nature neuroscience

Computational principles of synaptic memory consolidation

Marcus K Benna[1] & Stefano Fusi[1,2]

Memories are stored and retained through complex, coupled processes operating on multiple timescales. To understand the computational principles behind these intricate networks of interactions, we construct a broad class of synaptic models that efficiently harness biological complexity to preserve numerous memories by protecting them against the adverse effects of overwriting. The memory capacity scales almost linearly with the number of synapses, which is a substantial improvement over the square root scaling of previous models. This was achieved by combining multiple dynamical processes that initially store memories in fast variables and then progressively transfer them to slower variables. Notably, the interactions between fast and slow variables are bidirectional. The proposed models are robust to parameter perturbations and can explain several properties of biological memory, including delayed expression of synaptic modifications, metaplasticity, and spacing effects.

The complexity and diversity of the numerous biological mechanisms that underlie memory is both fascinating and disconcerting. The molecular machinery responsible for memory consolidation at the level of synaptic connections is believed to employ a complex network of diverse biochemical processes that operate on different timescales[1,2]. Understanding how these processes are orchestrated to preserve memories over a lifetime requires guiding principles to interpret the complex organization of the observed synaptic molecular interactions and explain its computational advantage. Here we present a class of synaptic models that can efficiently harness biological complexity to store and preserve a huge number of memories on long timescales, vastly outperforming all previous synaptic models of memory.

The models we construct solve a long-standing dilemma: on the one hand, in a memory system that is continually receiving and storing new information, synaptic strengths representing memories must be protected from being overwritten during the storage of new information. Failure to provide such protection results in memory lifetimes that are catastrophically low[3–5]. On the other hand, protecting old memories too rigidly causes memory traces of new information to be weak, being represented by small numbers of synapses. This is one aspect of the plasticity–rigidity dilemma[6–9]. Synapses that are highly plastic are good at storing new memories but poor at retaining old ones. Less plastic synapses are good at preserving memories but poor at storing new ones.

Previous theoretical work has estimated the consequences of the plasticity–rigidity dilemma on the memory performance for various synaptic models characterized by different degrees of complexity. Early memory models[10] suggested that networks of neurons connected by simple synapses can preserve a number of memories that scales linearly with the size of the network. However, subsequent theoretical analyses[3–5] revealed that ignoring the limits on synaptic strengths imposed on any real biological system, which had appeared to be a harmless assumption in the calculations, was actually a serious flaw. When these limits are included—for example, in the extreme case of binary synapses in which the weight takes only two distinct values—the memory capacity grows only logarithmically with the number of synapses N for highly plastic synapses, and as $\sqrt{N}$ for rigid synapses that can store only a small amount of information per memory.

A possible resolution of this dilemma is to make each synapse complex enough to contain both plastic and rigid components. In many models the plastic components are represented by fast biochemical processes, which can change rapidly to store new memories. This initial memory trace is strong but labile; it decays quickly when other memories are stored. Memories can be consolidated if they are progressively transferred to the slow components. This mechanism is widely used in artificial devices (for example, computer memories, which include fast RAM and hard drives). It was proposed to explain memory consolidation at the systems level[8,11] and incorporated into a cascade model of synaptic memory based on multiple biochemical processes that operate on different timescales[9]. This form of synaptic complexity allows extended memory lifetimes without sacrificing the initial memory strength, accounting for our remarkable ability to remember for long times a large number of details even when memories are learned in one shot[12]. The two quantities that characterize memory performance, memory lifetime and the strength of the initial memory trace, scale as $\sqrt{N}$ in the cascade model[9].

Here we show that these models can be markedly improved upon when the network of interactions between the multiple biochemical processes that control the synaptic dynamics is bidirectional and appropriately tuned. Indeed, the decay of the memory trace becomes substantially slower than in previous models, leading to a memory lifetime that scales almost linearly with the number of synapses N. Notably, in our model longer memory lifetimes do not require a systematic

[1]Center for Theoretical Neuroscience, College of Physicians and Surgeons, Columbia University, New York, New York, USA. [2]Mortimer B. Zuckerman Mind Brain Behavior Institute, College of Physicians and Surgeons, Columbia University, New York, New York, USA. Correspondence should be addressed to S.F. (sf2237@columbia.edu).

Received 6 January; accepted 1 September; published online 3 October 2016; doi:10.1038/nn.4401

reduction in the initial memory strength, which still scales approximately as $\sqrt{N}$. Although the proposed synaptic model requires some tuning, it is robust to noise and variation in its parameters. Moreover, we construct a broad class of synaptic models that are equivalent in terms of memory performance. These different models capture the complexity and diversity of biochemical processes believed to be involved in memory consolidation. Thanks to their complexity, they can also reproduce the rich phenomenology of a plethora of biology and psychology experiments, including power-law memory decay[13,14], synaptic metaplasticity[15], delayed expression of synaptic potentiation and depression, and spacing effects[16,17].

RESULTS

The memory benchmark

To study the process of storing multiple memories and compare memory models, we need to make assumptions about the nature of memories. Storage of new memories is likely to exploit similarities with previously stored information (consider, for example, semantic memories). In what follows, we focus on mechanisms responsible for storing new information that has already been preprocessed in this way and is thus incompressible. For this reason, we consider memories that are unstructured (random) and do not have any correlations with previously stored information (uncorrelated).

Consider an ensemble of N synapses that is exposed to an ongoing stream of modifications, each leading to the storage of a new memory defined by the pattern of N synaptic modifications. We will select arbitrarily one of these memories and track it over time. The selected memory is not special in any way, so the results for this particular memory apply equally to all the memories being stored.

To track the selected memory, we take the point of view of an ideal observer who knows the strengths of all the synapses[4,9]. In the brain the readout is implemented by complex neural circuitry, and the strength of the memory trace based on the ideal observer approach may be substantially larger than the memory trace that is actually usable by the neural circuits. However, given the remarkable memory capacity of biological systems, it is not unreasonable to assume that the readout circuits perform almost optimally. Moreover, we will show that the ideal observer approach predicts the correct scaling properties of the memory capacity of simple neural circuits that actually perform memory retrieval. More quantitatively, we define the memory signal as the overlap between the state of the synaptic ensemble and the pattern of synaptic modifications originally imposed by the event being remembered. Previously stored memories, which are assumed to be random and uncorrelated, make the memory trace noisy. Memories that are stored after the tracked one continuously degrade the memory signal and also contribute to its fluctuations. We will monitor the signal to noise ratio (SNR) of a memory, which is defined as the ratio between the overlap and its standard deviation (see Online Methods, "Formal definition of memory signal and noise").

One measure of memory performance is the memory lifetime, the maximal time since storage over which a memory can be detected; i.e., for which the SNR is larger than some threshold of order one (whose precise value does not affect the scaling properties of the memory performance). If new memories arrive at a constant rate, the lifetime is proportional to the memory capacity because memories that have been stored more recently than the tracked one will have a larger SNR, and hence if the tracked memory is likely to be retrievable, so are more recent ones.

Constructing the synaptic model

The value of a synaptic weight w_a at any given time t is typically the result of multiple synaptic modifications. To build an efficient synaptic model, it is instructive to start from an abstract memory model in which the present weight is expressed as a sum of synaptic modifications $\Delta w_a(t_l)$, weighted by a factor r that decreases with the age of the modification $t - t_l$. The signal of the corresponding memory would decay as $r(t - t_l)$, while the noise would be approximately proportional to the square root of the variance of $w_a(t)$

$$\mathrm{Var}\left(w_a(t)\right) = \sum_{l:t_l<t} \left\langle \Delta w_a\left(t_l\right)^2 \right\rangle r\left(t-t_l\right)^2 \qquad (1)$$

where we have assumed that the expectation value of $\Delta w_a(t_l)$ vanishes, which is equivalent to hypothesizing that synaptic potentiation and depression are balanced. A slowly decaying r would enable the synaptic weight to maintain a dependence on a large number of modifications, but it would also induce a large variance for $w_a(t)$, potentially arbitrarily large if the sum extends over arbitrarily many modifications. By contrast, fast decays would limit the number of synaptic modifications that are remembered. Therefore, the memory capacity and its growth as a function of N depend crucially on $r(t)$. From equation (1) it is apparent that, in the case of random and uncorrelated modifications, the slowest power-law decay one can afford while keeping w finite is approximately $r(t) \approx t^{-1/2}$ (see also Online Methods section "Abstract models with linear superpositions of memories"). In **Supplementary Note 1**, we show that under some conditions this is approximately the optimal solution among all possible decay functions (see also Discussion). As we will explain in detail below, the SNR

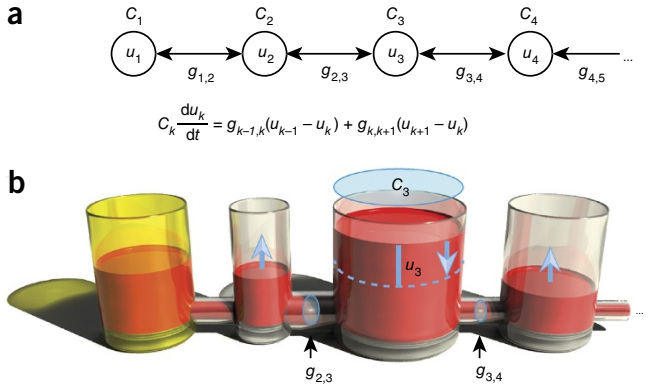

Figure 1 Model schematic. (**a**) Diagram of a simple synaptic plasticity model. The dynamical variables u_k represent different biochemical processes that are responsible for memory consolidation ($k = 1, \ldots, m$, where m is the total number of processes). They are arranged in a linear chain and interact only with their two nearest neighbors (see differential equation), except for the first and the last variable. The first one interacts only with the second one (and is also coupled to the input), while the last one interacts only with the penultimate one. Moreover, the last variable u_m has a leakage term that is proportional to its value (obtained by setting $u_{m+1} = 0$). The parameters $g_{k,k+1}$ are the strengths of the bidirectional interactions (double arrows). Together with the parameters C_k they determine the timescales on which each process operates. The first variable u_1 represents the strength of the synaptic weight. (**b**) The schematic model in **a** behaves like a set of communicating vessels. The u_k variables measure the deviation of the liquid level from equilibrium, shown in the third beaker as a blue dashed line. The C_k values represent the sizes (areas) of the beakers, and the coupling constants $g_{k,k+1}$ correspond to the cross-sections of the connecting tubes. The liquid level in the first beaker (yellow) represents the synaptic strength. The last beaker is connected to a reservoir whose liquid level is always at equilibrium. This interaction represents the leak in the differential equation of u_m.

is proportional to $\sqrt{N}$, and as a consequence a $t^{-1/2}$ decay would imply that the memory capacity scales linearly with N.

This abstract model reveals what kind of decay of the memory signal is desirable, but it does not explain how this behavior is achievable by synaptic dynamics. The next step is to construct a model that implements the desired power-law decay. One simple way would be to endow each synapse with a timer and introduce a mechanism to decrease the relative weight of each synaptic modification on the basis of the age of the modification[18], but this would just move the problem to the encoding and preservation of the memory age, which is potentially as difficult as the original memory problem we intend to solve. Fortunately, there is no need for a timer, as there are synaptic models in which the $1/\sqrt{t}$ decay emerges naturally from the interaction of multiple processes.

We will start with the construction of a simple chain model that captures and illustrates all the relevant scaling properties of more complex models. Then we will show how to generalize the model to incorporate less orderly interactions more similar to those observed in biological synapses. The simple chain model is characterized by multiple dynamical variables, each representing a different biochemical process (**Fig. 1a**). The first variable, which is the most plastic one, represents the strength of the synaptic weight. It is rapidly modified every time the conditions for synaptic potentiation or depression are met. The other dynamical variables represent biochemical processes that are affected by changes in the first variable. In the simplest configuration, these variables are arranged in a linear chain, and each variable interacts with its two nearest neighbors. These hidden variables tend to equilibrate around the weighted average of the neighboring variables. When the first variable is modified, the second variable tends to follow it. In this way a potentiation or depression is propagated downstream, through the chain of all variables. Importantly, the downstream variables also affect the upstream variables as the interactions are bidirectional.

To gain insight into the way this type of synapse works, it is useful to resort to an analogy with a set of communicating vessels, a more intuitive physical system (**Fig. 1b**). Each synaptic variable is represented by the level of liquid in a beaker. The interactions between variables are mediated by tubes that connect the beakers. The first beaker represents the synaptic weight. Potentiation of the synapse is implemented by pouring liquid into it, whereas depression is implemented by removing liquid. As the liquid level deviates from equilibrium, the fluid flow through the tubes will tend to balance the levels in all beakers. The balancing dynamics is fast when the beakers are small and the tubes large, but slow when the beakers are large and the tubes small. A single synaptic modification is remembered as long as the liquid levels remain significantly different from equilibrium.

We can construct the desired synaptic memory model by considering the analogous system of communicating vessels. An efficient memory system should have both long memory lifetimes (i.e., long relaxation times) and a large initial memory strength, obtained with a relatively small number m of variables (i.e., beakers). In a homogeneous chain (**Fig. 2a**), perturbations already decay with the desired $1/\sqrt{t}$ power law, but it requires a large m that grows as the square root of the memory lifetime. This problem can be circumvented by merging exponentially growing groups of beakers into larger ones of equivalent total area (**Fig. 2b**) and in addition reducing the sizes of the connecting tubes by exponentially increasing factors (**Fig. 2c**), which implies that the variables describing the system operate on different timescales that increase exponentially as one moves along the chain. This leads to a model with an approximately $1/\sqrt{t}$ decay of the

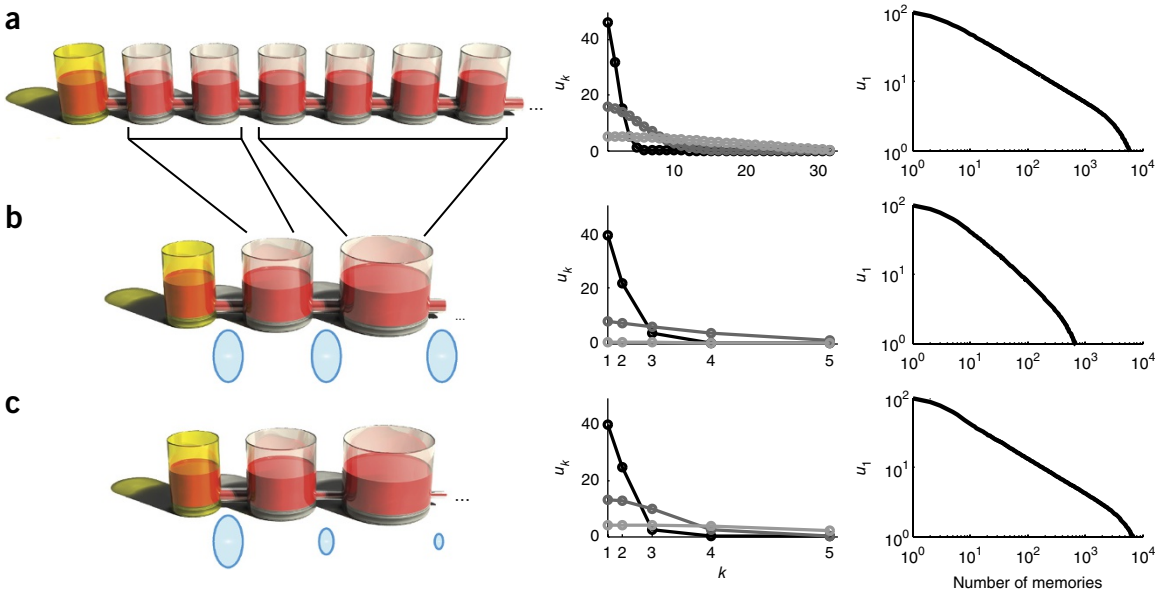

Figure 2 Model construction. (**a**) Relaxation dynamics in a set of 31 identical beakers connected by tubes of equal size ($C_k = 1$, $g_{k,k+1} = 1/8$). A perturbation of the liquid level of the first beaker diffuses to the others, slowly disappearing. The 31 u_k variables are shown in the middle at three different times. The decay of u_1, which approximates the desired $1/\sqrt{t}$ power law, is plotted on the right on a log–log scale. The number of beakers required in such a homogeneous system, however, grows as the square root of the number of stored memories. (**b**) A smaller set of beakers of progressively increasing sizes is obtained by merging those of **a**. The first beaker remains unchanged. The next two are merged into a larger beaker that contains the same volume of liquid as the two original ones. Then the next four beakers are combined, and so on, leading to successively larger ones ($C_k = 2^{k-1}$). The cross-sections of the tubes are still identical (indicated by blue ovals). While this merging procedure dramatically reduces the number of beakers, the convergence to equilibrium is now much faster than before ($\sim 1/t$). (**c**) We can recover the slow decay, without increasing the number of beakers, by tuning the cross-sections of the tubes connecting the communicating vessels. Their sizes are progressively reduced (by powers of two) to slow the decay ($g_{k,k+1} = 2^{-k-2}$), which now follows the desired $1/\sqrt{t}$ behavior over a time period that grows exponentially with the number of beakers.

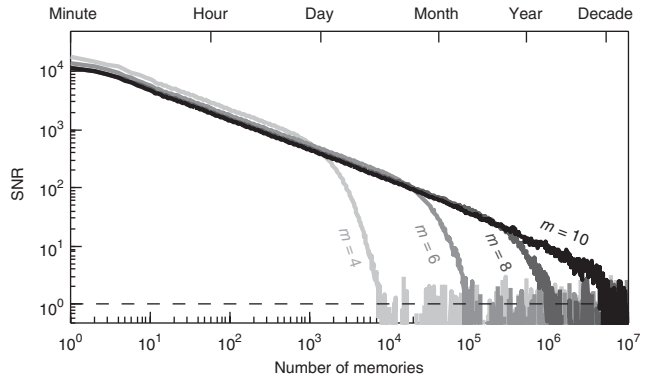

Figure 3 SNR of the synaptic model. Memory SNR as a function of the number of random uncorrelated memories that are stored after the tracked memory. The SNR is computed for a population of $N = 5.4 \times 10^9$ synapses. The scales of both axes are logarithmic. Different curves correspond to synaptic models with a different number m of dynamical variables (corresponding to the number of beakers in **Fig. 1**; here $m = 4$, 6, 8, 10). Each variable can vary on a discrete set of 40 equally spaced values. For all curves, the decay approximately follows a power law ($1/\sqrt{t}$) for a large number of memories and then becomes exponential where the curves visibly bend downwards. The range of the power-law decay increases exponentially with m, which is a measure of the complexity of the synapse. Memories are assumed to be stored at a constant rate of one new uncorrelated memory per unit time, which we choose to be 1 min here, so that the SNR decay can also be expressed as a function of time (upper horizontal axis). This choice of overall timescale is completely arbitrary, and time is considered only to help the reader appreciate the wide span of memory lifetimes. The memory lifetime is defined as the time elapsed since storage (or number of subsequently stored memories) at which the SNR falls below some arbitrary threshold (dashed line).

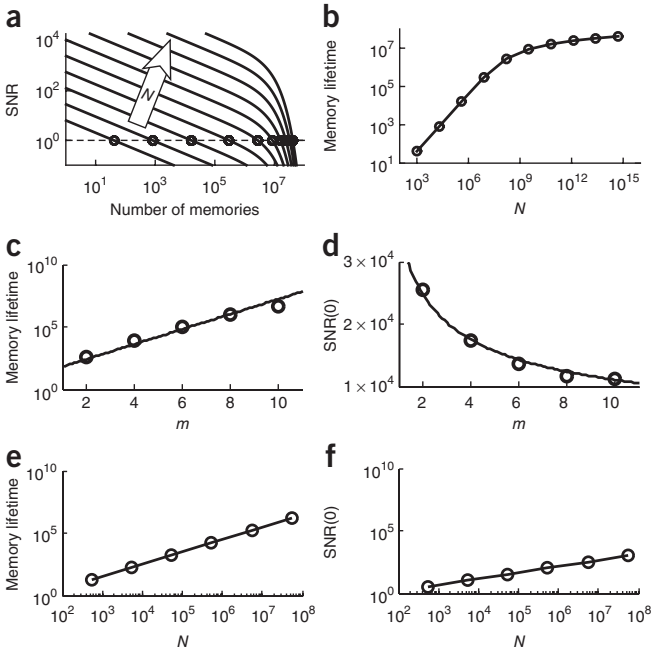

Figure 4 Scaling properties of the synaptic model. (**a**) Doubly logarithmic plots of a family of SNR curves versus time according to the approximate equation (2) for different values of N. All curves have the same shape but are displaced vertically, since the SNR is proportional to $\sqrt{N}$. The marked points of intersection with the retrieval threshold (dashed line) indicate the respective memory lifetimes. (**b**) Memory lifetime versus N corresponding to the curves in **a**, illustrating the connection between its scaling with N and the time dependence of the SNR. Initially the growth is linear, but it slows to a logarithmic increase when N is much larger than T. (**c**) Memory lifetime versus m. This and subsequent panels show simulation results from the model in **Figure 3**. The vertical axis is logarithmic, the horizontal one linear, so the line that fits the simulation points represents an exponential growth. (**d**) Initial SNR, denoted by SNR(0), versus m. Both axes are linear. As m increases, the initial SNR slowly degrades (as $\sim 1/\sqrt{m}$). $N = 5.4 \times 10^9$ in **c** and **d**. (**e**) Memory lifetime versus N on a log–log scale. The memory lifetime, which is proportional to the memory capacity (i.e., the total number of memories that can be recalled), increases linearly with N, as expected from the $1/\sqrt{t}$ decay of the SNR. This is a substantial improvement over previous synaptic models. (**f**) Initial SNR versus N on a log–log scale. SNR(0) grows as $\sqrt{N}$, as in the best previous synaptic models. Here and in **e** we set $m = 12$.

memory strength that requires only a number of variables that grows logarithmically with the memory lifetime (see also Online Methods, "Constructing models by coarse-graining random walks").

To understand more formally how the $1/\sqrt{t}$ decay is achieved, it is useful to consider the continuum limit of the equation in **Figure 1a**, in which the discrete index k can be replaced by a continuous variable x. For identical beakers and tubes (as in **Fig. 2a**), the differential equation governing the dynamics of the u_k variables becomes the well-known diffusion equation

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$$

with constant $D \propto g/C$. The fluid flow in the original discrete system of communicating vessels can be reinterpreted as the diffusion of particles or heat along the x-axis, with $u(x,t)$ representing the concentration or temperature at position x. Another familiar analogy is the cable equation, which governs the diffusion of voltage along axons or dendrites.

A unit perturbation introduced at time $t = 0$ and initially located at $x = 0$ spreads to neighboring locations in the shape of a continually broadening Gaussian peak. Its center remains at $x = 0$, where we read out the synaptic efficacy, and while the spatial extent of the perturbation grows as $\sqrt{t}$, the peak decays with time as the desired $1/\sqrt{t}$.

However, for this slow decay to continue at late times, the system requires an extended range of x, corresponding to a large number m of variables in the original system. Indeed, the above description of the shape of the perturbation holds only until it reaches the maximum value of x, which occurs at a time of $\mathcal{O}(m^2)$. One can dramatically reduce the number of required variables by considering an inhomogeneous diffusion process, with parameters $C(x)$ and $g(x)$ that depend exponentially on x, which leads to perturbations spreading only logarithmically with time (see Online Methods, "Continuum space limit and diffusion equation").

Discretization of the dynamical variables, scaling properties and memory capacity

It is unrealistic to assume that each dynamical variable u_k can vary over an unlimited range and be manipulated with arbitrary precision. Therefore, we discretize the u_k variables and impose rigid limits on them. The dynamics of the model is now described by stochastic transitions between a discrete set of levels for each variable, arranged to approximate the continuous-valued system constructed in **Figures 1** and **2**. At every time step the u_k values are first updated as in the case of continuous variables described above, but then each variable is discretized by setting it stochastically to one of the two values in the discrete set that are closest to the updated value. The probabilities of

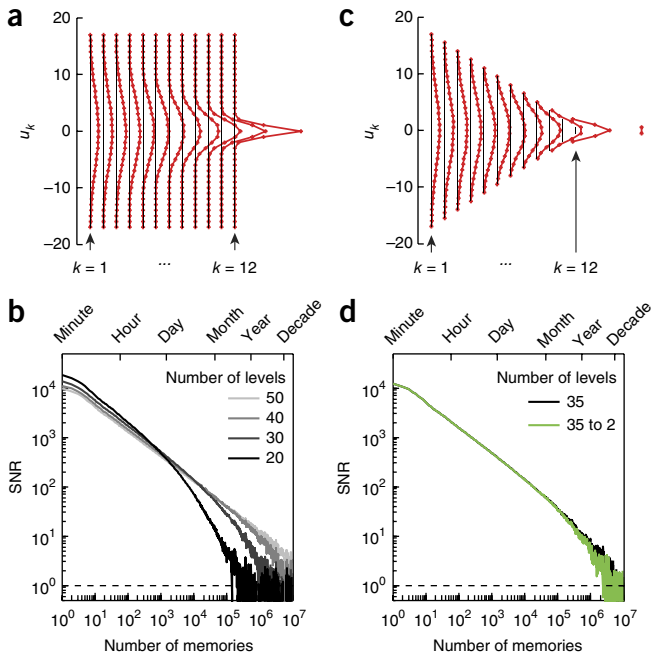

Figure 5 Effects of different discretization schemes of the dynamical variables on memory performance. (**a**) Distributions of the u_k variables for $k = 1, 2, \ldots, 12$ in a population of synapses at equilibrium. Each variable takes values in a discrete set of 35 equally spaced levels. All the distributions are approximately Gaussian (discretized and truncated) with a width that progressively decreases with k (see **Supplementary Note 2**). (**b**) SNR versus number of stored memories, as in **Figure 3**, for discretizations with different numbers of levels (namely 20, 30, 40 and 50). Since the values of almost all variables are well within the boundaries in **a**, the relaxation dynamics of the u_k variables is very similar to the unbounded case, and the SNR curve changes smoothly when we restrict the dynamical range further. (**c**) Distributions of the u_k variables when the number of discrete levels decreases with k. Because the distributions are narrower for the slower dynamical variables, one can reduce the range and the number of levels without affecting the memory performance appreciably. Here the number of equally spaced levels decreases linearly with k, and the distributions are very similar to those of **a**. The last variable has just two levels. (**d**) SNR versus number of stored memories for constant (black) and decreasing numbers of discrete levels (green). The performance remains almost unaffected by the reduction in the number of levels. This implies that the slower dynamical variables do not require as much precision as the faster ones and can operate with a surprisingly small number of levels.

ending up in each of the two values are chosen so that the average of the discretized u_k matches the continuous u_k (see also Online Methods, "Detailed description of models used in numerical simulations").

Assuming that memories are presented at a constant rate of one new uncorrelated memory per unit of time, we show in **Figure 3** the SNR as a function of time for memory models in which the complexity of the synapse (i.e., the number m of variables) progressively increases. The curves are plotted on a log–log scale, so a straight downwards line corresponds to a power-law decay. In all cases, the SNR decays approximately as $1/\sqrt{t}$ over a time interval that increases exponentially with the complexity of the synapse, before the decay accelerates and becomes exponential. These properties can be summarized by a simple approximate formula that expresses the SNR as a function of N, m and the age of the memory t:

$$\mathrm{SNR}(t) \approx \sqrt{\frac{N}{t}}\,\frac{e^{-t/T}}{\sqrt{\log T}} \qquad (2)$$

where T is the longest timescale of the synapse ($T \approx C_m/g_{m,m+1} = 2^{2m+1}$), which grows exponentially with m. This dependence on the parameters follows from the considerations in **Supplementary Notes 1** and **2**.

The memory lifetime of the tracked memory is the maximum time since storage for which the SNR is larger than some threshold that we set arbitrarily to 1 (**Fig. 3**). The SNR curves intersect the threshold when their decay is dominated by the exponential, and hence the memory lifetime is approximately equal to T, increasing exponentially with m (**Fig. 4c**). However, increasing T (by adding longer timescale variables) cannot improve the memory lifetime indefinitely, but only up to a limit of order N, the total number of synapses. If T is much larger than N, the SNR curve drops below the threshold in the power-law regime, where the exponential factor in equation (2) can be considered approximately constant. In this case the memory lifetime scales as N because the SNR is proportional to $\sqrt{N}$ and the threshold is reached when $\sqrt{N/t}$ is of order one.

These considerations can be illustrated by plotting the SNR of equation (2) as a function of time for different values of N (**Fig. 4a**). Graphically, changing N corresponds to shifting the SNR curves vertically in this doubly logarithmic plot. An upward shift moves the point of intersection with the threshold horizontally to the right and hence increases the memory lifetime. This increase is linear in N only if the SNR curve crosses the dashed line in the $1/\sqrt{t}$ power-law regime.

As the memory lifetime approaches the longest timescale T of the memory system the decay becomes exponential. In this regime, a shift of the SNR curve (which here bends downwards) due to an increase of N leads to only a modest (logarithmic) extension of the memory lifetime. We show explicitly the corresponding growth of the memory lifetime as a function of N, which is initially linear but then almost saturates (**Fig. 4b**). These scaling properties have been verified in simulations (**Fig. 4e**). We can avoid saturation by adjusting T appropriately (i.e., increasing m), which leads to a memory lifetime scaling as $N/\log N$ (see **Supplementary Note 1**).

The memory lifetime in previous models of complex synapses with bounded weights[9] scales at most as $\sqrt{N}$. A memory lifetime that scales (almost) linearly with the number of synapses constitutes a major improvement, especially in large neural systems. This improvement is achieved with a relatively small increase in the complexity of the synaptic machinery for memory consolidation, as m grows only logarithmically with the memory lifetime. Moreover, the initial SNR, which is related to the amount of information stored per memory, has the same scaling with N as in previous models (i.e., $\sqrt{N}$; **Fig. 4f**), and only decreases slowly with m (as $1/\sqrt{m}$; **Fig. 4d**).

Robustness of the model

The equilibrium distributions of the u_k variables are approximately Gaussian, and we can impose rigid boundaries (i.e., maximum and minimum values) on them[19] without perturbing their relaxation dynamics substantially compared to the unbounded case (**Fig. 5a**,**b**). The number of discrete levels required per variable scales only as $\sqrt{\log T}$. Since the width of the distributions decreases with k, longer timescale variables require fewer levels than u_1, down to only two states for u_m, consistent with bistable processes maintaining their state over very long time periods[20–25] (**Fig. 5c**,**d**).

The model not only is robust to discretization but also can tolerate surprisingly large perturbations of the optimal parameters. The SNR of the perturbed model clearly deviates from the SNR of the unperturbed model (**Fig. 6a**). However, the deviation increases smoothly with the amplitude of the perturbations and the SNR still decays approximately as $1/\sqrt{t}$. Notably, for long timescales there are still synapses that retain

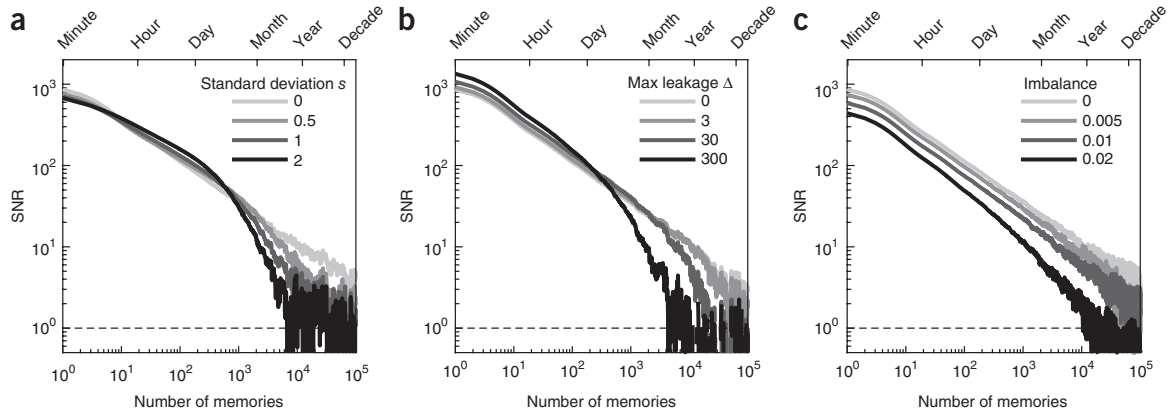

Figure 6 Robustness of the model. (**a**) SNR versus number of memories with perturbed model parameters. The effective coupling constants between neighboring dynamical variables are multiplied by stochastic variables drawn independently from a log-normal distribution (i.e., $g_{k,k+1}/C_k$ for forward and $g_{k,k+1}/C_{k+1}$ for backwards interactions between u_k and u_{k+1} are perturbed separately). The mean of these random variables is 1 and their standard deviation s. The SNR deviates from the $1/\sqrt{t}$ behavior, but the difference in memory performance is modest even for large standard deviations. Here $N = 2.5 \times 10^7$ with $m = 8$ variables per synapse and 37 levels per variable. (**b**) SNR versus number of memories for an ensemble of model synapses with additional leakage terms for all variables. We add to the right-hand side of the dynamical equations for $\mathrm{d}u_k/\mathrm{d}t$ additional decay terms proportional to $-u_k$ with random coefficients, which are $g_{m,m+1}/C_m$, the natural decay constant of the unperturbed model, multiplied by random numbers drawn independently for each variable from a uniform distribution between zero and Δ. When Δ is of order one the SNR curve is barely changed, but for large Δ the longest timescale T of the model is reduced by a factor of order $1/\Delta$ and the power-law decay breaks down sooner. (**c**) SNR versus number of memories when the rates of potentiation and depression are imbalanced. The power-law forgetting curves are very similar, but they are shifted downwards, even for fairly small imbalances. The memory performance is rather sensitive to an uncompensated imbalance between potentiation and depression.

the tracked memory, even when the SNR is below the threshold for retrievability. Indeed, the memory signal is still significantly different from zero in a subpopulation of synapses that happen to be well tuned. When reading out all synapses, this signal is too small compared to the noise. However, a smart selection mechanism[11] would enable the neural circuit to read out the memory even when the SNR of the synaptic population as a whole is too small.

The effects of another type of deformation are illustrated in **Figure 6b**, where we consider a more general class of synaptic models that, in the communicating vessels analogy, incorporate additional leakage pipes connecting every beaker to a reservoir held at the equilibrium level. Mathematically, this corresponds to extra decay terms proportional to $-u_k$ added to the right-hand side of the equations in **Figure 1a**, after dividing them by C_k. This is in contrast to the unperturbed model, in which only the last beaker has such a leak, with a decay coefficient inversely proportional to the longest timescale T. The deformed model is robust to extra decay terms that are smaller than or comparable in magnitude to that of the last variable of the unperturbed model. For larger leaks, however, the longest timescale of the model is reduced to approximately the inverse of the largest decay coefficient, and the onset of the exponential decay of the SNR correspondingly occurs sooner.

These results indicate that the model parameters do not need to be finely tuned. The model is less robust to perturbations in the input statistics, however. When the synaptic modifications are imbalanced (**Fig. 6c**), the decay remains almost unaltered, but the SNR curves shift downwards. The memory system is clearly sensitive to imbalances in the effective rates of potentiation and depression. However, even if the input statistics are imbalanced, the synapse may be able compensate for this by adjusting the relative magnitudes of the resulting plasticity steps (**Supplementary Note 3**). Malfunction of such a homeostatic mechanism could lead to memory decline, as observed in the early stages of Alzheimer's disease, when depression becomes more effective than potentiation[26].

Generalizations of the model

Above we considered synaptic models that can be represented by linear chains of dynamical variables. Their simplicity allowed us to illustrate the computational principles we used to design them. However, they appear too simple and orderly to accommodate the complexity and diversity of biological synapses. Here we construct a broad class of synaptic models that are equivalent to linear chains in terms of memory performance.

Models with arbitrarily complex networks of interactions can be constructed by starting from the undiscretized linear chain model depicted in **Figure 2** and then iteratively ramifying it by splitting off and merging branches (**Fig. 7a**). In **Supplementary Note 4** we show that with appropriately chosen parameters these complex models have the same dynamics for the first beaker and therefore the same memory performance as the linear chain models. We also note that they are robust to relatively large perturbations, such as the complete loss of one interaction pathway, which can be partially compensated by parallel branches. We can further generalize the model by considering plasticity events affecting longer timescale variables (rather than altering the synaptic efficacy directly), possibly different ones for potentiation and depression (see **Supplementary Note 3**).

Delayed expression of long-term potentiation and depression, metaplasticity and spacing

Our generalized synaptic models can readily reproduce various experimental observations, which include delayed expression of long-term potentiation (LTP) and depression (LTD), as well as metaplasticity[15,27], the dependence of plasticity on the history of previous synaptic modifications. There are several phenomenological models that can reproduce these observations[25]. However, here we show that this rich phenomenology can be captured by synaptic models that are also computationally efficient.

Metaplasticity is a natural consequence of the existence of hidden variables, represented by the beakers that are not directly read

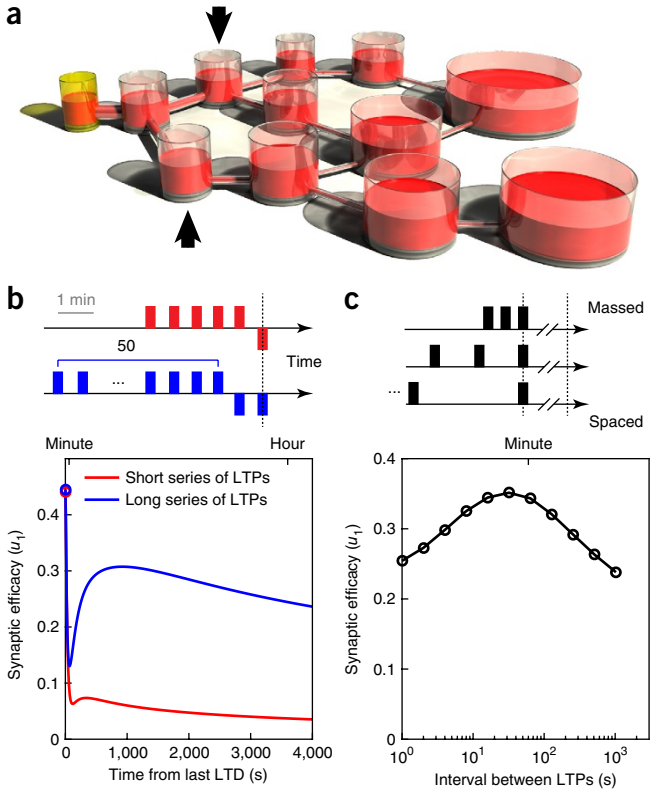

Figure 7 Generalizations and features of the model. (**a**) An example of a broad class of models in which variables can exhibit an arbitrary network of interactions (see **Supplementary Note 4**). With properly tuned parameters, the memory performance is the same as for the linear chain of beakers constructed in **Figure 2**. (**b**) When potentiation and depression act on intermediate-timescale beakers (indicated by up and down arrows, respectively, in **a**), we obtain delayed expression in addition to metaplasticity, the history dependence of the dynamics of the synaptic efficacy. We plot the efficacy versus the time elapsed since an LTD induction for two different protocols that lead to approximately equal initial efficacies. Red: LTD preceded by a short series of 5 LTP events. Depression is relatively stable and long-lasting. Blue: LTD preceded by a long series of 50 LTP events and another LTD event. Depression is more transient (despite two LTD events), revealing that the synapse is more resilient to long-term changes. (**c**) Synaptic efficacy 100 s (arbitrary units) after the end of a sequence of three LTP events versus spacing interval. Massed repetitions of LTP events (short intervals) and distributed repetitions (long intervals) are less effective than properly spaced ones. For long lags, the liquid added during potentiation has time to almost settle to equilibrium between repetitions, leading to little accumulation of synaptic modifications. For massed repetitions, conversely, one of the dynamical variables may hit its upper bound, corresponding to liquid spillover in our beaker analogy, reducing the overall effect of potentiation. The inverted U-shape of the plotted curve qualitatively reproduces observations of the spacing effect[29].

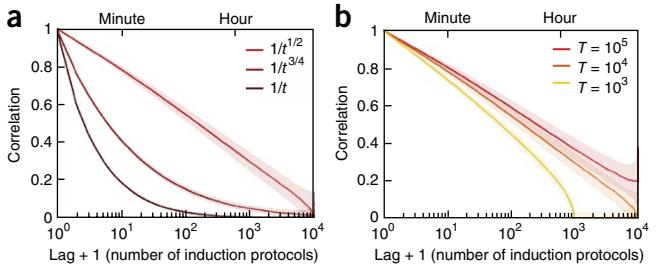

Figure 8 Testing the model in experiments. (**a**) Autocorrelation of the synaptic efficacy in a simulated experiment in which a synapse undergoes a long random series of 10,000 LTP and LTD protocols. Here we assumed a balanced input sequence with 10 protocols per minute. The scale of the time lag on the horizontal axis is logarithmic. The three curves represent the autocorrelation functions for three different models. Our proposed model (light red) has a distinctive decay, which appears almost as a straight line on a log–linear plot. Other models with faster decays of the SNR ($1/t^{3/4}$ and $1/t$ are shown) exhibit autocorrelations with a steeper falloff and prominent positive curvature. The shaded areas represent the standard error for 20 repetitions of the experiment. To measure the autocorrelation function of the synaptic efficacy, several technical issues must be overcome (see **Supplementary Note 7**). The duration of the experiment should be long enough for at least 1,000 brief induction protocols. Since one of the two protocols might be more effective than the other, some calibration is required to ensure LTP and LTD are suitably balanced. Unfortunately, the calibration procedure may require a time as long as the measurement period, as balance should be achieved on all timescales considered. (**b**) Dependence of the autocorrelation function on T. The different lines, again plotted on a log–linear scale, correspond to different longest timescales of our model. In the limit of very large timescales, this line would become horizontal.

out to determine the synaptic efficacy (see **Supplementary Note 5** for more general readout schemes). For example, synapses that undergo a long series of potentiating events become more resistant to depression[27]. **Figure 7b** illustrates these effects by comparing two different sequences of plasticity events. A long series of LTP induction events can increase the liquid levels in several beakers, making it more difficult to stabilize a subsequent synaptic depression. The different degrees of plasticity (despite equal efficacies immediately after the depression event) are determined by the states of the hidden variables, which were set by the previous history of synaptic modifications.

The model can also replicate the empirical phenomena known as spacing effects[16,17] (**Fig. 7c**). The stability of repeatedly stored memories is known to depend on the spacing between the times of memorization. This phenomenon has been observed in behavioral studies[16], but also in electrophysiology experiments on synaptic plasticity[28,29]. In these experiments, when the interval between repetitions is too short or too long, the memories are less stable than when the repetitions are properly spaced.

Testable predictions

A power-law decay of the memory SNR approximating $1/\sqrt{t}$ is a signature feature of our model. Although it is known that memory decay can be described by power laws in psychology experiments[13,14], the power varies substantially from experiment to experiment. This variability presumably occurs because the memories are not random and uncorrelated, as subjects often experience the same or similar episodes multiple times and can even internally rehearse previously stored memories. Consequently, the effective memory decay depends on the statistics of the memories, their relative importance, and the rate at which they are re-experienced, which we have not modeled in any generality.

A feasible experiment to test our theory would be to repeatedly modify a synapse and observe how the autocorrelation of the synaptic efficacy decays with time. A balanced, random series of LTP and LTD protocols can induce multiple changes in the synaptic efficacy. We expect that the observed autocorrelation would be very broad and its decay only logarithmic on long timescales (shorter than the memory lifetime; see **Supplementary Notes 6** and **7**). Such a logarithmic decay is a distinctive feature of models with a SNR that approximates $1/\sqrt{t}$. The autocorrelation is approximately a straight line when plotted against the logarithm of the time lag (**Fig. 8a**). We note that the slope

of the line depends on the longest timescale of the memory system under consideration (**Fig. 8b**).

DISCUSSION

We have presented a broad class of synaptic models that exhibit a huge memory capacity. These models show that complexity, which is widely observed in biological synapses, is important for achieving long memory lifetimes and strong initial memory traces. Complexity was shown to be beneficial in previous models, both for synaptic[9] and for systems level memory consolidation[11]. In both cases the memory traces were initially stored in fast variables and then progressively transferred to slower variables. Multiple timescales and memory transfer were the two key ingredients needed to achieve simultaneously slow decays of memory traces and strong initial signals. A $1/t$ decay, with t the age of the memory, led to initial memory traces and memory lifetimes whose magnitudes scale as $\sqrt{N}$, where N is the total number of synapses.

We show here that it is possible to combine the same key ingredients to markedly extend the memory lifetime without sacrificing the initial strength of the memory traces and without dramatically increasing the complexity of the synapse (for example, the number of dynamical variables). Indeed, the model presented here exhibits a substantially slower decay, approximately $1/\sqrt{t}$, which permits memory lifetimes that scale almost as N instead of $\sqrt{N}$ (see **Supplementary Note 8** for a direct comparison between models). When considering large systems such as the human brain, this is a huge improvement, obtained by introducing bidirectional interactions between fast and slow variables.

In our model the interactions between fast and slow variables are more important than in previous models. It is possible to build a system with noninteracting variables that exhibits a $1/\sqrt{t}$ decay (**Supplementary Note 9**). However, this requires disproportionately large populations of slow variables, which greatly reduce the initial SNR to $\mathcal{O}(N^{1/4})$, with memory lifetimes scaling only as $\sqrt{N}$. While for previous models interactions led to a considerable improvement[9], they did not substantially improve the scaling properties. Indeed, the model with noninteracting variables exhibited a $\sqrt{N}$ scaling for both the memory lifetime and the initial SNR, the same scaling obtained when fast and slow variables were interacting.

The proposed model synapse is complex, as it requires processes that operate on multiple timescales, but their number is relatively small and grows only logarithmically with the memory lifetime. Of note, for a given number of synapses there is an optimal number of synaptic variables (**Supplementary Note 8**), beyond which the memory performance slowly degrades. This implies that smaller nervous systems may do better with simpler synapses and larger nervous systems can benefit from more complex ones. For example, signaling complexity differs markedly between invertebrates and vertebrates[30], with an expansion of key synaptic components, notably receptors, adhesion and cytoskeletal proteins, and scaffold proteins. It is unclear how this measure of synaptic complexity relates to the number of synaptic variables in our model, but it illustrates that at least some forms of complexity grow with the number of synapses.

After discretizing the variables, our model has a finite number of states. The memory performance of any such model is bounded by the total number of internal states of the synapse[31]. Even though both the number of variables m and the number of discrete levels for each of them are small in our model, the space of all possible states of each synapse grows exponentially with m, which allows the slow memory decay it achieves (see **Supplementary Note 10**).

Optimality

The approximate $1/\sqrt{t}$ decay of the memory trace is the slowest allowed among power-law decays. Slower decays lead to synaptic efficacies that grow without bound. One can prove (see **Supplementary Note 1**) that the $1/\sqrt{t}$ decay maximizes the area between the log–log plot of the SNR and the threshold for memory retrieval (**Fig. 3**). This statement is true not only when one restricts the analysis to power laws, but also when all possible decay functions are considered.

Biological implementations of long timescales

One possible interpretation of the dynamical variables u_k is that they represent the deviations from equilibrium of chemical concentrations (see **Supplementary Note 11**). The timescales on which these variables change would then be determined by the equilibrium rates (and concentrations) of reversible chemical reactions. However, for the slowest variables, which vary on timescales on the order of years, it is probably necessary to consider biological implementations in which the u_k variables correspond to multistable processes. For example, we showed that the slowest variable can be discretized with only two levels, and hence it could be implemented by a bistable process, which would allow very long timescales[32–34]. For a small number of levels that is larger than two, one could combine multiple bistable processes or use slightly more complicated mechanisms[35].

These biochemical processes could be localized in individual synapses. However, these processes could also be distributed across different neurons in the same local circuit or even across multiple brain areas. The interaction between two coupled u_k variables could be mediated by neuronal activity, such as replay activity[11]. In the case of different brain areas, the synapses containing the fastest variables might be in the medial temporal lobe—for example, in the hippocampus—and the synapses with the slowest variables could reside in the long-range connections in the cortex. Our model would predict that the communication between these different areas should be bidirectional, which would imply that replay activity should be observed in both areas. This seems to be the case in at least two experimental studies[36,37].

Biological interpretations

Experiments on long-term synaptic modifications have revealed that synaptic consolidation is not a unitary phenomenon, but consists of multiple phases that involves different molecules. Directly mapping the variables u_k of our model to specific molecular processes would be interesting, but is probably difficult as our knowledge of the relevant biochemical processes is sparse and their dynamics characterized only incompletely. This is a known problem even in phenomenological models of synaptic processes.

One example involves models based on the synaptic tagging and capture (STC) hypothesis, which states that LTP consists of at least four steps[38,39]: first, the expression of synaptic potentiation with the setting of a local synaptic tag; second, the synthesis and distribution of plasticity-related proteins; third, the capture of these proteins by tagged synapses; and forth, the final stabilization of synaptic strength. Phenomenological models[24,25,40] of STC are characterized by four dynamical variables, whose dynamics may be triggered by neural inputs (see **Supplementary Note 12**), and can explain experiments on the induction of protein-synthesis-dependent late LTP. Two variables are related to tagging and probably involve calcium/calmodulin-dependent kinase II (CaMKII). The other two are important for protein-synthesis-based synaptic stabilization and may be related to protein kinase Mζ (PKMζ). However, each dynamical variable is likely to involve more than one molecular process. This suggests that they

should not be interpreted as concentrations of individual molecular species, but rather as 'reporters' indicating important changes in the molecular configuration of the synapse[25].

The situation is analogous for our u_k variables. Mapping them onto the dynamical variables of the phenomenological models would provide a different type of biological interpretation. This might require a complex transformation in which multiple u_k variables describe one or more nonlinear functions of the phenomenological variables. For example, in the STC model the synaptic efficacy is obtained as a sum of at least two components that depend on different dynamical variables. Although we discussed the case in which only u_1 is read out, our model can be extended to the case in which more than one dynamical variable determines the synaptic efficacy (see also **Supplementary Note 5**).

Memory retrieval

The ideal observer approach allowed us to analyze the scaling properties of memory systems with hardly any assumptions about the architecture of the neural circuit, the specific learning rule and the neural representations. However, it is important to test whether these scaling properties are preserved in specific neural circuits. We analyzed two simple cases of memory retrieval: a perceptron storing random patterns by one-shot learning and a fully connected recurrent neural network[10,41]. In both cases, the ideal observer approach predicts that the number of storable memories scales (almost) linearly with the number of neurons N_n (or synapses per neuron; see **Supplementary Note 13**). For the perceptron architecture, the linear scaling is verified in simulations. For attractor networks, the almost linear scaling predicted by the ideal observer approach well approximates the growth of the capacity with N_n observed in simulations and provides a consistently better description of the numerical results than the square root scaling that characterizes previous models of complex synapses[9] (**Supplementary Note 13**). This indicates that the approximately linear scaling of the ideal observer capacity may be preserved also with recurrent retrieval dynamics.

To study generalization, we trained a perceptron to classify random input patterns and then retrieved memories by imposing on the input neurons degraded versions of the stored patterns. The generalization ability can be expressed in terms of the minimum overlap between the input and the memory to be retrieved that can be tolerated (i.e., that produces the same response as the stored memory). This overlap is directly related to the SNR[42], and we show that it scales as 1/SNR. This demonstrates the importance of large SNRs, which allow better generalization ability.

We also presented the recurrent networks with degraded memory cues and tested that they could correctly retrieve the original memories. The scaling of the memory capacity with N_n was similar to that in the case of unperturbed cues, with the total number of retrievable memories decreasing smoothly with the level of degradation (**Supplementary Note 13**).

Sparseness of neural representations

Sparseness can increase the memory capacity both for synapses that can vary in a unlimited range[43] and for bounded, bistable synapses[3]. In both cases the number of storable memories scales almost quadratically with the number N_n of neurons when the representations are sparse enough (i.e., when f, the average fraction of active neurons, scales approximately as $1/N_n$). This is a notable improvement over the linear scaling obtained for dense representations. However, this capacity increase entails a reduction in the amount of information stored per memory.

The synaptic model we propose can also benefit from sparsification, with the SNR increasing by a factor of $\mathcal{O}\left(1/\sqrt{f}\right)$ (**Supplementary Note 14**). When $f \propto 1/N_n$ the memory capacity scales approximately quadratically with N_n, as in previous models[3], but with an initial SNR that is $\mathcal{O}(N_n)$ times larger. While an f of order $1/N_n$ may be compatible with electrophysiological data when N_n is the number of neurons of the local circuit, this is no longer true for much larger systems. Moreover, sparseness can also be beneficial for generalization, but only if f is not too small[44]. For these reasons, sparse representations are unlikely to be the sole solution to the memory problem.

Limitations of our approach

Our estimates of memory capacity are based on the ideal observer approach, and hence they only provide us with an upper bound on the memory signal. We validated our results in two local circuits, but it remains unclear how to perform this validation in large neural systems respecting the observed sparse connectivity and modular organization of the brain. Scalability has been studied only in specific cases[45,46] and is an important future direction for our work.

A second limitation, related to the first, arises from the assumption of random and uncorrelated synaptic modifications. Although it is reasonable to assume that the brain processes information to be stored so as to memorize only what is not already in memory, it is known that synaptic modifications are correlated, even when memories are not[3,47]. Fortunately, the disruptive effects of these correlations seem to disappear when neural representations are sparse[3] (see also **Supplementary Note 14**). Furthermore, the initiation of long-term synaptic modifications typically requires the coincidence of relatively rare events. It is not unreasonable to think that these mechanisms can also greatly contribute to the decorrelation of synaptic modifications. If this is the case, the theoretical framework that we developed will be applicable to a large number of memory systems.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to L.F. Abbott and U.S. Bhalla for many comments on the manuscript and for discussions. This work was supported by the Gatsby Charitable Foundation, the Simons Foundation, the Swartz Foundation, the Kavli Foundation, the Grossman Foundation and RISE, the Research Initiatives for Science and Engineering. The illustrations of the beakers were generated using the free ray tracing software POV-Ray.

AUTHOR CONTRIBUTIONS

M.K.B. conceived the original idea. M.K.B. and S.F. developed and analyzed the model, and wrote the article.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kandel, E., Swartz, J., Jessel, T., Siegelbaum, S. & Hudspeth, A.J. *Principles of Neural Science* (McGraw Hill, 2013).
2. Bhalla, U.S. Molecular computation in neurons: a modeling perspective. *Curr. Opin. Neurobiol.* **25**, 31–37 (2014).
3. Amit, D.J. & Fusi, S. Learning in neural networks with material synapses. *Neural Comput.* **6**, 957–982 (1994).
4. Fusi, S. Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biol. Cybern.* **87**, 459–470 (2002).
5. Fusi, S. & Abbott, L.F. Limits on the memory storage capacity of bounded synapses. *Nat. Neurosci.* **10**, 485–493 (2007).

6. McCloskey, M. & Cohen, N.J. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* **24**, 109–164 (1989).
7. Carpenter, G. & Grossberg, S. *Pattern Recognition by Self-Organizing Neural Networks* (MIT Press, 1991).
8. 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. *Psychol. Rev.* **102**, 419–457 (1995).
9. Fusi, S., Drew, P.J. & Abbott, L.F. Cascade models of synaptically stored memories. *Neuron* **45**, 599–611 (2005).
10. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558 (1982).
11. Roxin, A. & Fusi, S. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Comput. Biol.* **9**, e1003146 (2013).
12. Brady, T.F., Konkle, T., Alvarez, G.A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. USA* **105**, 14325–14329 (2008).
13. Wixted, J.T. & Ebbesen, E.B. On the form of forgetting. *Psychol. Sci.* **2**, 409–415 (1991).
14. Wixted, J.T. & Ebbesen, E.B. Genuine power curves in forgetting: a quantitative analysis of individual subject forgetting functions. *Mem. Cognit.* **25**, 731–739 (1997).
15. Abraham, W.C. Metaplasticity: tuning synapses and networks for plasticity. *Nat. Rev. Neurosci.* **9**, 387 (2008).
16. Anderson, John R. *Learning and Memory* (Wiley, 1995).
17. Smolen, P., Zhang, Y. & Byrne, J.H. The right time to learn: mechanisms and optimization of spaced learning. *Nat. Rev. Neurosci.* **17**, 77–88 (2016).
18. Wu, X.E. & Mel, B.W. Capacity-enhancing synaptic learning rules in a medial temporal lobe online learning model. *Neuron* **62**, 31–41 (2009).
19. Parisi, G. A memory which forgets. *J. Phys. A Math. Gen.* **19**, L617 (1986).
20. Lisman, J.E. A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. *Proc. Natl. Acad. Sci. USA* **82**, 3055–3057 (1985).
21. Fusi, S., Annunziato, M., Badoni, D., Salamon, A. & Amit, D.J. Spike-driven synaptic plasticity: theory, simulation, VLSI implementation. *Neural Comput.* **12**, 2227–2258 (2000).
22. Brader, J.M., Senn, W. & Fusi, S. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Comput.* **19**, 2881–2912 (2007).
23. Graupner, M. & Brunel, N. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proc. Natl. Acad. Sci. USA* **109**, 3991–3996 (2012).
24. Clopath, C., Ziegler, L., Vasilaki, E., Büsing, L. & Gerstner, W. Tag-trigger-consolidation: a model of early and late long-term-potentiation and depression. *PLoS Comput. Biol.* **4**, e1000248 (2008).
25. Ziegler, L., Zenke, F., Kastner, D.B. & Gerstner, W. Synaptic consolidation: from synapses to behavioral modeling. *J. Neurosci.* **35**, 1319–1334 (2015).
26. Shankar, G.M. *et al.* Amyloid-β protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. *Nat. Med.* **14**, 837–842 (2008).
27. O'Connor, D.H., Wittenberg, G.M. & Wang, S.S. Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proc. Natl. Acad. Sci. USA* **102**, 9679–9684 (2005).
28. Carew, T.J., Pinsker, H.M. & Kandel, E.R. Long-term habituation of a defensive withdrawal reflex in *Aplysia*. *Science* **175**, 451–454 (1972).
29. Zhou, Q., Tao, H.W. & Poo, M.M. Reversal and stabilization of synaptic modifications in a developing visual system. *Science* **300**, 1953–1957 (2003).
30. Emes, R.D. *et al.* Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nat. Neurosci.* **11**, 799–806 (2008).
31. Lahiri, S. & Ganguli, S. A memory frontier for complex synapses. *Adv. Neural Inf. Process. Syst.* **26**, 1034–1042 (2013).
32. Crick, F. Memory and molecular turnover. *Nature* **312**, 101 (1984).
33. Miller, P., Zhabotinsky, A.M., Lisman, J.E. & Wang, X.J. The stability of a stochastic CaMKII switch: dependence on the number of enzyme molecules and protein turnover. *PLoS Biol.* **3**, e107 (2005).
34. Si, K., Lindquist, S. & Kandel, E.R. A neuronal isoform of the *Aplysia* CPEB has prion-like properties. *Cell* **115**, 879–891 (2003).
35. Shouval, H.Z. Clusters of interacting receptors can stabilize synaptic efficacies. *Proc. Natl. Acad. Sci. USA* **102**, 14440–14445 (2005).
36. Ji, D. & Wilson, M.A. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100–107 (2007).
37. Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I. & Battaglia, F.P. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.* **12**, 919–926 (2009).
38. Reymann, K.G. & Frey, J.U. The late maintenance of hippocampal LTP: requirements, phases, 'synaptic tagging', 'late-associativity' and implications. *Neuropharmacology* **52**, 24–40 (2007).
39. Redondo, R.L. & Morris, R.G. Making memories last: the synaptic tagging and capture hypothesis. *Nat. Rev. Neurosci.* **12**, 17–30 (2011).
40. Barrett, A.B., Billings, G.O., Morris, R.G. & van Rossum, M.C. State based model of long-term potentiation and synaptic tagging and capture. *PLoS Comput. Biol.* **5**, e1000259 (2009).
41. Amit, D. *Modeling Brain Function* (Cambridge Univ. Press, 1989).
42. Krauth, W. & Mézard, M. Learning algorithms with optimal stability in neural networks. *J. Phys. A Math. Gen.* **20**, L745 (1987).
43. Tsodyks, M.V. & Feigel'man, M.V. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* **6**, 101–105 (1988).
44. Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
45. O'Kane, D. & Treves, A. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network* **3**, 379–384 (1992).
46. Roudi, Y. & Latham, P.E. A balanced memory network. *PLoS Comput. Biol.* **3**, 1679–1700 (2007).
47. Savin, C., Dayan, P. & Lengyel, M. Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area CA3. *PLoS Comput. Biol.* **10**, e1003489 (2014).

ONLINE METHODS

Formal definition of memory signal and noise. We assume that memories are stored through synaptic modification, with each new memory being encoded in a change in the efficacies of (a subset of) the synapses of a neural network. To formalize this problem, we will represent each memory as a random binary pattern $\Delta w_{ij}(t) = \pm 1$ of desired modifications (with +1 representing potentiation and −1 depression) of the synaptic weights between neurons labeled j and i. We will consider the components of $\Delta w_{ij}(t)$ to be uncorrelated (both across different memories and different synapses in a certain set), as would be the case if a suitable preprocessing step had decorrelated a stream of incoming patterns for optimal compression.

Note that we are not considering any particular network architecture and learning rule, but instead are working with synaptic modifications directly, thus sidestepping the learning rule that would determine them from the activities of pre- and postsynaptic neurons. This makes sense in the context of the ideal observer approach, where the underlying assumption is that all the information stored in the synaptic weights can be recovered, but of course it must be stressed that it is not obvious a priori whether there exists a network architecture that can in fact read out this information (see also the Discussion).

Nevertheless, classical memory models support the notion that the ideal observer approach correctly captures the scaling behavior of the achievable memory performance. For example, in the standard Hopfield model[10] the desirable modifications for a set of synapses that share a postsynaptic neuron would be uncorrelated (as assumed above), and a simple signal-to-noise analysis using the ideal observer approach correctly predicts a memory lifetime that scales linearly with the number of neurons.

If we index the set of N synapses under consideration by a (instead of i and j), the signal relevant for the retrieval of a particular memory that was stored at time t' is given by the overlap between the pattern of the associated (desirable) synaptic modifications Δw_a and the current state of the synaptic weights w_a at time t:

$$\mathcal{S}_{t'}(t) \equiv \frac{1}{N}\left\langle \sum_{a=1}^{N} w_a(t)\,\Delta w_a(t') \right\rangle \tag{3}$$

Here angle brackets indicate an average over the ensemble of random uncorrelated patterns that form the sequence of memories impinging on this set of synapses, and we have assumed for simplicity that the expectation value of Δw_a vanishes (i.e., the inputs are balanced); otherwise, a term proportional to this expectation value would have to be subtracted from the above.

Similarly, the corresponding (squared) noise term, again for the pattern stored at time t', is given by the variance of this overlap

$$\mathcal{N}^2_{t'}(t) \equiv \left\langle \frac{1}{N^2}\left(\sum_{a=1}^{N} w_a(t)\,\Delta w_a(t') \right)^2 \right\rangle - \mathcal{S}^2_{t'}(t)$$

The quotient of the signal and its standard deviation, the SNR, is the key quantity to consider when assessing the possibility and fidelity of recall of a previously stored memory. While we have considered a particular pattern stored at time t', we will assume that all memories are initially encoded with the same strength (though it is easy to generalize to a distribution of initial strengths), so that there is nothing special about any one memory. In this context, if the distribution of the synaptic weights reaches a steady state (as it does in the cases we are interested in), the SNR really only depends on the time $t - t'$ elapsed since storing the memory in question (i.e., the age of the memory). Accordingly, we will write it simply as a function of this time difference, which for a wide range of models will be monotonically decreasing.

A good memory system is one that has a large initial SNR, such that recent memories can easily be retrieved (using only a small—i.e., potentially highly corrupted—cue), and a long memory lifetime. The latter is defined as the time elapsed until the SNR drops below a certain retrieval threshold, the minimum value of $\mathcal{S}/\mathcal{N}$ at which recall is still possible. The precise value of this threshold will depend on the details of the network architecture and the retrieval dynamics, but as long as it is of order unity this will not affect the scaling results derived below, and thus in what follows we will simply set it to one unless otherwise noted. If the rate of memory storage is constant, the memory lifetime is proportional to the capacity of the system—i.e., the total number of memories that can be recalled at a given time. The tradeoff between the two goals of large initial signal and long memory lifetime will be discussed in detail below and will eventually lead us to optimizing an appropriately defined area under the signal-to-noise curve that captures the joint target of having as large a SNR as possible for as long as possible.

Desiderata for a useful synaptic memory model. Our aim here is to build a model of long-term memory that exhibits a number of properties we consider essential. We would like our model synapse to be able to learn online (one pattern at a time) and to forget gradually and smoothly without a phase transition such as the catastrophic forgetting in standard Hopfield-type models[41]. In addition to exhibiting a large initial SNR and long memory lifetime, the synaptic weights should reach a steady state distribution (given constant input statistics) that has support in only a small range of values (i.e., that does not allow arbitrarily large weights or, equivalently, weights in a finite range that must be read out with arbitrarily high precision). Note that one can easily obtain a model with bounded synaptic weights by restricting (hard-limiting) the range of a standard unbounded synapse (with plasticity events of unit magnitude) to values of order $\sqrt{N}$, which is still an unrealistically large number[5,19]. We will consider much more tightly bounded synaptic weights. Finally, all this has to be achieved while keeping the complexity of the model relatively small, avoiding overly baroque internal mechanisms involving too many variables.

Abstract models with linear superpositions of memories. *Basic assumptions.* To build an efficient synapse with bounded weights, we will start by considering a continuous synaptic variable with an additive plasticity rule and a time-dependent kernel $r(t - t')$, which we take to be the same for all synapses and plasticity events (i.e., across all stored patterns):

$$w_{ij}(t) = \sum_{t'<t} \Delta w_{ij}(t')\, r(t - t') \tag{4}$$

By additive plasticity rule we simply mean that the efficacy w_{ij} is a weighted sum over past plasticity events, which we take to be of fixed magnitude $\Delta w_{ij}(t) = \pm 1$ (with a plus sign for potentiation and minus for depression). The $\Delta w_{ij}(t)$ may be computed from the neural activations ξ_i corresponding to the patterns we want to store. For example, they could be determined according to a covariance rule $\Delta w_{ij}(t) \propto \left(\xi_i - \langle \xi_i \rangle\right)\left(\xi_j - \langle \xi_j \rangle\right)$, where the $\xi_i = \pm 1$ patterns are binary with equal probability for both values (such that $\langle \xi_i \rangle = 0$). Recall could be achieved by the network dynamics of an autoassociative Hopfield-type network[10] that completes the stored pattern of neural activations from a partial (or corrupted) cue $\tilde{\xi}_i$.

However, we deliberately divorce our analysis from the choice of learning rule and the network dynamics by focusing on a subset of synapses that receive statistically independent inputs and taking an ideal observer approach. Successful retrieval of a previously stored memory then requires the SNR of this set of synapses to be larger than a certain threshold (which we will set to 1).

We are assuming that potentiation and depression events are equally likely and are uncorrelated between different synapses and memories. In other words, we consider storing random patterns of synaptic modifications in which each bit of each memory can be thought of as determined independently by the flip of an unbiased coin. If this was not the case, a homeostatic mechanism would be needed to adjust the relative magnitude of these types of plasticity events to achieve a steady state without introducing any explicit bounds on the synaptic variables w_{ij}. (More generally, one could imagine a distribution over magnitudes of plasticity events, and again the existence of an equilibrium without explicit bounds on the weights would require a balance condition: namely, that the expectation value of the initial size of plasticity events vanishes. Another conceivable generalization would be to introduce different kernels for potentiation and depression events.)

Signal-to-noise ratio. We have introduced a time-dependent kernel $r(t - t')$ above since otherwise the synaptic weight would grow without bound as more and more patterns are stored. This can avoided, however, if $r(t - t')$ decays sufficiently fast as a function of the age of the corresponding memory (i.e., the time elapsed since storage).

Following the definition in equation (3), the signal (at time t) associated with a particular memory is given by the overlap of the corresponding pattern

of synaptic modifications (stored at time t') with the current synaptic weights, which using the ansatz (4) leads to

$$\mathcal{S}(t-t') = \frac{1}{N}\left\langle \sum_{a=1}^{N} w_a(t)\,\Delta w_a(t') \right\rangle = r(t-t')$$

where the neuronal indices i and j have now been replaced by a single synaptic index a ranging over the set of synapses under consideration. Combining this with the corresponding noise term, we obtain the SNR

$$\mathcal{S}/\mathcal{N}(t-t') = \sqrt{\frac{Nr^2(t-t')}{\sum_{t''<t,t''\neq t'} r^2(t-t'')}} \qquad (5)$$

It will be convenient in what follows to approximate the sum in the denominator by an integral over the full range of past t'' values (see also **Supplementary Note 6** for details), neglecting the small correction that arises from the fact that there is a term corresponding to $t'' = t'$ missing in the sum (since this term is the signal, rather than part of the noise). The noise will then be represented by an integral of the form

$$\int_1^\infty r^2(t)\,\mathrm{d}t$$

and thus if the decay kernel is a power law $r(t) = t^{-\gamma}$ then we must have $\gamma > 1/2$ or else this integral will not converge. Crucially, the divergence of this noise integral also implies that the variance of the synaptic weight would blow up, so that even if we regularized the integral appropriately for $\gamma < 1/2$, the resulting range of synaptic efficacies would be large. Therefore, the slowest power-law decay we can afford is $r(t) \approx t^{-1/2}$, which is the critical case in which the synaptic variance just starts to diverge (see also **Supplementary Note 1**).

Constructing models by coarse-graining random walks. Here we describe the procedure for building a model of a complex synapse that implements the required forgetting curve ($1/\sqrt{t}$) in a natural and parsimonious fashion. We will begin with general considerations of random walks and diffusion processes, and then refine as well as generalize the model step by step, in the following sections and **Supplementary Note 4**.

The present section serves primarily to provide a more systematic background for the model construction steps leading from **Figure 2a** to **Figure 2c** and furnish some mathematical details. Reducing the analogy of fluid flowing through a system of communication vessel to its most basic ingredients, we will consider a random walk of particles (which can be thought of as the molecules in the liquid) along a chain of discrete sites (which correspond to the beakers). Even though more abstract and general, this construction is equivalent to that of the main text in the particular case discussed there. See the next section for an alternative point of view using the (approximately equivalent) language of diffusion processes, which leads to a particularly simple description of the proposed synaptic dynamics that allows analytical derivations of some important properties of the model.

Linear chain models. Consider a random walk of particles on a semi-infinite chain in discrete time steps. We denote the number of particles at location j at time t by $v_j(t)$ for $j = 1 \ldots \infty$. (Note that this number can be negative; we can think of the particle number as being measured relative to a constant background density.) At every time step each particle has a finite probability of moving one step to the left or to the right. This probability is the same for both directions and for all locations except $j = 1$, which has no left neighbor. For such a stochastic process the time derivative of the particle numbers is equal to a discrete Laplacian: $\mathrm{d}v_j/\mathrm{d}t \propto v_{j-1} - 2v_j + v_{j+1}$ for $j > 1$. In other words, we have a spatially discretized diffusion process with constant diffusivity (see **Supplementary Note 4** for an illustration of a similar construction).

To make contact with systems of exponentially varying diffusivities that we are interested in, we will now consider discretizing the above random walk even further, on a coarser scale. We introduce a new set of coarse-grained variables u_i that are located at positions $j = 2^{i-1}$ on top of $v_{2^{i-1}}$; i.e., they are exponentially spaced. Our goal is to find an effective, approximate description of the system in terms of the u variables alone, where we think of each u_i as reflecting the average behavior of the system in the interval between its own location and that of its right neighbor u_{i+1}.

We can achieve this by assuming that the particle density profile is piecewise linear, with kinks located only at positions $j = 2^{i-1}$, such that all the curvature (which drives diffusion) is concentrated there. We can then use simple linear interpolation to eliminate all the v_j values from the equations of motion except those that coincide with the u_i. This would lead to the following expressions:

$$\frac{\mathrm{d}v_{2^{i-1}}}{\mathrm{d}t} \propto 2^{-i+2}\left(v_{2^{i-2}} - v_{2^{i-1}}\right) - 2^{-i+1}\left(v_{2^{i-1}} - v_{2^i}\right)$$

for $i = 2, 3, 4 \ldots$, while the time derivatives of the other v_j variables (for which j is not a power of two) would vanish because they are situated in regions of linear particle density.

However, for the piecewise linear approximation to be self-consistent (i.e., still applicable at the next time step), changing the particle number at the end of a line segment must be accompanied by an appropriate change everywhere along the segment to maintain linearity. In other words, the time derivative of the endpoint $v_{2^{i-1}}$ must be distributed among all variables along the line segment. Thus if our effective variable u_i is proportional to $v_{2^{i-1}}$, its time derivative must be proportional to the average derivative along the line segment to its right. (The details of this coarse-graining procedure are a matter of choice. The mathematically inclined reader might find it more appealing to have a symmetric prescription in which we average over the left and right line segments or, alternatively, to think of the u_i variables as living in the middle of a line segment. This would merely change the overall timescale of variation, which is unimportant here, so for ease of illustration we stick with a one-sided prescription.)

There are 2^{i-1} variables on this line segment, and denoting the constant of proportionality by $\alpha/2$ this leads to $\mathrm{d}u_i/\mathrm{d}t = 2^{-2i+2}\alpha(u_{i-1} - u_i) - 2^{-2i+1}\alpha(u_i - u_{i+1})$, which describes a discretized diffusion process on a logarithmic scale (i.e., as if viewed on a plot in which the spatial axis is logarithmic). In such a random walk model a plasticity event would correspond to adding or removing a particle from the leftmost location, which modifies the equation for $i = 1$. If we denote this time-dependent input of unit magnitude (and sign discriminating potentiation from depression) by $\mathcal{I}$, we find $\mathrm{d}u_1/\mathrm{d}t = \mathcal{I} - 2^{-1}\alpha(u_1 - u_2)$. Similarly, if the chain is not semi-infinite the equation for the last (mth) variable will only contain a coupling to its (sole) left-hand neighbor, but we can add a leak (exponential decay) term to it to render the variances of all particle numbers finite. This is easily achieved by simply setting the value of the (nonexistent) right-hand neighbor to zero, such that $\mathrm{d}u_m/\mathrm{d}t = 2^{-2m+2}\alpha(u_{m-1} - u_m) - 2^{-2m+1}\alpha u_m$.

Models with different ratios of timescales. While above and in the main text we have chosen parameters that vary as powers of two for ease of illustration, this can easily be generalized to arbitrary exponents

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = n^{-2i+2}\alpha(u_{i-1} - u_i) - n^{-2i+1}\alpha(u_i - u_{i+1}) \qquad (6)$$

which still approximates the desired $1/\sqrt{t}$ behavior of the Green's function for arbitrary real-valued $n > 1$, with ratios of successive timescales of $\mathcal{O}(n^2)$. The tradeoff in the choice of n is that for large n this approximation is not very good (since a superposition of a small number of exponentials leads to a rather bumpy surrogate for a power law), while for n only slightly bigger than unity a large number m of variables are needed to cover a given range of timescales, say between 1 and T, namely $m \approx \log T/(2\log n)$.

Note that even within the space of linear (and first order in time) equations with nearest neighbor interactions on a chain, we could generalize equation (6) even further by introducing different ratios of successive timescales instead of just one global parameter n, while still approximating the inverse square root Green's function.

Continuum space limit and diffusion equation. In the preceding section we discussed a set of first-order differential equations describing a random walk of a large number of particles (or, equivalently, the flow of water between connected beakers). In this construction space was discrete from the beginning (represented by a number of sites or beakers), but we could have chosen instead to step back even further and start from a model in which space is continuous. This even simpler model, which is highly instructive and allows an intuitive explanation of important properties such as the $1/\sqrt{t}$ decay, connects the proposed synaptic dynamics to heat diffusion on a line (for example, along a thermally insulated wire).

doi:10.1038/nn.4401

Consider the one-dimensional diffusion equation (with $u(x',t)$ interpreted as the temperature profile along a homogeneous rod)

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x'^2} \tag{7}$$

Its Green's function for a δ-function input (of one unit of heat energy) at time $t = 0$

$$G_u(x',t) = \frac{1}{\sqrt{4\pi Dt}} e^{-x'^2/4Dt} \tag{8}$$

decays as $1/\sqrt{t}$ at the origin (i.e., at $x' = 0$, where the δ-function is located). Thus if we represent the input to the system by such an instantaneous pulse, the correct decay of the signal is already built in, as long as we read out the synaptic weight at $x' = 0$. Since the equation is linear, we can simply superimpose Green's functions for a sequence of such inputs (positive for potentiation and negative for depression) and they will behave as required by equation (4).

Even though the Green's function we wrote here is for an infinite line, it is symmetric around the origin, and thus we can simply fold the system in half (leading to a Neumann boundary condition) and use the same Green's function (up to a factor of 2) on the semi-infinite line. A δ-function input at the origin will then evolve into a half Gaussian bump that will gradually spread, the peak remaining at the origin, with a standard deviation that grows in proportion to $\sqrt{t}$.

To revert back to the system of communicating beakers described above, we simply have to spatially discretize this diffusion process by chopping up the rod into finite chunks and considering the resulting interactions of the average temperatures of those chunks. The piece closest to the origin corresponds to the synaptic weight, while the other ones give rise to the hidden variables. If all those chunks have the same (say, unit) size, this will lead to the system shown in **Figure 2a**. While it has the correct decay behavior, the system cannot be of infinite extent. There will be some finite number m of separate chunks, and when the width of the Gaussian bump becomes comparable to the total size of the system, the $1/\sqrt{t}$ decay of the Green's function (equation (8)), which assumes an infinite system, will break down. In other words, if there is a second boundary, we have to choose a boundary condition there, which will modify the power-law decay on a timescale $T \propto m^2$. Thus if want to achieve an extensive memory lifetime $T \propto N$, the number of variables that would be required is $m \propto \sqrt{N}$, which is unrealistically large.

Note that we have assumed that the system is purely diffusive and free of any drift term. If that were not the case, the situation would be even worse, since the peak of the Green's function would move at a finite velocity and hit the second boundary at a time $T \propto m$, so that we would need even more variables ($m \propto N$) to obtain an extensive memory lifetime.

Fortunately, drastically reducing the number of required variables while maintaining a close approximation to power-law decay is not difficult. Recall that the (thermal) diffusivity D in equation (7) in general is a ratio of a thermal conductivity $g(x)$ and a heat capacity $C(x)$ and that those can vary spatially, which leads to the more general diffusion equation

$$\frac{\partial u}{\partial t} = \frac{1}{C(x)}\frac{\partial}{\partial x}\left(g(x)\frac{\partial u}{\partial x}\right)$$

If we break the homogeneity of the system by introducing exponentially varying parameters $C(x) \propto e^{\beta x}$ and $g(x) \propto e^{-\beta x}$, we obtain the differential equation

$$\frac{\partial u}{\partial t} = \frac{D}{\beta^2 e^{\beta x}}\frac{\partial}{\partial x}\left(e^{-\beta x}\frac{\partial u}{\partial x}\right) \tag{9}$$

parameterized by positive constants D and β, which has a Green's function given by

$$G_u(x,t) = \frac{1}{\sqrt{4\pi Dt}} e^{-(e^{\beta x}-1)^2/4Dt}$$

It describes a signal (in the form of a temperature difference) that propagates only very slowly toward larger x. This is because the thermal conductivity decreases exponentially while the heat capacity increases with x, and thus an input given by a certain amount of heat energy at $x = 0$ will lead to a noticeable temperature difference at finite x only at exponentially large times, when $t \approx (e^{\beta x} - 1)^2/4D$. Therefore, to reach an extensive memory lifetime, the largest value of x we need to consider, which is proportional to m, will now only scale as log$N$.

Throughout this diffusion process, the heat energy $Q = \int dx\ C(x)\ u(x,t)$ is a conserved quantity, modulo a leakage term potentially introduced by the second boundary condition at $x \propto m$. Spatially discretizing this system as above leads to the model of communicating vessels shown in **Figure 2c**, which achieves the correct power-law decay and extensive memory lifetime with only a logarithmic number of variables.

Note that the two continuum models (7) and (9) we have discussed in this section are in fact equivalent under the change of variables $x' = e^{\beta x} - 1$. This implies that there is another way of arriving at the simple linear chain model we want. We can start from a homogeneous diffusion process (constant diffusivity), but instead of discretizing space on a linear scale (into chunks of equal length) we can discretize on a logarithmic scale (i.e., divide the system into chunks of exponentially increasing size). This is precisely in the spirit in which we have described the transition from the homogeneous random walk (communicating vessels of constant size) to the desired linear chain model in **Figure 2** and the previous section. Both are approximations to a simple one-dimensional diffusion process, but spatially discretized in different ways, with the latter being much more efficient in terms of the number of variables needed.

Detailed description of models used in numerical simulations. While above we have written equations for a continuous time system, it is a simple matter to discretize time, as is appropriate for an incoming stream of temporally discrete patterns representing different experiences to be stored. We will choose one time step to correspond to one such memory and write

$$u_i(t+1) = u_i(t) + n^{-2i+2}\alpha\big(u_{i-1}(t) - u_i(t)\big) - n^{-2i+1}\alpha\big(u_i(t) - u_{i+1}(t)\big) \tag{10}$$

Again, the last equation (for $i = m$) is obtained from this by simply setting $u_{m+1} = 0$ for all times (thus introducing an exponential decay term on very long timescales), while the first equation (for $i = 1$) is modified by introducing the binary input of unit magnitude $\mathcal{I}(t)$:

$$u_1(t+1) = u_1(t) + \mathcal{I}(t) - n^{-1}\alpha\big(u_1(t) - u_2(t)\big) \tag{11}$$

α is a free parameter in these equations that determines the overall timescale of the dynamics (but its value should be chosen small enough such that the transition matrix on the right side of these equations has no negative eigenvalues, which could lead to oscillations). We will take $\alpha = 1/4$ below and in all numerical experiments.

Having discretized time, we are now left with a model of a complex synapse consisting of a small number m of coupled variables operating in discontinuous time steps, one step per incoming memory, according to the deterministic (given $\mathcal{I}(t)$) dynamical equations (10) and (11). However, the values of these variables are still continuous, and in the next step we will discretize those as well, thus turning the model into a Markov chain with inputs given by the random patterns to be stored and stochastic transition dynamics for the u_i.

To achieve this discretization, we will simply declare that every variable can take only one of a finite number of values (which we will refer to as levels) at every time step. We assume that these levels are integer-spaced and distributed symmetrically around zero (such that for an odd number of levels the allowed values are integers, while for an even number they are odd multiples of one half), though the algorithm described below can easily be generalized to arbitrary choices of discrete levels.

For every time step, we first compute the right sides of equations (10) and (11), with the $u_i(t)$ from the previous time step entering as (half) integers (and similarly the input $\mathcal{I}(t)$ as ±1). If the resulting $u_i(t+1)$ happens to coincide with one of the quantization levels there is nothing further to be done, but in general the result will fall between two levels, and in that case we must decide which of the two neighboring levels will be the new state of that variable. This can be done by independently flipping a biased coin for each such decision, with the odds ratio of the coin (corresponding to one or the other level being chosen) equal to the inverse ratio of the distances from the desired (unquantized) value to the respective levels, such that the closer one of the neighboring levels will be more likely.

The number of levels for each variable is finite, and if the right sides of equations (10) and (11) lead to a desired update for any variable u_i that would

cause it to become larger than the value of its highest level, we set it to this level with probability one (and similarly for the lower end of its dynamical range).

This is the fully discretized, stochastic model that we use for simulations (with $n = 2$), in particular those shown in **Figures 3–6**. It should be stressed, however, that the quantization of the variables is neither necessary for the model to work nor required for a plausible biophysical implementation. In fact the SNR will be somewhat higher without the additional noise introduced by the stochasticity of the random choices between nearby levels (though the scaling behavior appears to be the same). However, even though we do not need stable, discrete levels, we do perform this quantization in order to emphasize that the variables never need to be kept track of with high precision (as long as there is no systematic drift) and that there is no biologically implausible information hidden in exactly read out continuous variables.

For the simulations of **Figure 7**, by contrast, we use continuous-valued variables to better illustrate the transient dynamics of the synaptic efficacy. There we also generalize to a broad class of complex synaptic models in which each dynamical variable is coupled to two or more other variables, forming more complicated networks of interactions. These models are constructed iteratively starting from the linear chain of beakers of **Figure 2**. For example, the second beaker could be connected to two identical beakers on its right instead of one, splitting the chain into two. Each of the two beakers would then be connected to a series of progressively larger ones. Pairs of corresponding beakers would have the same total capacity as the associated single beaker of the original chain. This ramification process can be iterated an arbitrary number of times, with any choice of relative importance weights assigned to different branches. Furthermore, such branches can merge again, leading to complex networks of interactions, as described in detail in **Supplementary Note 4**. When the cross-sections of the tubes are properly tuned, the memory performance of the model is the same as for the original linear chain of beakers.

Another generalization we introduce in **Figure 7b** to model delayed plasticity is inputs affecting longer timescale variables. Until now we have considered models in which the synaptic efficacy is instantaneously modified by adding or removing liquid from the first beaker. The long-term memory performance remains basically unaltered when liquid is added or removed from other beakers instead, but the expression of a synaptic modification is delayed by the time it takes the liquid to flow into the beaker representing the efficacy. This suggests that LTP and LTD induction protocols may affect distinct biochemical processes that correspond to different beakers in the model and do not need to operate directly on the same variable, as discussed further in **Supplementary Note 3**.

For the simulated experiments comparing synapses with different decay functions in **Figure 8**, we use the abstract model of equation (4) above.

Code availability. The code for the simulations was written in C++ and Matlab, and is available upon request.

A **Supplementary Methods Checklist** is available.

doi:10.1038/nn.4401