# Optimal synaptic strategies for different timescales of memory

**Lahiri *et al.*** **www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX**

This manuscript was compiled on March 19, 2018

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

## 1. 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 (1) with some ideas from (2).

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

$$\mathbf{A}^{\mathrm{dg}}_{ij} \equiv \delta_{ij}\mathbf{A}_{ij}, \qquad \overline{\mathbf{A}} \equiv \mathbf{A} - \mathbf{A}^{\mathrm{dg}}. \tag{1}$$

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

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \mathbf{p}(t)\mathbf{Q}, \tag{2}$$

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

The equilibrium probabilities, $\boldsymbol{\pi}$, satisfy

$$\boldsymbol{\pi}\mathbf{Q} = 0, \qquad \boldsymbol{\pi}\mathbf{e} = 1. \tag{3}$$

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

$$\boldsymbol{\Lambda} \equiv (-\mathbf{Q}^{\mathrm{dg}})^{-1}, \qquad \mathbf{P} \equiv \mathbf{I} + \boldsymbol{\Lambda}\mathbf{Q}. \tag{4}$$

It can be shown that $\boldsymbol{\Lambda}_{ii}$ 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 $j$:

$$\boldsymbol{\Lambda}_{ii} = \frac{1}{\sum_{j\neq j}\mathbf{Q}_{ij}}, \qquad \mathbf{P}_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \frac{\mathbf{Q}_{ij}}{\sum_{k\neq j}\mathbf{Q}_{ik}} & \text{otherwise.} \end{cases} \tag{5}$$

Furthermore, we also define

$$\mathbf{D} \equiv \mathrm{diag}(\boldsymbol{\pi})^{-1}, \qquad \Longrightarrow \qquad \boldsymbol{\pi}\mathbf{D} = \mathbf{e}^{\mathrm{T}}. \tag{6}$$

**B. 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 (3). For continuous time, we define:

$$\mathbf{Z} \equiv (-\mathbf{Q} + \mathbf{e}\boldsymbol{\xi})^{-1}, \tag{7}$$

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

Note that the canonical choice for the discrete time version, $\boldsymbol{\xi} = \boldsymbol{\pi}$, is not available here due to problems with units. It will be helpful to choose $\boldsymbol{\xi}$ to be independent of $\mathbf{Q}$, e.g. $\boldsymbol{\xi} = \mathbf{e}^{\mathrm{T}}/(n\tau)$. 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. $(-\mathbf{Q} + \mathbf{e}\boldsymbol{\xi})$ is invertible.

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

$$(-\mathbf{Q} + \mathbf{e}\boldsymbol{\xi})\mathbf{x} = 0. \tag{8}$$

Multiplying from the left with $\boldsymbol{\pi}$ gives

$$\boldsymbol{\xi}\mathbf{x} = 0. \tag{9}$$

Substituting back into Eq. [8] 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 Eq. [9] gives

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

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

**Corollary 2:**

$$\boldsymbol{\xi}\mathbf{Z} = \boldsymbol{\pi}, \tag{10}$$

$$\mathbf{Z}\mathbf{e} = \tau\mathbf{e}, \tag{11}$$

$$\mathbf{I} + \mathbf{Q}\mathbf{Z} = \mathbf{e}\boldsymbol{\pi}, \tag{12}$$

$$\mathbf{I} + \mathbf{Z}\mathbf{Q} = \tau\mathbf{e}\boldsymbol{\xi}. \tag{13}$$

*Proof.* We can deduce Eqs. [10] and [11] be pre/post-multiplying the following equations by $\mathbf{Z}$:

$$\boldsymbol{\pi}(-\mathbf{Q} + \mathbf{e}\boldsymbol{\xi}) = \boldsymbol{\xi},$$

$$(-\mathbf{Q} + \mathbf{e}\boldsymbol{\xi})\mathbf{e} = \frac{\mathbf{e}}{\tau}.$$

We can then deduce Eqs. [12] and [13] by substituting these into

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

□

### C. 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

$$\mathbf{T} = (\mathbf{E}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z} + \boldsymbol{\Lambda})\mathbf{D}, \tag{14}$$

see (4) for a proof. We can separate this into its diagonal and off-diagonal pieces.

The recurrence times are given by

$$\mathbf{T}^{\mathrm{dg}} = \boldsymbol{\Lambda}\mathbf{D}. \tag{15}$$

or in component form

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

The extra factor of $\boldsymbol{\Lambda}_{ii}$, compared to the discrete case (1, 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

$$\overline{\mathbf{T}} = (\mathbf{E}\mathbf{Z}^{\mathrm{dg}} - \mathbf{Z})\mathbf{D}. \tag{16}$$

or in component form:

$$\overline{\mathbf{T}}_{ij} = \frac{\mathbf{Z}_{jj} - \mathbf{Z}_{ij}}{\boldsymbol{\pi}_j}. \tag{17}$$

### D. Mixing time (Kemeny's constant).

**Theorem 3:**
The quantity

$$\eta \equiv \sum_j \overline{\mathbf{T}}_{ij}\boldsymbol{\pi}_j \tag{18}$$

is independent of $i$.

*Proof.* For discrete time, a proof can be found in (1, Th.4.4.10). For continuous time, we use Eq. [16], Eq. [11] and the transpose of Eq. [6]:

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

which proves Eq. [18] 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 $\eta$ by 1.

**E. Sensitivity of equilibrium distribution.** Suppose that the Markov process, defined by $\mathbf{Q}$, depends on some parameter $\alpha$. Differentiating Eq. [7] gives

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}\alpha} = \mathbf{Z}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z}. \qquad [19]$$

We can substitute this into the derivative of Eq. [10]:

$$\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha} = \boldsymbol{\xi}\mathbf{Z}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z} = \boldsymbol{\pi}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\alpha}\mathbf{Z}. \qquad [20]$$

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

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\pi}_k}{\mathrm{d}\alpha} &= \sum_{i,j}\boldsymbol{\pi}_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\mathbf{Z}_{jk} \\ &= \sum_{i\neq j}\boldsymbol{\pi}_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\mathbf{Z}_{jk} + \sum_{i}\boldsymbol{\pi}_i\frac{\mathrm{d}\mathbf{Q}_{ii}}{\mathrm{d}\alpha}\mathbf{Z}_{ik} \\ &= \sum_{i\neq j}\boldsymbol{\pi}_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}(\mathbf{Z}_{jk} - \mathbf{Z}_{ik}) \\ &= \sum_{i\neq j}\boldsymbol{\pi}_i\frac{\mathrm{d}\mathbf{Q}_{ij}}{\mathrm{d}\alpha}\boldsymbol{\pi}_k(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk}).\end{aligned} \qquad [21]$$

This is a generalization of a result of (5) 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.

**F. 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:

$$\left(\mathbf{I}^{\mathcal{A}}\right)_{ij} = \begin{cases}1 & \text{if } i = j \in \mathcal{A}, \\ 0 & \text{otherwise.}\end{cases} \qquad [22]$$

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

$$\begin{aligned}&\boldsymbol{\xi}^{\mathcal{A}} = \boldsymbol{\xi}\mathbf{I}^{\mathcal{A}}, \quad \mathbf{M}^{\cdot\mathcal{A}} = \mathbf{M}\mathbf{I}^{\mathcal{A}}, \quad \mathbf{M}^{\mathcal{A}\cdot} = \mathbf{I}^{\mathcal{A}}\mathbf{M}, \qquad \mathbf{x}^{\mathcal{A}} = \mathbf{I}^{\mathcal{A}}\mathbf{x}, \\ &\boldsymbol{\xi}_{\mathcal{A}} = \boldsymbol{\xi}\mathbf{e}^{\mathcal{A}}, \quad \mathbf{M}_{\cdot\mathcal{A}} = \mathbf{M}\mathbf{e}^{\mathcal{A}}, \quad \mathbf{M}_{\mathcal{A}\cdot} = \left(\mathbf{e}^{\mathcal{A}}\right)^{\mathrm{T}}\mathbf{M}, \quad \mathbf{x}_{\mathcal{A}} = \left(\mathbf{e}^{\mathcal{A}}\right)^{\mathrm{T}}\mathbf{x},\end{aligned} \qquad [23]$$

where $\boldsymbol{\xi}$ 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:

$$\boldsymbol{\Phi} = \mathbf{D}^{-1}\mathbf{Q}, \qquad \boldsymbol{\Phi}_{ij} = \boldsymbol{\pi}_i\mathbf{Q}_{ij}. \qquad [24]$$

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:

$$\boldsymbol{\Phi}_{\mathcal{A}\mathcal{B}} = \boldsymbol{\pi}^{\mathcal{A}}\mathbf{Q}\mathbf{e}^{\mathcal{B}}. \qquad [25]$$

One can show that

$$\boldsymbol{\Phi}_{\mathcal{A}\mathcal{A}^{\mathrm{c}}} = \boldsymbol{\Phi}_{\mathcal{A}^{\mathrm{c}}\mathcal{A}} = -\boldsymbol{\Phi}_{\mathcal{A}\mathcal{A}} = -\boldsymbol{\Phi}_{\mathcal{A}^{\mathrm{c}}\mathcal{A}^{\mathrm{c}}} \qquad [26]$$

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

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

$$\bigcup_\alpha \mathcal{A}_\alpha = \mathcal{S}, \qquad \mathcal{A}_\alpha \cap \mathcal{A}_\beta = \delta_{\alpha\beta}\mathcal{A}_\alpha. \tag{27}$$

We will use $\alpha$ 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

$$\sigma(i) = \alpha \qquad \iff \qquad i \in \mathcal{A}_\alpha. \tag{28}$$

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 (1, §6.3) for the discrete time version and (6, 7) for continuous time):

$$\sigma(i) = \sigma(j) \quad \implies \quad \mathbf{Q}_{i\alpha} = \mathbf{Q}_{j\alpha} \equiv \sum_{k\in\mathcal{A}_\alpha} \mathbf{Q}_{jk}, \tag{29}$$

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

$$U_{\alpha i} = \frac{\delta_{\alpha\sigma(i)}}{|\mathcal{A}_\alpha|}, \qquad V_{i\alpha} = \delta_{\sigma(i)\alpha}. \tag{30}$$

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:

$$U_{\alpha i} = \frac{\boldsymbol{\pi}_i{}^\alpha}{\boldsymbol{\pi}_\alpha}. \tag{31}$$

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:

$$VU\mathbf{x} = \mathbf{x} \qquad \iff \qquad \mathbf{x} = \sum_\alpha x_\alpha \mathbf{e}^\alpha. \tag{32}$$

This allows us to write the lumpability condition Eq. [29], and the transition matrix for the lumped process compactly:

$$VU\mathbf{Q}V = \mathbf{Q}V, \qquad \widehat{\mathbf{Q}} = U\mathbf{Q}V. \tag{33}$$

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

$$VU\mathbf{Q}^n V = \mathbf{Q}^n V, \qquad \widehat{\mathbf{Q}}^n = U\mathbf{Q}^n V, \tag{34}$$

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

$$VU\mathrm{e}^{t\mathbf{Q}}V = \mathrm{e}^{t\mathbf{Q}}V, \qquad \mathrm{e}^{t\widehat{\mathbf{Q}}} = U\mathrm{e}^{t\mathbf{Q}}V. \tag{35}$$

The equilibrium distribution of the lumped process is given by

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}V. \tag{36}$$

## 2. 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.

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

$$\mathbf{M}^{\text{pot/dep}} \equiv \mathbf{I} + \mathbf{W}^{\text{pot/dep}}, \qquad \mathbf{M}^{\text{pot/dep}}\mathbf{e} = \mathbf{e}, \qquad \mathbf{M}^{\text{pot/dep}}_{ij} \in [0,1]. \tag{37}$$

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:

$$\mathbf{Q} = r\mathbf{W}^{\text{F}} \equiv r\left(f^{\text{pot}}\mathbf{W}^{\text{pot}} + f^{\text{dep}}\mathbf{W}^{\text{dep}}\right). \tag{38}$$

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 Eq. [3]. During the initial signal creation, a fraction $f^{\text{pot}}$ will change to $\boldsymbol{\pi}\mathbf{M}^{\text{pot}}$ and a fraction $f^{\text{dep}}$ will change to $\boldsymbol{\pi}\mathbf{M}^{\text{dep}}$. After this, probabilities will evolve according to Eq. [2].

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

$$\mathrm{SNR}(t) = \frac{\langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty) \rangle}{\sqrt{\mathrm{Var}(\vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty))}}. \qquad [39]$$

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

$$\begin{aligned} \mathrm{Var}(\vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty)) &= \sum_{\alpha\beta} \left\langle \vec{w}^{\alpha}_{\mathrm{ideal}} \vec{w}^{\alpha}(\infty) \vec{w}^{\beta}_{\mathrm{ideal}} \vec{w}^{\beta}(\infty) \right\rangle - \left( \sum_{\alpha} \langle \vec{w}^{\alpha}_{\mathrm{ideal}} \vec{w}^{\alpha}(\infty) \rangle \right)^2 \\ &= \sum_{\alpha} \left\langle (\vec{w}^{\alpha}_{\mathrm{ideal}})^2 (\vec{w}^{\alpha}(\infty))^2 \right\rangle + \sum_{\alpha \neq \beta} \langle \vec{w}^{\alpha}_{\mathrm{ideal}} \vec{w}^{\alpha}(\infty) \rangle \langle \vec{w}^{\beta}{}_{\mathrm{ideal}} \vec{w}^{\beta}(\infty) \rangle - \left( \sum_{\alpha} \langle \vec{w}^{\alpha}_{\mathrm{ideal}} \vec{w}^{\alpha}(\infty) \rangle \right)^2 \\ &= N \langle 1 \rangle + N(N-1) \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(\infty) \right\rangle^2 - N^2 \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(\infty) \right\rangle^2 \\ &= N(1 - \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(\infty) \right\rangle^2), \end{aligned} \qquad [40]$$

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

For the numerator, we can write

$$\langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(t) \rangle = \sum_{\alpha} \langle \vec{w}^{\alpha}_{\mathrm{ideal}} \vec{w}^{\alpha}(t) \rangle = N \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(t) \right\rangle, \qquad [41]$$

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

$$\begin{aligned} \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(t) \right\rangle &= f^{\mathrm{pot}} \left\langle \vec{w}^{1}(t) \right\rangle_{\mathrm{pot},t=0} - f^{\mathrm{dep}} \left\langle \vec{w}^{1}(t) \right\rangle_{\mathrm{dep},t=0} \\ &= f^{\mathrm{pot}} \sum_i P(\mathrm{state} = i, t \mid \mathrm{pot}, 0) \mathbf{w}_i - f^{\mathrm{dep}} \sum_i P(\mathrm{state} = i, t \mid \mathrm{dep}, 0) \mathbf{w}_i. \end{aligned} \qquad [42]$$

From the previous section,

$$P(\mathrm{state} = i, t \mid \mathrm{pot/dep}, 0) = \left[ \boldsymbol{\pi} \mathbf{M}^{\mathrm{pot/dep}} \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \right]_i, \qquad [43]$$

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

$$\begin{aligned} \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(t) \right\rangle &= \boldsymbol{\pi} (f^{\mathrm{pot}} \mathbf{M}^{\mathrm{pot}} - f^{\mathrm{dep}} \mathbf{M}^{\mathrm{dep}}) \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w}, \\ \left\langle \vec{w}^{1}_{\mathrm{ideal}} \vec{w}^{1}(\infty) \right\rangle &= \boldsymbol{\pi} (f^{\mathrm{pot}} \mathbf{M}^{\mathrm{pot}} - f^{\mathrm{dep}} \mathbf{M}^{\mathrm{dep}}) \, \mathbf{e} \boldsymbol{\pi} \mathbf{w} \\ &= \boldsymbol{\pi} (f^{\mathrm{pot}} \mathbf{e} - f^{\mathrm{dep}} \mathbf{e}) \, \boldsymbol{\pi} \mathbf{w} \\ &= (f^{\mathrm{pot}} - f^{\mathrm{dep}}) \, \boldsymbol{\pi} \mathbf{w} \\ &= (f^{\mathrm{pot}} - f^{\mathrm{dep}}) \, \boldsymbol{\pi} \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w}. \end{aligned} \qquad [44]$$

Combining these allows us to write the numerator as

$$\begin{aligned} \langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty) \rangle &= N \boldsymbol{\pi} (f^{\mathrm{pot}} (\mathbf{M}^{\mathrm{pot}} - \mathbf{I}) - f^{\mathrm{dep}} (\mathbf{M}^{\mathrm{dep}} - \mathbf{I})) \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w} \\ &= N \boldsymbol{\pi} (f^{\mathrm{pot}} (\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{F}}) - f^{\mathrm{dep}} (\mathbf{W}^{\mathrm{dep}} - \mathbf{W}^{\mathrm{F}})) \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w} \\ &= N (2 f^{\mathrm{pot}} f^{\mathrm{dep}}) \boldsymbol{\pi} (\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}}) \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w}. \end{aligned} \qquad [45]$$

where we used $\boldsymbol{\pi} \mathbf{W}^{\mathrm{F}} = 0$ in going from the first to second lines. Combining with Eq. [40] gives

$$\mathrm{SNR}(t) = \frac{\sqrt{N} (2 f^{\mathrm{pot}} f^{\mathrm{dep}}) \boldsymbol{\pi} (\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}}) \, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}} \mathbf{w}}{\sqrt{1 - (f^{\mathrm{pot}} - f^{\mathrm{dep}})^2 (\boldsymbol{\pi}_+ - \boldsymbol{\pi}_-)^2}}. \qquad [46]$$

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 $f^{\mathrm{pot}} = f^{\mathrm{dep}}$ or $\boldsymbol{\pi}_+ = \boldsymbol{\pi}_-$. We can set the denominator to 1 without changing any of our results.

[*] Note that expanding the exponential gives

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

Thus, evolving according to Eq. [2] results in averaging over all sequences of plasticity events, as we only need linear expectations of $\vec{w}(t)$ in the end.

This results in our final formula:

$$\mathrm{SNR}(t) = \sqrt{N}(2f^{\mathrm{pot}}f^{\mathrm{dep}})\,\boldsymbol{\pi}(\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}})\,\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\,\mathbf{w}. \tag{47}$$

The factor of $\boldsymbol{\pi}$ describes the synapses being in the steady-state distribution before the memory is encoded. The factor of $(\mathbf{M}^{\mathrm{pot}} - \mathbf{M}^{\mathrm{dep}})$ comes from the encoding of the memory at $t = 0$, with $\vec{w}_{\mathrm{ideal}}$ being $\pm 1$ in synapses that are potentiated/depotentiated. The factor of $\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}$ 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:

$$\mathbf{W}(\alpha) = \alpha\mathbf{W}^{\mathrm{pot}} + (1-\alpha)\mathbf{W}^{\mathrm{dep}}, \qquad \implies \qquad \begin{aligned} \mathbf{W}^{\mathrm{F}} &= \mathbf{W}(f^{\mathrm{pot}}), \\ \mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}} &= \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\alpha}, \\ \boldsymbol{\pi}\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\alpha} &= -\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha}\mathbf{W}^{\mathrm{F}}. \end{aligned} \tag{48}$$

Then Eq. [47] becomes

$$\mathrm{SNR}(t) = \sqrt{N}(2f^{\mathrm{pot}}f^{\mathrm{dep}})\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha}(-\mathbf{W}^{\mathrm{F}})\,\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\,\mathbf{w}. \tag{49}$$

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

$$VU\mathbf{W}^{\mathrm{pot/dep}}V = \mathbf{W}^{\mathrm{pot/dep}}V, \qquad VU\mathbf{w} = \mathbf{w}. \tag{50}$$

We can define a new synapse with

$$\widehat{\mathbf{W}}^{\mathrm{pot/dep}} = U\mathbf{W}^{\mathrm{pot/dep}}V, \qquad \widehat{\mathbf{w}} = U\mathbf{w}, \qquad \widehat{\mathbf{p}}^{\infty} = \boldsymbol{\pi}V. \tag{51}$$

This synapse has an SNR curve:

$$\begin{aligned} \frac{\mathrm{SNR}(t)}{\sqrt{N}(2f^{\mathrm{pot}}f^{\mathrm{dep}})} &= \widehat{\mathbf{p}}^{\infty}(\widehat{\mathbf{W}}^{\mathrm{pot}} - \widehat{\mathbf{W}}^{\mathrm{dep}})\mathrm{e}^{rt\widehat{\mathbf{W}}^{F}}\,\widehat{\mathbf{w}}. \\ &= \boldsymbol{\pi}VU(\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}})VU\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}VU\mathbf{w}. \\ &= \boldsymbol{\pi}(\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}})VU\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}VU\mathbf{w}. \\ &= \boldsymbol{\pi}(\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}})\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}VU\mathbf{w}. \\ &= \boldsymbol{\pi}(\mathbf{W}^{\mathrm{pot}} - \mathbf{W}^{\mathrm{dep}})\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}. \end{aligned} \tag{52}$$

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

**D. Initial SNR and flux.** Using $\boldsymbol{\pi}\mathbf{W}^{\mathrm{F}} = 0$ and the first line of Eq. [45], we can write the initial SNR as

$$\frac{\mathrm{SNR}(0)}{\sqrt{N}} = I = (\boldsymbol{\pi}^{+} + \boldsymbol{\pi}^{-})(f^{\mathrm{pot}}\mathbf{W}^{\mathrm{pot}} - f^{\mathrm{dep}}\mathbf{W}^{\mathrm{dep}})(\mathbf{e}^{+} - \mathbf{e}^{-}). \tag{53}$$

Using $\mathbf{W}^{\mathrm{pot/dep}}(\mathbf{e}^{+} + \mathbf{e}^{-}) = 0$ and Eq. [26]:

$$r\boldsymbol{\pi}^{-}(f^{\mathrm{pot}}\mathbf{W}^{\mathrm{pot}} + f^{\mathrm{dep}}\mathbf{W}^{\mathrm{dep}})\mathbf{e}^{+} = \boldsymbol{\Phi}_{-+} = \boldsymbol{\Phi}_{+-} = r\boldsymbol{\pi}^{+}(f^{\mathrm{pot}}\mathbf{W}^{\mathrm{pot}} + f^{\mathrm{dep}}\mathbf{W}^{\mathrm{dep}})\mathbf{e}^{-},$$

we can rewrite Eq. [7] as

$$I = \frac{4\boldsymbol{\Phi}_{-+}}{r} - 4\boldsymbol{\pi}^{+}f^{\mathrm{pot}}\mathbf{W}^{\mathrm{pot}}\mathbf{e}^{-} - 4\boldsymbol{\pi}^{-}f^{\mathrm{dep}}\mathbf{W}^{\mathrm{dep}}\mathbf{e}^{+}. \tag{54}$$

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

$$\mathrm{SNR}(0) \leq \frac{4\sqrt{N}\boldsymbol{\Phi}_{-+}}{r}. \tag{55}$$

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.

## 3. Area maximisation

In this section we will find an upper bound on the area under the signal-to-noise curve. As in Section 2, we will only consider ergodic Markov chains. We will see in Section D 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.

**A. Area under signal-to-noise curve.** The signal-to-noise curve is given by Eq. [49]. The area is computed by integrating this

$$\begin{aligned} A &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha}\left[-\mathrm{e}^{rt\mathbf{W}^{\text{F}}}\right]_0^\infty \mathbf{w} \\ &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha}(\mathbf{I} - \mathbf{e}\boldsymbol{\pi})\mathbf{w} \\ &= \frac{\sqrt{N}(2f^{\text{pot}}f^{\text{dep}})}{r}\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\alpha}\mathbf{w}. \end{aligned} \tag{56}$$

We can rewrite this using Eq. [21], with $A = \sqrt{N}(2f^{\text{pot}}f^{\text{dep}})\hat{A}$ and $\mathbf{q}_{ij} \equiv \frac{\mathrm{d}\mathbf{W}^{\text{F}}_{ij}}{\mathrm{d}\alpha} = \mathbf{W}^{\text{pot}}_{ij} - \mathbf{W}^{\text{dep}}_{ij}$

$$\hat{A} = \sum_{i,j,k} \boldsymbol{\pi}_i \mathbf{q}_{ij}(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk})\boldsymbol{\pi}_k \mathbf{w}_k. \tag{57}$$

**Definition 3: Partial mixing times**

We define the ± mixing times as

$$\begin{aligned} \eta^\pm_i \equiv \sum_k \overline{\mathbf{T}}_{ik}\boldsymbol{\pi}_k\left(\frac{1 \pm \mathbf{w}_k}{2}\right) \qquad &= \sum_{k\in\pm} \overline{\mathbf{T}}_{ik}\boldsymbol{\pi}_k \\ = \sum_k (\mathbf{Z}_{kk} - \mathbf{Z}_{ik})\left(\frac{1 \pm \mathbf{w}_k}{2}\right) \qquad &= \sum_{k\in\pm} (\mathbf{Z}_{kk} - \mathbf{Z}_{ik}). \end{aligned} \tag{58}$$

We can think of $\eta^+_i$ as a measure of the "distance" to the $\mathbf{w}_k = +1$ states and $\eta^-_i$ as the "distance" to the $\mathbf{w}_k = -1$ states.

Using Eq. [18], we can write:

$$\begin{aligned} \eta^+_i + \eta^-_i &= \eta, \\ 2(\eta^+_i - \eta^+_j) &= \sum_k (\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk})\boldsymbol{\pi}_k\mathbf{w}_k = \sum_k (\mathbf{Z}_{jk} - \mathbf{Z}_{ik})\mathbf{w}_k. \end{aligned} \tag{59}$$

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

We can rewrite Eq. [57] as

$$\begin{aligned} \hat{A} = 2\sum_{i,j}\mathbf{q}_{ij}\boldsymbol{\pi}_i(\eta^+_i - \eta^+_j) \qquad &= -2\sum_{i,j}\mathbf{q}_{ij}\boldsymbol{\pi}_i\eta^+_j \\ = 2\sum_{i,j}\mathbf{q}_{ij}\boldsymbol{\pi}_i(\eta^-_j - \eta^-_i) \qquad &= 2\sum_{i,j}\mathbf{q}_{ij}\boldsymbol{\pi}_i\eta^-_j. \end{aligned} \tag{60}$$

We can also express it in terms of the fundamental matrix Eq. [7] as

$$\hat{A} = \sum_{i,j,k,l} \mathbf{q}_{ij}\boldsymbol{\xi}_l\mathbf{Z}_{li}(\mathbf{Z}_{jk} - \mathbf{Z}_{ik})\mathbf{w}_k = \boldsymbol{\xi}\mathbf{Z}q\mathbf{Z}\mathbf{w}. \tag{61}$$

It is also helpful to define the following quantities:

$$\begin{aligned} c_k &= \frac{\mathrm{d}\ln\boldsymbol{\pi}_k}{\mathrm{d}\alpha} = \sum_{ij}\boldsymbol{\pi}_i\mathbf{q}_{ij}\left(\overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk}\right) = -\left(\boldsymbol{\pi}\mathbf{q}\overline{\mathbf{T}}\right)_k = \frac{(\boldsymbol{\pi}\mathbf{q}\mathbf{Z})_k}{\boldsymbol{\pi}_k}, \\ a_i &= \sum_j \mathbf{q}_{ij}\boldsymbol{\pi}_i(\eta^+_i - \eta^+_j), \\ \implies \hat{A} &= \sum_k c_k\boldsymbol{\pi}_k\mathbf{w}_k = 2\sum_i a_i. \end{aligned} \tag{62}$$

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

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

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

$$\frac{\partial\mathbf{W}^{\text{F}}_{ij}}{\partial\mathbf{W}^{\text{pot/dep}}_{gh}} = f^{\text{pot/dep}}\delta_{gi}(\delta_{hj} - \delta_{ij}), \qquad \frac{\partial\mathbf{q}_{ij}}{\partial\mathbf{W}^{\text{pot/dep}}_{gh}} = \pm\delta_{gi}(\delta_{hj} - \delta_{ij}). \tag{63}$$

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

In particular, differentiating Eq. [7],

$$\frac{\partial \mathbf{Z}_{ij}}{\partial \mathbf{W}^{\text{pot/dep}}_{gh}} = r f^{\text{pot/dep}} \mathbf{Z}_{ig}(\mathbf{Z}_{hj} - \mathbf{Z}_{gj}). \tag{64}$$

We can then differentiate expression Eq. [61] to get

$$\frac{\partial \hat{A}}{\partial \mathbf{W}^{\text{pot/dep}}_{gh}} = 2 r f^{\text{pot/dep}} \boldsymbol{\pi}_g \left[ \sum_i a_i (\overline{\mathbf{T}}_{gi} - \overline{\mathbf{T}}_{hi}) + c_g(\eta^+_g - \eta^+_h) \right] \pm 2\boldsymbol{\pi}_g(\eta^+_g - \eta^+_h). \tag{65}$$

where $a_i$ and $c_k$ were defined in Eq. [62].

It is sometimes useful to consider the following derivatives:

$$\frac{\partial}{\partial \mathbf{W}^{\text{F}}_{gh}} \equiv \frac{\partial}{\partial \mathbf{W}^{\text{pot}}_{gh}} + \frac{\partial}{\partial \mathbf{W}^{\text{dep}}_{gh}}, \qquad \frac{\partial}{\partial \mathbf{q}_{gh}} \equiv f^{\text{dep}} \frac{\partial}{\partial \mathbf{W}^{\text{pot}}_{gh}} - f^{\text{pot}} \frac{\partial}{\partial \mathbf{W}^{\text{dep}}_{gh}}. \tag{66}$$

Each of these derivatives behaves as their names suggest:

$$\frac{\partial \mathbf{W}^{\text{F}}_{ij}}{\partial \mathbf{W}^{\text{F}}_{gh}} = \frac{\partial \mathbf{q}_{ij}}{\partial \mathbf{q}_{gh}} = \delta_{gi}(\delta_{hj} - \delta_{ij}), \qquad \frac{\partial \mathbf{q}_{ij}}{\partial \mathbf{W}^{\text{F}}_{gh}} = \frac{\partial \mathbf{W}^{\text{F}}_{ij}}{\partial q_{gh}} = 0. \tag{67}$$

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

***B.1. Scaling mode.*** Consider the following differential operator:

$$\Delta \equiv \sum_{g,h} \mathbf{W}^{\text{pot}}_{gh} \frac{\partial}{\partial \mathbf{W}^{\text{pot}}_{gh}} + \mathbf{W}^{\text{dep}}_{gh} \frac{\partial}{\partial \mathbf{W}^{\text{dep}}_{gh}}. \tag{68}$$

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:

$$\begin{aligned} \Delta \mathbf{Z} &= \tau \mathbf{e} \boldsymbol{\pi} - \mathbf{Z}, & \Delta \boldsymbol{\pi} &= 0, & \Delta \mathbf{T} &= -\mathbf{T}, \\ \Delta \eta^\pm_i &= -\eta^\pm_i, & \Delta \mathbf{q}_{ij} &= \mathbf{q}_{ij}, & \Delta \hat{A} &= 0, \end{aligned} \tag{69}$$

The anomalous bit in the scaling of $\mathbf{Z}$ is due to the lack of dependence of $\boldsymbol{\xi}$ and $\tau$ 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.

**C. Kuhn-Tucker conditions.** Consider the Lagrangian

$$\mathcal{L} = \hat{A} + \sum_{\text{pot/dep}} \sum_{i \neq j} \mu^{\text{pot/dep}}_{ij} \mathbf{W}^{\text{pot/dep}}_{ij}. \tag{70}$$

Necessary conditions for an extremum are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{pot/dep}}_{gh}} = 0, \qquad \mu^{\text{pot/dep}}_{gh} \geq 0, \quad \mathbf{W}^{\text{pot/dep}}_{gh} \geq 0, \quad \mu^{\text{pot/dep}}_{gh} \mathbf{W}^{\text{pot/dep}}_{gh} = 0. \tag{71}$$

with $g \neq h$. This enforces the positivity constraints on the off-diagonal elements, but not the diagonals. As discussed in Section B.1, that can be enforced after finding the maximum using the null scaling degree of freedom.

***C.1. Triangularity.*** Here we describe the perturbations corresponding to Fig. 3(a,b) of the main paper.

Consider

$$\frac{\partial \mathcal{L}}{\partial \mathbf{q}_{gh}} = f^{\text{dep}} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{pot}}_{gh}} - f^{\text{pot}} \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{\text{dep}}_{gh}} = (f^{\text{dep}} \mu^{\text{pot}}_{gh} - f^{\text{pot}} \mu^{\text{dep}}_{gh}) + 2\boldsymbol{\pi}_g(\eta^+_g - \eta^+_h) = 0. \tag{72}$$

This corresponds to the shift

$$\mathbf{W}^{\text{pot}}_{ij} \to \mathbf{W}^{\text{pot}}_{ij} + f^{\text{dep}} \epsilon_{ij}, \qquad \mathbf{W}^{\text{dep}}_{ij} \to \mathbf{W}^{\text{dep}}_{ij} - f^{\text{pot}} \epsilon_{ij}, \qquad \sum_j \epsilon_{ij} = 0, \tag{73}$$

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

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

$$f^{\text{dep}}\mu^{\text{pot}}_{gh} - f^{\text{pot}}\mu^{\text{dep}}_{gh} < 0 \qquad \Longrightarrow \qquad \mu^{\text{dep}}_{gh} > 0 \qquad \Longrightarrow \qquad \mathbf{W}^{\text{dep}}_{gh} = 0. \tag{74}$$

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

$$f^{\text{dep}}\mu^{\text{pot}}_{gh} - f^{\text{pot}}\mu^{\text{dep}}_{gh} > 0 \qquad \Longrightarrow \qquad \mu^{\text{pot}}_{gh} > 0 \qquad \Longrightarrow \qquad \mathbf{W}^{\text{pot}}_{gh} = 0. \tag{75}$$

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

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

***C.2. Shortcuts.*** In this subsection we will define perturbations corresponding to Fig. 3(c) of the main paper.

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

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m} \equiv \left[\sum_{k=0}^{m-1} \frac{1}{\boldsymbol{\pi}_{g\pm k}} \left(\frac{\partial}{\partial \mathbf{W}^{\text{pot/dep}}_{g\pm k, g\pm(k+1)}}\right)\right] - \frac{1}{\boldsymbol{\pi}_g}\left(\frac{\partial}{\partial \mathbf{W}^{\text{pot/dep}}_{g,g\pm m}}\right). \tag{76}$$

Once again, they are only well defined if all the states have non-zero equilibrium probabilities (see the comment in Section C.1 about this being satisfied for ergodic chains).

One can show that the equilibrium probabilities, $\boldsymbol{\pi}$, are invariant under these operators Eq. [21]:

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m}\boldsymbol{\pi}_i = 0, \tag{77}$$

which makes it possible to integrate the perturbation:

$$\mathbf{W}^{\text{pot/dep}} \to \mathbf{W}^{\text{pot/dep}} + \mathbf{D}\boldsymbol{\epsilon}^{\pm(g,m)}, \qquad \begin{aligned} \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g,g\pm m} &= -\epsilon, \\ \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g\pm k,g\pm(k+1)} &= \epsilon \qquad \forall\, k \in [0, m-1], \\ \left(\boldsymbol{\epsilon}^{\pm(g,m)}\right)_{g\pm k,g\pm k} &= -\epsilon \qquad \forall\, k \in [1, m-1]. \end{aligned} \tag{78}$$

But more interestingly for our purposes:

$$\widetilde{\Delta}^{\text{pot/dep}}_{g,m}\mathcal{L} = \left[\sum_{k=0}^{m-1} \frac{\mu^{\text{pot/dep}}_{g\pm k,g\pm(k+1)}}{\boldsymbol{\pi}_{g\pm k}} - \frac{\mu^{\text{pot/dep}}_{g,g\pm m}}{\boldsymbol{\pi}_g}\right] + 2rf^{\text{pot/dep}}\sum_{k=0}^{m-1}\left(\eta^+_{g\pm k} - \eta^+_{g\pm(k+1)}\right)(c_{g\pm k} - c_g)\,, \tag{79}$$

In the section below, we will show that the $c_k$ are non-decreasing, if we put the states in order of decreasing $\eta^+_k$. This implies that the last term of the final expression in Eq. [79] 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.

***C.3. Increasing $c_k$.*** In the previous section we defined perturbations corresponding to Fig. 3(c) of the main paper. 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 $\eta^+_k$.

Consider the following combinations of derivatives:

$$\Delta_{gh} \equiv \frac{1}{\boldsymbol{\pi}_g}\left(\frac{\partial}{\partial \mathbf{W}^{\text{F}}_{gh}}\right) + \frac{1}{\boldsymbol{\pi}_h}\left(\frac{\partial}{\partial \mathbf{W}^{\text{F}}_{hg}}\right), \tag{80}$$

$$\tag{81}$$

Note that they are only well defined if all the states have non-zero equilibrium probabilities (see the comment in Section C.1 about this being satisfied for ergodic chains).

One can show that the equilibrium probabilities, $\boldsymbol{\pi}$, are invariant under these operators using Eq. [21]:

$$\Delta_{gh}\boldsymbol{\pi}_i = 0, \tag{82}$$

which makes it possible to integrate the perturbation:

$$\mathbf{W}^{\text{pot/dep}} \to \mathbf{W}^{\text{pot/dep}} + \mathbf{D}\boldsymbol{\epsilon}, \qquad \begin{aligned} \boldsymbol{\epsilon} &= \boldsymbol{\epsilon}^{\text{T}}, \\ \boldsymbol{\epsilon}\mathbf{e} &= 0. \end{aligned} \tag{83}$$

But more interestingly:

$$\Delta_{gh}\mathcal{L} = \frac{\mu^{\text{pot}}_{gh} + \mu^{\text{dep}}_{gh}}{\boldsymbol{\pi}_g} + \frac{\mu^{\text{pot}}_{hg} + \mu^{\text{dep}}_{hg}}{\boldsymbol{\pi}_h} + 2r\,(c_g - c_h)\left(\eta^+_g - \eta^+_h\right), \tag{84}$$

$$\tag{85}$$

where $c_k$ were defined in Eq. [62].

Using the non-negativity of the Kuhn-Tucker multipliers, $\mu_{ij}^{\text{pot/dep}}$, Eq. [84] tells us that if we arrange the states in order of decreasing $\eta_i^+$, the optimal process will have non-decreasing $c_k$ (if any of the $\eta_k^+$ are degenerate, we can choose their order to ensure this).

Note that, according to Section C.1, either $\mathbf{W}_{gh}^{\text{pot}}$ or $\mathbf{W}_{gh}^{\text{dep}}$ will be zero at the maximum, therefore we can expect one of $\mu_{gh}^{\text{pot}} + \mu_{gh}^{\text{dep}}$ to be non-zero. This would rule out degeneracy of the $c_k$ or $\eta_k^+$. Looking at Eq. [72] closely, the only way $\mu_{gh}^{\text{pot}} + \mu_{gh}^{\text{dep}}$ could be zero is if $\eta_g^+ = \eta_h^+$ or $\boldsymbol{\pi}_g = 0$.

***C.4. Summary.*** Using the Kuhn-Tucker formalism, we have shown that, with the states arranged in order of non-increasing $\eta_i^+$:

- 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 $\eta_k^+$ are also degenerate.
- If the $c_k$ increase and the $\eta_i^+$ 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 Section D, or there is some degeneracy, which would allow shortcuts provided that they do not bypass an entire degenerate set (see Eq. [79]).

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 Section G). 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 Section C), we would need only consider the multi-state topology.

**D. 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 (8–10)):

$$\mathbf{W}_{ij}^{\text{pot}} = q_i^{\text{pot}}(\delta_{i+1,j} - \delta_{ij}), \qquad \mathbf{W}_{ij}^{\text{dep}} = q_j^{\text{dep}}(\delta_{i,j+1} - \delta_{ij}). \tag{86}$$

Because it has no shortcuts, it saturates various inequalities:

$$\begin{aligned} \overline{\mathbf{T}}_{ik} - \overline{\mathbf{T}}_{jk} &= \begin{cases} \overline{\mathbf{T}}_{ij}, & \text{if} \quad i \leq j \leq k \quad \text{or} \quad i \geq j \geq k, \\ -\overline{\mathbf{T}}_{ji}, & \text{if} \quad j \leq i \leq k \quad \text{or} \quad j \geq i \geq k, \end{cases} \\ r\boldsymbol{\pi}_i \mathbf{W}_{ij}^{\text{F}} \left(\overline{\mathbf{T}}_{ij} + \overline{\mathbf{T}}_{ji}\right) &= 1 \quad \text{if} \quad i = j \pm 1, \end{aligned} \tag{87}$$

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

$$f^{\text{pot}} q_i^{\text{pot}} \boldsymbol{\pi}_i = f^{\text{dep}} q_i^{\text{dep}} \boldsymbol{\pi}_{i+1}, \tag{88}$$

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

This allows us to calculate the $c_k$'s:

$$\begin{aligned} c_k &= \sum_{i<k} \mathbf{T}_{i,i+1} \left(\boldsymbol{\pi}_i q_i^{\text{pot}} + \boldsymbol{\pi}_{i+1} q_i^{\text{dep}}\right) - \sum_{i \geq k} \mathbf{T}_{i+1,i} \left(\boldsymbol{\pi}_i q_i^{\text{pot}} + \boldsymbol{\pi}_{i+1} q_i^{\text{dep}}\right), \\ c_{k+1} - c_k &= \left(\mathbf{T}_{k,k+1} + \mathbf{T}_{k+1,k}\right) \left( \frac{\boldsymbol{\pi}_k \mathbf{W}_{k,k+1}^{\text{F}}}{f^{\text{pot}}} + \frac{\boldsymbol{\pi}_{k+1} \mathbf{W}_{k+1,k}^{\text{F}}}{f^{\text{dep}}} \right) = \frac{1}{r f^{\text{pot}} f^{\text{dep}}}, \\ \sum_k c_k \boldsymbol{\pi}_k &= \sum_{ij} \boldsymbol{\pi}_i \mathbf{q}_{ij} (\eta - \eta) = 0, \\ \implies c_k &= \frac{k - \sum_j j \boldsymbol{\pi}_j}{r f^{\text{pot}} f^{\text{dep}}}, \end{aligned} \tag{89}$$

where we used Eq. [87] to derive the first two equations respectively and Theorem 3 to derive the third. This allows us to write the area as

$$A = \frac{2\sqrt{N}}{r} \sum_k \left[ k - \sum_j j \boldsymbol{\pi}_j \right] \boldsymbol{\pi}_k \mathbf{w}_k = \frac{2\sqrt{N}}{r} \sum_k \left| k - \sum_j j \boldsymbol{\pi}_j \right| \boldsymbol{\pi}_k, \tag{90}$$

where we used $\mathbf{w}_k = \text{sgn}(c_k)$, as discussed after Eq. [62]. This reproduces Eq. [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 Eq. [90] at fixed $\boldsymbol{\pi}_\pm = \sum_k \boldsymbol{\pi}_k \left(\frac{1 \pm \mathbf{w}_k}{2}\right)$. Clearly this will happen when we put all of the probability at the ends: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_-$ and $\boldsymbol{\pi}_n = \boldsymbol{\pi}_+$ are the only non-zero $\boldsymbol{\pi}_k$. This gives an area of

$$A \leq \frac{\sqrt{N}}{r}(M-1)\left(4\boldsymbol{\pi}_+\boldsymbol{\pi}_-\right). \qquad [91]$$

This is maximised at $\boldsymbol{\pi}_+ = \boldsymbol{\pi}_- = \frac{1}{2}$:

$$A \leq \frac{\sqrt{N}}{r}(M-1). \qquad [92]$$

This yields the area bound of equation Eq. [16] of the main paper.

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 (11) in a slightly different situation.

1. Kemeny J, Snell J (1960) *Finite markov chains.* (Springer).
2. 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.
3. Kemeny JG (1981) Generalization of a fundamental matrix. *Linear Algebra and its Applications* 38(0):193 – 206.
4. Yao DD (1985) First-passage-time moments of markov processes. *Journal of Applied Probability* 22(4):pp. 939–945.
5. Cho G, Meyer C (2000) Markov chain sensitivity measured by mean first passage times. *Linear Algebra and its Applications* 316(1-3):21–28.
6. Burke C, Rosenblatt M (1958) A markovian function of a markov chain. *Ann. Math. Statist.* 29(4):1112–1122.
7. Ball F, Yeo GF (1993) Lumpability and marginalisability for continuous-time markov chains. *J. Appl. Probab.* 30(3):518–528.
8. Amit DJ, Fusi S (1994) Learning in neural networks with material synapses. *Neural Comput.* 6(5):957–982.
9. Fusi S, Abbott LF (2007) Limits on the memory storage capacity of bounded synapses. *Nat. Neurosci.* 10(4):485–493.
10. Leibold C, Kempter R (2008) Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. *Cereb. Cortex* 18(1):67–77.
11. Barrett AB, van Rossum MC (2008) Optimal learning rules for discrete synapses. *PLoS Comput. Biol.* 4(11):e1000230.