Optimal synaptic strategies for different timescales of memory

Subhaneil Lahiri [*] **and Surya Ganguli** [*]

[*] Department of Applied Physics, Stanford University, Stanford CA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

An incredible gulf separates theoretical models of synapses, often described solely by a single scalar value denoting the size of a postsynaptic potential, from the immense complexity of molecular signaling pathways underlying real synapses. To understand the functional contribution of such molecular complexity to learning and memory, it is essential to expand our theoretical conception of a synapse from a single scalar to an entire dynamical system with many internal molecular functional states. Moreover, theoretical considerations alone demand such an expansion; network models with scalar synapses assuming finite numbers of distinguishable synaptic strengths have strikingly limited memory capacity. This raises the fundamental question, how does synaptic complexity give rise to memory? To address this, we develop new mathematical theorems elucidating the relationship between the structural organization and memory properties of complex synapses that are themselves molecular networks. Moreover, in proving such theorems, we uncover a framework, based on first passage time theory, to impose an order on the internal states of complex synaptic models, thereby simplifying the relationship between synaptic structure and function.

Overall, we uncover general design principles governing the functional organization of complex molecular networks, and suggest new experimental observables in synaptic physiology, based on first

synaptic plasticity | learning | memory | neuroscience

Abbreviations: SNR, signal-to-noise ratio; LTP, long term potentiation; LTD, long term depression

It is widely thought that our very ability to remember the past over long time scales depends crucially on our ability to modify synapses in our brain in an experience dependent manner. Classical models of synaptic plasticity model synaptic efficacy as an analog scalar value, denoting the size of a post-synaptic potential injected into one neuron from another. Theoretical work has shown that such models have a reasonable, extensive memory capacity, in which the number of long term associations that can be stored by a neuron is proportional its number of afferent synapses [1, 2, 3]. However, recent experimental work has shown that many synapses are more digital than analog; they cannot robustly assume an infinite continuum of analog values, but rather can only take on a finite number of distinguishable strengths, a number than can be as small as two [4, 5, 6] (though see [7]). This one simple modification leads to a catastrophe in memory capacity: classical models with digital synapses, when operating in a palimpsest mode in which the ongoing storage of new memories can overwrite previous memories, have a memory capacity proportional to the logarithm of the number of synapses [8, 9]. Intuitively, when synapses are digital, the storage of a new memory can flip a population of synaptic switches, thereby rapidly erasing previous memories stored in the same synaptic population. This result indicates that the dominant theoretical basis for the storage of long term memories in modifiable synaptic switches is flawed.

Recent work [10, 11, 12] has suggested that a way out of this logarithmic catastrophe is to expand our theoretical conception of a synapse from a single scalar value to an entire stochastic dynamical system in its own right. This conceptual expansion is further necessitated by the experimental reality that synapses contain within them immensely complex molecular signaling pathways, with many internal molecular functional states (e.g. see [4, 13, 14]). While externally, synaptic efficacy could be digital, candidate patterns of electrical activity leading to potentiation or depression could yield transitions between these internal molecular states without necessarily inducing an associated change in synaptic efficacy. This form of synaptic change, known as metaplasticity [15, 16], can allow the probability of synaptic potentiation or depression to acquire a rich dependence on the history of prior changes in efficacy, thereby potentially improving memory capacity.

Theoretical studies of complex, metaplastic synapses have focused on analyzing the memory performance of a limited number of very specific molecular dynamical systems, characterized by a number of internal states in which potentiation and depression each induce a specific set of allowable transitions between states (e.g. see Fig. 1 below). While these models can vastly outperform simple binary synaptic switches, these analyses leave open several deep and important questions. For example, how does the structure of a synaptic dynamical system determine its memory performance? What are the fundamental limits of memory performance over the space of all possible synaptic dynamical systems? What is the structural organization of synaptic dynamical systems that achieve these limits? Moreover, from an experimental perspective, it is unlikely that all synapses can be described by a single canonical synaptic model; just like the case of neurons, there is an incredible diversity of molecular networks underlying synapses both across species and across brain regions within a single organism [17]. In order to elucidate the functional contribution of this diverse molecular complexity to learning and memory, it is essential to move beyond the analysis of specific models and instead develop a general theory of learning and memory for complex synapses. Moreover, such a general theory of complex synapses could aid in development of novel artificial memory storage devices.

Here we initiate such a general theory by proving upper bounds on the memory curve associated with any synaptic dynamical system, within the well established ideal observer framework of [10, 11, 18]. Along the way we develop principles based on first passage time theory to order the structure of synaptic dynamical systems and relate this structure

Reserved for Publication Footnotes

to memory performance. We summarize our main results in the discussion section.

Overall framework: synaptic models and their memory curves

In this section, we describe the class of models of synaptic plasticity that we are studying and how we quantify their memory performance. In the subsequent sections, we will find upper bounds on this performance.

We use a well established formalism for the study of learning and memory with complex synapses (see [10, 11, 18]). In this approach, electrical patterns of activity corresponding to candidate potentiating and depressing plasticity events occur randomly and independently at all synapses at a Poisson rate r. These events reflect possible synaptic changes due to either spontaneous network activity, or the storage of new memories. We let f^{pot} and f^{dep} denote the fraction of these events that are candidate potentiating or depressing events respectively. Furthermore, we assume our synaptic model has M internal molecular functional states, and that a candidate potentiating (depotentiating) event induces a stochastic transition in the internal state described by an $M \times M$ discrete time Markov transition matrix $\mathbf{M}^{\text{pot}}$ ($\mathbf{M}^{\text{dep}}$). In this framework, the states of different synapses will be independent, and the entire synaptic population can be fully described by the probability distribution across these states, which we will indicate with the row-vector $\partial(t)$. Thus the i'th component of $\partial(t)$ denotes the fraction of the synaptic population in state i. Furthermore, each state i has its own synaptic weight, $\mathbf{w}_i$, which we take, in the worst case scenario, to be restricted to two values. After shifting and scaling these two values, we can assume they are ±1, without loss of generality.

We also employ an "ideal observer" approach to the memory readout, where the synaptic weights are read directly. This provides an upper bound on the quality of any readout using neural activity.

For any single memory, stored at time $t = 0$, we assume there will be an ideal pattern of synaptic weights across a population of N synapses, the N-element vector $\vec{w}_{\text{ideal}}$, that is $+1$ at all synapses that experience a candidate potentiation event, and -1 at all synapses that experience a candidate depression event at the time of memory storage. We assume that any pattern of synaptic weights close to $\vec{w}_{\text{ideal}}$ is sufficient to recall the memory. However, the actual pattern of synaptic weights at some later time, t, will change to $\vec{w}(t)$ due to further modifications from the storage of subsequent memories. We can use the overlap between these, $\vec{w}_{\text{ideal}} \cdot \vec{w}(t)$, as a measure of the quality of the memory. As $t \to \infty$, the system will return to its steady state distribution which will be uncorrelated with the memory stored at $t = 0$. The probability distribution of the quantity $\vec{w}_{\text{ideal}} \cdot \vec{w}(\infty)$ can be used as a "null model" for comparison.

The extent to which the memory has been stored is described by a signal-to-noise ratio (SNR) [10, 11]:

$$\mathrm{SNR}(t) = \frac{\langle \vec{w}_{\text{ideal}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\text{ideal}} \cdot \vec{w}(\infty) \rangle}{\sqrt{\mathrm{Var}(\vec{w}_{\text{ideal}} \cdot \vec{w}(\infty))}}. \qquad [\mathbf{1}]$$

The noise in the denominator is essentially $\sqrt{N}$. There is a correction when potentiation and depression are imbalanced, but this will not affect the upper bounds that we will discuss below and will be ignored in the subsequent formulae.

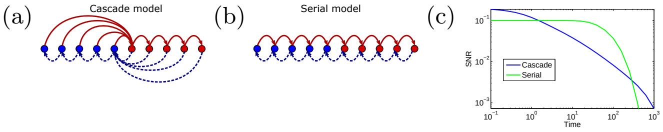

Fig. 1. Models of complex synapses. (**??**) The cascade model of [10], showing transitions between states of high/low synaptic weight (red/blue circles) due to potentiation/depression (solid red/dashed blue arrows). (**??**) The serial model of [12]. () The memory curves of these two models, showing the decay of the signal-to-noise ratio (to be defined below) as subsequent memories are stored.

A simple average memory curve can be derived as follows. All of the preceding plasticity events, prior to $t = 0$, will put the population of synapses in its steady-state distribution, $\mathbf{p}^\infty$. The memory we are tracking at $t = 0$ will change the internal state distribution to $\mathbf{p}^\infty \mathbf{M}^{\text{pot}}$ (or $\mathbf{p}^\infty \mathbf{M}^{\text{dep}}$) in those synapses that experience a candidate potentiation (or depression) event. As the potentiating/depressing nature of the subsequent memories is independent of $\vec{w}_{\text{ideal}}$, we can average over all sequences, resulting in the evolution of the probability distribution:

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = r\mathbf{p}(t)\mathbf{W}^{\mathrm{F}}, \qquad \text{where} \quad \mathbf{W}^{\mathrm{F}} = f^{\text{pot}}\mathbf{M}^{\text{pot}} + f^{\text{dep}}\mathbf{M}^{\text{dep}} - \mathbf{I}. \qquad [\mathbf{2}]$$

Here $\mathbf{W}^{\mathrm{F}}$ is a continuous time transition matrix that models the process of forgetting the memory stored at time $t = 0$ due to random candidate potentiation/depression events occurring at each synapse due to the storage of subsequent memories. Its stationary distribution is $\mathbf{p}^\infty$.

This results in the following SNR

$$\mathrm{SNR}(t) = \sqrt{N}(2f^{\text{pot}}f^{\text{dep}})\mathbf{p}^\infty \left[\mathbf{M}^{\text{pot}} - \mathbf{M}^{\text{dep}}\right] \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}. \qquad [\mathbf{3}]$$

A detailed derivation of this formula can be found in the supplementary material. We will frequently refer to this function as the memory curve. It can be thought of as the excess fraction of synapses (relative to equilibrium) that maintain their ideal synaptic strength at time t, as dictated by the stored memory at time $t = 0$.

Much of the previous work on these types of complex synaptic models has focused on understanding the memory curves of specific models, or choices of $\mathbf{M}^{\text{pot/dep}}$. Two examples of these models are shown in Fig. 1. We see that they have different memory properties. The serial model performs relatively well at one particular timescale, but it performs poorly at other times. The cascade model does not perform quite as well at that time, but it maintains its performance over a wider range of timescales.

In this work, rather than analyzing specific models, we take a different approach, in order to obtain a more general theory. We consider the *entire* space of these models and find upper bounds on the memory capacity of *any* of them. The space of models with a fixed number of internal states M is parameterized by the pair of $M \times M$ discrete time stochastic transition matrices $\mathbf{M}^{\text{pot}}$ and $\mathbf{M}^{\text{dep}}$, in addition to $f^{\text{pot/dep}}$. The parameters must satisfy the following constraints:

$$\begin{aligned} &\mathbf{M}_{ij}^{\text{pot/dep}} \in [0,1], && \sum_j \mathbf{M}_{ij}^{\text{pot/dep}} = 1, \\ &f^{\text{pot/dep}} \in [0,1], && f^{\text{pot}} + f^{\text{dep}} = 1, \qquad [\mathbf{4}] \\ &\mathbf{p}^\infty \mathbf{W}^{\mathrm{F}} = 0, \qquad \sum_i \mathbf{p}_i^\infty = 1, && \mathbf{w}_i = \pm 1. \end{aligned}$$

The upper bounds on $\mathbf{M}_{ij}^{\text{pot/dep}}$ and $f^{\text{pot/dep}}$ follow automatically from the other constraints.

The critical question is: what do these constraints imply about the space of achievable memory curves in **[3]**? To answer this question, especially for limits on achievable memory at finite times, it will be useful to employ the eigenmode decomposition:

$$\mathbf{W}^{\mathrm{F}} = \sum_a -q_a \mathbf{u}^a \mathbf{v}^a, \qquad \mathbf{v}^a \mathbf{u}^b = \delta_{ab}, \qquad \textbf{[5]}$$
$$\mathbf{W}^{\mathrm{F}} \mathbf{u}^a = -q_a \mathbf{u}^a, \qquad \mathbf{v}^a \mathbf{W}^{\mathrm{F}} = -q_a \mathbf{v}^a.$$

Here q_a are the negative of the eigenvalues of the forgetting process $\mathbf{W}^{\mathrm{F}}$, $\mathbf{u}^a$ are the right (column) eigenvectors and $\mathbf{v}^a$ are the left (row) eigenvectors. This decomposition allows us to write the memory curve as a sum of exponentials,

$$\mathrm{SNR}(t) = \sqrt{N} \sum_a \mathcal{I}_a \mathrm{e}^{-rt/\tau_a}, \qquad \textbf{[6]}$$

where $\mathcal{I}_a = (2 f^{\mathrm{pot}} f^{\mathrm{dep}}) \mathbf{p}^\infty (\mathbf{M}^{\mathrm{pot}} - \mathbf{M}^{\mathrm{dep}}) \mathbf{u}^a \mathbf{v}^a \mathbf{w}$ and $\tau_a = 1/q_a$. We can then ask the question: what are the constraints on these quantities, namely eigenmode initial SNR's, $\mathcal{I}_a$, and time constants, τ_a, implied by the constraints in **[4]**? We will derive some of these constraints in the next section.

Upper bounds on achievable memory capacity

In the previous section, in **[3]** we have described an analytic expression for a memory curve as a function of the structure of a synaptic dynamical system, described by the pair of stochastic transition matrices $\mathbf{M}^{\mathrm{pot/dep}}$. Since the performance measure for memory is an entire memory curve, and not just a single number, there is no universal scalar notion of optimal memory in the space of synaptic dynamical systems. Instead there are tradeoffs between storing proximal and distal memories; often attempts to increase memory at late (early) times by changing $\mathbf{M}^{\mathrm{pot/dep}}$, incurs a performance loss in memory at early (late) times in specific models considered so far [10, 11, 12]. Thus our end goal, achieved below, is to derive an envelope memory curve in the SNR-time plane, or a curve that forms an upper-bound on the *entire* memory curve of *any* model. In order to achieve this goal, in this section, we must first derive upper bounds, over the space of all possible synaptic models, on two different scalar functions of the memory curve: its initial SNR, and the area under the memory curve. In the process of upper-bounding the area, we will develop an essential framework to organize the structure of synaptic dynamical systems based on first passage time theory.

Bounding initial SNR. We now give an upper bound on the initial SNR,

$$\mathrm{SNR}(0) = \sqrt{N} \left(2 f^{\mathrm{pot}} f^{\mathrm{dep}} \right) \mathbf{p}^\infty \left(\mathbf{M}^{\mathrm{pot}} - \mathbf{M}^{\mathrm{dep}} \right) \mathbf{w}, \qquad \textbf{[7]}$$

over *all* possible models and also find the class of models that saturate this bound. A useful quantity is the equilibrium probability flux between two disjoint sets of states, $\mathcal{A}$ and $\mathcal{B}$:

$$\mathbf{\Phi}_{\mathcal{AB}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} r \mathbf{p}_i^\infty \mathbf{W}_{ij}^{\mathrm{F}}. \qquad \textbf{[8]}$$

The initial SNR is closely related to the flux from the states with $\mathbf{w}_i = -1$ to those with $\mathbf{w}_j = +1$ (see supplementary material):

$$\mathrm{SNR}(0) \leq \frac{4\sqrt{N} \mathbf{\Phi}_{-+}}{r}. \qquad \textbf{[9]}$$

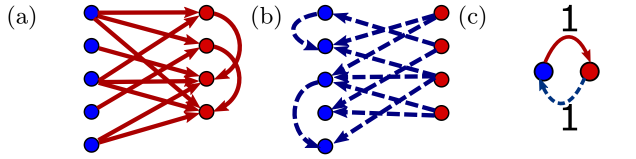

Fig. 2. Synaptic models that maximize initial SNR. (**??**) For potentiation, all transitions starting from a weak state lead to a strong state, and the probabilities for all transitions leaving a given weak state sum to 1. (**??**) Depression is similar to potentiation, but with strong and weak interchanged. (13) The equivalent two state model, with transition probabilities under potentiation and depression equal to one.

This inequality becomes an equality if potentiation never decreases the synaptic weight and depression never increases it, which should be a property of any sensible model.

To maximize this flux, potentiation from a weak state must be guaranteed to end in a strong state, and depression must do the reverse. An example of such a model is shown in Fig. 2(**??**,**??**). These models have a property known as "lumpability" (see [19, §6.3] for the discrete time version and [20, 21] for continuous time). They are completely equivalent (i.e. have the same memory curve) as a two state model with transition probabilities equal to 1, as shown in Fig. 2(13).

This two state model has the equilibrium distribution $\mathbf{p}^\infty = (f^{\mathrm{dep}}, f^{\mathrm{pot}})$ and its flux is given by $\mathbf{\Phi}_{-+} = r f^{\mathrm{pot}} f^{\mathrm{dep}}$. This is maximized when $f^{\mathrm{pot}} = f^{\mathrm{dep}} = \frac{1}{2}$, leading to the upper bound:

$$\mathrm{SNR}(0) \leq \sqrt{N}. \qquad \textbf{[10]}$$

We note that while this model has high initial SNR, it also has very fast memory decay – with a timescale $\tau \sim \frac{1}{r}$. As the synapse is very plastic, the initial memory is encoded very easily, but the subsequent memories also overwrite it rapidly. This is one example of the tradeoff between optimizing memory at early versus late times.

Imposing order on internal states through first passage times. Our goal of understanding the relationship between structure and function in the space of all possible synaptic models is complicated by the fact that this space contains many different possible network topologies, encoded in the nonzero matrix elements of $\mathbf{M}^{\mathrm{pot/dep}}$. To systematically analyze this entire space, we develop an important organizing principle using the theory of first passage times in the stochastic process of forgetting, described by $\mathbf{W}^{\mathrm{F}}$. The mean first passage time matrix, $\overline{\mathbf{T}}_{ij}$, is defined as the average time it takes to reach state j for the first time, starting from state i. The diagonal elements are defined to be zero.

A remarkable theorem we will exploit is that the quantity

$$\eta \equiv \sum_j \overline{\mathbf{T}}_{ij} \mathbf{p}_j^\infty, \qquad \textbf{[11]}$$

known as Kemeny's constant (see [19, §4.4]), is independent of the starting state i. Intuitively, **[11]** states that the average time it takes to reach any state, weighted by its equilibrium probability, is independent of the starting state, implying a hidden constancy inherent in any stochastic process.

In the context of complex synapses, we can define the partial sums

$$\eta_i^+ = \sum_{j \in +} \overline{\mathbf{T}}_{ij} \mathbf{p}_j^\infty, \qquad \eta_i^- = \sum_{j \in -} \overline{\mathbf{T}}_{ij} \mathbf{p}_j^\infty. \qquad \textbf{[12]}$$

[1] Note that we do not need to worry about the order of the $\eta_i^\pm$ changing during the optimization: necessary conditions for a maximum only require that there is no infinitesimal perturbation that increases the area. Therefore we need only consider an infinitesimal neighborhood of the model, in which the order will not change.

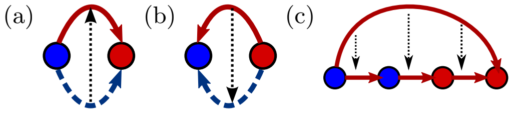

Fig. 3. Perturbations that increase the area. (**??**) Perturbations that increase elements of $\mathbf{M}^{\text{pot}}$ above the diagonal and decrease the corresponding elements of $\mathbf{M}^{\text{dep}}$. It can no longer be used when $\mathbf{M}^{\text{dep}}$ is lower triangular, i.e. depression must move synapses to “more depressed” states. (**??**) Perturbations that decrease elements of $\mathbf{M}^{\text{pot}}$ below the diagonal and increase the corresponding elements of $\mathbf{M}^{\text{dep}}$. It can no longer be used when $\mathbf{M}^{\text{pot}}$ is upper triangular, i.e. potentiation must move synapses to “more potentiated” states. () Perturbation that decreases “shortcut” transitions and increases the bypassed “direct” transitions. It can no longer be used when there are only nearest-neighbor “direct” transitions.

These can be thought of as the average time it takes to reach the strong/weak states respectively. Using these definitions, we can then impose an order on the states by arranging them in order of decreasing η_i^+ or increasing η_i^-. Because $\eta_i^+ + \eta_i^- = \eta$ is independent of i, the two orderings are the same. In this order, which depends sensitively on the structure of $\mathbf{M}^{\text{pot/dep}}$, states later (to the right in figures below) can be considered to be more potentiated than states earlier (to the left in figures below), despite the fact that they have the same synaptic efficacy. In essence, in this order, a state is considered to be more potentiated if the average time it takes to reach all the strong efficacy states is shorter. We will see that synaptic models that optimize various measures of memory have an exceedingly simple structure when, and only when, their states are arranged in this order.[1]

Bounding area. Now consider the area under the memory curve:

$$A = \int_0^\infty \mathrm{d}t \; \mathrm{SNR}(t). \qquad [13]$$

We will find an upper bound on this quantity as well as the model that saturates this bound.

First passage time theory introduced in the previous section becomes useful because the area has a simple expression in terms of quantities introduced in [**12**] (see supplementary material):

$$\begin{aligned} A &= \sqrt{N}(4 f^{\text{pot}} f^{\text{dep}}) \sum_{ij} \mathbf{p}_i^\infty \left[\mathbf{M}_{ij}^{\text{pot}} - \mathbf{M}_{ij}^{\text{dep}} \right] (\eta_i^+ - \eta_j^+) \\ &= \sqrt{N}(4 f^{\text{pot}} f^{\text{dep}}) \sum_{ij} \mathbf{p}_i^\infty \left[\mathbf{M}_{ij}^{\text{pot}} - \mathbf{M}_{ij}^{\text{dep}} \right] (\eta_j^- - \eta_i^-). \end{aligned} \qquad [14]$$

With the states in the order described above, we can find perturbations of $\mathbf{M}^{\text{pot/dep}}$ that will always increase the area, whilst leaving the equilibrium distribution, $\mathbf{p}^\infty$, unchanged. Some of these perturbations are shown in Fig. 3, see supplementary material for details. For example, in Fig. 3(**??**), for two states i on the left and j on the right, with j being more “potentiated” than i (i.e. $\eta_i^+ > \eta_j^+$), we have proven that increasing $\mathbf{M}_{ij}^{\text{pot}}$ and decreasing $\mathbf{M}_{ij}^{\text{dep}}$ leads to an increase in area. The only thing that can prevent these perturbations from increasing the area is when they require the decrease of a matrix element that has already been set to 0. This determines the topology (non-zero transition probabilities) of the model with maximal area. It is of the form shown in Fig. 4(),with potentiation moving one step to the right and depression moving one step to the left. Any other topology would allow some class of perturbations (e.g. in Fig. 3) to further increase the area.

As these perturbations do not change the equilibrium distribution, this means that the area of *any* model is bounded by that of a linear chain with the same equilibrium distribution. The area of a linear chain model can be expressed directly in terms of its equilibrium state distribution, $\mathbf{p}^\infty$, yielding the following upper bound on the area of any model with the same $\mathbf{p}^\infty$ (see supplementary material):

$$\begin{aligned} A &\leq \frac{2\sqrt{N}}{r} \sum_k \left[k - \sum_j j \mathbf{p}_j^\infty \right] \mathbf{p}_k^\infty \mathbf{w}_k \\ &= \frac{2\sqrt{N}}{r} \sum_k \left| k - \sum_j j \mathbf{p}_j^\infty \right| \mathbf{p}_k^\infty, \end{aligned} \qquad [15]$$

where we chose $\mathbf{w}_k = \operatorname{sgn}[k - \sum_j j \mathbf{p}_j^\infty]$. We can then maximize this by pushing all of the equilibrium distribution symmetrically to the two end states. This can be done by reducing the transition probabilities out of these states, as in Fig. 4(). This makes it very difficult to exit these states once they have been entered. The resulting area is

$$A \leq \frac{\sqrt{N}(M-1)}{r}. \qquad [16]$$

This analytical result is similar to a numerical result found in [18] under a slightly different information theoretic measure of memory performance.

The “sticky” end states result in very slow decay of memory, but they also make it difficult to encode the memory in the first place, since a small fraction of synapses are able to change synaptic efficacy during the storage of a new memory. Thus models that maximize area optimize memory at late times, at the expense of early times.

Memory curve envelope

Now we will look at the implications of the upper bounds found in the previous section for the SNR at finite times. As argued in [**6**], the memory curve can be written in the form

$$\mathrm{SNR}(t) = \sqrt{N} \sum_a \mathcal{I}_a \mathrm{e}^{-rt/\tau_a}. \qquad [17]$$

The upper bounds on the initial SNR, [**10**], and the area, [**16**], imply the following constraints on the parameters $\{\mathcal{I}_a, \tau_a\}$:

$$\sum_a \mathcal{I}_a \leq 1, \qquad \sum_a \mathcal{I}_a \tau_a \leq M - 1. \qquad [18]$$

We are not claiming that these are a complete set of constraints: not every set $\{\mathcal{I}_a, \tau_a\}$ that satisfies these inequalities will actually be achievable by a synaptic model. However, any set that violates either inequality will definitely not be achievable.

Now we can pick some fixed time, t_0, and maximize the SNR at that time wrt. the parameters $\{\mathcal{I}_a, \tau_a\}$, subject to the constraints above. This always results in a single nonzero $\mathcal{I}_a$; in essence, optimizing memory at a single time requires a single exponential. The resulting optimal memory curve, along with the achieved memory at the chosen time, depends on t_0 as follows:

$$\begin{aligned} t_0 \leq \frac{M-1}{r} \quad &\implies \quad \mathrm{SNR}(t) = \sqrt{N} \mathrm{e}^{-rt/(M-1)} \\ &\implies \quad \mathrm{SNR}(t_0) = \sqrt{N} \mathrm{e}^{-rt_0/(M-1)}, \\ t_0 \geq \frac{M-1}{r} \quad &\implies \quad \mathrm{SNR}(t) = \frac{\sqrt{N}(M-1)\mathrm{e}^{-t/t_0}}{rt_0} \\ &\implies \quad \mathrm{SNR}(t_0) = \frac{\sqrt{N}(M-1)}{\mathrm{e}rt_0}. \end{aligned} \qquad [19]$$

Both the initial SNR bound and the area bound are saturated at early times. At late times, only the area bound is saturated. The function $\mathrm{SNR}(t_0)$, the green curve in Fig. 4(**??**), above forms a memory curve envelope with late-time power-law decay $\sim t_0^{-1}$. No synaptic model can have an SNR that is greater than this at any time. We can use this to find an upper bound on the memory lifetime, $\tau(\epsilon)$, by finding the point at which the envelope crosses ϵ:

$$\tau(\epsilon) \leq \frac{\sqrt{N}(M-1)}{\epsilon e r}, \qquad [20]$$

where we assume $N > (\epsilon e)^2$. Intriguingly, both the lifetime and memory envelope expand linearly with the number of internal states M, and increase as the square root of the number of synapses N.

This leaves the question of whether this bound is achievable. At any time, can we find a model whose memory curve touches the envelope? The red curves in Fig. 4(**??**) show the closest we have come to the envelope with actual models, by repeated numerical optimization of $\mathrm{SNR}(t_0)$ over $\mathbf{M}^{\mathrm{pot/dep}}$ with random initialization and by hand designed models.

We see that at early, but not late times, there is a gap between the upper bound that we can prove and what we can achieve with actual models. There may be other models we haven't found that could beat the ones we have, and come closer to our proven envelope. However, we suspect that the area constraint is not the bottleneck for optimizing memory at times less than $\mathcal{O}(\frac{M}{r})$. We believe there is some other constraint that prevents models from approaching the envelope, and currently are exploring several mathematical conjectures for the precise form of this constraint in order to obtain a potentially tighter envelope. Nevertheless, we have proven rigorously that no model's memory curve can ever exceed this envelope, and that it is at least tight for late times, longer than $\mathcal{O}(\frac{M}{r})$, where models of the form in Fig. 4()can come close to the envelope.

Discussion

We have initiated the development of a general theory of learning and memory with complex synapses, allowing for an exploration of the entire space of complex synaptic models, rather than analyzing individual models one at a time. In doing so, we have obtained several new mathematical results delineating the functional limits of memory achievable by synaptic complexity, and the structural characterization of synaptic dynamical systems that achieve these limits. In particular, operating within the ideal observer framework of [10, 11, 18], we have shown that for a population of N synapses with M internal states, (a) the initial SNR of any synaptic model cannot exceed $\sqrt{N}$, and any model that achieves this bound is equivalent to a binary synapse, (b) the area under the memory curve of any model cannot exceed that of a linear chain model with the same equilibrium distribution, (c) both the area and memory lifetime of any model cannot exceed $\mathcal{O}(\sqrt{N}M)$, and the model that achieves this limit has a linear chain topology with only nearest neighbor transitions, (d) we have derived an envelope memory curve in the SNR-time plane that cannot be exceeded by the memory curve of any model, and models that approach this envelope for times greater than $\mathcal{O}(\frac{M}{r})$ are linear chain models, and (e) this late-time envelope is a power-law proportional to $\mathcal{O}(\sqrt{N}M/rt)$, indicating that synaptic complexity can strongly enhance the limits of achievable memory.

This theoretical study opens up several avenues for further inquiry. In particular, the tightness of our envelope for early times, less than $\mathcal{O}(\frac{M}{r})$, remains an open question, and we are currently pursuing several conjectures. We have also derived memory constrained envelopes, by asking in the space of models that achieve a given SNR at a given time, what is the maximal SNR achievable at other times. If these two times are beyond a threshold separation, optimal constrained models require two exponentials. It would be interesting to systematically analyze the space of models that achieve good memory at multiple times, and understand their structural organization, and how they give rise to multiple exponentials, leading to power law memory decays.

Finally, it would be interesting to design physiological experiments in order to perform optimal systems identification of potential Markovian dynamical systems hiding within biological synapses, given measurements of pre and post-synaptic spike trains along with changes in post-synaptic potentials. Then given our theory, we could match this measured synaptic model to optimal models to understand for which timescales of memory, if any, biological synaptic dynamics may be tuned.

In summary, we hope that a deeper theoretical understanding of the functional role of synaptic complexity, initiated here, will help advance our understanding of the neurobiology of learning and memory, aid in the design of engineered memory circuits, and lead to new mathematical theorems about stochastic processes.

ACKNOWLEDGMENTS. We thank Sloan, Genenetech, Burroughs-Wellcome, and Swartz foundations for support. We thank Larry Abbott, Marcus Benna, Stefano Fusi, Jascha Sohl-Dickstein and David Sussillo for useful discussions.

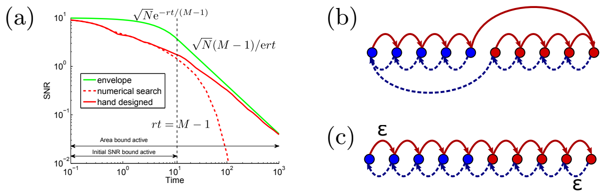

Fig. 4. The memory curve envelope for $N = 100$, $M = 12$. (**??**) An upper bound on the SNR at any time is shown in green. The red dashed curve shows the result of numerical optimization of synaptic models with random initialization. The solid red curve shows the highest SNR we have found with hand designed models. At early times these models are of the form shown in (**??**) with different numbers of states, and all transition probabilities equal to 1. At late times they are of the form shown in () with different values of ε. The model shown in () also saturates the area bound **[16]** in the limit $\varepsilon \to 0$.

1. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. 79(8):2554–2558.
2. Amit DJ, Gutfreund H, Sompolinsky H (1985) Spin-glass models of neural networks. Phys. Rev. A 32:1007–1018.
3. Gardner E (1988) The space of interactions in neural network models. Journal of Physics A: Mathematical and General 21(1):257.
4. Bliss TVP, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361:31–39.
5. Petersen CCH, Malenka RC, Nicoll RA, Hopfield JJ (1998) All-or-none potentiation at ca3-ca1 synapses. Proc. Natl. Acad. Sci. U.S.A. 95(8):4732–4737.
6. O'Connor DH, Wittenberg GM, Wang SSH (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. Proc. Natl. Acad. Sci. U.S.A. 102(27):9679–9684.
7. Enoki R, ling Hu Y, Hamilton D, Fine A (2009) Expression of long-term plasticity at individual synapses in hippocampus is graded, bidirectional, and mainly presynaptic: Optical quantal analysis. Neuron 62(2):242 – 253.
8. Amit DJ, Fusi S (1992) Constraints on learning in dynamic synapses. Network: Computation in Neural Systems 3(4):443–464.
9. Amit DJ, Fusi S (1994) Learning in neural networks with material synapses. Neural Computation 6(5):957–982.

10. Fusi S, Drew PJ, Abbott LF (2005) Cascade models of synaptically stored memories. Neuron 45(4):599–611.
11. Fusi S, Abbott LF (2007) Limits on the memory storage capacity of bounded synapses. Nat. Neurosci. 10(4):485–493.
12. Leibold C, Kempter R (2008) Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. Cerebral Cortex 18(1):67–77.
13. Bredt DS, Nicoll RA (2003) Ampa receptor trafficking at excitatory synapses. Neuron 40(2):361 – 379.
14. Coba MP et al. (2009) Neurotransmitters drive combinatorial multistate postsynaptic density networks. Sci Signal 2(68):ra19.
15. Abraham WC, Bear MF (1996) Metaplasticity: the plasticity of synaptic plasticity. Trends in Neurosciences 19(4):126 – 130.
16. Montgomery JM, Madison DV (2002) State-dependent heterogeneity in synaptic depression between pyramidal cell pairs. Neuron 33(5):765 – 777.
17. Emes RD, Grant SG (2012) Evolution of synapse complexity and diversity. Annual Review of Neuroscience 35(1):111–131.
18. Barrett AB, van Rossum MC (2008) Optimal learning rules for discrete synapses. PLoS Comput. Biol. 4(11):e1000230.
19. Kemeny J, Snell J (1960) Finite markov chains. (Springer).
20. Burke C, Rosenblatt M (1958) A markovian function of a markov chain. The Annals of Mathematical Statistics 29(4):1112–1122.
21. Ball F, Yeo GF (1993) Lumpability and marginalisability for continuous-time markov chains. Journal of Applied Probability 30(3):518–528.
22. Hunter J (2000) A survey of generalized inverses and their use in stochastic modelling. Research Letters in the Information and Mathematical Sciences 1(1):25–36.
23. Kemeny JG (1981) Generalization of a fundamental matrix. Linear Algebra and its Applications 38(0):193 – 206.
24. Yao DD (1985) First-passage-time moments of markov processes. Journal of Applied Probability 22(4):pp. 939–945.
25. Cho G, Meyer C (2000) Markov chain sensitivity measured by mean first passage times. Linear Algebra and its Applications 316(1-3):21–28.

Supplementary material

Here we provide more details underlying the derivations of results in the main paper.

Continuous time Markov processes

In this section we'll provide a summary of all the relevant properties of ergodic Markov chains in continuous time to define notation. It is a generalization of material that can be found in [19] with some ideas from [22].

Notation. For any matrix $\mathbf{A}$, we define matrices $\mathbf{A}^{\mathrm{dg}}$ and $\overline{\mathbf{A}}$ as

$$\mathbf{A}^{\mathrm{dg}}_{ij} \equiv \delta_{ij}\mathbf{A}_{ij}, \qquad \overline{\mathbf{A}} \equiv \mathbf{A} - \mathbf{A}^{\mathrm{dg}}. \tag*{[21]}$$

We let $\mathbf{e}$ denote a column-vector of ones and $\mathbf{E} = \mathbf{e}\mathbf{e}^{\mathrm{T}}$ denote a matrix of ones.

A continuous time Markov process is described by a matrix of transitions rates, $\mathbf{Q}_{ij}$, from state i to j with row sums equal to zero ($\mathbf{Q}\mathbf{e} = 0$). The probabilities of being in each state at time t, the row-vector $\mathbf{p}(t)$, evolve according to

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \mathbf{p}(t)\mathbf{Q}, \tag*{[22]}$$

where $\mathbf{p}(t)\mathbf{e} = 1$.

The equilibrium probabilities, $\mathbf{p}^{\infty}$, satisfy

$$\mathbf{p}^{\infty}\mathbf{Q} = 0, \qquad \mathbf{p}^{\infty}\mathbf{e} = 1. \tag*{[23]}$$

As we assume an ergodic process, this eigenvalue is non-degenerate. If all other eigenvalues have strictly negative real parts, the process is regular (aperiodic).

We define additional matrices

$$\mathbf{\Lambda} \equiv (-\mathbf{Q}^{\mathrm{dg}})^{-1}, \qquad \mathbf{P} \equiv \mathbf{I} + \mathbf{\Lambda}\mathbf{Q}. \tag*{[24]}$$

It can be shown that $\mathbf{\Lambda}_{ii}$ is the mean time it takes to leave state i and $\mathbf{P}_{ij}$ is the probability the the next transition from state i goes to state j:

$$\mathbf{\Lambda}_{ii} = \frac{1}{\sum_{j\neq j}\mathbf{Q}_{ij}}, \qquad \mathbf{P}_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \frac{\mathbf{Q}_{ij}}{\sum_{k\neq j}\mathbf{Q}_{ik}} & \text{otherwise.} \end{cases} \tag*{[25]}$$

Furthermore, we also define

$$\mathbf{D} \equiv \operatorname{diag}(\mathbf{p}^{\infty})^{-1}, \qquad \Longrightarrow \qquad \mathbf{p}^{\infty}\mathbf{D} = \mathbf{e}^{\mathrm{T}}. \tag*{[26]}$$

Fundamental matrix. For our results below regarding the integral of the memory curve, it can be useful to invert the stochastic transition matrix, $\mathbf{Q}$. However, since $\mathbf{Q}$ has a zero eigenvalue, it cannot be inverted. For this reason, the fundamental matrix arises as a useful surrogate for the inverse of $\mathbf{Q}$. It is related to the first passage times, as we will see in the next subsection. Here we define the fundamental matrix and review its properties.

Definition 1: Fundamental matrix
For discrete time, the generalized fundamental matrix was defined in [23]. For continuous time, we define:

$$\mathbf{Z} \equiv (-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi})^{-1}, \tag*{[27]}$$

where $\boldsymbol{\pi}$ is any row-vector with $\boldsymbol{\pi}\mathbf{e} = 1/\tau \neq 0$.

Note that the canonical choice for the discrete time version, $\boldsymbol{\pi} = \mathbf{p}^{\infty}$, is not available here due to problems with units. It will be helpful to choose $\boldsymbol{\pi}$ to be independent of $\mathbf{Q}$, e.g. $\boldsymbol{\pi} = \mathbf{e}^{\mathrm{T}}/(n\tau)$. All quantities that we calculate using $\mathbf{Z}$ below will be independent of this choice.

Theorem 1:
The definition of $\mathbf{Z}$ is valid, i.e. $(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi})$ is invertible.

Proof. Assume there exists an $\mathbf{x}$ such that

$$(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi})\mathbf{x} = 0. \tag*{[28]}$$

Multiplying from the left with $\mathbf{p}^{\infty}$ gives

$$\boldsymbol{\pi}\mathbf{x} = 0. \tag*{[29]}$$

Substituting back into [**28**] gives

$$\mathbf{Q}\mathbf{x} = 0.$$

As we assume an ergodic process, the zero eigenvalue is non-degenerate. Therefore, $\mathbf{x} = \lambda\mathbf{e}$. Substituting this into [**29**] gives

$$\lambda\boldsymbol{\pi}\mathbf{e} = \frac{\lambda}{\tau} = 0.$$

As we defined $\boldsymbol{\pi}$ such that $1/\tau \neq 0$, this means $\lambda = 0 \implies \mathbf{x} = 0$. □

Corollary 2:

$$\boldsymbol{\pi}\mathbf{Z} = \mathbf{p}^\infty, \qquad [\mathbf{30}]$$
$$\mathbf{Z}\mathbf{e} = \tau\mathbf{e}, \qquad [\mathbf{31}]$$
$$\mathbf{I} + \mathbf{Q}\mathbf{Z} = \mathbf{e}\mathbf{p}^\infty, \qquad [\mathbf{32}]$$
$$\mathbf{I} + \mathbf{Z}\mathbf{Q} = \tau\mathbf{e}\boldsymbol{\pi}. \qquad [\mathbf{33}]$$

Proof. We can deduce [**30**] and [**31**] be pre/post-multiplying the following equations by $\mathbf{Z}$:

$$\mathbf{p}^\infty(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi}) = \boldsymbol{\pi},$$
$$(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi})\mathbf{e} = \frac{\mathbf{e}}{\tau}.$$

We can then deduce [**32**] and [**33**] by substituting these into

$$(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi})\mathbf{Z} = \mathbf{Z}(-\mathbf{Q} + \mathbf{e}\boldsymbol{\pi}) = \mathbf{I}.$$

□

First passage times.

Definition 2: First passage time matrix
We define $\overline{\mathbf{T}}_{ij}$ as the mean time it takes the process to reach state j for the first time, starting from state i. We also define $\mathbf{T}^{\mathrm{dg}}_{ii}$ as the mean time it takes the process to return to state i. As usual, $\mathbf{T} = \overline{\mathbf{T}} + \mathbf{T}^{\mathrm{dg}}$.

This matrix is given by

$$\mathbf{T} = (\mathbf{E}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z} + \boldsymbol{\Lambda})\mathbf{D}, \qquad [\mathbf{34}]$$

see [24] for a proof. We can separate this into its diagonal and off-diagonal pieces.

The recurrence times are given by

$$\mathbf{T}^{\mathrm{dg}} = \boldsymbol{\Lambda}\mathbf{D}. \qquad [\mathbf{35}]$$

or in component form

$$\mathbf{p}^\infty_i \boldsymbol{\Lambda}^{-1}_{ii} \mathbf{T}^{\mathrm{dg}}_{ii} = 1.$$

The extra factor of $\boldsymbol{\Lambda}_{ii}$, compared to the discrete case [19, Th.4.4.5], occurs because in this case we are demanding that the process leaves the initial state once before returning, whereas in the discrete case we only measure the time it takes to go to the initial state after the first time-step.

The off-diagonal mean first passage times are given by

$$\overline{\mathbf{T}} = (\mathbf{E}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z})\mathbf{D}. \qquad [\mathbf{36}]$$

or in component form:

$$\overline{\mathbf{T}}_{ij} = \frac{\mathbf{Z}_{jj} - \mathbf{Z}_{ij}}{\mathbf{p}^\infty_j}. \qquad [\mathbf{37}]$$

Mixing time (Kemeny's constant).

Theorem 3:
The quantity

$$\eta \equiv \sum_j \overline{\mathbf{T}}_{ij}\mathbf{p}^\infty_j \qquad [\mathbf{38}]$$

is independent of i.

Proof. For discrete time, a proof can be found in [19, Th.4.4.10]. For continuous time, we use [**36**], [**31**] and the transpose of [**26**]:

$$\begin{aligned}\overline{\mathbf{T}}(\mathbf{p}^\infty)^{\mathrm{T}} &= (\mathbf{E}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z})\mathbf{D}(\mathbf{p}^\infty)^{\mathrm{T}} \\ &= (\mathbf{e}\mathbf{e}^{\mathrm{T}}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z})\mathbf{e} \\ &= (\mathbf{e}^{\mathrm{T}}\mathbf{Z}^{\mathrm{dg}}\mathbf{e})\mathbf{e} - \mathbf{Z}\mathbf{e} \\ &= (\operatorname{tr}\mathbf{Z} - \tau)\mathbf{e}.\end{aligned}$$

which proves [**38**] with $\eta = \operatorname{tr}\mathbf{Z} - \tau$. □

Note that it is essential that we use $\overline{\mathbf{T}}$ and not $\mathbf{T}$ here, as that would lead to $\eta_i = \eta + \boldsymbol{\Lambda}_{ii}$, unlike the discrete time version, where this would only shift η by 1.

Sensitivity of equilibrium distribution. Suppose that the Markov process, defined by $\mathbf{Q}$, depends on some parameter α. Differentiating $[\mathbf{27}]$ gives

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}\alpha} = \mathbf{Z}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z}. \qquad [\mathbf{39}]$$

We can substitute this into the derivative of $[\mathbf{30}]$:

$$\frac{\mathrm{d}\mathbf{p}^\infty}{\mathrm{d}\alpha} = \boldsymbol{\pi}\mathbf{Z}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z} = \mathbf{p}^\infty\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z}. \qquad [\mathbf{40}]$$

We can rewrite this in component form and use the fact that $\mathbf{Q}_{ii} = -\sum_{i\neq j}\mathbf{Q}_{ij}$:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{p}^\infty_k}{\mathrm{d}\alpha} &= \sum_{i,j}\mathbf{p}^\infty_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\mathbf{Z}_{jk} \\ &= \sum_{i\neq j}\mathbf{p}^\infty_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\mathbf{Z}_{jk} + \sum_{i}\mathbf{p}^\infty_i\frac{\mathrm{d}\mathbf{Q}_{ii}}{\mathrm{d}\alpha}\mathbf{Z}_{ik} \\ &= \sum_{i\neq j}\mathbf{p}^\infty_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}(\mathbf{Z}_{jk} - \mathbf{Z}_{ik}) \\ &= \sum_{i\neq j}\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\mathbf{p}^\infty_i\mathbf{p}^\infty_k(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk}).\end{aligned} \qquad [\mathbf{41}]$$

This is a generalization of a result of [25] from discrete to continuous time that we will need below. Note that the summand vanishes for $i = j$, so we can drop the restriction $i \neq j$ from the range of the sum.

Subsets and flux. Let us denote the set of states by $\mathcal{S}$. Consider a subset $\mathcal{A} \subset \mathcal{S}$. We can define a projection operator onto this subset:

$$\left(\mathbf{I}^\mathcal{A}\right)_{ij} = \begin{cases} 1 & \text{if } i = j \in \mathcal{A}, \\ 0 & \text{otherwise.}\end{cases} \qquad [\mathbf{42}]$$

We will use superscripts/subscripts to denote projection onto/summation over a subset:

$$\begin{aligned} &\boldsymbol{\pi}^\mathcal{A} = \boldsymbol{\pi}\mathbf{I}^\mathcal{A}, && \mathbf{M}^{\cdot\mathcal{A}} = \mathbf{M}\mathbf{I}^\mathcal{A}, && \mathbf{M}^{\mathcal{A}\cdot} = \mathbf{I}^\mathcal{A}\mathbf{M}, && \mathbf{x}^\mathcal{A} = \mathbf{I}^\mathcal{A}\mathbf{x}, \\ &\boldsymbol{\pi}_\mathcal{A} = \boldsymbol{\pi}\mathbf{e}^\mathcal{A}, && \mathbf{M}_{\cdot\mathcal{A}} = \mathbf{M}\mathbf{e}^\mathcal{A}, && \mathbf{M}_{\mathcal{A}\cdot} = \left(\mathbf{e}^\mathcal{A}\right)^\mathrm{T}\mathbf{M}, && \mathbf{x}_\mathcal{A} = \left(\mathbf{e}^\mathcal{A}\right)^\mathrm{T}\mathbf{x}, \end{aligned} \qquad [\mathbf{43}]$$

where $\boldsymbol{\pi}$ is a row vector, $\mathbf{M}$ is a matrix and $\mathbf{x}$ is a column vector.

We can define a flux matrix, a.k.a. ergodic flow:

$$\boldsymbol{\Phi} = \mathbf{D}^{-1}\mathbf{Q}, \qquad \boldsymbol{\Phi}_{ij} = \mathbf{p}^\infty_i\mathbf{Q}_{ij}. \qquad [\mathbf{44}]$$

This measures the flow of probability between states in the equilibrium distribution. Detailed balance, a.k.a. reversibility, is equivalent to $\boldsymbol{\Phi} = \boldsymbol{\Phi}^\mathrm{T}$.

The flux between two subsets is a particularly useful quantity:

$$\boldsymbol{\Phi}_{\mathcal{A}\mathcal{B}} = \mathbf{p}^{\infty\mathcal{A}}\mathbf{Q}\mathbf{e}^\mathcal{B}. \qquad [\mathbf{45}]$$

One can show that

$$\boldsymbol{\Phi}_{\mathcal{A}\mathcal{A}^\mathrm{c}} = \boldsymbol{\Phi}_{\mathcal{A}^\mathrm{c}\mathcal{A}} = -\boldsymbol{\Phi}_{\mathcal{A}\mathcal{A}} = -\boldsymbol{\Phi}_{\mathcal{A}^\mathrm{c}\mathcal{A}^\mathrm{c}} \qquad [\mathbf{46}]$$

using $\left(\mathbf{p}^{\infty\mathcal{A}} + \mathbf{p}^{\infty\mathcal{A}^\mathrm{c}}\right)\mathbf{Q} = 0$ and $\mathbf{Q}\left(\mathbf{e}^\mathcal{A} + \mathbf{e}^{\mathcal{A}^\mathrm{c}}\right) = 0$.

Lumpability. Suppose we have partitioned the states into disjoint subsets, $\{\mathcal{A}_\alpha\}$:

$$\bigcup_\alpha \mathcal{A}_\alpha = \mathcal{S}, \qquad \mathcal{A}_\alpha \cap \mathcal{A}_\beta = \delta_{\alpha\beta}\mathcal{A}_\alpha. \qquad [\mathbf{47}]$$

We will use α instead of $\mathcal{A}_\alpha$ in superscripts and subscripts in the following. The fact that these subsets are disjoint and exhaustive allows us to define the function

$$\sigma(i) = \alpha \qquad \Longleftrightarrow \qquad i \in \mathcal{A}_\alpha. \qquad [\mathbf{48}]$$

We can use this partition to define a new stochastic process associated with the original Markov chain. At time t, if the state of the original process is i, the state of the new process is $\sigma(i)$.

One may ask if this new process is a Markov chain. The answer is yes, if the original Markov chain has a property called lumpability wrt. the partition (see [19, §6.3] for the discrete time version and [20, 21] for continuous time):

$$\sigma(i) = \sigma(j) \quad \Longrightarrow \quad \mathbf{Q}_{i\alpha} = \mathbf{Q}_{j\alpha} \equiv \sum_{k\in\mathcal{A}_\alpha}\mathbf{Q}_{jk}, \qquad [\mathbf{49}]$$

i.e. the total transition rate from some state to some subset is the same for all starting states within the same subset. This common value is the transition rate for the new lumped Markov chain.

This can be rewritten with the aid of two matrices

$$U_{\alpha i} = \frac{\delta_{\alpha\sigma(i)}}{|\mathcal{A}_\alpha|}, \qquad V_{i\alpha} = \delta_{\sigma(i)\alpha}. \tag{50}$$

Left multiplication by U averages over subsets, right multiplication by V sums over subsets. For U, we chose the uniform measure in each subset. Any measure would work equally well, e.g. one proportional to the equilibrium distribution:

$$U_{\alpha i} = \frac{\mathbf{p}^{\infty\,\alpha}_i}{\mathbf{p}^\infty_\alpha}. \tag{51}$$

One can show that $(UV)_{\alpha\beta} = \delta_{\alpha\beta}$. The matrix VU is also interesting. It has a block diagonal structure, with each block corresponding to a subset. Each block is a discrete-time ergodic Markov matrix (it is an independent trials process with probabilities given by the measure chosen for U). This means that the right eigenvectors with eigenvalue 1 will be those that are constant in each subset:

$$VU\mathbf{x} = \mathbf{x} \qquad \Longleftrightarrow \qquad \mathbf{x} = \sum_\alpha x_\alpha \mathbf{e}^\alpha. \tag{52}$$

This allows us to write the lumpability condition **[49]**, and the transition matrix for the lumped process compactly:

$$VU\mathbf{Q}V = \mathbf{Q}V, \qquad \widehat{\mathbf{Q}} = U\mathbf{Q}V. \tag{53}$$

By induction, one can show that similar relations hold for all powers:

$$VU\mathbf{Q}^n V = \mathbf{Q}^n V, \qquad \widehat{\mathbf{Q}}^n = U\mathbf{Q}^n V, \tag{54}$$

and, via the Taylor series, for the exponential as well:

$$VU\mathrm{e}^{t\mathbf{Q}} V = \mathrm{e}^{t\mathbf{Q}} V, \qquad \mathrm{e}^{t\widehat{\mathbf{Q}}} = U\mathrm{e}^{t\mathbf{Q}} V. \tag{55}$$

The equilibrium distribution of the lumped process is given by

$$\widehat{\mathbf{p}}^\infty = \mathbf{p}^\infty V. \tag{56}$$

Signal-to-Noise ratio (SNR)

In this section we will look at the signal-to-noise curve, and put an upper bound on its initial value. We need only consider ergodic Markov chains. Transient states would be unoccupied in equilibrium and would not be accessed by the signal creation process, therefore they could be removed from the analysis. Absorbing chains are degenerate cases: they have zero initial signal but infinite decay times, so they can only be approached as the limit of a sequence of ergodic chains.

Framework. The individual potentiation/depression events will be described by *discrete*-time Markov chains:

$$\mathbf{M}^{\text{pot/dep}} \equiv \mathbf{I} + \mathbf{W}^{\text{pot/dep}}, \qquad \mathbf{M}^{\text{pot/dep}}\mathbf{e} = \mathbf{e}, \qquad \mathbf{M}^{\text{pot/dep}}_{ij} \in [0,1]. \tag{57}$$

The initial signal creation event occurs at time $t = 0$, but all subsequent potentiation/depression events occur at random times according to Poisson processes with rates $rf^{\text{pot/dep}}$, where $f^{\text{pot}} + f^{\text{dep}} = 1$ are the fraction of plasticity events that are potentiating/depressing respectively. This means that the "forgetting" process will be described by the *continuous*-time Markov chain:

$$\mathbf{Q} = r\mathbf{W}^{\text{F}} \equiv r\left(f^{\text{pot}}\mathbf{W}^{\text{pot}} + f^{\text{dep}}\mathbf{W}^{\text{dep}}\right). \tag{58}$$

We only require that this Markov chain is ergodic. The Markov chains described by $\mathbf{M}^{\text{pot/dep}}$ need not be.

We assume that the probability distribution starts in the equilibrium distribution **[23]**. During the initial signal creation, a fraction f^{pot} will change to $\mathbf{p}^\infty\mathbf{M}^{\text{pot}}$ and a fraction f^{dep} will change to $\mathbf{p}^\infty\mathbf{M}^{\text{dep}}$. After this, probabilities will evolve according to **[22]**.

SNR curve. As discussed in the main text, the signal-to-noise ratio is given by

$$\text{SNR}(t) = \frac{\langle \vec{w}_{\text{ideal}} \cdot \vec{w}(t)\rangle - \langle \vec{w}_{\text{ideal}} \cdot \vec{w}(\infty)\rangle}{\sqrt{\text{Var}(\vec{w}_{\text{ideal}} \cdot \vec{w}(\infty))}}. \tag{59}$$

First, let's look at the denominator, remembering that the states and plasticity events of each synapse are independent and identically distributed:

$$\begin{aligned}
\text{Var}(\vec{w}_{\text{ideal}} \cdot \vec{w}(\infty)) &= \sum_{\alpha\beta} \left\langle \vec{w}^\alpha_{\text{ideal}} \vec{w}^\alpha(\infty) \vec{w}^\beta_{\text{ideal}} \vec{w}^\beta(\infty) \right\rangle - \left(\sum_\alpha \left\langle \vec{w}^\alpha_{\text{ideal}} \vec{w}^\alpha(\infty) \right\rangle \right)^2 \\
&= \sum_\alpha \left\langle (\vec{w}^\alpha_{\text{ideal}})^2 (\vec{w}^\alpha(\infty))^2 \right\rangle + \sum_{\alpha\neq\beta} \left\langle \vec{w}^\alpha_{\text{ideal}} \vec{w}^\alpha(\infty) \right\rangle \left\langle \vec{w}^\beta{}_{\text{ideal}} \vec{w}^\beta(\infty) \right\rangle \\
&\qquad - \left(\sum_\alpha \left\langle \vec{w}^\alpha_{\text{ideal}} \vec{w}^\alpha(\infty) \right\rangle \right)^2 \\
&= N\langle 1 \rangle + N(N-1)\left\langle \vec{w}^1_{\text{ideal}} \vec{w}^1(\infty) \right\rangle^2 - N^2 \left\langle \vec{w}^1_{\text{ideal}} \vec{w}^1(\infty) \right\rangle^2 \\
&= N(1 - \left\langle \vec{w}^1_{\text{ideal}} \vec{w}^1(\infty) \right\rangle^2),
\end{aligned} \tag{60}$$

where we used $\vec{w}^\alpha = \pm 1$.

For the numerator, we can write

$$\langle \vec{w}_\text{ideal} \cdot \vec{w}(t) \rangle = \sum_\alpha \langle \vec{w}^\alpha_\text{ideal} \vec{w}^\alpha(t) \rangle = N \left\langle \vec{w}^1_\text{ideal} \vec{w}^1(t) \right\rangle, \qquad \textbf{[61]}$$

Noting that $\vec{w}_\text{ideal} = \pm 1$ with probability $f^\text{pot/dep}$,

$$\begin{aligned} \left\langle \vec{w}^1_\text{ideal} \vec{w}^1(t) \right\rangle &= f^\text{pot} \left\langle \vec{w}^1(t) \right\rangle_{\text{pot},t=0} - f^\text{dep} \left\langle \vec{w}^1(t) \right\rangle_{\text{dep},t=0} \\ &= f^\text{pot} \sum_i P(\text{state} = i, t \mid \text{pot}, 0) \mathbf{w}_i - f^\text{dep} \sum_i P(\text{state} = i, t \mid \text{dep}, 0) \mathbf{w}_i. \end{aligned} \qquad \textbf{[62]}$$

From the previous section,

$$P(\text{state} = i, t \mid \text{pot/dep}, 0) = \left[\mathbf{p}^\infty \mathbf{M}^\text{pot/dep} \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \right]_i, \qquad \textbf{[63]}$$

which describes the synapses starting in the equilibrium distribution, changing state due to the plasticity event at $t = 0$ and subsequent evolution according to **[22]** due to plasticity events uncorrelated with $\vec{w}_\text{ideal}$.[2] This results in

$$\begin{aligned} \left\langle \vec{w}^1_\text{ideal} \vec{w}^1(t) \right\rangle &= \mathbf{p}^\infty (f^\text{pot} \mathbf{M}^\text{pot} - f^\text{dep} \mathbf{M}^\text{dep}) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}, \\ \left\langle \vec{w}^1_\text{ideal} \vec{w}^1(\infty) \right\rangle &= \mathbf{p}^\infty (f^\text{pot} \mathbf{M}^\text{pot} - f^\text{dep} \mathbf{M}^\text{dep}) \, \mathbf{e} \mathbf{p}^\infty \mathbf{w} \\ &= \mathbf{p}^\infty (f^\text{pot} \mathbf{e} - f^\text{dep} \mathbf{e}) \, \mathbf{p}^\infty \mathbf{w} \\ &= (f^\text{pot} - f^\text{dep}) \, \mathbf{p}^\infty \mathbf{w} \\ &= (f^\text{pot} - f^\text{dep}) \, \mathbf{p}^\infty \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}. \end{aligned} \qquad \textbf{[64]}$$

Combining these allows us to write the numerator as

$$\begin{aligned} \langle \vec{w}_\text{ideal} \cdot \vec{w}(t) \rangle - \langle \vec{w}_\text{ideal} \cdot \vec{w}(\infty) \rangle &= N \mathbf{p}^\infty (f^\text{pot} (\mathbf{M}^\text{pot} - \mathbf{I}) - f^\text{dep} (\mathbf{M}^\text{dep} - \mathbf{I})) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w} \\ &= N \mathbf{p}^\infty (f^\text{pot} (\mathbf{W}^\text{pot} - \mathbf{W}^\text{F}) - f^\text{dep} (\mathbf{W}^\text{dep} - \mathbf{W}^\text{F})) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w} \\ &= N (2 f^\text{pot} f^\text{dep}) \mathbf{p}^\infty (\mathbf{W}^\text{pot} - \mathbf{W}^\text{dep}) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}. \end{aligned} \qquad \textbf{[65]}$$

where we used $\mathbf{p}^\infty \mathbf{W}^\text{F} = 0$ in going from the first to second lines. Combining with **[60]** gives

$$\text{SNR}(t) = \frac{\sqrt{N} (2 f^\text{pot} f^\text{dep}) \mathbf{p}^\infty (\mathbf{W}^\text{pot} - \mathbf{W}^\text{dep}) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}}{\sqrt{1 - (f^\text{pot} - f^\text{dep})^2 (\mathbf{p}^\infty_+ - \mathbf{p}^\infty_-)^2}}. \qquad \textbf{[66]}$$

The denominator will not play any role in what follows, as the models that maximize the various measures of memory performance all have some sort of balance between potentiation and depression, either with $f^\text{pot} = f^\text{dep}$ or $\mathbf{p}^\infty_+ = \mathbf{p}^\infty_-$. We can set the denominator to 1 without changing any of our results.

This results in our final formula:

$$\text{SNR}(t) = \sqrt{N} (2 f^\text{pot} f^\text{dep}) \, \mathbf{p}^\infty (\mathbf{W}^\text{pot} - \mathbf{W}^\text{dep}) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}. \qquad \textbf{[67]}$$

The factor of $\mathbf{p}^\infty$ describes the synapses being in the steady-state distribution before the memory is encoded. The factor of $(\mathbf{M}^\text{pot} - \mathbf{M}^\text{dep})$ comes from the encoding of the memory at $t = 0$, with $\vec{w}_\text{ideal}$ being ± 1 in synapses that are potentiated/depotentiated. The factor of $\mathrm{e}^{rt\mathbf{W}^\text{F}}$ describes the subsequent evolution of the probability distribution, averaged over all sequences of plasticity events, and the factor of $\mathbf{w}$ indicates the readout via the synaptic weight.

We can express this in terms of the one parameter family of transition matrices:

$$\mathbf{W}(\alpha) = \alpha \mathbf{W}^\text{pot} + (1 - \alpha) \mathbf{W}^\text{dep}, \qquad \Longrightarrow \qquad \begin{aligned} \mathbf{W}^\text{F} &= \mathbf{W}(f^\text{pot}), \\ \mathbf{W}^\text{pot} - \mathbf{W}^\text{dep} &= \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\alpha}, \\ \mathbf{p}^\infty \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\alpha} &= -\frac{\mathrm{d}\mathbf{p}^\infty}{\mathrm{d}\alpha} \mathbf{W}^\text{F}. \end{aligned} \qquad \textbf{[68]}$$

Then **[67]** becomes

$$\text{SNR}(t) = \sqrt{N} (2 f^\text{pot} f^\text{dep}) \frac{\mathrm{d}\mathbf{p}^\infty}{\mathrm{d}\alpha} (-\mathbf{W}^\text{F}) \, \mathrm{e}^{rt\mathbf{W}^\text{F}} \mathbf{w}. \qquad \textbf{[69]}$$

[2] Note that expanding the exponential gives

$$\mathrm{e}^{rt\mathbf{W}^\text{F}} = \sum_{n=0}^\infty \frac{(rt)^n \, \mathrm{e}^{-rt}}{n!} \sum_{m=0}^n (f^\text{pot})^m (f^\text{dep})^{n-m} \left[\mathbf{M}^\text{pot} \mathbf{M}^\text{dep} \mathbf{M}^\text{pot} \mathbf{M}^\text{pot} \ldots + \text{permutations} \right].$$

Thus, evolving according to **[22]** results in averaging over all sequences of plasticity events, as we only need linear expectations of $\vec{w}(t)$ in the end.

Lumpability and the SNR curve. Suppose that we have a partition such that $\mathbf{W}^{\text{pot}}$ and $\mathbf{W}^{\text{dep}}$ are simultaneously lumpable, and that all the states in each subset have the same synaptic strength (see §):

$$VU\mathbf{W}^{\text{pot/dep}}V = \mathbf{W}^{\text{pot/dep}}V, \qquad VU\mathbf{w} = \mathbf{w}. \tag{70}$$

We can define a new synapse with

$$\widehat{\mathbf{W}}^{\text{pot/dep}} = U\mathbf{W}^{\text{pot/dep}}V, \qquad \widehat{\mathbf{w}} = U\mathbf{w}, \qquad \widehat{\mathbf{p}}^{\infty} = \mathbf{p}^{\infty}V. \tag{71}$$

This synapse has an SNR curve:

$$\begin{aligned} \frac{\text{SNR}(t)}{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})} &= \widehat{\mathbf{p}}^{\infty}(\widehat{\mathbf{W}}^{\text{pot}} - \widehat{\mathbf{W}}^{\text{dep}})\mathrm{e}^{rt\widehat{\mathbf{W}}^{F}}\,\widehat{\mathbf{w}}. \\ &= \mathbf{p}^{\infty}VU(\mathbf{W}^{\text{pot}} - \mathbf{W}^{\text{dep}})VU\mathrm{e}^{rt\mathbf{W}^{\text{F}}}VU\mathbf{w}. \\ &= \mathbf{p}^{\infty}(\mathbf{W}^{\text{pot}} - \mathbf{W}^{\text{dep}})VU\mathrm{e}^{rt\mathbf{W}^{\text{F}}}VU\mathbf{w}. \\ &= \mathbf{p}^{\infty}(\mathbf{W}^{\text{pot}} - \mathbf{W}^{\text{dep}})\mathrm{e}^{rt\mathbf{W}^{\text{F}}}VU\mathbf{w}. \\ &= \mathbf{p}^{\infty}(\mathbf{W}^{\text{pot}} - \mathbf{W}^{\text{dep}})\mathrm{e}^{rt\mathbf{W}^{\text{F}}}\mathbf{w}. \end{aligned} \tag{72}$$

i.e. the lumped process has exactly the same SNR as the original one.

Initial SNR and flux. Using $\mathbf{p}^{\infty}\mathbf{W}^{\text{F}} = 0$ and the first line of [**65**], we can write the initial SNR as

$$\frac{\text{SNR}(0)}{\sqrt{N}} = I = (\mathbf{p}^{\infty+} + \mathbf{p}^{\infty-})(f^{\text{pot}}\mathbf{W}^{\text{pot}} - f^{\text{dep}}\mathbf{W}^{\text{dep}})(\mathbf{e}^{+} - \mathbf{e}^{-}). \tag{73}$$

Using $\mathbf{W}^{\text{pot/dep}}(\mathbf{e}^{+} + \mathbf{e}^{-}) = 0$ and [**46**]:

$$r\mathbf{p}^{\infty-}(f^{\text{pot}}\mathbf{W}^{\text{pot}} + f^{\text{dep}}\mathbf{W}^{\text{dep}})\mathbf{e}^{+} = \mathbf{\Phi}_{-+} = \mathbf{\Phi}_{+-} = r\mathbf{p}^{\infty+}(f^{\text{pot}}\mathbf{W}^{\text{pot}} + f^{\text{dep}}\mathbf{W}^{\text{dep}})\mathbf{e}^{-},$$

we can rewrite [**73**] as

$$I = \frac{4\mathbf{\Phi}_{-+}}{r} - 4\mathbf{p}^{\infty+}f^{\text{pot}}\mathbf{W}^{\text{pot}}\mathbf{e}^{-} - 4\mathbf{p}^{\infty-}f^{\text{dep}}\mathbf{W}^{\text{dep}}\mathbf{e}^{+}. \tag{74}$$

The last two terms are guaranteed to be negative, as the diagonal parts of $\mathbf{W}^{\text{pot/dep}}$ cannot contribute. Therefore

$$\text{SNR}(0) \leq \frac{4\sqrt{N}\mathbf{\Phi}_{-+}}{r}. \tag{75}$$

This inequality is saturated if potentiation never takes it from a + state to a − state and depression never takes it from a − state to a + state.

Area maximisation

In this section we will find an upper bound on the area under the signal-to-noise curve. As in §, we will only consider ergodic Markov chains. We will see in § that the optimal chain is absorbing, so it lies on the boundary of the (open) set of ergodic chains, but it still puts an upper bound on the area for any chain in the interior.

Area under signal-to-noise curve. The signal-to-noise curve is given by [**69**]. The area is computed by integrating this

$$\begin{aligned} A &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\mathbf{p}^{\infty}}{\mathrm{d}\alpha}\left[-\mathrm{e}^{rt\mathbf{W}^{\text{F}}}\right]_{0}^{\infty}\mathbf{w} \\ &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\mathbf{p}^{\infty}}{\mathrm{d}\alpha}(\mathbf{I} - \mathbf{e}\mathbf{p}^{\infty})\mathbf{w} \\ &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\mathbf{p}^{\infty}}{\mathrm{d}\alpha}\mathbf{w}. \end{aligned} \tag{76}$$

We can rewrite this using [**41**], with $A = \sqrt{N}(2f^{\text{pot}}f^{\text{dep}})\hat{A}$ and $\mathbf{q}_{ij} \equiv \frac{\mathrm{d}\mathbf{W}^{\text{F}}_{ij}}{\mathrm{d}\alpha} = \mathbf{W}^{\text{pot}}_{ij} - \mathbf{W}^{\text{dep}}_{ij}$

$$\hat{A} = \sum_{i,j,k}\mathbf{p}^{\infty}_{i}\mathbf{q}_{ij}(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk})\mathbf{p}^{\infty}_{k}\mathbf{w}_{k}. \tag{77}$$

Definition 3: Partial mixing times

We define the ± mixing times as

$$\begin{aligned} \eta^{\pm}_{i} \equiv \sum_{k}\overline{\mathbf{T}}_{ik}\mathbf{p}^{\infty}_{k}\left(\frac{1 \pm \mathbf{w}_{k}}{2}\right) &= \sum_{k\in\pm}\overline{\mathbf{T}}_{ik}\mathbf{p}^{\infty}_{k} \\ = \sum_{k}\left(\mathbf{Z}_{kk} - \mathbf{Z}_{ik}\right)\left(\frac{1 \pm \mathbf{w}_{k}}{2}\right) &= \sum_{k\in\pm}\left(\mathbf{Z}_{kk} - \mathbf{Z}_{ik}\right). \end{aligned} \tag{78}$$

We can think of η^{+}_{i} as a measure of the “distance” to the $\mathbf{w}_{k} = +1$ states and η^{-}_{i} as the “distance” to the $\mathbf{w}_{k} = -1$ states.

Using **[38]**, we can write:

$$\begin{aligned}\eta_i^+ + \eta_i^- &= \eta,\\ 2(\eta_i^+ - \eta_j^+) &= \sum_k (\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk})\mathbf{p}_k^\infty \mathbf{w}_k = \sum_k (\mathbf{Z}_{jk} - \mathbf{Z}_{ik})\mathbf{w}_k.\end{aligned} \qquad \textbf{[79]}$$

We could arrange the states in order of decreasing η^+, which is the same as the order of increasing η^-.

We can rewrite **[77]** as

$$\begin{aligned}\hat{A} &= 2\sum_{i,j} \mathbf{q}_{ij}\mathbf{p}_i^\infty (\eta_i^+ - \eta_j^+) &&= -2\sum_{i,j} \mathbf{q}_{ij}\mathbf{p}_i^\infty \eta_j^+\\ &= 2\sum_{i,j} \mathbf{q}_{ij}\mathbf{p}_i^\infty (\eta_j^- - \eta_i^-) &&= 2\sum_{i,j} \mathbf{q}_{ij}\mathbf{p}_i^\infty \eta_j^- .\end{aligned} \qquad \textbf{[80]}$$

We can also express it in terms of the fundamental matrix **[27]** as

$$\hat{A} = \sum_{i,j,k,l} \mathbf{q}_{ij}\boldsymbol{\pi}_l \mathbf{Z}_{li}(\mathbf{Z}_{jk} - \mathbf{Z}_{ik})\mathbf{w}_k = \boldsymbol{\pi}\mathbf{Z}q\mathbf{Z}\mathbf{w}. \qquad \textbf{[81]}$$

It is also helpful to define the following quantities:

$$\begin{aligned}c_k &= \frac{\mathrm{d}\ln \mathbf{p}_k^\infty}{\mathrm{d}\alpha} = \sum_{ij} \mathbf{p}_i^\infty \mathbf{q}_{ij}\left(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk}\right) = -\left(\mathbf{p}^\infty \mathbf{q}\overline{\mathbf{T}}\right)_k = \frac{(\mathbf{p}^\infty \mathbf{q}\mathbf{Z})_k}{\mathbf{p}_k^\infty},\\ a_i &= \sum_j \mathbf{q}_{ij}\mathbf{p}_i^\infty (\eta_i^+ - \eta_j^+),\\ \implies \hat{A} &= \sum_k c_k \mathbf{p}_k^\infty \mathbf{w}_k = 2\sum_i a_i.\end{aligned} \qquad \textbf{[82]}$$

Note that the optimal choice of $\mathbf{w}$ is $\mathbf{w}_k = \mathrm{sgn}(c_k)$.

Derivatives wrt. $\mathbf{W}^{\mathbf{pot/dep}}$. In the following, we will mathematically define the classes of perturbations pictorially described in Figure 3 of the main paper. In order to do so, we will need to consider expressions for derivatives of various quantities with respect to $\mathbf{W}_{ij}^{\mathrm{pot/dep}}$.

As discussed in the main text, we will regard the off-diagonal elements of $\mathbf{W}_{ij}^{\mathrm{pot/dep}}$ to be the independent variables, with $\mathbf{W}_{ii}^{\mathrm{pot/dep}} = -\sum_{j\neq i} \mathbf{W}_{ij}^{\mathrm{pot/dep}}$ imposed by hand. Thus,

$$\frac{\partial \mathbf{W}_{ij}^{\mathrm{F}}}{\partial \mathbf{W}_{gh}^{\mathrm{pot/dep}}} = f^{\mathrm{pot/dep}}\delta_{gi}(\delta_{hj} - \delta_{ij}), \qquad \frac{\partial \mathbf{q}_{ij}}{\partial \mathbf{W}_{gh}^{\mathrm{pot/dep}}} = \pm\delta_{gi}(\delta_{hj} - \delta_{ij}). \qquad \textbf{[83]}$$

The implicit $g \neq h$ that comes with all derivatives is unnecessary, as the derivatives above vanish when $g = h$.

In particular, differentiating **[27]**,

$$\frac{\partial \mathbf{Z}_{ij}}{\partial \mathbf{W}_{gh}^{\mathrm{pot/dep}}} = rf^{\mathrm{pot/dep}}\mathbf{Z}_{ig}(\mathbf{Z}_{hj} - \mathbf{Z}_{gj}). \qquad \textbf{[84]}$$

We can then differentiate expression **[81]** to get

$$\frac{\partial \hat{A}}{\partial \mathbf{W}_{gh}^{\mathrm{pot/dep}}} = 2rf^{\mathrm{pot/dep}}\mathbf{p}_g^\infty \left[\sum_i a_i(\overline{\mathbf{T}}_{gi} - \overline{\mathbf{T}}_{hi}) + c_g(\eta_g^+ - \eta_h^+)\right] \pm 2\mathbf{p}_g^\infty(\eta_g^+ - \eta_h^+). \qquad \textbf{[85]}$$

where a_i and c_k were defined in **[82]**.

It is sometimes useful to consider the following derivatives:

$$\frac{\partial}{\partial \mathbf{W}_{gh}^{\mathrm{F}}} \equiv \frac{\partial}{\partial \mathbf{W}_{gh}^{\mathrm{pot}}} + \frac{\partial}{\partial \mathbf{W}_{gh}^{\mathrm{dep}}}, \qquad \frac{\partial}{\partial \mathbf{q}_{gh}} \equiv f^{\mathrm{dep}}\frac{\partial}{\partial \mathbf{W}_{gh}^{\mathrm{pot}}} - f^{\mathrm{pot}}\frac{\partial}{\partial \mathbf{W}_{gh}^{\mathrm{dep}}}. \qquad \textbf{[86]}$$

Each of these derivatives behaves as their names suggest:

$$\frac{\partial \mathbf{W}_{ij}^{\mathrm{F}}}{\partial \mathbf{W}_{gh}^{\mathrm{F}}} = \frac{\partial \mathbf{q}_{ij}}{\partial \mathbf{q}_{gh}} = \delta_{gi}(\delta_{hj} - \delta_{ij}), \qquad \frac{\partial \mathbf{q}_{ij}}{\partial \mathbf{W}_{gh}^{\mathrm{F}}} = \frac{\partial \mathbf{W}_{ij}^{\mathrm{F}}}{\partial q_{gh}} = 0. \qquad \textbf{[87]}$$

This is because we could treat $\mathbf{W}^{\mathrm{F}}$ and q as the independent variables. However, the boundaries of the allowed region are more easily expressed in terms of $\mathbf{W}^{\mathrm{pot/dep}}$.

Scaling mode

Consider the following differential operator:

$$\Delta \equiv \sum_{g,h} \mathbf{W}^{\text{pot}}_{gh} \frac{\partial}{\partial \mathbf{W}^{\text{pot}}_{gh}} + \mathbf{W}^{\text{dep}}_{gh} \frac{\partial}{\partial \mathbf{W}^{\text{dep}}_{gh}}. \qquad [88]$$

This corresponds to the scaling, $\mathbf{W}^{\text{pot/dep}} \to (1+\epsilon)\mathbf{W}^{\text{pot/dep}}$. Intuitively, this has two effects: it scales up the initial potentiation/depression and it scales down all timescales. This intuition is confirmed by the following results:

$$\begin{aligned} \Delta\mathbf{Z} &= \tau\mathbf{e}\mathbf{p}^\infty - \mathbf{Z}, & \Delta\mathbf{p}^\infty &= 0, & \Delta\mathbf{T} &= -\mathbf{T}, \\ \Delta\eta^\pm_i &= -\eta^\pm_i, & \Delta\mathbf{q}_{ij} &= \mathbf{q}_{ij}, & \Delta\hat{A} &= 0, \end{aligned} \qquad [89]$$

The anomalous bit in the scaling of $\mathbf{Z}$ is due to the lack of dependence of $\boldsymbol{\pi}$ and τ on $\mathbf{W}^{\text{pot/dep}}$.

As the area is invariant under this scaling, we can consider the $\mathbf{W}^{\text{pot/dep}}$ to be projective coordinates. Therefore we don't need to enforce the lower bound on the diagonal matrix elements while looking for the maximum area, as we can use this null-mode to enforce it afterwards without changing the area.

Kuhn-Tucker conditions. Consider the Lagrangian

$$\mathcal{L} = \hat{A} + \sum_{\text{pot/dep}} \sum_{i \neq j} \mu^{\text{pot/dep}}_{ij} \mathbf{W}^{\text{pot/dep}}_{ij}. \qquad [90]$$

Necessary conditions for an extremum are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{pot/dep}}_{gh}} = 0, \qquad \mu^{\text{pot/dep}}_{gh} \geq 0, \quad \mathbf{W}^{\text{pot/dep}}_{gh} \geq 0, \quad \mu^{\text{pot/dep}}_{gh} \mathbf{W}^{\text{pot/dep}}_{gh} = 0. \qquad [91]$$

with $g \neq h$. This enforces the positivity constraints on the off-diagonal elements, but not the diagonals. As discussed in §, that can be enforced after finding the maximum using the null scaling degree of freedom.

Triangularity

Here we describe the perturbations corresponding to Figure 3(a,b) of the main paper.

Consider

$$\frac{\partial \mathcal{L}}{\partial \mathbf{q}_{gh}} = f^{\text{dep}} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{pot}}_{gh}} - f^{\text{pot}} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{dep}}_{gh}} = (f^{\text{dep}} \mu^{\text{pot}}_{gh} - f^{\text{pot}} \mu^{\text{dep}}_{gh}) + 2\mathbf{p}^\infty_g (\eta^+_g - \eta^+_h) = 0. \qquad [92]$$

This corresponds to the shift

$$\mathbf{W}^{\text{pot}}_{ij} \to \mathbf{W}^{\text{pot}}_{ij} + f^{\text{dep}} \epsilon_{ij}, \qquad \mathbf{W}^{\text{dep}}_{ij} \to \mathbf{W}^{\text{dep}}_{ij} - f^{\text{pot}} \epsilon_{ij}, \qquad \sum_j \epsilon_{ij} = 0, \qquad [93]$$

which leaves $\mathbf{W}^{\text{F}}$ unchanged, and therefore $\mathbf{p}^\infty$, $\mathbf{T}$ and $\eta^\pm$ as well.

Assume $\eta^+_g > \eta^+_h$. Then

$$f^{\text{dep}} \mu^{\text{pot}}_{gh} - f^{\text{pot}} \mu^{\text{dep}}_{gh} < 0 \qquad \Longrightarrow \qquad \mu^{\text{dep}}_{gh} > 0 \qquad \Longrightarrow \qquad \mathbf{W}^{\text{dep}}_{gh} = 0. \qquad [94]$$

Similarly, if $\eta^+_g < \eta^+_h$, then

$$f^{\text{dep}} \mu^{\text{pot}}_{gh} - f^{\text{pot}} \mu^{\text{dep}}_{gh} > 0 \qquad \Longrightarrow \qquad \mu^{\text{pot}}_{gh} > 0 \qquad \Longrightarrow \qquad \mathbf{W}^{\text{pot}}_{gh} = 0. \qquad [95]$$

Thus, if we arrange the states in order of decreasing η^+, $\mathbf{W}^{\text{pot}}$ is upper-triangular and $\mathbf{W}^{\text{dep}}$ is lower triangular.

We have ignored the possibility that $\mathbf{p}^\infty_g = 0$, as this would imply that $\mathbf{T}_{ig} = \infty$, which would in turn imply that the Markov process is not ergodic.

Shortcuts

In this subsection we will define perturbations corresponding to Figure 3(c) of the main text.

Now consider the following combinations of derivatives for $m > 1$:

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m} \equiv \left[\sum_{k=0}^{m-1} \frac{1}{\mathbf{p}^\infty_{g \pm k}} \left(\frac{\partial}{\partial \mathbf{W}^{\text{pot/dep}}_{g \pm k, g \pm (k+1)}} \right) \right] - \frac{1}{\mathbf{p}^\infty_g} \left(\frac{\partial}{\partial \mathbf{W}^{\text{pot/dep}}_{g, g \pm m}} \right). \qquad [96]$$

Once again, they are only well defined if all the states have non-zero equilibrium probabilities (see the comment in § about this being satisfied for ergodic chains).

One can show that the equilibrium probabilities, $\mathbf{p}^\infty$, are invariant under these operators [**41**]:

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m}\mathbf{p}^\infty_i = 0, \tag{97}$$

which makes it possible to integrate the perturbation:

$$\mathbf{W}^{\text{pot/dep}} \to \mathbf{W}^{\text{pot/dep}} + \mathbf{D}\boldsymbol{\epsilon}^{\pm(g,m)}, \qquad \begin{aligned} \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g,g\pm m} &= -\epsilon, \\ \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g\pm k,g\pm(k+1)} &= \epsilon \qquad \forall\, k \in [0, m-1], \\ \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g\pm k,g\pm k} &= -\epsilon \qquad \forall\, k \in [1, m-1]. \end{aligned} \tag{98}$$

But more interestingly for our purposes:

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m}\mathcal{L} = \left[\sum_{k=0}^{m-1} \frac{\mu^{\text{pot/dep}}_{g\pm k,g\pm(k+1)}}{\mathbf{p}^\infty_{g\pm k}} - \frac{\mu^{\text{pot/dep}}_{g,g\pm m}}{\mathbf{p}^\infty_g}\right] + 2rf^{\text{pot/dep}} \sum_{k=0}^{m-1} \left(\eta^+_{g\pm k} - \eta^+_{g\pm(k+1)}\right)\left(c_{g\pm k} - c_g\right), \tag{99}$$

In the section below, we will show that the c_k are non-decreasing, if we put the states in order of decreasing η^+_k. This implies that the last term of the final expression in [**99**] is non-negative. If it is non-zero (there would need to be a lot of degeneracy for it to be zero), this would imply that $\mu^{\text{pot/dep}}_{g,g\pm m} > 0$, which in turn implies that $\mathbf{W}^{\text{pot/dep}}_{g,g\pm m} = 0$. This would tell us that the process with the maximal area has to have a multi-state topology.

Increasing c_k

In the previous subsection we defined perturbations corresponding to Figure 3(c) of the main text. In order to show that those perturbations increase the area, we must now show that the c_k are non-decreasing, if we put the states in order of decreasing η^+_k.

Consider the following combinations of derivatives:

$$\Delta_{gh} \equiv \frac{1}{\mathbf{p}^\infty_g}\left(\frac{\partial}{\partial \mathbf{W}^{\text{F}}_{gh}}\right) + \frac{1}{\mathbf{p}^\infty_h}\left(\frac{\partial}{\partial \mathbf{W}^{\text{F}}_{hg}}\right), \tag{100}$$

$$\tag{101}$$

Note that they are only well defined if all the states have non-zero equilibrium probabilities (see the comment in § about this being satisfied for ergodic chains).

One can show that the equilibrium probabilities, $\mathbf{p}^\infty$, are invariant under these operators using [**41**]:

$$\Delta_{gh}\mathbf{p}^\infty_i = 0, \tag{102}$$

which makes it possible to integrate the perturbation:

$$\mathbf{W}^{\text{pot/dep}} \to \mathbf{W}^{\text{pot/dep}} + \mathbf{D}\boldsymbol{\epsilon}, \qquad \begin{aligned} \boldsymbol{\epsilon} &= \boldsymbol{\epsilon}^{\text{T}}, \\ \boldsymbol{\epsilon}\mathbf{e} &= 0. \end{aligned} \tag{103}$$

But more interestingly:

$$\Delta_{gh}\mathcal{L} = \frac{\mu^{\text{pot}}_{gh} + \mu^{\text{dep}}_{gh}}{\mathbf{p}^\infty_g} + \frac{\mu^{\text{pot}}_{hg} + \mu^{\text{dep}}_{hg}}{\mathbf{p}^\infty_h} + 2r\left(c_g - c_h\right)\left(\eta^+_g - \eta^+_h\right), \tag{104}$$

$$\tag{105}$$

where c_k were defined in [**82**].

Using the non-negativity of the Kuhn-Tucker multipliers, $\mu^{\text{pot/dep}}_{ij}$, [**104**] tells us that if we arrange the states in order of decreasing η^+_i, the optimal process will have non-decreasing c_k (if any of the η^+_k are degenerate, we can choose their order to ensure this).

Note that, according to §, either $\mathbf{W}^{\text{pot}}_{gh}$ or $\mathbf{W}^{\text{dep}}_{gh}$ will be zero at the maximum, therefore we can expect one of $\mu^{\text{pot}}_{gh} + \mu^{\text{dep}}_{gh}$ to be non-zero. This would rule out degeneracy of the c_k or η^+_k. Looking at [**92**] closely, the only way $\mu^{\text{pot}}_{gh} + \mu^{\text{dep}}_{gh}$ could be zero is if $\eta^+_g = \eta^+_h$ or $\mathbf{p}^\infty_g = 0$.

Summary

Using the Kuhn-Tucker formalism, we have shown that, with the states arranged in order of non-increasing η^+_i:

- There can be no ergodic maximum for which $\mathbf{W}^{\text{pot}}$ contains backwards transitions or $\mathbf{W}^{\text{dep}}$ contains forwards transitions.
- There can be no ergodic maximum with the c_k decreasing.
- The c_k may only be degenerate at an ergodic maximum if the corresponding η^+_k are also degenerate.

- If the c_k increase and the η_i^+ decrease, there can be no ergodic maximum with shortcuts.

These were shown by finding allowed perturbations that increase the area.

This leaves two possibilities for the maximum area Markov chain. Either there is no degeneracy and no shortcuts, which implies the Multi-state/serial topology that we'll discuss in §, or there is some degeneracy, which would allow shortcuts provided that they do not bypass an entire degenerate set (see **[99]**).

Degeneracy tends to be very delicate. It is usually hard to arrange without some symmetry relating degenerate states. Such a symmetry would imply lumpability (see §). The lumped chain would not have any shortcuts, as an entire degenerate set cannot be bypassed. As this lumped chain has the same area (see §), we would need only consider the multi-state topology.

Multi-state/Serial topology. The previous results indicate that the area under the memory curve of any model is bounded by the area under the memory curve of a model with the serial/multistate topology having the same equilibrium distribution. Here we compute this area, which we will see depends only on this equilibrium distribution.

The multi-state/serial topology is defined by (see [9, 11, 12]):

$$\mathbf{W}^{\text{pot}}_{ij} = q^{\text{pot}}_i \delta_{i+1,j}, \qquad \mathbf{W}^{\text{dep}}_{ij} = q^{\text{dep}}_j \delta_{i,j+1}. \qquad \textbf{[106]}$$

Because it has no shortcuts, it saturates various inequalities:

$$\begin{aligned} \overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk} &= \begin{cases} \overline{\mathbf{T}}_{ij}, & \text{if} \quad i \leq j \leq k \quad \text{or} \quad i \geq j \geq k, \\ -\overline{\mathbf{T}}_{ji}, & \text{if} \quad j \leq i \leq k \quad \text{or} \quad j \geq i \geq k, \end{cases} \\ r\mathbf{p}^\infty_i \mathbf{W}^{\text{F}}_{ij} \left(\overline{\mathbf{T}}_{ij} + \overline{\mathbf{T}}_{ji}\right) &= 1 \quad \text{if} \quad i = j \pm 1, \end{aligned} \qquad \textbf{[107]}$$

and it satisfies detailed balance (a.k.a. reversibility a.k.a. $\mathcal{L}^2_{\mathbf{p}^\infty}$ self-adjointness):

$$f^{\text{pot}} q^{\text{pot}}_i \mathbf{p}^\infty_i = f^{\text{dep}} q^{\text{dep}}_i \mathbf{p}^\infty_{i+1}, \qquad \textbf{[108]}$$

which means we can always choose the transition rates, $q_i^{\text{pot/dep}}$, to give any desired equilibrium probabilities, $\mathbf{p}^\infty_i$.

This allows us to calculate the c_k's:

$$\begin{aligned} c_k &= \sum_{i<k} \mathbf{T}_{i,i+1} \left(\mathbf{p}^\infty_i q^{\text{pot}}_i + \mathbf{p}^\infty_{i+1} q^{\text{dep}}_i\right) - \sum_{i \geq k} \mathbf{T}_{i+1,i} \left(\mathbf{p}^\infty_i q^{\text{pot}}_i + \mathbf{p}^\infty_{i+1} q^{\text{dep}}_i\right), \\ c_{k+1} - c_k &= \left(\mathbf{T}_{k,k+1} + \mathbf{T}_{k+1,k}\right) \left(\frac{\mathbf{p}^\infty_k \mathbf{W}^{\text{F}}_{k,k+1}}{f^{\text{pot}}} + \frac{\mathbf{p}^\infty_{k+1} \mathbf{W}^{\text{F}}_{k+1,k}}{f^{\text{dep}}}\right) = \frac{1}{r f^{\text{pot}} f^{\text{dep}}}, \\ \sum_k c_k \mathbf{p}^\infty_k &= \sum_{ij} \mathbf{p}^\infty_i \mathbf{q}_{ij} (\eta - \eta) = 0, \\ \implies c_k &= \frac{k - \sum_j j\mathbf{p}^\infty_j}{r f^{\text{pot}} f^{\text{dep}}}, \end{aligned} \qquad \textbf{[109]}$$

where we used **[107]** to derive the first two equations respectively and Th.3 to derive the third. This allows us to write the area as

$$A = \frac{2\sqrt{N}}{r} \sum_k \left[k - \sum_j j\mathbf{p}^\infty_j\right] \mathbf{p}^\infty_k \mathbf{w}_k = \frac{2\sqrt{N}}{r} \sum_k \left|k - \sum_j j\mathbf{p}^\infty_j\right| \mathbf{p}^\infty_k, \qquad \textbf{[110]}$$

where we used $\mathbf{w}_k = \operatorname{sgn}(c_k)$, as discussed after **[82]**. This reproduces equation (15) of the main paper.

In order to obtain an upper bound on the area under the memory curve of any model, we now maximise the area of the serial model with respect to its equilibrium distribution. First let us maximise **[110]** at fixed $\mathbf{p}^\infty_\pm = \sum_k \mathbf{p}^\infty_k \left(\frac{1 \pm \mathbf{w}_k}{2}\right)$. Clearly this will happen when we put all of the probability at the ends: $\mathbf{p}^\infty_1 = \mathbf{p}^\infty_-$ and $\mathbf{p}^\infty_n = \mathbf{p}^\infty_+$ are the only non-zero $\mathbf{p}^\infty_k$. This gives an area of

$$A \leq \frac{\sqrt{N}}{r} (M-1) \left(4\mathbf{p}^\infty_+ \mathbf{p}^\infty_-\right). \qquad \textbf{[111]}$$

This is maximised at $\mathbf{p}^\infty_+ = \mathbf{p}^\infty_- = \frac{1}{2}$:

$$A \leq \frac{\sqrt{N}}{r} (M-1). \qquad \textbf{[112]}$$

This yields the area bound of equation (16) of the main text.

Note that the chain that achieves this is not ergodic, the two states at each end are absorbing. This is similar to the results found numerically in [18] in a slightly different situation.