output follows a systematic mapping between input and output. To fix a similar scale across different relationships, we often consider the case where the variance of the teacher's output is one,

$$\langle y^2 \rangle = \sigma^2_{\bar{w}} + \sigma^2_\epsilon = 1, \tag{3}$$

such that the SNR fixes the variances,

$$\sigma^2_{\bar{w}} = \tfrac{\mathcal{S}}{1+\mathcal{S}}, \tag{4}$$

$$\sigma^2_\epsilon = \tfrac{1}{1+\mathcal{S}}. \tag{5}$$

Conceptually, the teacher provides a simple generative model of the environment. We emphasize that taking the teacher to be a simple neural network does not reflect an assumption that the environment is a neural network. Rather, the teacher network can be thought of as containing the optimal synaptic weights for approximating the true generative model in the environment, which may reflect diverse causal processes (such as the physics of the world and the neural circuits that generate input and output activity patterns). In this sense, the teacher is the goal or target configuration for the student network, not an actual mechanistic theory of the environment. We discuss further interpretations of the teacher in Section 9.

Student Network. The goal of the "student" network is to learn to approximate the relationship defined by the teacher. Here we take the student to have the same architecture as the teacher, that is, it is a shallow linear network that receives an N-dimensional input x and produces a predicted output $\hat{y}$ according to

$$\hat{y} = wx, \tag{6}$$

where $w \in R^{1\times N}$ are the student weights. These weights are learned using gradient descent on a loss function $\mathcal{L}(w)$

$$\tau \frac{d}{dt} w = -\frac{\partial}{\partial w} \mathcal{L}(w), \tag{7}$$

here formulated in continuous time (also known as gradient flow) with time constant τ. We take the loss function to be the mean squared error over a set of patterns

$$\mathcal{L}(w) = \frac{1}{P}\sum_{\mu=1}^{P}(y^\mu - \hat{y}^\mu)^2, \tag{8}$$

where y^μ is the scalar target output and $\hat{y}^\mu$ is the scalar network prediction in response to input vector x^μ. Here $\mu = 1, \cdots, P$ indexes examples. As described in more detail subsequently, the target patterns that drive learning can have multiple sources–they may come directly from the teacher, or from notebook-mediated replay of the past.

The performance of the student can be measured in two ways. First, its predictions can be evaluated on the specific examples $\mu = 1, \cdots, P$ seen during training, which we refer to as the *memory* error $\mathcal{E}_m$ (also known as the "training error" in machine learning contexts),

$$\mathcal{E}_m = \frac{1}{P}\sum_{\mu=1}^{P}(y^\mu - \hat{y}^\mu)^2, \tag{9}$$

where here we have not indicated the time dependence of $\hat{y}^\mu$ for notational simplicity. Second, the student's predictions can be evaluated on novel input-output pairs drawn from the teacher, which we refer to as the *generalization* error $\mathcal{E}_g$ (also known as the "test error" in machine learning contexts),

$$\begin{aligned} \mathcal{E}_g &= \langle (y - \hat{y})^2 \rangle, & (10)\\ &= \langle (\bar{w}x + \epsilon - wx)^2 \rangle & (11)\\ &= \left\langle ((\bar{w} - w)x + \epsilon)^2 \right\rangle & (12)\\ &= ||\bar{w} - w||_2^2 + \sigma_\epsilon^2, & (13) \end{aligned}$$

where $\langle \cdot \rangle$ denotes the average over the teacher input distribution and output noise distribution, and we have used the fact that these distributions are independent.

Notebook Network. Finally, the job of the notebook is to faithfully memorize experienced patterns as attractors of neural network dynamics, making possible later recall and replay. We consider a notebook of M neurons recurrently connected through the $M \times M$ weight matrix J. The activity $h \in R^M$ in this network evolves according to

$$h(u) = f(Jh(u-1) - \theta), \tag{14}$$

where f is the Heaviside step function, u denotes discrete time steps of synchronous activity propagation, and θ is a threshold which can be dynamically adapted to maintain a desired sparsity of activity (as described subsequently).

The notebook represents memorized patterns as binary vectors of zeros and ones by embedding these vectors as fixed points of the dynamics in Eq. (14).

In particular, to store input-output pairs $\{x^\mu, y^\mu\},\ \mu = 1, \cdots, P$, the notebook first chooses P binary (0/1) vectors of length M, uniformly at random from the set of vectors with sparsity a (i.e. with exactly aM nonzero entries). These binary patterns of activity in the notebook act as distinctive neural codes (or indexes) to be associated with each pattern.

Stacking the binary patterns into the columns of the $M \times P$ matrix ξ, and similarly stacking the input and output patterns into the $N \times P$ and $1 \times P$ matrices X and Y respectively, the weights within the notebook and between the notebook and student are given through a Hebbian scheme,

$$J_{ij} = \begin{cases} \left(\frac{(\xi-a)(\xi-a)^T}{Ma(1-a)} - \frac{\gamma}{aM} \right)_{ij} & \text{for } i \neq j \\ 0 \quad \text{otherwise} \end{cases}, \tag{15}$$

$$U^{S_x \to N} = (\xi - a)X^T, \tag{16}$$

$$U^{S_y \to N} = (\xi - a)Y^T, \tag{17}$$

$$V^{N \to S_x} = \frac{X(\xi^T - a)}{Ma(1-a)}, \tag{18}$$

$$V^{N \to S_y} = \frac{Y(\xi^T - a)}{Ma(1-a)}. \tag{19}$$

Here $U^{S_x \to N} \in R^{M \times N}$ and $U^{S_y \to N} \in R^{M \times 1}$ map from the student inputs x and output y to the notebook activity h, and the matrices $V^{N \to S_x} \in R^{N \times M}$ and $V^{N \to S_y} \in R^{1 \times M}$ perform the reverse mapping from the notebook activity back to the student input and output. The parameter γ in Eqn. 15 implements global all-to-all inhibition, which causes activity that is far from stored patterns to decay to a silent state [8]. In simulations, we take $\gamma = 0.6$, which lies in the theoretically derived operating regime for this model [8]. For simplicity and tractability, we take all neurons to be linear, save those in the notebook (because attractor networks require nonlinearity to avoid having an abundance of stable mixture states). These pathways allow diverse interactions between notebook and student, and we describe a number of specific interaction patterns subsequently.

The mean subtraction and normalization in these updates have been chosen to aid performance, as derived subsequently in Section 5.1 for connections from notebook neurons. In essence, the notebook generates distinct, pattern-separated activity patterns, stabilizes these as attractors of its recurrent dynamics, and links these bidirectionally to the student's input and output neurons to facilitate later reactivation and replay.

2 Learning setting

The Teacher-Student-Notebook framework can allow for diverse learning settings in which examples from the teacher arrive at different times and in different quantities. Here we usually characterize memorization and generalization performance in a simple setting: the *single-batch, high-dimensional* regime. That is, we consider a scenario where an organism receives P training experiences up front in a short time window, and memory and generalization performance are evaluated subsequently over longer periods of time. For instance, a human subject might learn a task in a single hour long session, but then be tested after several weeks' delay; or a rodent might perform several trials in a water maze on one day, and be tested on the next. In our framework, these P experiences are drawn *i.i.d.* from the teacher and constitute one single batch for learning and consolidation. For convenience, we can collect this batch of samples into the $N \times P$ matrix X with columns $x^\mu,\ \mu = 1, \cdots, P$, and the $1 \times P$ row vector Y with elements $Y_\mu = y^\mu,\ \mu = 1, \cdots, P$.

Given abundant training experience ($P >> N$), many different learning schemes can converge to similar performance. However, real world learning is often severely data limited. Animals may receive only one or two foot shocks. A human subject may need to learn a new visual discrimination (possibly dependent on millions of pixels) from just a few blocks of training trials. Real world settings therefore place a premium on learning from limited experience. Moreover, neuronal networks in the brain are typically very large relative to the amount of training experience. Even a simple visual discrimination may engage a network of millions or billions of neurons interconnected by billions or trillions of adjustable synapses. To address this large network, limited data setting, we analyze the *high-dimensional* regime, in which the size of the student network and the number of training samples both tend to infinity ($N \to \infty, P \to \infty$), but their ratio $\alpha = P/N$ remains finite. The "load" parameter α is a key parameter of our setting, and it measures the amount of experience relative to the number of tunable synapses in the student network. For $\alpha < 1$, the network has more tunable parameters than training experiences, allowing analysis of highly overparametrized learning settings. For $\alpha >> 1$, the network has many more training experience than tunable parameters, reflecting the more standard classical regime of statistics.

While in this paper we emphasize this single-batch, high-dimensional learning setting, future work in the teacher-student-notebook framework could investigate more complex scenarios where examples continue to arrive over time.

3 Interaction policies & performance

The single-batch learning setting still allows diverse possible interaction policies between the modules in the teacher-student-notebook framework. These interaction policies specify which modules undergo learning, from what activity patterns (e.g. from the teacher, or from replay from the student), and which modules are used to answer queries for new experiences. We consider four interaction policies, meant to typify common approaches to learning and consolidation.

Online Student. Only the student is trained, without any replay. Each example drives one update of error-corrective learning and is never revisited.

This strategy provides a reference point for performance of a system based on gradient descent learning, without replay.

Online Notebook. Only the notebook is used. Each example is stored in the notebook with Hebbian updates, and predictions for novel inputs are generated using the notebook only. This strategy provides a reference point for performance of a system based on Hebbian memorization, without replay-guided learning.

Memory-optimized Replay. This strategy initially stores all experiences in the notebook and trains the student using notebook-driven reactivations until the student has fully memorized all examples in a manner similar to standard systems consolidation theory.

Generalization-optimized Replay. This novel strategy, proposed in this work, initially stores all experiences in the notebook but only trains the student using notebook-driven reactivations so long as generalization performance improves.

The next four sections of the supplement sequentially characterize the memorization and generalization performance of each of these interaction policies.

4 Online Student Policy

In the online student policy, each example $x^\mu \in R^N, \mu = 1, \cdots, P$, in the batch is visited in order and a single step of error corrective gradient descent learning is applied with a (possibly example-dependent) learning rate η^μ. In this section we characterize the average generalization error dynamics under this scheme; and to ensure a robust normative comparison to other policies, we derive the globally optimal learning rate function that maximizes generalization performance after all updates.

4.1 Generalization dynamics with time-dependent learning rate

Upon receiving each example $\mu = 1, \cdots, P$, the student weights are updated according to

$$w_{\mu+1} = w_\mu + \eta^\mu e^\mu x^{\mu^T}, \tag{20}$$

where $w^{\mu+1}$ is the weight vector resulting from the μth learning step, x^μ is the μth input example, η^μ is the time-dependent learning rate, and $e^\mu = y^\mu - \hat{y}^\mu$ is the error between the network's output and the target output for this example. We assume that the initial weights $w^1 = 0$. Using the teacher model, $y^\mu = \bar{w}x^\mu + \epsilon^\mu$, we have

$$\begin{aligned} w^{\mu+1} &= w^\mu + \eta^\mu \left(\bar{w}x^\mu + \epsilon^\mu - w^\mu x^\mu\right) x^{\mu^T} \\ &= w^\mu + \eta^\mu \left(\bar{w} - w^\mu\right) x^\mu x^{\mu^T} + \eta^\mu \epsilon^\mu x^{\mu^T}. \end{aligned} \tag{21}$$

In contrast to Eqn. (13), which expresses the generalization error $\mathcal{E}_g$ for a specific student and teacher, here we ask what the expected generalization error is for a randomly drawn teacher by averaging over the teacher weight distribution as well. That is, we track the expected generalization error $E_g = \langle \mathcal{E}_g \rangle$ where the average is over the teacher weight distribution. In the high-dimensional regime, the generalization error is self-averaging, such that any specific realization will closely track this expected generalization error, as verified by the close match between single simulations and the average dynamics we derive in the following.

The expected generalization error before example μ is

$$\begin{aligned}
E_g[\mu] &= \langle (y-\hat{y})^2 \rangle \\
&= \langle ((\bar{w}-w^\mu)\, x + \epsilon^\mu)^2 \rangle \\
&= \mathrm{Tr}\langle (\bar{w}-w^\mu)\, x x^T\, (\bar{w}-w^\mu)^T \rangle + \langle \epsilon^\mu (\bar{w}-w^\mu) x \rangle + \langle (\epsilon^\mu)^2 \rangle \\
&= \mathrm{Tr}\langle (\bar{w}-w^\mu)^T\, (\bar{w}-w^\mu)\, x x^T \rangle + \sigma_e^2 \\
&= \frac{1}{N} \langle \|\bar{w}-w^\mu\|^2 \rangle + \sigma_e^2 \\
&= \langle ((\bar{w})_i - (w^\mu)_i)^2 \rangle + \sigma_e^2 \qquad (22)
\end{aligned}$$

where the index i is arbitrary and is used to replace the average of a vector norm with a simpler average over a single component. Next, note that after example μ, the expected generalization error becomes

$$E_g[\mu+1] = \frac{1}{N} \langle \left\| \bar{w} - w^{\mu+1} \right\|^2 \rangle + \sigma_e^2. \qquad (23)$$

Substituting Eqn. 21, we have

$$\begin{aligned}
E_g[\mu+1] &= \frac{1}{N} \left\langle \left\| \bar{w} - w^\mu - \eta^\mu (\bar{w}-w^\mu) x^\mu x^{\mu^T} - \eta^\mu \epsilon^\mu x^{\mu^T} \right\|^2 \right\rangle + \sigma_e^2, \\
&= \frac{1}{N} \langle \mathrm{Tr} \Big[\Big((\bar{w}-w^\mu)(1-\eta^\mu x^\mu x^{\mu^T}) - \eta^\mu \epsilon^\mu x^{\mu^T} \Big) \\
&\quad \times \Big((\bar{w}-w^\mu)(1-\eta^\mu x^\mu x^{\mu^T}) - \eta^\mu \epsilon^\mu x^{\mu^T} \Big)^T \Big] \rangle + \sigma_e^2 \\
&= \frac{1}{N} \langle \mathrm{Tr}\ (\bar{w}-w^\mu)(1-\eta^\mu x^\mu x^{\mu^T})^2 (\bar{w}-w^\mu)^T \rangle \\
&\quad - \frac{2}{N} \langle \mathrm{Tr}\ \eta \epsilon^\mu x^{\mu^T} (1-\eta^\mu x^\mu x^{\mu^T})(\bar{w}-w^\mu)^T \rangle \qquad (24) \\
&\quad + \frac{1}{N} \langle \mathrm{Tr}\ (\eta^\mu)^2 (\epsilon^\mu)^2 x^{\mu^T} x^\mu \rangle + \sigma_e^2. \qquad (25)
\end{aligned}$$

Because $\langle \epsilon^\mu \rangle = 0$ and ϵ^μ is independent of all other terms, the term in Eqn. 24 is zero. Using the fact that x^μ is normal with variance $\frac{1}{N} I$, we have Tr $x^{\mu^T} x^\mu = 1$

and the last term is $\frac{(\eta^\mu)^2 \sigma_e^2}{N}$. Hence

$$\begin{aligned} E_g[\mu+1] &= \frac{1}{N}\langle \mathrm{Tr}\ (\bar{w}-w^\mu)(1-\eta^\mu x^\mu x^{\mu^T})^2(\bar{w}-w^\mu)^T\rangle \\ &\quad + \left(1+\frac{(\eta^\mu)^2}{N}\right)\sigma_e^2 \\ &= \frac{1}{N}\langle \|\bar{w}-w^\mu\|\rangle - \frac{2}{N}\langle \mathrm{Tr}\ (\bar{w}-w^\mu)\eta^\mu x^\mu x^{\mu^T}(\bar{w}-w^\mu)^T\rangle \\ &\quad + \frac{1}{N}\langle \mathrm{Tr}\ (\bar{w}-w^\mu)(\eta^\mu x^\mu x^{\mu^T})^2(\bar{w}-w^\mu)^T\rangle \\ &\quad + \left(1+\frac{(\eta^\mu)^2}{N}\right)\sigma_e^2 \end{aligned} \tag{26}$$

Using the fact that $\langle x^\mu x^{\mu^T} \rangle = \frac{1}{N} I$ and $\langle (x^\mu x^{\mu^T})^2 \rangle = \frac{2}{N^2} I + \frac{1}{N} I$ we have,

$$
\begin{aligned}
E_g[\mu+1] &= \left[1 - \frac{2\eta^\mu}{N} + (\eta^\mu)^2 \frac{2+N}{N^2}\right] \frac{1}{N} \langle \| \bar{w} - w^\mu \| \rangle \\
&\quad + \left(1 + \frac{(\eta^\mu)^2}{N}\right) \sigma_e^2 \\
&= \left[1 - \frac{2\eta^\mu}{N} + (\eta^\mu)^2 \frac{2+N}{N^2}\right] E_g[\mu] \\
&\quad - \left[1 - \frac{2\eta^\mu}{N} + (\eta^\mu)^2 \frac{2+N}{N^2}\right] \sigma_e^2 \\
&\quad + \left(1 + \frac{(\eta^\mu)^2}{N}\right) \sigma_e^2 \\
&= \left[1 - 2\frac{\eta^\mu}{N} + \left(\frac{\eta^\mu}{N}\right)^2 (2+N)\right] E_g[\mu] \\
&\quad + \left[1 + \frac{(\eta^\mu)^2}{N} - 1 + \frac{2\eta^\mu}{N} - (\eta^\mu)^2 \frac{2+N}{N^2}\right] \sigma_e^2 \\
&= \left[1 - 2\frac{\eta^\mu}{N} + \left(\frac{\eta^\mu}{N}\right)^2 (2+N)\right] E_g[\mu] \\
&\quad + \left[\frac{(\eta^\mu)^2 N + 2\eta^\mu N - 2(\eta^\mu)^2 - (\eta^\mu)^2 N}{N^2}\right] \sigma_e^2 \\
&= \left[1 - 2\frac{\eta^\mu}{N} + \left(\frac{\eta^\mu}{N}\right)^2 (2+N)\right] E_g[\mu] \\
&\quad + \left[2\frac{\eta^\mu N - (\eta^\mu)^2}{N^2}\right] \sigma_e^2 \\
&= \left[1 - 2\frac{\eta^\mu}{N} + \left(\frac{\eta^\mu}{N}\right)^2 (2+N)\right] E_g[\mu] \\
&\quad + \left[2\frac{\eta^\mu (N - \eta^\mu)}{N^2}\right] \sigma_e^2 \\
&= \left[1 - 2\frac{\eta^\mu}{N} + \left(\frac{\eta^\mu}{N}\right)^2 (2+N)\right] E_g[\mu] \\
&\quad + 2\left[\frac{\eta^\mu}{N}\left(1 - \frac{\eta^\mu}{N}\right)\right] \sigma_e^2
\end{aligned} \tag{27}
$$

Now passing to the limit $N >> 1$, we have

$$
E_g[\mu+1] = \left[1 - \frac{\eta^\mu (2 - \eta^\mu)}{N}\right] E_g[\mu] + 2\frac{\eta^\mu}{N} \sigma_e^2. \tag{28}
$$

We then enter the high dimensional regime where $\alpha = P/N$ and consider the new continuous variables $E_g(\alpha) \approx E_g[\alpha N]$ and $\eta(\alpha) \approx \eta^{\alpha N}$. We wish to calculate an equivalent differential equation,

$$\frac{dE_g(\alpha)}{d\alpha} = \frac{E_g(\alpha + d\alpha) - E_g(\alpha)}{d\alpha} \tag{29}$$

where we take $d\alpha = 1/N$ which is infinitesimal in the limit $N \to \infty$. Thus

$$\begin{aligned} \frac{E_g(\alpha + d\alpha) - E_g(\alpha)}{d\alpha} &= N(E_g[\alpha N + 1] - E_g[\alpha N]) \\ &= -\eta^{\alpha N}(2 - \eta^{\alpha N})E_g[\alpha N] + 2\eta^{\alpha N}\sigma_e^2. \end{aligned} \tag{30}$$

We thus have the ordinary linear differential equation

$$\frac{d}{d\alpha}E_g(\alpha) = -\eta(\alpha)(2 - \eta(\alpha))E_g(\alpha) + 2\eta(\alpha)\sigma_e^2. \tag{31}$$

The solution can be found through the method of integrating factors. Define

$$H(\alpha) = \int_0^\alpha \eta(\alpha')(2 - \eta(\alpha'))d\alpha'. \tag{32}$$

Then

$$E_g(\alpha) = E_g(0)e^{-H(\alpha)} + 2\sigma_e^2 e^{-H(\alpha)} \int_0^\alpha \eta(\tau)e^{H(\tau)}d\tau. \tag{33}$$

4.2 Optimal online learning rate

Equation (33) yields the expected generalization error for arbitrary learning rate functions. To ensure a fair normative comparison to other methods, we now compute the optimal learning rate as a function of example, $\eta^*(\mu)$, which minimizes the expected generalization error on example $T = \alpha N$. Let

$$f(x, \eta) = \left[1 - \frac{\eta(2 - \eta)}{N}\right] x + 2\frac{\eta}{N}\sigma_e^2 \tag{34}$$

be the discrete time dynamics update from Eqn. (28), that is, the generalization error on example $\mu + 1$ if the generalization error on example μ is x and the learning rate used on example μ is η.

At the penultimate example $T - 1$ before the deadline, because there is only one update left, the best learning rate is given by greedily optimizing f,

$$\eta^*(T - 1) = \text{argmin}_\eta f(E_g(T - 1), \eta). \tag{35}$$

We directly perform the minimization by differentiating with respect to η and setting this derivative to zero,

$$\begin{aligned} \frac{\partial}{\partial\eta}f(x, \eta) &= \eta x/N - (2 - \eta)x/N + 2\sigma_e^2/N \\ 0 &= 2\eta^* x - 2x + 2\sigma_e^2 \\ \eta^* &= 1 - \frac{\sigma_e^2}{x}, \end{aligned} \tag{36}$$

which yields the optimal $\eta^*(T-1) = 1 - \frac{\sigma_e^2}{E_g(T-1)}$.

The final generalization error as a function of the penultimate generalization error x is

$$\begin{aligned} g(x) &\equiv \min_\eta f(x,\eta) \\ &= \left[1 - \frac{\left(1-\frac{\sigma_e^2}{x}\right)(1+\frac{\sigma_e^2}{x})}{N}\right]x + 2\frac{1-\frac{\sigma_e^2}{x}}{N}\sigma_e^2 \\ &= \left[1 - \frac{\left(1-\frac{\sigma_e^4}{x^2}\right)}{N}\right]x + 2\frac{1-\frac{\sigma_e^2}{x}}{N}\sigma_e^2 \\ &= (1-1/N)x - \frac{\sigma_e^4}{Nx} + 2\frac{\sigma_e^2}{N}. \end{aligned} \tag{37}$$

Differentiating with respect to x, we have

$$\frac{\partial}{\partial x}g(x) = 1 - 1/N + \frac{\sigma_e^4}{Nx^2}, \tag{38}$$

which is strictly positive for $N \geq 1, x > 0$, indicating that the function $g(x)$ is strictly increasing.

Let $v_\mu(x)$ denote the optimal final generalization error on example T, starting from an error of x at step μ and choosing the optimal learning rate thereafter. We have shown that $v_{T-1}(x) = g(x)$, and it is strictly increasing. Now for the inductive step, assume that $v_{\mu+1}(x)$ is strictly increasing. Then

$$\begin{aligned} \eta^*(\mu) &= \text{argmin}_\eta v_{\mu+1}(f(x,\eta)) \\ &= \text{argmin}_\eta f(x,\eta). \end{aligned} \tag{39}$$

Therefore the optimal learning rate is again selected by greedily minimizing $f(x,n)$. Finally, we note that $v_\mu(x) = v_{\mu+1}(g(x))$ is the composition of strictly increasing functions, and therefore strictly increasing. By induction this yields the optimal learning rate function for all examples $\eta^*(\mu) = 1 - \frac{\sigma_e^2}{E_g(\mu)}$.

In the high-dimensional regime, the optimal learning rate is thus

$$\eta^*(\alpha) = 1 - \frac{\sigma_e^2}{E_g(\alpha)}. \tag{40}$$

Inserting this optimal learning rate function back into Eqn. (31) yields the following optimal generalization error dynamics,

$$\frac{d}{d\alpha}E_g(\alpha) = 2\sigma_e^2 - E_g(\alpha) - \frac{\sigma_e^4}{E_g(\alpha)}. \tag{41}$$

5 Online Notebook Policy

In the online notebook policy, each example is stored in the notebook according to the Hebbian scheme in Eqns. (15)-(19). The notebook is then used to make predictions even for novel inputs, by allowing the notebook to converge to an attractor and reading off the predicted output.

In particular, an input x arriving at the student from the teacher can be used to seed recurrent pattern completion in the notebook, by letting $h(0) = f(U^{S_x \to N} x)$ and then running the notebook dynamics. In the simulations in the main text, rather than run the recurrent dynamics to convergence, we use the pattern obtained after 9 updates. At each update, the neurons are ranked by net input and the threshold θ is chosen so that the top aM are active (in the case of ties, slightly more neurons can be active). After the network dynamics have settled on some pattern $\tilde{\xi}$, a predicted output can be generated (using just the notebook) as $\tilde{y} = V^{N \to S_y} \tilde{\xi}$.

This section shows that, in the high-dimensional setting considered here, the notebook attains low memorization error (i.e. error on already-experienced examples) but is incapable of generalization.

5.1 Hebbian learning rule scale factor and offset

The memorization ability of recurrent attractor networks, as well as the performance of Hebbian plasticity rules in mapping from notebook activity patterns to student activity patterns, is known to depend on the statistics of the patterns and the specific form of the learning rule used to configure the weights [10, 4, 5, 16]. We begin by justifying the scaling and subtractive offsets in Eqns. (15)-(19), as an approximate implementation of the pseudo-inverse learning rule given our sparse pattern statistics.

5.1.1 Recurrent weights

The job of the notebook is to faithfully memorize example patterns as attractors of neural network dynamics. The pseudo-inverse learning rule is a flexible mechanism to memorize these patterns, wherein the $M \times M$ matrix of recurrent notebook connections would be

$$J = \xi \xi^{+} = \xi (\xi^T \xi)^{-1} \xi^T, \tag{42}$$

where ξ^{+} is the pseudo-inverse of ξ, and we assumed that $P \leq M$. Suppose that the neural network dynamics have the form $h(u) = f(Jh(u-1))$, where h is the pattern of notebook activity. Assuming that $f(0) = 0$ and $f(1) = 1$ (e.g. f may be linear, threshold-linear, or binary), then these weights would successfully memorize all P patterns as steady-states of the network dynamics. In particular, note that

$$f(J\xi) = f(\xi(\xi^T \xi)^{-1} \xi^T \xi) = f(\xi) = \xi, \tag{43}$$

so that the network dynamics map each memorized pattern back onto itself. It is instructive to expand the pseudo-inverse weights in terms of the stored patterns,

$$J_{ij} = \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} \xi_{i\mu} (\xi^T \xi)^{-1}_{\mu\nu} \xi_{j\nu}. \tag{44}$$

This reveals a practical problem with the pseudo-inverse learning rule, as the storage prescription for each pattern depends on the other stored patterns through the inverse pattern correlation, $(\xi^T \xi)^{-1}_{\mu\nu}$.

The Hopfield model can be viewed as a solution to this problem that assumes simple random statistics for ξ in order to simplify the necessary structure of the learning rule. In particular, suppose that each memory randomly assigns aM neurons to the 1-state and $(1-a)M$ neurons to the 0-state. Thus, a quantifies the fraction of 1-states in the memorized patterns, and we refer to a as the sparseness parameter. We also assume that the memorized patterns are statistically independent from each other. These statistics imply that

$$\begin{aligned} \langle (\xi^T \xi)_{\mu\nu} \rangle = \sum_{i=1}^{M} \langle \xi_{i\mu} \xi_{i\nu} \rangle = \sum_{i=1}^{M} (a\delta_{\mu\nu} + a^2(1-\delta_{\mu\nu})) \\ = Ma^2 + Ma(1-a)\delta_{\mu\nu}. \end{aligned} \tag{45}$$

In matrix notation, this implies that

$$\langle \xi^T \xi \rangle = Ma(1-a) I_P + Ma^2 1_P 1_P^T, \tag{46}$$

where I_P is the $P \times P$ identity matrix, and 1_P is the P-vector of ones. This form allows us to use the Sherwood-Morrison formula,

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}, \tag{47}$$

with $A = Ma(1-a)I_P$, $u = Ma^2 1_P$, and $v = 1_P$ to obtain

$$\begin{aligned} \langle \xi^T \xi \rangle^{-1} &= \frac{1}{Ma(1-a)} I_P - \frac{Ma^2/(Ma(1-a))^2 1_P 1_P^T}{1 + MPa^2/(Ma(1-a))} \\ &= \frac{1}{Ma(1-a)} I_P - \frac{1/(M(1-a)^2)}{1 + Pa/(1-a)} 1_P 1_P^T \\ &= \frac{1}{Ma(1-a)} I_P - \frac{1}{M(1-a)^2 + MPa(1-a)} 1_P 1_P^T \\ &\approx \frac{1}{Ma(1-a)} I_P - \frac{1}{M^2 \beta a(1-a)} 1_P 1_P^T, \end{aligned} \tag{48}$$

where the final approximation used $P = \beta M$, $\beta = O(1)$, and $M \gg 1$. The Hopfield model approximates the pseudo-inverse learning rule by replacing $(\xi^T \xi)^{-1}$ by $\langle \xi^T \xi \rangle^{-1}$, To see what this means, we need to do a bit more algebra

$$J \approx \xi \langle \xi^T \xi \rangle^{-1} \xi^T = \frac{\xi \xi^T}{Ma(1-a)} - \frac{(\xi 1_P)(\xi 1_P)^T}{M^2 \beta a(1-a)}. \tag{49}$$

The Hopfield model also approximates $\xi 1_P$ by $\langle \xi 1_P \rangle = Pa1_M$, where 1_M is the M-vector of ones, such that

$$J \approx \frac{\xi\xi^T}{Ma(1-a)} - \frac{a^2P^2 1_M 1_M^T}{M^2\beta a(1-a)} = \frac{1}{Ma(1-a)}\xi\xi^T - \frac{\beta a}{1-a}1_M 1_M^T. \tag{50}$$

To compare this to the Hopfield model, we first consider a general Hebbian weight matrix of the form,

$$J_{ij} = \sum_{\mu=1}^{P} B(\xi_{i\mu} - b)(\xi_{j\mu} - b), \tag{51}$$

where B and b are constants that scale and center the learning rule. Again using the approximation that $\xi 1_P \approx \langle \xi 1_P \rangle = Pa1_M$, we find

$$\begin{aligned} J &= B\xi\xi^T - Bb1_M 1_P^T \xi^T - Bb\xi 1_P 1_M^T + Bb^2 1_M 1_P^T 1_P 1_M^T \\ &\approx B\xi\xi^T + (-2abBP + b^2BP)1_M 1_M^T = B\xi\xi^T + bBP(-2a+b)1_M 1_M^T. \end{aligned} \tag{52}$$

Comparing Eqs. (50) and (52), we see that the two correspond when

$$B = \frac{1}{Ma(1-a)} \tag{53}$$

and

$$\begin{aligned} -\frac{\beta a}{1-a} = bBP(-2a+b) = \frac{bP(b-2a)}{Ma(1-a)} = \frac{b\beta(b-2a)}{a(1-a)} \\ \implies 0 = b^2 - 2ba + a^2 = (b-a)^2 \implies b = a. \end{aligned} \tag{54}$$

Therefore, the pseudo-inverse rule can be approximated by the Hebbian rule,

$$J_{ij} = \sum_{\mu=1}^{P} \frac{(\xi_{i\mu} - a)(\xi_{j\mu} - a)}{Ma(1-a)}, \tag{55}$$

which is the weight matrix of the Hopfield model, and is the first term in Eqn. (15) of the notebook learning rules.

5.1.2 Notebook-to-Student weights

Similar to the Hopfield storage prescription used to store binary indices as fixed points of the recurrent notebook dynamics, here we assume Hebbian connectivity between the notebook and student. In particular, we can form the $(N+1)\times P$ matrix Z by vertically stacking the matrices X and Y, such that Z represents the combined student input-output activity to be stored. We also define the $(N+1)\times M$ matrix V by vertically stacking the matrices $V^{N\to S_x}$ and $V^{N\to S_y}$, which represents the mapping from notebook activity to student activity. In this setting the relevant pseudo-inverse learning rule for the weights from notebook to student neurons is

$$V = Z\xi^{+} = Z(\xi^T\xi)^{-1}\xi^T, \tag{56}$$

and N is the number of student input neurons. The same approximations used in the previous section lead to

$$V \approx \frac{Z\xi^T}{Ma(1-a)} - \frac{(Z1_P)(\xi 1_P)^T}{M^2\beta a(1-a)}. \tag{57}$$

Replacing $\xi 1_P$ by $\langle \xi 1_P\rangle = Pa1_M$, we find

$$V \approx \frac{Z\xi^T}{Ma(1-a)} - \frac{Za1_P1_M^T}{Ma(1-a)} = \frac{Z(\xi^T - a1_P1_M^T)}{Ma(1-a)}, \tag{58}$$

or

$$V_{ij} \approx \sum_{\mu=1}^{P} \frac{Z_{i\mu}(\xi_{j\mu} - a)}{Ma(1-a)}. \tag{59}$$

This is the Hebbian learning rule that we use to connect the notebook to the student for purposes of pattern reactivation (Eqns. (18)-(19)).

5.2 Notebook memory error

With these Hebbian learning prescriptions in hand, we now characterize their performance. In this section, we consider the statistics by which the notebook reactivates stored patterns of student activity, in order to understand its typical memory error. Previous studies of the Hopfield model [4, 5, 16] imply that large notebooks can accurately recall each random index if the number of stored patterns does not exceed the capacity of the network $P_c = \beta_c M$. Here we assume that $M \gg 1$ and $P < P_c$, such that erroneous index retrieval by the notebook is rare. Once a notebook memory index is accurately retrieved by the notebook's dynamics, the notebook can generate a predicted output using the Hebbian weights from notebook to student output ($V^{N\to S_y}$). The memory error of the notebook can thus be approximated as the typical error of this prediction.

As in the previous section, let Z be a $(N+1)\times P$ matrix that groups together all input and output neuron responses for all memorized patterns. Then the notebook reactivated student pattern is

$$\hat{Z}_{i\mu} = \sum_{j=1}^{M} V_{ij}\xi_{j\mu} = \sum_{j=1}^{M}\sum_{\nu=1}^{P} \frac{1}{Ma(1-a)} Z_{i\nu}(\xi_{j\nu} - a)\xi_{j\mu}. \tag{60}$$

We first consider how well the notebook reactivates the student on average. In particular, averaging this expression over all possible notebook indices gives

$$\begin{aligned} \langle \hat{Z}_{i\mu}\rangle \quad &= \tfrac{1}{Ma(1-a)} \textstyle\sum_{j=1}^{M}\sum_{\nu=1}^{P} Z_{i\nu}\langle(\xi_{j\nu} - a)\xi_{j\mu}\rangle \\ &= \tfrac{1}{Ma(1-a)} \textstyle\sum_{j=1}^{M}\sum_{\nu=1}^{P} Z_{i\nu} a(1-a)\delta_{\mu\nu} = Z_{i\mu}. \end{aligned} \tag{61}$$

Therefore, the Hebbian learning rule is unbiased, and it on average reactivates all student neuron responses accurately.

However, the randomness of notebook indices does cause notebook-driven student reactivations to fluctuate away from these average values. To determine the magnitude of notebook memory error quantitatively, first note that the training error of the notebook is

$$\mathcal{E}_m = \frac{1}{P}\sum_{\mu=1}^{P}(Y_\mu - \hat{Y}_\mu)^2 = \frac{1}{P}\sum_{\mu=1}^{P}(Y_\mu^2 - 2Y_\mu\hat{Y}_\mu + \hat{Y}_\mu^2). \tag{62}$$

Averaging over possible notebook indices, we find

$$E_m = \langle \mathcal{E}_m \rangle = \frac{1}{P}\sum_{\mu=1}^{P}(Y_\mu^2 - 2Y_\mu^2 + \langle \hat{Y}_\mu^2 \rangle) = \frac{1}{P}\sum_{\mu=1}^{P}\mathrm{Var}(\hat{Y}_\mu). \tag{63}$$

This variance term can be written

$$\mathrm{Var}(\hat{Y}_\mu) = \left\langle \sum_{\nu=1}^{P}\sum_{j=1}^{M}\frac{Y_\nu(\xi_{j\nu}-a)}{Ma(1-a)}\xi_{j\mu}\sum_{\rho=1}^{P}\sum_{k=1}^{M}\frac{Y_\rho(\xi_{k\rho}-a)}{Ma(1-a)}\xi_{k\mu}\right\rangle - Y_\mu^2. \tag{64}$$

This expression shows that the exact value of the notebook training error depends on the specific realizations of the student outputs. However, for practical purposes, it will be good enough to average over these possibilities. Letting $\langle\langle\cdot\rangle\rangle$ denote the average over student patterns, and noting that $\langle\langle Y_\mu Y_\nu\rangle\rangle = \delta_{\mu\nu}$, we find

$$\begin{aligned}\langle\langle \mathrm{Var}(\hat{Y}_\mu)\rangle\rangle &= \frac{1}{(Ma(1-a))^2}\sum_{\nu=1}^{P}\sum_{j=1}^{M}\sum_{k=1}^{M}\langle(\xi_{j\nu}-a)\xi_{j\mu}(\xi_{k\nu}-a)\xi_{k\mu}\rangle - 1\\ &= \frac{1}{(Ma(1-a))^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\Big(\langle(\xi_{j\mu}-a)\xi_{j\mu}(\xi_{k\mu}-a)\xi_{k\mu}\rangle\\ &\qquad + \sum_{\nu\neq\mu}\langle(\xi_{j\nu}-a)\xi_{j\mu}(\xi_{k\nu}-a)\xi_{k\mu}\rangle\Big) - 1. \end{aligned} \tag{65}$$

It is straightforward to evaluate the first expectation as

$$\begin{aligned}\langle(\xi_{j\mu}-a)\xi_{j\mu}(\xi_{k\mu}-a)\xi_{k\mu}\rangle &= \delta_{jk}(1-a)^2P(\xi_{j\mu}=1)\\ &+(1-\delta_{jk})(1-a)^2P(\xi_{j\mu}=1)P(\xi_{k\mu}=1|\xi_{j\mu}=1)\\ &= \delta_{jk}(1-a)^2a + (1-\delta_{jk})(1-a)^2a\frac{aM-1}{M-1}\\ = \delta_{jk}a(1-a)^2 + (1-\delta_{jk})&\left(Ma^2(1-a)^2/(M-1) - a(1-a)^2/(M-1)\right). \end{aligned} \tag{66}$$

Because $\mu \neq \nu$ in the second expectation, it straightforwardly separates into the product of two terms:

$$\langle(\xi_{j\nu}-a)\xi_{j\mu}(\xi_{k\nu}-a)\xi_{k\mu}\rangle = \langle(\xi_{j\nu}-a)(\xi_{k\nu}-a)\rangle\langle\xi_{j\mu}\xi_{k\mu}\rangle. \tag{67}$$

First,

$$\begin{aligned}\langle(\xi_{j\nu}-a)(\xi_{k\nu}-a)\rangle &= \delta_{jk}a(1-a)+(1-\delta_{jk})\Big((1-a)^2P(\xi_{j\nu}=1)P(\xi_{k\nu}=1|\xi_{j\nu}=1)\\ &\quad -a(1-a)P(\xi_{j\nu}=1)P(\xi_{k\nu}=0|\xi_{j\nu}=1)-a(1-a)P(\xi_{j\nu}=0)P(\xi_{k\nu}=1|\xi_{j\nu}=0)\\ &\quad +a^2P(\xi_{j\nu}=0)P(\xi_{k\nu}=1|\xi_{j\nu}=0)\Big)\\ &= \delta_{jk}a(1-a)+(1-\delta_{jk})\Big((1-a)^2a\frac{aM-1}{M-1}-a(1-a)a\frac{(1-a)M}{M-1}\\ &\quad -a(1-a)(1-a)\frac{aM}{M-1}+a^2(1-a)\frac{(1-a)M-1}{M-1}\Big)\\ &= \delta_{jk}a(1-a)+(1-\delta_{jk})\Big(-a(1-a)^2\frac{1}{M-1}-a^2(1-a)\frac{1}{M-1}\Big)\\ &= \delta_{jk}a(1-a)-(1-\delta_{jk})\frac{a(1-a)}{M-1}.\end{aligned} \tag{68}$$

Second,

$$\begin{aligned}\langle\xi_{j\mu}\xi_{k\mu}\rangle &= \delta_{jk}a+(1-\delta_{jk})P(\xi_{j\mu}=1)P(\xi_{k\mu}=1|\xi_{j\mu}=1)\\ &= \delta_{jk}a+(1-\delta_{jk})a\frac{aM-1}{M-1}.\end{aligned} \tag{69}$$

Combining these two terms, we find,

$$\langle(\xi_{j\nu}-a)\xi_{j\mu}(\xi_{k\nu}-a)\xi_{k\mu}\rangle = \delta_{jk}a^2(1-a)-(1-\delta_{jk})a^2(1-a)\frac{aM-1}{(M-1)^2} \tag{70}$$

for $\mu\neq\nu$.

Plugging these expressions back into the expression for $\langle\langle \mathrm{Var}(\hat{Y}_\mu)\rangle\rangle$, we find

$$\begin{aligned}
\langle\langle \mathrm{Var}(\hat{Y}_\mu)\rangle\rangle = \frac{1}{\big(Ma(1-a)\big)^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\Bigg(\delta_{jk}a(1-a)^2 \\
+(1-\delta_{jk})\big(Ma^2(1-a)^2/(M-1) - a(1-a)^2/(M-1)\big) \\
+\sum_{\nu\neq\mu}\left(\delta_{jk}a^2(1-a) - (1-\delta_{jk})a^2(1-a)\frac{aM-1}{(M-1)^2}\right)\Bigg) - 1 \\
= \frac{1}{\big(Ma(1-a)\big)^2}\Big(Ma(1-a)^2 \\
+M(M-1)\big(Ma^2(1-a)^2/(M-1) - a(1-a)^2/(M-1)\big) \\
+(P-1)Ma^2(1-a) - (P-1)a^2(1-a)M(aM-1)/(M-1)\Big) - 1 \\
= \frac{Ma(1-a)^2 + M^2a^2(1-a)^2 - Ma(1-a)^2 - M^2a^2(1-a)^2}{M^2a^2(1-a)^2} \\
+(P-1)\frac{Ma^2(1-a) - a^3(1-a)M^2/(M-1) + a^2(1-a)M/(M-1)}{M^2a^2(1-a)^2} \\
= (P-1)\frac{M-1-aM+1}{M(1-a)(M-1)} \\
= \frac{P-1}{M-1}
\end{aligned} \tag{71}$$

We therefore find that the training error of the notebook is simply

$$\langle\langle\langle \mathcal{E}_m\rangle\rangle\rangle = \frac{P-1}{M-1}, \tag{72}$$

where the triple average, $\langle\langle\langle\cdot\rangle\rangle\rangle$, denotes an average over the notebook patterns followed by the average over teacher patterns. Remarkably, note that this expression is independent of the notebook's sparseness. If $P = \beta M$ and $M \gg 1$, this implies that

$$\langle\langle\langle \mathcal{E}_m\rangle\rangle\rangle \approx \beta. \tag{73}$$

We thus see that the expected memorization error of the notebook scales with the number of memories stored in the system and can become significant when the loading is large. Note that this expression only makes sense if $\beta < \beta_c$, because we've assumed faithful index reactivation within the notebook itself.

5.3 Notebook generalization error

Next we examine the expected error when the notebook is used to predict the teacher output on a novel example. Because we operate the Hopfield network below capacity, it successfully embeds all patterns as fixed points, and does so with relatively large basins of attraction. We therefore model the notebook as a nearest neighbor algorithm, that operates by returning the output associated to the nearest stored pattern to any given input.

Let $x^\mu, 1, \cdots, P$ be the N-dimensional column vectors of stored inputs, and $y^\mu, 1, \cdots, P$ the associated outputs. For a novel input $x \in R^N$, we find the nearest neighbor as

$$\mu^* = \operatorname{argmin}_\mu \|x - x^\mu\|^2 . \tag{74}$$

With the nearest neighbor identified, the prediction is $\hat{y} = y^{\mu^*} = \bar{w}x^{\mu^*} + \epsilon^{\mu^*}$. The expected generalization error is thus

$$\begin{aligned} E_g &= \langle (y-\hat{y})^2 \rangle \\ &= \left\langle \left(\bar{w}\left(x - x^{\mu^*}\right) + \epsilon - \epsilon^{\mu^*} \right)^2 \right\rangle \\ &= \mathrm{Tr}\left\langle \bar{w}\left(x - x^{\mu^*}\right)\left(x - x^{\mu^*}\right)^T \bar{w}^T \right\rangle + \langle (\epsilon - \epsilon^{\mu^*})\bar{w}(x - x^{\mu^*})\rangle \\ &\quad + \langle (\epsilon - \epsilon^{\mu^*})^2 \rangle \\ &= \sigma_{\bar{w}}^2 \mathrm{Tr}\left\langle \left(x - x^{\mu^*}\right)^T \left(x - x^{\mu^*}\right) \right\rangle + 2\sigma_e^2 \\ &= \sigma_{\bar{w}}^2 \left\langle \left\| x - x^{\mu^*} \right\|_2^2 \right\rangle + 2\sigma_e^2 . \end{aligned} \tag{75}$$

We note that $x - x^\mu \sim \mathcal{N}(0, \frac{2}{N} I)$ for all μ, and so $z^\mu = \sqrt{\frac{N}{2}}(x - x^\mu) \sim \mathcal{N}(0, I)$. By the Gaussian Annulus Theorem (see e.g. Thm 2.9, pg 15 of [7]),

$$\begin{aligned} P\left(\left| \|z^\mu\|_2^2 - N \right| \geq \beta\sqrt{N} \right) &\leq 3e^{-c\beta^2}, \\ P\left(\left| \frac{N}{2} \|x - x^\mu\|_2^2 - N \right| \geq \beta\sqrt{N} \right) &\leq 3e^{-c\beta^2}, \\ P\left(\left| \|x - x^\mu\|_2^2 - 2 \right| \geq 2\beta/\sqrt{N} \right) &\leq 3e^{-c\beta^2}, \end{aligned} \tag{76}$$

where $c > 0$ is a constant independent of N. By the union bound, the probability that one or more patterns among the $\mu = 1, \cdots, \alpha N$ fails to concentrate is no more than

$$P\left(\bigcup_{\mu=1}^{\alpha N} \left\{ \left| \|x - x^\mu\|_2^2 - 2 \right| \geq 2\beta/\sqrt{N} \right\} \right) \leq 3\alpha N e^{-c\beta^2} . \tag{77}$$

Therefore the minimum over all patterns will fail to concentrate with probability no more than

$$P\left(\left| \left\| x - x^{\mu^*} \right\|_2^2 - 2 \right| \geq 2\beta/\sqrt{N} \right) \leq 3\alpha N e^{-c\beta^2} . \tag{78}$$

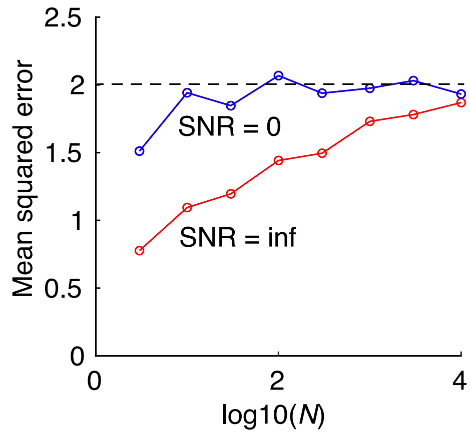

Figure 1: Numerical simulation of the generalization error of the nearest neighbor algorithm, for $\alpha = 1$ and SNR $\mathcal{S} = \infty$ (red) and $\mathcal{S} = 0$ (blue). As input dimension N approaches infinity, generalization error in both cases approaches 2, consistent with our analytical derivation.

Choosing $\beta = N^{1/4}$ we have,

$$P\left(\left|\left\|x - x^{\mu^*}\right\|_2^2 - 2\right| \geq 2N^{-1/4}\right) \leq 3\alpha N e^{-c\sqrt{N}}, \tag{79}$$

such that as $N \to \infty$, the minimum concentrates near 2 with probability one. Substituting back into the expression for the expected generalization error, in the high dimensional limit with high probability we have

$$E_g \quad = \quad 2\sigma_{\bar{w}}^2 + 2\sigma_e^2. \tag{80}$$

For our standard scaling where $\sigma_{\bar{w}}^2 + \sigma_e^2 = 1$, the error is therefore 2 regardless of the SNR. We note that this result applies in the high-dimensional limit where $N, P \to \infty$ and their ratio is $\alpha = P/N$. In finite size simulations, the generalization error can modestly differ, as shown in Fig. 1.

In essence, in the high dimensional regime, the nearest neighbor is typically very far away from the new sample, such that generalization fails completely. In fact, it is so poor that always predicting zero would be better (attaining generalization error of 1 rather than 2 for our setting). This finding strongly motivates the need for a trained student, but we note that notebook-mediated generalization could be better in different settings where, for instance, input examples arise from a low number of clusters [9].

6 Memorization-optimized Replay Policy

In the memorization-optimized replay policy, each example is stored in the notebook according to the Hebbian scheme in Eqns. (15)-(19). These patterns can then be reactivated offline to drive learning. In the simulations reported in the main text, offline notebook reactivations undergo a two-step retrieval process:

1. A random binary pattern is used to seed the reactivation event. Starting at this random state, the notebook updates through the recurrent dynamics 9 times synchronously to retrieve a stored pattern. On each update, the threshold θ is chosen to enforce a sparsity of a (up to ties, which can cause slightly more neurons to be active). Without this adaptive threshold, a silent attractor dominates retrieval.

2. The notebook then uses the retrieved pattern from (1) to seed a second round of pattern completion using a fixed threshold $\theta = -0.15$, which in combination with the global inhibition parameter $\gamma = 0.6$ provides good retrieval alongside the possibility of retrieving a silent state (see [8] for detailed derivation of performance as a function of these parameters). This two step process enables retrieval of patterns that are not forced to have a fixed sparseness, and a "silent state" attractor can be retrieved when the seeding pattern lies far away from any of the encoded patterns.

This models a simple form of replay. Supposing that the notebook pattern at convergence is $\tilde{\xi}$, the student input and target output are then reconstructed based on the Hebbian connectivity as $\tilde{x} = V^{N \to S_x}\tilde{\xi}$ and $\tilde{y} = V^{N \to S_y}\tilde{\xi}$. This provides an $\{\tilde{x}, \tilde{y}\}$ sample from which the student can learn using gradient descent.

The policy is memory-optimized, in the sense that this replay continues indefinitely, such that all samples stored in the notebook are eventually learned by the student. This section characterizes the memory and generalization performance of the student resulting from this replay process. If reactivations perfectly reconstructed the stored examples, this replay strategy would be similar to 'batch' learning strategies in machine learning, in which the same stored dataset is repeatedly revisited to update network weights. However, errors in reactivation could in principle degrade the learning process. In Section 6.1 we show that although reactivations introduce errors, remarkably, these errors are correlated in such a way that learning still proceeds like batch learning from perfectly recalled examples up to a rescaling of the learning rate. Using this fact, in Section 6.2 we provide the expected memory and generalization errors, based on results known in prior work [11, 3].

In this policy, both the notebook and student learn potentially beneficial information, and in principle either could be used to answer a specific query for a point x. We take the normative assumption that the best system is selected to make the prediction. Often, this means that the output for a previously stored input will be predicted by the notebook, while that for a novel input will be predicted by the student. However, in Section 6.1 we show that there are conditions under which the student memory error in fact surpasses the notebook, and the student would be used to make predictions for previously stored inputs.

6.1 Accurate learning despite errors in reactivation

How do reactivation errors influence learning dynamics in the student? One hint that learning from reactivations can be effective comes from Fig. 2 of the main text. Given that the notebook is specifically designed to rapidly store memories, it often has a lower memory error than the student. Surprisingly, however, Figs. 2a-h of the main paper show that the student's training error can fall below that of the notebook. How could it be that the student learns to accurately produce a memory that was imperfectly memorized by the notebook? Our key theoretical observation is that although the notebook imperfectly activates the output of the student, it also imperfectly activates the inputs of the student. These errors are correlated between input and output neurons in a way that does not harm student learning. We demonstrate this fact in this section.

Reactivations have subtly different statistics to the original samples. In particular, when the notebook settles on a pattern ξ^μ (one column of the matrix ξ) that was associated with an original sample x^μ, y^μ from the teacher, this results in reactivated student activity input and output patterns $\tilde{x}^\mu = V^{N\to S_x}\xi^\mu$ and $\tilde{y}^\mu = V^{N\to S_y}\xi^\mu$, respectively. Horizontally concatenating the input and output reactivations into the matrices $\tilde{X} \in R^{N\times P}$ and $\tilde{Y} \in R^{1\times P}$, this reactivation leads the weights in the student network to change (in the reactivated gradient direction) by the amount,

$$\tilde{\Delta}_\mu w_i = -\lambda \frac{\partial}{\partial w_i}\left(\sum_j w_j \tilde{X}_{j\mu} - \tilde{Y}_\mu\right)^2 = -2\lambda\left(\sum_j w_j \tilde{X}_{j\mu} - \tilde{Y}_\mu\right)\tilde{X}_{i\mu}$$
$$= -2\lambda\left(\sum_j w_j \tilde{X}_{i\mu}\tilde{X}_{j\mu} - \tilde{X}_{i\mu}\tilde{Y}_\mu\right). \quad (81)$$

Therefore, the change expected from gradient descent learning with a random notebook index is

$$\langle\tilde{\Delta}_\mu w_i\rangle = -2\lambda\left(\sum_j w_j \langle\tilde{X}_{i\mu}\tilde{X}_{j\mu}\rangle - \langle\tilde{X}_{i\mu}\tilde{Y}_\mu\rangle\right). \quad (82)$$

To evaluate these expectations, we form the matrix $\tilde{Z}$ by vertically stacking $\tilde{X}$ and $\tilde{Y}$, then note that

$$\langle\tilde{Z}_{i\mu}\tilde{Z}_{j\mu}\rangle = \frac{1}{M^2a^2(1-a)^2}\sum_{\nu=1}^{P}\sum_{k=1}^{M}\sum_{\rho=1}^{P}\sum_{l=1}^{M} Z_{i\nu}Z_{j\rho}\langle(\xi_{k\nu}-a)\xi_{k\mu}(\xi_{l\rho}-a)\xi_{l\mu}\rangle. \quad (83)$$

When $\mu \neq \nu \neq \rho$, the statistical independence of memories allows us to factor out $\langle\xi_{k\nu} - a\rangle$, which is zero and causes the whole term to vanish. Similarly, we get no contributions if $\mu \neq \rho \neq \nu$. This implies that both the ν and ρ indices must either pair with each other or with μ, and the only terms that contribute are thus $\nu = \rho = \mu$ and $\nu = \rho \neq \mu$.

$$\langle \tilde{Z}_{i\mu}\tilde{Z}_{j\mu}\rangle = \frac{1}{M^2a^2(1-a)^2}\sum_{k=1}^{M}\sum_{l=1}^{M}\Big(Z_{i\mu}Z_{j\mu}\langle(\xi_{k\mu}-a)\xi_{k\mu}(\xi_{l\mu}-a)\xi_{l\mu}\rangle + \sum_{\nu\neq\mu} Z_{i\nu}Z_{j\nu}\langle(\xi_{k\nu}-a)\xi_{k\mu}(\xi_{l\nu}-a)\xi_{l\mu}\rangle\Big). \quad (84)$$

Both of these expectations have been calculated en route to calculating the notebook's training error. Plugging Eqs. 66 and 70 into the above expression, we find,

$$\begin{aligned}\langle \tilde{Z}_{i\mu}\tilde{Z}_{j\mu}\rangle &= \frac{1}{M^2a^2(1-a)^2}\sum_{k=1}^{M}\sum_{l=1}^{M}\Bigg(Z_{i\mu}Z_{j\mu}\left(\delta_{kl}a(1-a)^2\right.\\ &\quad \left.+(1-\delta_{kl})\left(Ma^2(1-a)^2/(M-1) - a(1-a)^2/(M-1)\right)\right)\\ &\quad + \sum_{\nu\neq\mu} Z_{i\nu}Z_{j\nu}\left(\delta_{kl}a^2(1-a) - (1-\delta_{kl})a^2(1-a)\frac{aM-1}{(M-1)^2}\right)\Bigg)\\ &= \frac{1}{M^2a^2(1-a)^2}\Bigg(Z_{i\mu}Z_{j\mu}\left(Ma(1-a)^2 + M^2a^2(1-a)^2 - Ma(1-a)^2\right)\\ &\quad + \sum_{\nu\neq\mu} Z_{i\nu}Z_{j\nu}\left(Ma^2(1-a) - Ma^2(1-a)\frac{aM-1}{M-1}\right)\Bigg)\\ &= Z_{i\mu}Z_{j\mu} + \sum_{\nu\neq\mu} Z_{i\nu}Z_{j\nu}\frac{(M-1)a - a(aM-1)}{(M-1)Ma(1-a)}\\ &= Z_{i\mu}Z_{j\mu} + \sum_{\nu\neq\mu}\frac{Z_{i\nu}Z_{j\nu}}{M-1}. \quad (85)\end{aligned}$$

Therefore,

$$\begin{aligned}\langle \tilde{\Delta}_\mu w_i\rangle &= -2\lambda\left(\sum_{j=1}^{M} w_j\left(X_{i\mu}X_{j\mu} + \sum_{\nu\neq\mu}\frac{X_{i\nu}X_{j\nu}}{M-1}\right) - X_{i\mu}Y_\mu - \sum_{\nu\neq\mu}\frac{X_{i\nu}Y_\nu}{M-1}\right)\\ &= \Delta_\mu w_i + \frac{1}{M-1}\sum_{\nu\neq\mu}\Delta_\nu w_i \quad (86)\end{aligned}$$

where $\Delta_\mu w_i$ is the weight update that would occur if the student were perfectly reactivated by the notebook pattern μ. Equivalently, $\Delta_\mu w_i$ is the weight update that would occur from online learning to the teacher's example. Importantly, all contributions to $\langle\tilde{\Delta} w_i\rangle$ are in the gradient direction of one of the teacher examples. Rearranging this expression slightly, we find:

$$\langle \tilde{\Delta}_\mu w_i\rangle = \left(1 - \frac{1}{M-1}\right)\Delta_\mu w_i + \frac{1}{M-1}\sum_{\nu=1}^{P}\Delta_\nu w_i. \quad (87)$$

Therefore, each notebook reactivation of pattern μ is equivalent to a mini-batch update for that particular pattern with effective learning rate $\lambda\left(1 - \frac{1}{M-1}\right)$, plus a batch update for all stored patterns with effective learning rate $\frac{\lambda}{M-1}$. Similarly, the learning expected by sequential notebook reactivation of all P patterns is

$$\langle \tilde{\Delta} w_i \rangle \equiv \sum_{\mu=1}^{P} \langle \tilde{\Delta}_\mu w_i \rangle = \left(1 + \frac{P-1}{M-1}\right) \sum_{\mu=1}^{P} \Delta_\mu w_i \tag{88}$$

This is equivalent to batch learning with an effective learning rate of

$$\tilde{\lambda} = \lambda \left(1 + \frac{P-1}{M-1}\right) \tag{89}$$

In sum, the notebook's imperfect reactivation patterns hurt notebook memory performance, but they do not harm the student's ability to learn from past memories if the learning rate is appropriately controlled.

6.2 Student memory and generalization error from replay

As shown in Sections 5.2 and 6.1, notebook reactivations closely recapitulate stored student activity patterns when run below a critical capacity, and reactivation errors are correlated in such a way as to preserve the relevant statistics for student learning. In this regime, when replay events are random and the learning rate is small, the student effectively learns from the whole batch of samples. We therefore leverage known solutions to the batch learning dynamics of student-teacher models in our high-dimensional setting [11, 3]. Batch learning dynamics differ fundamentally from online learning dynamics, because in the batch setting the noise associated with each example is repeatedly revisited. This difference raises the danger of overfitting to the specific batch of stored data, rather than learning the general rule.

The average memory error is (see Section 2 of [3])

$$E_m(t) = \frac{1}{\alpha} \int \rho^{MP}(\lambda) \left(\frac{1+\lambda\mathcal{S}}{1+\mathcal{S}}\right) e^{-\frac{2\lambda t}{\tau}} d\lambda + \left(1 - \frac{1}{\alpha}\right) \frac{1}{1+\mathcal{S}} \mathbb{1}\{\alpha > 1\}, \tag{90}$$

and the generalization error is

$$E_g(t) = \frac{\mathcal{S}}{1+\mathcal{S}} \int \rho^{\mathrm{MP}}(\lambda) \left[e^{-\frac{2\lambda t}{\tau}} + \frac{1}{\lambda\mathcal{S}}(1 - e^{-\frac{\lambda t}{\tau}})^2\right] d\lambda + \frac{1}{1+\mathcal{S}}, \tag{91}$$

where here t measures time in units of epochs, such that as t goes from 0 to 1, and each stored example will be replayed once. The density $\rho^{\mathrm{MP}}(\cdot)$ denotes the Marchenko-Pastur distribution [13, 12], which describes the eigenvalue distribution of the input correlations XX^T in the high-dimensional regime. It has the form

$$\rho^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} + \mathbb{1}\{\alpha < 1\}(1-\alpha)\delta(\lambda) \tag{92}$$

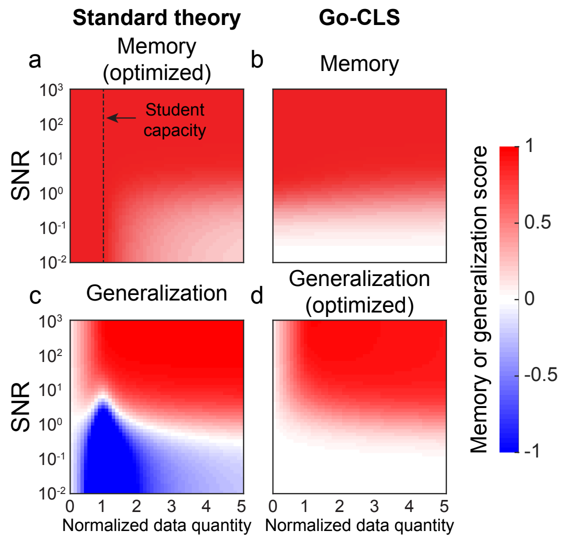

Figure 2: Heatmaps of student memorization performance (a, b) and generalization performance (c, d) as a function of SNR and α, when optimized for student memorization (a, c) or generalization (b, d).

for $\lambda = 0$ or $\lambda \in [\lambda_-, \lambda_+]$, and is zero elsewhere. Here $\mathbb{1}\{\cdot\}$ is an indicator function that is 1 when the argument is true and zero otherwise. The distribution comprises a delta function "spike" at zero (corresponding to input directions with zero variance that occur when there are fewer samples than the input dimension, i.e., $\alpha < 1$), and a "bulk" with upper and lower limits $\lambda_\pm = (\sqrt{\alpha} \pm 1)^2$ that depend on the load α.

We call this strategy memory-optimized, because Eqn. (90) is strictly decreasing in time, so to optimize student memory, replay should be continued indefinitely. However, while sustained replay optimizes student memory, Eqn. (91) shows that this strategy can cause catastrophic overfitting at the "student capacity" or interpolation threshold $\alpha = 1$, where the number of samples is equal to the input dimension and training error can just reach zero at long times. This overfitting behavior emerges in the high-dimensional regime, and better performance can be obtained for larger and smaller α, a finding known as the "double descent" phenomenon [12, 11, 3, 6]. The behavior of this strategy for a range of SNRs and loads α is depicted in Supplementary Fig. 2a,c. While memorization performance is good throughout this space, generalization suffers for low SNRs and loads near one.

6.3 Weight norm dynamics

While memory and generalization error are two key measures of learning progress, we can also ask how the strength of student weights changes throughout learning. This quantity could enable certain experimental links, for instance, as a proxy for functional connectivity in the context of the Sweegers et al. [15] experiment discussed in the main text.

A straightforward modification to the derivation in Section 2.1 of [3] yields the time-dependent average student weight norm as

$$\begin{aligned}\langle ||w(t)||_2^2\rangle &= \int \rho^{MP}(\lambda)\left[\sigma_w^2 e^{-2\lambda t/\tau} + \left(\sigma_{\bar{w}}^2 + \frac{\sigma_\epsilon^2}{\lambda}\right)\left(1 - e^{-\lambda t/\tau}\right)^2\right] d\lambda, \\ &= \int \rho^{MP}(\lambda)\left[\sigma_w^2 e^{-2\lambda t/\tau} + \frac{1+\lambda\mathcal{S}}{\lambda(1+\mathcal{S})}\left(1 - e^{-\lambda t/\tau}\right)^2\right] d\lambda, \qquad (93)\end{aligned}$$

where σ_w^2 denotes the initialization variance of the student weights, i.e., $w(0)_i \sim \mathcal{N}(0, \sigma_w^2)$. For large σ_w^2, this equation can describe an initial decrease in norm, followed by an increase in norm as weights align with the teacher. While in most of the supplement we have assumed $w(0) = 0$, we note that solutions for memory and generalization error dynamics for other weight initialization variances are well known [11, 3].

7 Generalization-optimized Replay Policy

The generalization-optimized replay policy is similar to the memory-optimized replay policy. Samples are stored in the notebook and replayed to the student to drive learning. When it comes time to make a prediction, the system with the best error is used. The key difference, however, is that replay is not continued indefinitely. Instead, replay is terminated when generalization error stops improving and starts to worsen due to overfitting. That is, this strategy regulates replay to maximize generalization error.

7.1 Student memory and generalization error with regulated replay

In detail, this strategy continues replay until the optimal early stopping time t^*, defined as

$$t^* = \text{argmin}_t E_g^{\text{MO}}(t), \qquad (94)$$

where $E_g^{\text{MO}}(t)$ is the memorization-optimized generalization error trajectory (i.e. unregulated trajectory) from Eqn. (91). The student memory and generalization error therefore have the piece-wise form

$$E_m(t) = \begin{cases} E_m^{\text{MO}}(t) & t < t^* \\ E_m^{\text{MO}}(t^*) & t \geq t^* \end{cases} \qquad (95)$$

$$E_g(t) = \begin{cases} E_g^{\text{MO}}(t) & t < t^* \\ E_g^{\text{MO}}(t^*) & t \geq t^* \end{cases} \qquad (96)$$

where $E_m^{\text{MO}}(t)$ denotes the memorization-optimized memory error trajectory from Eqn. (90). Crucially, under this regulated strategy, the student memory error can remain large indefinitely. Conversely, regulation avoids potentially catastrophic overfitting. The performance of the student under this strategy is depicted in Fig. 2b,d as a function of SNR $\mathcal{S}$ and load α. Finally, we note that, for this strategy, weight norm dynamics have a similar piece-wise form, such that the dynamics follow Eqn. 93 for $t < t^*$, at which point they stop.

7.2 Properties of early stopping

To gain a better understanding of this strategy, we can ask how the optimal stopping time depends on dataset parameters. While there is no closed form expression for t^*, some intuition can be obtained by computing the optimal stopping time for one fixed value of λ in the integral of Eqn. (91) (a strategy that would be exact if the MP distribution were a delta function at a single value of λ). The optimal stopping time is then (see Sec 2.2 of [3])

$$t^* = \frac{\tau}{\lambda}\log(\lambda\mathcal{S} + 1), \tag{97}$$

which shows that replay can continue longer for higher SNR relationships, though the relationship is logarithmic.

Early stopping is only one out of a variety of regularization strategies that can combat overfitting. Another possibility is to explicitly penalize large weight values. The L_2 regularization strategy sets the student weights according to

$$w^{L_2} = \mathrm{argmin}_w \mathcal{E}_m(w) + \frac{\omega}{2}||w||_2^2, \tag{98}$$

where ω denotes the regularization strength. The optimal L_2 regularization strength for our setting is known to be inversely proportional to SNR, $\omega^{\mathrm{opt}} = 1/\mathcal{S}$ (see [1, 2, 3]). Further, for the specific teacher and student regression problem we consider here, this regularization is known to be Bayes optimal, such that no algorithm can outperform it [1, 2]. It therefore can serve as a normative standard of comparison for early stopping. Prior work has shown that, in our setting, early stopping closely approximates the effect of explicit L_2 regularization (see, e.g., Fig. 5a of [3]), providing a normative basis for the early stopping strategy.

Finally, we can exploit the similar performance of early stopping and optimal L_2 regularization to obtain an explicit (but approximate) expression for the performance of the generalization-optimized replay strategy after the early stopping time. In particular, for $t > t^*$ we have

$$\begin{aligned} E_g(t) &\approx E_g^{L_2} \quad \text{for } t \geq t^* \\ &= \frac{\mathcal{S}}{2(1+\mathcal{S})}\left(1 - \alpha - 1/\mathcal{S} + \sqrt{(1/\mathcal{S} + \alpha - 1)^2 + 4/\mathcal{S}}\right) \\ &\quad + \frac{1}{1+\mathcal{S}}, \end{aligned} \tag{99}$$

where the latter step is the known generalization error of optimal L_2 regularization on this problem [1, 2, 3].

Using a similar approach, we can approximate the weight norm at the optimal stopping time as the weight norm of the optimal L_2 regularized solution (see Eqn. 66 [3]),

$$\begin{aligned} \left\langle \left\| w_{\text{opt}}^{L_2} \right\|_2^2 \right\rangle &= \sigma_{\bar{w}}^2 \int \rho^{\text{MP}}(\lambda) \frac{\lambda}{\lambda + 1/\mathcal{S}} d\lambda, \\ &= \int \rho^{\text{MP}}(\lambda) \frac{\lambda \mathcal{S}^2}{(1+\mathcal{S})(1+\lambda\mathcal{S})} d\lambda, \end{aligned} \tag{100}$$

which we note limits to 1 as $\mathcal{S} \to \infty$ and 0 as $\mathcal{S} \to 0$, such that high-SNR relationships have larger weight norms than low-SNR relationships at the optimal stopping time.

8 Example of generalization non-limiting unpredictability

The main text provides several examples of generalization-limiting unpredictability, with the canonical example being a teacher with output noise. However, not all sources of unpredictability are generalization limiting. For example, suppose that the teacher generates noiseless data,

$$y = \bar{w}x, \tag{101}$$

but the student has internal noise in its input neurons that affects its predictions

$$\hat{y} = \bar{w}(x + \eta). \tag{102}$$

Averaging over the input and noise distributions (but not the teacher weights), the generalization error of the student is

$$\begin{aligned} \mathcal{E}_g = \langle (y - \hat{y})^2 \rangle = \langle (y - \sum_i w_i(x_i + \eta_i))^2 \rangle \\ = \langle y^2 \rangle - 2\sum_i w_i \langle y(x_i + \eta_i) \rangle + \sum_i \sum_j w_i w_j \langle (x_i + \eta_i)(x_j + \eta_j) . \rangle \end{aligned} \tag{103}$$

Assuming that x, y, and η are zero mean random variables, and that η is uncorrelated with x and y, this is equal to

$$\mathcal{E}_g = \sigma_y^2 - 2C_{yx}w + w^T(C_{xx} + C_{\eta\eta})w. \tag{104}$$

Setting the derivative with respect to w equal to zero,

$$0 = \frac{\partial \mathcal{E}_g}{\partial w} = -2C_{yx} + 2w^T(C_{xx} + C_{\eta\eta}) \tag{105}$$

we find that the student weights that optimize generalization are

$$w^* = (C_{xx} + C_{\eta\eta})^{-1} C_{xy} \tag{106}$$

In contrast, the teacher weights satisfy

$$C_{yx} = \bar{w}^T C_{xx} \Longrightarrow \bar{w} = C_{xx}^{-1} C_{xy}. \tag{107}$$

Since $w^* \neq \bar{w}$, the generalization-optimized student is statistically biased,

$$\langle \hat{y} \rangle_\eta - y = (w^* - \bar{w})^T x \neq 0, \tag{108}$$

and the generalization error is nonzero. The teacher is unpredictable by the student.

Nevertheless, this type of unpredictability does not require strongly regulated systems consolidation. For example, suppose that the notebook perfectly memorizes P input-output patterns of the student. Then, the memory error averaged over student neuron noise is

$$\begin{aligned}\langle \mathcal{E}_m \rangle_\eta &= \frac{1}{P} \sum_\mu \left\langle \left(y_\mu - \sum_i w_i (x_{i\mu} + \eta_i) \right)^2 \right\rangle \\ &= \frac{1}{P} \sum_\mu \left(y_\mu^2 - 2 \sum_i w_i y_\mu x_{i\mu} + \sum_i \sum_j w_i w_j (x_{i\mu} x_{j\mu} + \langle \eta_i \eta_j \rangle) \right) \\ &= \frac{1}{P} \left(y^T y - 2 y X^T w^T + w^T (X X^T + C_{\eta\eta}) w \right).\end{aligned} \tag{109}$$

Setting the derivative with respect to w equal to 0,

$$0 = \frac{\partial \langle \mathcal{E}_m \rangle_\eta}{\partial w} = \frac{1}{P} \left(-2 y X^T + 2 w (X X^T + C_{\eta\eta}) \right) \tag{110}$$

we find that the weights minimizing the training error are

$$\hat{w} = y X^T (X X^T + C_{\eta\eta})^{-1}. \tag{111}$$

Noting that XX^T and yX^T are (proportional to) estimates of C_{xx} and C_{xy} given the P teacher examples, we see that this is the same basic form as the weights that minimize the generalization error.

In terms of the learning dynamics, the role played by eigenvalues of XX^T is now played by eigenvalues of $XX^T + C_{\eta\eta}$, which are lower bounded by the minimum eigenvalue of $C_{\eta\eta}$. For white noise, this is just σ_η^2. Overfitting was previously due to eigenvalues near 0, but those have now been shifted up to σ_η^2. The student input noise regularizes the learning process.

9 Complex teacher

Here we show that a mismatch between teacher and student, such that the teacher is deterministic but more complex than the student, is a form of generalization-limiting predictability that behaves similarly to observing a teacher with noise.

Because of the importance of this fact, we include a derivation in our notation for completeness. Our derivation follows Appendix C of [3]. Suppose the teacher generates inputs independently from some distribution $x \sim p(x)$, and labels them using the possibly nonlinear, stochastic function $y = g(x)$. The best possible linear student (i.e., the student trained on infinite data) will have weights

$$\hat{w}_{\text{OPT}} = C_{yx} C_{xx}^{\dagger} \tag{112}$$

where $C_{yx} = \langle yx^T \rangle$ is the input-output correlation matrix, $C_{xx} = \langle xx^T \rangle$ is the input correlation matrix, and $\dagger$ denotes the pseudoinverse.

We can rewrite the teacher output as the prediction of this optimal student and a residual,

$$y = C_{yx} C_{xx}^{\dagger} x + \delta y, \tag{113}$$

where the residual is $\delta y = g(x) - C_{yx} C_{xx}^{\dagger} x$.

Next, we consider learning the student weights from a finite batch of data with P examples, given in matrices Y, X with examples in the columns. The student weights are

$$\hat{w}_{\text{LS}} = YX^T \left(XX^T\right)^{\dagger} \tag{114}$$

$$= \hat{w}_{\text{OPT}} + \delta Y X^T \left(XX^T\right)^{\dagger} \tag{115}$$

where $\delta Y = Y - C_{yx} C_{xx}^{\dagger} X$ is the matrix of residuals. This formulation clearly separates contributions to the student weights into an optimal component and an overfitting component. Notably, the overfitting term $\delta Y X^T \left(XX^T\right)^{\dagger}$ has the same form as for additive noise.

10 Testable predictions of the theory

Future experiments could test our core theoretical prediction that the brain regulates the amount of systems consolidation based on the predictability of experience. This requires a behavioral paradigm that can vary the degree of predictability and a means to measure systems consolidation as a function of time. Sweegers et al. [15] performed a closely-related psychology experiment that associated visual stimuli with spatial locations on a computer screen, and they concluded that systems consolidation increased functional connectivity in human neocortex only when the stimulus-location association was predictable (Figs. 3l). However, the experimental details of Sweegers et al. were tailored for human participants, and an adapted paradigm that works in rodents would permit better characterizations of neural mechanism. In addition, although the different functional connectivity between brain areas was qualitatively observed for rule vs no-rule tasks in Sweegers et al., no rigorous statistical analysis was performed. It would be beneficial to verify this results in future work, including rodent studies.

Interestingly, Reinert et al. [14] recently showed that mice can learn predictable associations between visually oriented grating stimuli and appropriate behavior. A rodent-tailored paradigm could therefore present oriented grating stimuli in a cue location and require the animal to lick for water reward in one of two stimulus-instructed goal locations. As in Sweegers et al., the associations between stimuli and goal-locations could be assigned according to either a predictable rule or randomly. Memories of past associations could be assessed by measuring behavioral performance at various times after task acquisition, and generalization performance could be quantified by comparing the time required to learn new associations from the predictable and random rules.

We predict that systems consolidation would gradually allow mice to recall associations from the general rule without hippocampal involvement, but not associations from the random rule. This could be tested directly by measuring memory performance during reversible chemogenetic inactivation of the hippocampus at various times after memory acquisition. It could also be informative to extend the paradigm to intermediate predictability levels by associating stimuli with locations through partially predictable rules. Note that it would also be possible to train humans on an adapted version of this task. This would permit a study of human neural correlates using functional magnetic resonance imaging, but precisely timed and reversible inactivation of hippocampus is not possible in humans.

References

[1] M. Advani and S. Ganguli. An equivalence between high dimensional bayes optimal inference and m-estimation. *Advances in Neural Information Processing Systems*, 2016.

[2] M. Advani and S. Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

[3] M.S. Advani, A.M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[4] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530–1533, 1985.

[5] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Information storage in neural networks with low levels of activity. *Physical Review A*, 35(5):2293–2303, 1987.

[6] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, August 2019.

[7] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

[8] J. Buhmann, R. Divko, and K. Schulten. Associative memory with high information content. *Physical Review. A, General Physics*, 39(5):2689–2692, March 1989.

[9] J.F. Fontanari. Generalization in a hopfield network. *J Phys France*, 51:2421–2430, 1990.

[10] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982.

[11] A. Krogh and J.A. Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25:1135–1147, 1992.

[12] Y. LeCun, I. Kanter, and S.A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18):2396, 1991.

[13] V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114:507–536, 1967.

[14] S. Reinert, M. Hübener, T. Bonhoeffer, and P. M. Goltstein. Mouse prefrontal cortex represents learned rules for categorization. *Nature*, 593(7859):411–417, May 2021.

[15] C.C.G. Sweegers, A. Takashima, G. Fernández, and L.M. Talamini. Neural mechanisms supporting the extraction of general knowledge across episodic memories. *NeuroImage*, 87:138–146, February 2014.

[16] M.V. Tsodyks and M.V. Feigelman. The enhanced storage capacity in neural networks with low-level activity. *Europhysics Letters*, 6(2), 1988.