Laplacian envelope

Subhaneil Lahiri

July 15, 2020

Abstract

We try to find the continuous time Markov process that has the maximal Laplace transformed signal-to-noise curve.

Contents

1 Framework **2**
1.1 Recognition memory . . . 2
1.2 Signal-to-noise ratio . . . 3
1.3 Markov models of synapses . . . 4

2 Laplace transform **7**
2.1 Fundamental matrix etc. . . . 7
2.2 Laplace transform of SNR curve . . . 8
2.3 Derivatives . . . 9

3 Upper bounds **10**
3.1 Initial SNR . . . 10
3.2 Area bound . . . 10
3.3 Envelope . . . 10

4 Finite time **12**
4.1 Shifted problem . . . 14

5 Nearly uniform serial models **15**
5.1 Uniform serial model . . . 17
5.2 Shortened serial model . . . 19
5.3 Sticky serial model . . . 21
5.4 Heuristic envelope . . . 23

6 Conclusions **24**

List of Figures

1 Proven envelope for the signal-to-noise ratio 12
2 Numerical envelope for the signal-to-noise ratio 14
3 Optimal models 16
4 Heuristic optimal models 17
5 Signal-to-noise ratio for optimal shortened models 21
6 Signal-to-noise ratio for optimal sticky models 22
7 Heuristic envelope for the signal-to-noise ratio 24

1 Framework

c:framework

1.1 Recognition memory

sec:recog

We will be trying to store patterns in a set od N synaptic weights, $\vec{w}$. Every time we try to store a pattern, these synapses are subjected to a plasticity event where each synapse is either potentiated or depressed, depending on the pattern. we will assume that these patterns are spatially and temporally independent.

At some time, suppose we wish to determine if a given pattern is one of those that we previously attempted to store. We wish to answer this question by looking at the synaptic weights directly (ideal observer). For that given pattern there will be an ideal set of synaptic weights, $\vec{w}_{\mathrm{id}}$, where those synapses that were supposed to be potentiated are maximised and those that were supposed to be depressed are minimised. Suppose that the given pattern was actually seen at time 0 and we are observing the synapses at time t. The actual set of synaptic weights we see, $\vec{w}(t)$, will be a vector of random variables that differs from $\vec{w}_{\mathrm{id}}$ due to the stochasticity of the pattern encoding and all of the other (uncorrelated) pattern that are stored after it. As $t \to \infty$, the synaptic weights will become independent of the patter stored at $t = 0$. Thus, the vector of random variables $\vec{w}(\infty)$ also describes the synaptic weights under the null hypothesis – if that given pattern had never been stored.

We can test if the pattern had been previously stored by computing $\vec{w}_{\mathrm{id}}\vec{w}$ and comparing it to some threshold [1]. For large N, this quantity will have a Gausssian distribution. There will be a ROC curve as a function of this threshold:

$$\begin{aligned}
\mathrm{TPR} = \Phi_c\left(\frac{\Phi_c^{-1}(\mathrm{FPR}) - \mathrm{SNR}}{\mathrm{NNR}}\right), \quad \text{where} \quad \Phi_c(x) &= \int_x^\infty \frac{\mathrm{e}^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}z, \\
\mathrm{SNR} &= \frac{\langle \vec{w}_{\mathrm{id}}\cdot\vec{w}(t)\rangle - \langle \vec{w}_{\mathrm{id}}\cdot\vec{w}(\infty)\rangle}{\sqrt{\mathrm{Var}(\vec{w}_{\mathrm{id}}\cdot\vec{w}(\infty))}}, \qquad (1) \\
\mathrm{NNR} &= \sqrt{\frac{\mathrm{Var}(\vec{w}_{\mathrm{id}}\cdot\vec{w}(t))}{\mathrm{Var}(\vec{w}_{\mathrm{id}}\cdot\vec{w}(\infty))}},
\end{aligned}$$

{eq:ROC}

and TPR/FPR are the true/false positive rates. When the signal-to-noise ratio, SNR is larger than $\Phi_c^{-1}(\mathrm{FPR})$, it is beneficial to decrease the noise-to-noise ratio, NNR. In the

other case, it is beneficial to increase it. The expectations and variances are over the probability distribution of the synaptic states given the sequence of plasticity events, the probability distribution of the sequence of plasticity events themselves and the probability distribution of the pattern we are testing, $\vec{w}_{\mathrm{id}}$.

There are other measures of memory one could consider, e.g. asymptotic error exponents. For the Neyman-Pearson approach to hypothesis testing, the error exponent is the KL divergence:

$$D_{\mathrm{KL}}\left(P_0 \| P_1\right) = \ln \mathrm{NNR} + \frac{1+\mathrm{SNR}^2}{2\,\mathrm{NNR}^2} - \frac{1}{2}.$$

For Bayesian hypothesis testing, the error exponent is the Chernoff distance:

$$D^*\left(P_0 \| P_1\right) = \max_{\alpha} \left\{ \frac{1}{2} \ln \left[\frac{1+\alpha(\mathrm{NNR}^2-1)}{\mathrm{NNR}^{2\alpha}} \right] + \frac{\alpha(1-\alpha)\,\mathrm{SNR}^2}{2[1+\alpha(\mathrm{NNR}^2-1)]} \right\}.$$

The formulae above assumed that we know the time between storage and recognition. We should also compute the expectations and variances over probability distribution of the recall time, t, as well. If we only know average time, τ, a natural choice for this distribution is

$$P(t|\tau) = \frac{\mathrm{e}^{-t/\tau}}{\tau}. \tag{2}$$

{eq:recogtim

Different parts of the brain, that store memories for different timescales, could be characterised by different values of τ.

1.2 Signal-to-noise ratio

sec:snr

We will model the synapses as having two possible synaptic weights. As we are looking at the information contained in the synaptic weights, rather than modelling the readout via electrical activity of the neurons, what values the synaptic weights actually take is irrelevant. It will be convenient to call these two values ± 1.

As discussed in §1.1, the signal-to-noise ratio is given by

$$\mathrm{SNR}(t) = \frac{\langle \vec{w}_{\mathrm{id}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\mathrm{id}} \cdot \vec{w}(\infty) \rangle}{\sqrt{\mathrm{Var}(\vec{w}_{\mathrm{id}} \cdot \vec{w}(\infty))}}, \tag{3}$$

{eq:SNRdef}

Where the averages and variances are over the probability distribution of the synaptic states given the sequence of plasticity events, the probability distribution of the sequence of plasticity events themselves and the probability distribution of the pattern we are testing, $\vec{w}_{\mathrm{id}}$. If we also average over the time between storage and recognition, we have

$$\overline{\mathrm{SNR}}(\tau) = \int_0^\infty \mathrm{d}t\, P(t|\tau)\, \mathrm{SNR}(t) = \frac{1}{\tau} \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-t/\tau}\, \mathrm{SNR}(t), \tag{4}$$

{eq:snrb}

This is similar to the average SNR up to time τ, but with the hard cut-off replaced with a smooth exponential. This can be expressed in terms of he Laplace transform of the SNR curve in (3):

$$A(s) = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-st}\, \mathrm{SNR}(t), \qquad \overline{\mathrm{SNR}}(\tau) = \frac{A(1/\tau)}{\tau}. \tag{5}$$

{eq:laplaced

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

$$\begin{aligned}
\mathrm{Var}(\vec{w}_{\mathrm{id}}\cdot\vec{w}(t)) &= \sum_{\alpha\beta}\left\langle \vec{w}^{\alpha}_{\mathrm{id}}\vec{w}^{\alpha}(t)\vec{w}^{\beta}_{\mathrm{id}}\vec{w}^{\beta}(t)\right\rangle - \Big(\sum_{\alpha}\langle \vec{w}^{\alpha}_{\mathrm{id}}\vec{w}^{\alpha}(t)\rangle\Big)^2 \\
&= \sum_{\alpha}\left\langle (\vec{w}^{\alpha}_{\mathrm{id}})^2(\vec{w}^{\alpha}(t))^2\right\rangle + \sum_{\alpha\neq\beta}\langle \vec{w}^{\alpha}_{\mathrm{id}}\vec{w}^{\alpha}(t)\rangle\,\langle \vec{w}^{\beta}_{\mathrm{id}}\vec{w}^{\beta}(t)\rangle - \Big(\sum_{\alpha}\langle \vec{w}^{\alpha}_{\mathrm{id}}\vec{w}^{\alpha}(t)\rangle\Big)^2 \\
&= N\,\langle 1\rangle + N(N-1)\left\langle \vec{w}^{1}_{\mathrm{id}}\vec{w}^{1}(t)\right\rangle^2 - N^2\left\langle \vec{w}^{1}_{\mathrm{id}}\vec{w}^{1}(t)\right\rangle^2 \\
&= N(1-\left\langle \vec{w}^{1}_{\mathrm{id}}\vec{w}^{1}(t)\right\rangle^2),
\end{aligned} \tag{6}$$

{eq:noise}

where we used $\vec{w}^{\alpha}=\pm 1$. We can compute $\mathrm{Var}(\vec{w}_{\mathrm{id}}\cdot\vec{w}(\infty))$ by taking $t\to\infty$.

For the numerator, we can write

$$\langle \vec{w}_{\mathrm{id}}\cdot\vec{w}(t)\rangle = \sum_{\alpha}\langle \vec{w}^{\alpha}_{\mathrm{id}}\vec{w}^{\alpha}(t)\rangle = N\left\langle \vec{w}^{1}_{\mathrm{id}}\vec{w}^{1}(t)\right\rangle, \tag{7}$$

{eq:overlape

If the elements of $\vec{w}_{\mathrm{id}}$ take values ± 1 with probability $f^{\mathrm{pot/dep}}$,

$$\left\langle \vec{w}^{1}_{\mathrm{id}}\vec{w}^{1}(t)\right\rangle = f^{\mathrm{pot}}\left\langle \vec{w}^{1}(t)\right\rangle_{\mathrm{pot},t=0} - f^{\mathrm{dep}}\left\langle \vec{w}^{1}(t)\right\rangle_{\mathrm{dep},t=0} \tag{8}$$

{eq:overlapo

To compute this quantity, we will need to discuss how we model individual synapses.

1.3 Markov models of synapses

rkovsynapse

We model a synapse as having M internal states. The synaptic weight depends deterministically on the state, given by the M-element column vector $\mathbf{w}$. We denote the probability distribution across these states by the row vector $\mathbf{p}(t)$. We denote a column vector of ones by $\mathbf{e}$, so the normalisation condition is $\mathbf{p}(t)\,\mathbf{e}=1$.

When the synapse is subjected to a plasticity event the internal state will change stochastically, described by the matrices $\mathbf{M}^x n$ whose i,j elements are the transition probabilities from state i to state j, and ν labels the diffrerent types of plasticity event. Potentiating events have $\nu=$pot, depressing events have $\nu=$dep. We also use the notation

$$(-1)^{\nu} = \begin{cases} 1, & \nu=\mathrm{pot}, \\ -1, & \nu=\mathrm{dep}. \end{cases}$$

The matrices $\mathbf{M}^{\nu}$ are transition matrices of discrete time Markov processes. Their elements must satisfy the following constraints

$$\mathbf{M}^{\nu}_{ij}\geq 0 \quad \forall\,\nu,i,j, \qquad \sum_{j}\mathbf{M}^{\nu}_{ij}=1 \quad \forall\,\nu,i. \tag{9}$$

{eq:constr}

The upper bounds $\mathbf{M}^{\nu}_{ij}\leq 1$ follow automatically from these.

If we treat the off diagonal elements as the independent degrees of freedom, with the off diagonal elements determined by the row sum constraints (second column of (9)), the constraints are

$$\mathbf{M}^{\nu}_{ij} \geq 0 \quad \forall\, \nu, i, j : i \neq j, \qquad \sum_{j:j\neq i} \mathbf{M}^{\nu}_{ij} \leq 1 \quad \forall\, \nu, i. \tag{10}$$

{eq:constri}

The second of these will be referred to as the "diagonal" constraint.

Given a sequence of plasticity events at times $t_1, t_2, t_3, \ldots$, the probability distribution across the internal states at time t_n, mariginalising over states at earlier times, is

$$\mathbf{p}(t_n) = \mathbf{p}(t_0)\mathbf{M}^{\rm pot}\mathbf{M}^{\rm dep}\mathbf{M}^{\rm dep}\mathbf{M}^{\rm pot}\mathbf{M}^{\rm dep}\mathbf{M}^{\rm pot}\mathbf{M}^{\rm pot}\mathbf{M}^{\rm dep}\ldots\mathbf{M}^{\rm pot}. \tag{11}$$

{eq:plastseq

If the probability of a given plasticity event being potentiating/depressing is $f^{\rm pot/dep}$, the probability of the sequence above (conditioned on the number of events) is given by $f^{\rm pot}f^{\rm dep}f^{\rm dep}f^{\rm pot}f^{\rm dep}f^{\rm pot}f^{\rm pot}f^{\rm dep}\ldots f^{\rm pot}$.

In computing the expectation in (8) the plasticity event that occurs at $t = 0$ is fixed, but all of the previous and subsequent events are marginalised over. If we marginalise over the types (potentiating/depressing) of these other events (conditioned on the number of events) we find

$$\begin{aligned}\mathbf{p}(t_n) = \mathbf{p}(t_0)&(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})\ldots(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})\\ &\mathbf{M}^{\rm pot/dep}(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})\ldots(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep}),\end{aligned} \tag{12}$$

{eq:plastseq

where the factor of $\mathbf{M}^{\rm pot/dep}$ is the event at $t = 0$. If we expand out all of the parenthetical factors, we will get a sum of terms of the form (11) weighted by their probabilities.

If $t_0 \ll 0$, the effect of the large number of earlier events will be to put the synapse in the steady state distribution, $\boldsymbol{\pi}$, of the stochastic process $(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})$:

$$\mathbf{p}(t_n) = \boldsymbol{\pi}\mathbf{M}^{\rm pot/dep}(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})^n, \tag{13}$$

{eq:plastseq

where n is the number of plasticity events that took place.

If the plasticity events occur at Poisson rate r, the probability of n events occurring in time t is given by a Poisson distribution with mean rt. If we marginalise over n as well

$$\begin{aligned}P(\text{state} = i, t \mid \text{pot/dep}, 0) &= \sum_n \frac{(rt)^n}{n!}\mathrm{e}^{-rt}\boldsymbol{\pi}\mathbf{M}^{\rm pot/dep}(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep})^n\\ &= \left[\boldsymbol{\pi}\mathbf{M}^{\rm pot/dep}\,\mathrm{e}^{rt(f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep} - \mathbf{I})}\right]_i,\end{aligned} \tag{14}$$

{eq:plastseq

where $\boldsymbol{\pi}$ is the steady state distribution of the continuous time transition matrix that appears in the exponential:

$$\boldsymbol{\pi}\mathbf{W}^{\rm F} = 0, \qquad \text{where} \quad \mathbf{W}^{\rm F} = \sum_{\nu} f^{\nu}\mathbf{M}^{\nu} - \mathbf{I} = f^{\rm pot}\mathbf{M}^{\rm pot} + f^{\rm dep}\mathbf{M}^{\rm dep} - \mathbf{I}. \tag{15}$$

{eq:eq}

This results in

$$\begin{aligned}
\left\langle \vec{w}^1_{\mathrm{id}} \vec{w}^1(t) \right\rangle &= \boldsymbol{\pi}(f^{\mathrm{pot}}\mathbf{M}^{\mathrm{pot}} - f^{\mathrm{dep}}\mathbf{M}^{\mathrm{dep}})\, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}, \\
\left\langle \vec{w}^1_{\mathrm{id}} \vec{w}^1(\infty) \right\rangle &= \boldsymbol{\pi}(f^{\mathrm{pot}}\mathbf{M}^{\mathrm{pot}} - f^{\mathrm{dep}}\mathbf{M}^{\mathrm{dep}})\, \mathbf{e}\boldsymbol{\pi}\mathbf{w} \\
&= \boldsymbol{\pi}(f^{\mathrm{pot}}\mathbf{e} - f^{\mathrm{dep}}\mathbf{e})\, \boldsymbol{\pi}\mathbf{w} \\
&= (f^{\mathrm{pot}} - f^{\mathrm{dep}})\, \boldsymbol{\pi}\mathbf{w} \\
&= (f^{\mathrm{pot}} - f^{\mathrm{dep}})\, \boldsymbol{\pi}\mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}.
\end{aligned} \tag{16}$$

{eq:overlapw

Combining these allows us to write the numerator of (3) as

$$\begin{aligned}
&\langle \vec{w}_{\mathrm{id}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\mathrm{id}} \cdot \vec{w}(\infty) \rangle = N\boldsymbol{\pi}\mathbf{K}\, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}, \\
&\qquad \text{where} \qquad \mathbf{K} = \sum_{\nu} (-1)^{\nu} f^{\nu}(\mathbf{M}^{\nu} - \mathbf{I}) = f^{\mathrm{pot}}(\mathbf{M}^{\mathrm{pot}} - \mathbf{I}) - f^{\mathrm{dep}}(\mathbf{M}^{\mathrm{dep}} - \mathbf{I})
\end{aligned} \tag{17}$$

{eq:signal}

Combining with (6) gives

$$\mathrm{SNR}(t) = \frac{\sqrt{N}\boldsymbol{\pi}\mathbf{K}\, \mathrm{e}^{rt\mathbf{W}^{\mathrm{F}}}\mathbf{w}}{\sqrt{1 - (f^{\mathrm{pot}} - f^{\mathrm{dep}})^2(\boldsymbol{\pi}\mathbf{w})^2}}. \tag{18}$$

{eq:SNRcurve

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 perform the maximisation in two steps. First maximise the numerator at fixed $\boldsymbol{\pi}\mathbf{w}$. Then maximise the ratio wrt. $\boldsymbol{\pi}\mathbf{w}$. This will allow us to ignore the denominator until the very end.

We can then treat the off-diagonal elements of $\mathbf{M}^{\nu}$ as the independent degrees of freedom, subject to these constraints, when maximising the SNR. The diagonal elements are determined by setting the row sums to 1. We then have

$$\frac{\partial \mathbf{M}^{\mu}_{ij}}{\partial \mathbf{M}^{\nu}_{mn}} = \delta_{\mu\nu}\, \delta_{im}(\delta_{jn} - \delta_{jm}), \tag{19}$$

{eq:derivpd}

It will also be convenient to define the differential operators

$$\frac{\partial}{\partial \mathbf{W}^{\mathrm{F}}_{ij}} = \frac{1}{2f^{\mathrm{pot}}}\frac{\partial}{\partial \mathbf{M}^{\mathrm{pot}}_{ij}} + \frac{1}{2f^{\mathrm{dep}}}\frac{\partial}{\partial \mathbf{M}^{\mathrm{dep}}_{ij}}, \qquad \frac{\partial}{\partial \mathbf{K}_{ij}} = \frac{1}{2f^{\mathrm{pot}}}\frac{\partial}{\partial \mathbf{M}^{\mathrm{pot}}_{ij}} - \frac{1}{2f^{\mathrm{dep}}}\frac{\partial}{\partial \mathbf{M}^{\mathrm{dep}}_{ij}}, \tag{20}$$

{eq:pertfe}

these satisfy

$$\frac{\partial \mathbf{W}^{\mathrm{F}}_{ij}}{\partial \mathbf{W}^{\mathrm{F}}_{mn}} = \frac{\partial \mathbf{K}_{ij}}{\partial \mathbf{K}_{mn}} = \delta_{im}(\delta_{jn} - \delta_{jm}), \qquad \frac{\partial \mathbf{W}^{\mathrm{F}}_{ij}}{\partial \mathbf{K}_{mn}} = \frac{\partial \mathbf{K}_{ij}}{\partial \mathbf{W}^{\mathrm{F}}_{mn}} = 0. \tag{21}$$

{eq:derivfe}

If we wish to use the off-diagonal elements of $\mathbf{W}^{\mathrm{F}}$ and $\mathbf{K}$ as the degrees of freedom, the constraints are

$$\begin{aligned}
&\mathbf{W}^{\mathrm{F}}_{i\neq j} \geq |\mathbf{K}_{i\neq j}|, \qquad \sum_{j\neq i} \mathbf{W}^{\mathrm{F}}_{ij} + (-1)^{\mu}\mathbf{K}_{ij} \leq f^{\mu}, \\
&\sum_{j} \mathbf{W}^{\mathrm{F}}_{ij} = 0, \qquad \sum_{j} \mathbf{K}_{ij} = 0.
\end{aligned} \tag{22}$$

{eq:constrfe

2 Laplace transform

sec:laplace

2.1 Fundamental matrix etc.

sec:lfund

In analogy to the generalised fundamental matrix of a Markov chain [2], define

$$\mathbf{Z}(s) = (s\mathbf{I} + \mathbf{e}\boldsymbol{\xi} - \mathbf{Q})^{-1}, \tag{23}$$

{eq:lfund}

where $\boldsymbol{\xi}$ is an arbitrary row vector satisfying $\boldsymbol{\xi}\mathbf{e} = \frac{1}{\tau_0} \neq 0$. This reduces to the fundamental matrix at $s = 0$. We can see that $\mathbf{e}$ is an eigenvector:

$$\mathbf{Z}(s)\mathbf{e} = \frac{\tau_0}{1 + s\tau_0}\mathbf{e}. \tag{24}$$

{eq:lfundrow

Suppose we have some row vector $\mathbf{a}$ such that $\mathbf{ae} = 0$. Then $\mathbf{aZ}(s)$ is independent of $\boldsymbol{\xi}$:

$$\frac{\partial(\mathbf{aZ}(s))_i}{\partial\boldsymbol{\xi}_k} = -(\mathbf{aZ}(s)\mathbf{e})\mathbf{Z}_{ki}(s) = -\frac{\tau_0}{1 + s\tau_0}(\mathbf{ae})\mathbf{Z}_{ki}(s) = 0. \tag{25}$$

{eq:alfund}

Then we also define

$$\overline{\mathbf{T}}(s) = (\mathbf{EZ}^{\mathrm{dg}}(s) - \mathbf{Z}(s))\mathbf{D}, \qquad \overline{\mathbf{T}}_{ij}(s) = \frac{\mathbf{Z}_{jj}(s) - \mathbf{Z}_{ij}(s)}{\boldsymbol{\pi}_j}. \tag{26}$$

{eq:lfptb}

This reduces to the mean first-passage time matrix at $s = 0$.

At other values of s can they be interpreted as the mean first-passage times of a different process? In order to have eq. (23) as its fundamental matrix, the extra bit must draw probability from each state at a rate s. In other words, it should have the form $s(\mathbf{e}\boldsymbol{\zeta} - \mathbf{I})$ for some $\boldsymbol{\zeta}$ (which can be absorbed into $\boldsymbol{\xi}$ in eq. (23)). This other process must have the same steady-state distribution, so that its mean first-passage times are given by eq. (26), so the probability drawn from each state must be redistributed proportional to $\boldsymbol{\pi}$. This other stochastic process is

$$\widehat{\mathbf{Q}}(s) = \mathbf{Q} + s(\mathbf{e}\boldsymbol{\pi} - \mathbf{I}). \tag{27}$$

{eq:lshift}

This is a valid continuous-time Markov process. Its fundamental matrix is given by eq. (23) with $\boldsymbol{\xi}_0 = \boldsymbol{\xi} + s\boldsymbol{\pi}$.

This means that the $\overline{\mathbf{T}}(s)$ have all the same properties as in the case $s = 0$. I particular, we see that the quantity

$$\eta(s) = \sum_j \overline{\mathbf{T}}_{ij}(s)\boldsymbol{\pi}_j \tag{28}$$

{eq:lkemeny}

is independent of the starting state i, but now $\eta(s) = \operatorname{tr}\mathbf{Z}(s) - \frac{\tau_0}{1+s\tau_0}$. We can then define

$$\eta_i^{\pm}(s) = \sum_j \overline{\mathbf{T}}_{ij}(s)\boldsymbol{\pi}_j\left(\frac{1 \pm \mathbf{w}_j}{2}\right) = \sum_{j\in\pm} \overline{\mathbf{T}}_{ij}(s)\boldsymbol{\pi}_j. \tag{29}$$

{eq:lpmixtim

We can arrange the states in order of decreasing $\eta_i^+(s)$ or increasing $\eta_i^-(s)$. It will be more convenient to use the quantities

$$\boldsymbol{\eta}_i^w(s) = \sum_j \overline{\mathbf{T}}_{ij}(s)\boldsymbol{\pi}_j\mathbf{w}_j = \eta_i^+(s) - \eta_i^-(s) = 2\eta_i^+(s) - \eta(s) = \eta(s) - 2\eta_i^-(s). \tag{30}$$

{eq:lwpmixti

Arranging the states in order of decreasing $\boldsymbol{\eta}_i^w(s)$ is the same as the order of decreasing $\eta_i^+(s)$ or increasing $\eta_i^-(s)$. The $\boldsymbol{\eta}_i^w(s)$ can still be used when $\mathbf{w}$ takes more than two values.

2.2 Laplace transform of SNR curve

laplaceSNR

Consider the Laplace transform of the evolution operator:

$$\mathbf{G}(s) = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{(\mathbf{Q}-s)t}. \tag{31}$$

{eq:levol}

For $\mathfrak{Re}\, s > 0$, we have

$$(s-\mathbf{Q})\mathbf{G}(s) = \int_0^\infty \mathrm{d}t\, (s-\mathbf{Q})\mathrm{e}^{(\mathbf{Q}-s)t} = \left[-\mathrm{e}^{(\mathbf{Q}-s)t}\right]_0^\infty = \mathbf{I}. \tag{32}$$

{eq:levolcal

As $(s-\mathbf{Q})$ is invertible for $\mathfrak{Re}\, s > 0$, because the real part of all eigenvalues of $\mathbf{Q}$ are nonpositive, we have

$$\mathbf{G}(s) = (s-\mathbf{Q})^{-1}. \tag{33}$$

{eq:levolres

For $s = 0$, we can avoid problems by replacing $\mathbf{G}(s) \to \mathbf{Z}(s)$.

Now consider the Laplace transform of the SNR curve (18)

$$\begin{aligned} A(s) = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-st}\, \mathrm{SNR}(t) &= \int_0^\infty \mathrm{d}t\, \frac{\sqrt{N}\boldsymbol{\pi}\mathbf{K}\mathrm{e}^{(r\mathbf{W}^{\mathrm{F}}-s)t}\mathbf{w}}{\sqrt{1-(f^{\mathrm{pot}}-f^{\mathrm{dep}})^2(\boldsymbol{\pi}\mathbf{w})^2}} \\ &= \frac{\sqrt{N}\boldsymbol{\pi}\mathbf{K}\left(s\mathbf{I}-r\mathbf{W}^{\mathrm{F}}\right)^{-1}\mathbf{w}}{\sqrt{1-(f^{\mathrm{pot}}-f^{\mathrm{dep}})^2(\boldsymbol{\pi}\mathbf{w})^2}}. \end{aligned} \tag{34}$$

{eq:lsnr}

This expression is ill-behaved at $s = 0$. Thanks to (25), we can solve this by the replacement $\mathbf{G}(s) \to \mathbf{Z}(s)$, as $\mathbf{K}\mathbf{e} = 0$.

$$\begin{aligned} \hat{A}(s) = A(s)\sqrt{\frac{1-(f^{\mathrm{pot}}-f^{\mathrm{dep}})^2(\boldsymbol{\pi}\mathbf{w})^2}{N}} &= \boldsymbol{\pi}\mathbf{K}\mathbf{Z}(s)\mathbf{w} \\ &= \sum_{ijk} \boldsymbol{\pi}_i\mathbf{K}_{ij}\left(\overline{\mathbf{T}}_{ik}(s) - \overline{\mathbf{T}}_{jk}(s)\right)\boldsymbol{\pi}_k\mathbf{w}_k \\ &= \sum_{ij} \boldsymbol{\pi}_i\mathbf{K}_{ij}(\boldsymbol{\eta}_i^w(s) - \boldsymbol{\eta}_j^w(s)). \end{aligned} \tag{35}$$

{eq:larea}

Note that $A(0) = A$, the area, and $\lim_{s\to\infty}\{sA(s)\} = \mathrm{SNR}(0)$, the initial SNR.

We can also express this in terms of the shifted process in eq. (35). We can choose how to distribute the shift across $\mathbf{M}^{\text{pot}}$ and $\mathbf{M}^{\text{dep}}$. Two choices from a continuum[1] are:

$$\widehat{\mathbf{M}}^{\nu}(s) = \mathbf{M}^{\nu} + \frac{s}{2rf^{\nu}}(\mathbf{e}\boldsymbol{\pi} - \mathbf{I}) \qquad \text{or} \qquad \widehat{\mathbf{M}}^{\nu}(s) = \mathbf{M}^{\nu} + \frac{s}{r}(\mathbf{e}\boldsymbol{\pi} - \mathbf{I}). \tag{36}$$

{eq:lshiftpd

The first choice leaves $\widehat{\mathbf{K}} = \mathbf{K}$, so it only affects the $\boldsymbol{\eta}^{w}$ piece of eq. (35). The second choice does not leave $\mathbf{K}$ invariant (in fact $\widehat{\mathbf{K}}(s) = \mathbf{K} + (f^{\text{pot}} - d^{\text{dep}})(\mathbf{e}\boldsymbol{\pi} - \mathbf{I})$), but because $\boldsymbol{\pi}(\mathbf{e}\boldsymbol{\pi} - \mathbf{I}) = 0$ the shift does not make any contribution to eq. (35). Either way:

$$\hat{A}\left(s; \mathbf{M}^{\nu}\right) = \hat{A}\left(0; \widehat{\mathbf{M}}^{\nu}(s)\right). \tag{37}$$

{eq:lareashi

In other words, the laplace transform of the SNR is the area under the SNR for the shifted process.

Note that while eq. (27) produces a valid continuous-time Markov process, eq. (36) does not always produce valid discrete-time Markov processes, as the diagonal elements can become negative. However, $r(\widehat{\mathbf{M}}^{\nu} - \mathbf{I})$ is a valid continuous-time Markov process.

In analogy with the case $s = 0$, we define

$$\begin{aligned}
\mathbf{c}_k(s) &= \sum_{ij} \boldsymbol{\pi}_i \mathbf{K}_{ij} \left(\overline{\mathbf{T}}_{ik}(s) - \overline{\mathbf{T}}_{jk}(s)\right), \\
\mathbf{a}_j(s) &= \sum_{k} \mathbf{K}_{jk}(\boldsymbol{\eta}^{w}_{j}(s) - \boldsymbol{\eta}^{w}_{k}(s)), \qquad \Longrightarrow \hat{A}(s) = \sum_{k} \boldsymbol{\pi}_k \mathbf{c}_k(s) \mathbf{w}_k = \boldsymbol{\pi}\mathbf{a}(s). \\
\boldsymbol{\theta}_i(s) &= \sum_{j} \overline{\mathbf{T}}_{ij}(0)\, \boldsymbol{\pi}_j \mathbf{a}_j(s),
\end{aligned} \tag{38}$$

{eq:lareacoe

2.3 Derivatives

sec:lderiv

Using the same approach as the area bound in [3] we find that the derivatives are

$$\frac{\partial \hat{A}(s)}{\partial \mathbf{M}^{\mu}_{mn}} = f^{\mu}\boldsymbol{\pi}_m \left(r\left[\boldsymbol{\theta}_m(s) - \boldsymbol{\theta}_n(s)\right] + \left(r\mathbf{c}_m(s) + (-1)^{\mu}\right)\left[\boldsymbol{\eta}^{w}_{m}(s) - \boldsymbol{\eta}^{w}_{n}(s)\right]\right). \tag{39}$$

{eq:lAdiff}

The term with $\Theta(s)$ comes from differentiating $\boldsymbol{\pi}$, the term with $c(s)$ comes from differentiating $\boldsymbol{\eta}^{w}(s)$ and the term with ± 1 comes from differentiating $\mathbf{K}$ in (35).

To write down the Hessian, we will first define some notation. We define the matrix $\mathbf{X}(s) = \mathbf{Z}(0)\mathbf{K}\mathbf{Z}(s)$. Given a four-index array H_{ijkl}, we define $H^{\mathrm{T}}_{ijkl} = H_{klij}$. Then we denote

$$\begin{aligned}
R_{ijkl} = \frac{\partial^2 \hat{A}}{\partial \mathbf{W}^{\mathrm{F}}_{ij} \circ \partial \mathbf{W}^{\mathrm{F}}_{kl}} &= \boldsymbol{\pi}_i \left\{\mathbf{Z}_{ik}(0) - \mathbf{Z}_{jk}(0)\right\}\left\{\boldsymbol{\theta}_k(s) - \boldsymbol{\theta}_l(s)\right\} \\
&\quad + \boldsymbol{\pi}_i \left\{\mathbf{X}_{ik}(s) - \mathbf{X}_{jk}(s)\right\}\left\{\boldsymbol{\eta}^{w}_{k}(s) - \boldsymbol{\eta}^{w}_{l}(s)\right\} \\
&\quad + \boldsymbol{\pi}_i \mathbf{c}_i(s) \left\{\mathbf{Z}_{ik}(s) - \mathbf{Z}_{jk}(s)\right\}\left\{\boldsymbol{\eta}^{w}_{k}(s) - \boldsymbol{\eta}^{w}_{l}(s)\right\}, \\
S_{ijkl} = \frac{\partial^2 \hat{A}}{\partial \mathbf{W}^{\mathrm{F}}_{ij} \partial \mathbf{K}_{kl}} &= \boldsymbol{\pi}_i \left\{\mathbf{Z}_{ik}(0) - \mathbf{Z}_{jk}(0)\right\}\left\{\boldsymbol{\eta}^{w}_{k}(s) - \boldsymbol{\eta}^{w}_{l}(s)\right\} \\
&\quad + \boldsymbol{\pi}_k \left\{\mathbf{Z}_{ki}(s) - \mathbf{Z}_{li}(s)\right\}\left\{\boldsymbol{\eta}^{w}_{i}(s) - \boldsymbol{\eta}^{w}_{j}(s)\right\},
\end{aligned} \tag{40}$$

{eq:hesspart

[1] We could choose $\widehat{\mathbf{M}}^{\nu}(s) = \mathbf{M}^{\nu} + \frac{sg^{\nu}}{rf^{\nu}}(\mathbf{e}\boldsymbol{\pi} - \mathbf{I})$, for any g^{ν} with $\sum_{\nu} g^{\nu} = 1$. The two possibilities in eq. (36) come from $g^{\nu} = \frac{1}{2}$ and $g^{\nu} = f^{\nu}$ respectively.

where R contains only those terms where $\mathbf{W}^{\mathrm{F}}_{ij}$ precedes $\mathbf{W}^{\mathrm{F}}_{kl}$ in the matrix products from the expression for $\hat{A}$ in eq. (35). The other terms sum to R^{T}.

Then we can write the Hessian as

$$\frac{\partial^2 \hat{A}(s)}{\partial \mathbf{M}^{\mu}_{ij} \partial \mathbf{M}^{\nu}_{kl}} = f^{\mu} f^{\nu} [R + R^{\mathrm{T}} + (-1)^{\mu} S^{\mathrm{T}} + (-1)^{\nu} S]_{ijkl}. \tag{41}$$

{eq:hessian}

This is useful in numerical maximisation.

3 Upper bounds

:upperbnds

3.1 Initial SNR

ec:initial

The initial SNR is:

$$\mathrm{SNR}(0) = \overline{\mathrm{SNR}}(0) = \lim_{s \to \infty} s A(s). \tag{42}$$

{eq:initdef}

As shown in [3], this satisfies the inequality

$$\mathrm{SNR}(0) \leq \sqrt{N}, \tag{43}$$

{eq:initbnd}

The model that saturates this bound is (equivalent to) a two-state model with deterministic transitions and $f^{\mathrm{pot/dep}} = \frac{1}{2}$.

3.2 Area bound

sec:area

The area under the SNR curve is:

$$\int_0^{\infty} \mathrm{d}t \; \mathrm{SNR}(t) = \lim_{\tau \to \infty} \tau \overline{\mathrm{SNR}}(\tau) = A(0). \tag{44}$$

{eq:areadef}

As shown in [3], this satisfies the inequality

$$A(0) \leq \frac{\sqrt{N}(M-1)}{r}. \tag{45}$$

{eq:areabnd}

The model that saturates this bound is one of serial topology with $\boldsymbol{\pi}$ concentrated symmetrically at the two end states.

3.3 Envelope

ec:envelope

Now we introduce an eigenvector decomposition for $-\mathbf{W}^{\mathrm{F}}$:

$$\mathbf{W}^{\mathrm{F}} = -\sum_a q_a \mathbf{u}^a \boldsymbol{\eta}^a, \quad \boldsymbol{\eta}^a \mathbf{u}^b = \delta_{ab}, \quad \mathbf{W}^{\mathrm{F}} \mathbf{u}^a = -q_a \mathbf{u}^a, \quad \boldsymbol{\eta}^a \mathbf{W}^{\mathrm{F}} = -q_a \boldsymbol{\eta}^a. \tag{46}$$

{eq:eigendec

Then we can write

$$\mathrm{SNR}(t) = \sqrt{N} \sum_a \mathcal{I}_a \mathrm{e}^{-rt/\tau_a}, \qquad \overline{\mathrm{SNR}}(\tau) = \sqrt{N} \sum_a \frac{\mathcal{I}_a}{1 + r\tau/\tau_a}, \tag{47}$$ {eqsnreig}

where

$$\mathcal{I}_a = (\boldsymbol{\pi}\widetilde{\mathbf{K}}\mathbf{u}^a)(\boldsymbol{\eta}^a\mathbf{w}), \qquad \tau_a = \frac{1}{q_a}. \tag{48}$$ {eq:snrcoeff

From (45) and (43), we see that they are subject to the constraints

$$\sum_a \mathcal{I}_a \tau_a \leq M - 1, \qquad \sum_a \mathcal{I}_a \leq 1. \tag{49}$$ {eq:coeffcon

No doubt there are many other constraints that these quantities must satisfy, but we can see what these ones imply for the $\overline{\mathrm{SNR}}(\tau)$ by maximising it subject to these constraints.

Consider the Lagrangian

$$\mathcal{L} = \sum_a \frac{\mathcal{I}_a}{1 + r\tau/\tau_a} + \mu_{\mathcal{I}} \left(1 - \sum_a \mathcal{I}_a\right) + \mu_{\mathcal{A}} \left((M-1) - \sum_a \mathcal{I}_a \tau_a\right). \tag{50}$$ {eq:envlagra

The Karush-Kuhn-Tucker necessary conditions for a maximum are

$$\frac{\partial \mathcal{L}}{\partial \mathcal{I}_a} = 0, \qquad \frac{\partial \mathcal{L}}{\partial \tau_a} = 0, \qquad \mu_\alpha \geq 0, \qquad \frac{\partial \mathcal{L}}{\partial \mu_\alpha} \geq 0, \qquad \mu_\alpha \frac{\partial \mathcal{L}}{\partial \mu_\alpha} = 0. \tag{51}$$ {eq:envKTcon

These are solved by

$$\begin{gathered} \mathcal{I}_1 = 1, \qquad \mathcal{I}_{a>1} = 0, \qquad \tau_1 = M - 1, \\ \mu_{\mathcal{I}} = \frac{(M-1)^2}{(r\tau + (M-1))^2}, \qquad \mu_{\mathcal{A}} = \frac{r\tau}{(r\tau + (M-1))^2}. \end{gathered} \tag{52}$$ {eq:envsol}

This leads to the envelope

$$\overline{\mathrm{SNR}}(\tau) \leq \frac{\sqrt{N}(M-1)}{r\tau + (M-1)}. \tag{53}$$ {eq:env}

This is the envelope ("memory frontier") that we can prove. Finding more constraints would lower it. Clearly there have to be more constraints, as (52) indicates that this would be achieved by the same model at all times. This model would have to saturate both the initial SNR bound (43) and the area bound (45), whereas we saw that they were achieved by different models, which perform well at short/long timescales respectively.

This envelope is shown in Figure 1.

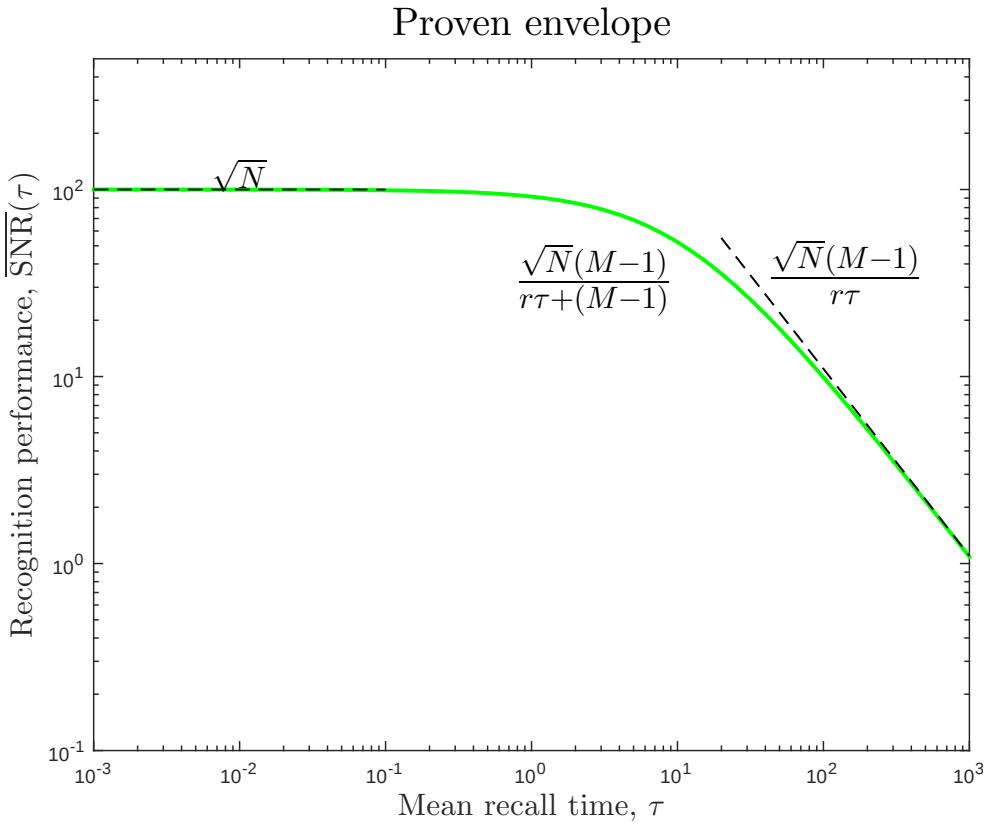

Figure 1: The memory curve envelope that we can prove, (53). The number of synapses $N = 10^4$ and the number of states $M = 12$. The dashed lines indicate the initial SNR bound, (43), and the area bound, (45), from [3].

:envproven

4 Finite time

sec:finite

Now we'll maximise $\overline{\mathrm{SNR}}(\tau)$ for some fixed τ, subject to the constraints (9). First, we will maximise the numerator of (34) holding $\boldsymbol{\pi}\mathbf{w}$ fixed. We can maximise the ration wrt. $\boldsymbol{\pi}\mathbf{w}$ later. Consider the Lagrangian

$$\mathcal{L} = \sum_{ij} \boldsymbol{\pi}_i \mathbf{K}_{ij} (\boldsymbol{\eta}_i^w(s) - \boldsymbol{\eta}_j^w(s)) + \sum_{\nu ij} \left(f^\nu \mu_{ij}^\nu \mathbf{M}_{ij}^\nu \right) \quad + \lambda \left(\Delta\mathbf{p} - \sum_i \boldsymbol{\pi}_i \mathbf{w}_i \right), \tag{54}$$

{eq:finitela

where $\Delta\mathbf{p}$ is the constant value we are holding $\boldsymbol{\pi}\mathbf{w}$ at. The Karush-Kuhn-Tucker necessary conditions for a maximum are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}_{ij}^\nu} = 0, \quad \mu_{ij}^\nu \geq 0, \quad \frac{\partial \mathcal{L}}{\partial \mu_{ij}^\nu} \geq 0, \quad \mu_{ij}^\nu \frac{\partial \mathcal{L}}{\partial \mu_{ij}^\nu} = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0. \tag{55}$$

{eq:KTfinite

We have scaled the KKT multipliers by f^ν for later convenience.

The derivatives, for $m \neq n$ are

$$\frac{1}{f^\nu}\frac{\partial \mathcal{L}}{\partial \widetilde{\mathbf{M}}^\nu_{mn}} = \boldsymbol{\pi}_m \left(r\left[\boldsymbol{\theta}_m(s) - \boldsymbol{\theta}_n(s)\right] + \left(r\mathbf{c}_m(s) + (-1)^\nu\right)\left[\boldsymbol{\eta}^w_m(s) - \boldsymbol{\eta}^w_n(s)\right]\right) \\ + \mu^\nu_{mn} - \mu^\nu_{mm} + r\lambda\boldsymbol{\pi}_m(\boldsymbol{\eta}^w_m(0) - \boldsymbol{\eta}^w_n(0)) = 0. \tag{56}$$

{eq:finitede

We can solve for the KKT multipliers by noting that

$$\mu^\nu_{ij}\mathbf{M}^\nu_{ij} = 0, \qquad \text{and} \qquad \sum_{j:j\neq i}\mu^\nu_{ii}\mathbf{M}^\nu_{ij} = \mu^\nu_{ii}(1 - \mathbf{M}^\nu_{ii}) = \mu^\nu_{ii}.$$

Then, with some new vectors defined:

$$\begin{aligned} \mu^\nu_{mm} &= -\boldsymbol{\pi}_m\left(r\boldsymbol{\alpha}^\nu_m(s) + r\lambda\boldsymbol{\omega}^\nu_i(0) + \left(r\mathbf{c}_m(s) + (-1)^\nu\right)\boldsymbol{\omega}^\nu_m(s)\right), \\ \mu^\nu_{mn} &= \boldsymbol{\pi}_m\Big(r\left[\boldsymbol{\theta}_n(s) - \boldsymbol{\theta}_m(s) - \boldsymbol{\alpha}^\nu_m(s)\right] + r\lambda\left[\boldsymbol{\eta}^w_n(0) - \boldsymbol{\eta}^w_m(0) - \boldsymbol{\omega}^\nu_m(0)\right] \\ &\qquad + \left[rc_m(s) + (-1)^\nu\right]\left[\boldsymbol{\eta}^w_n(s) - \boldsymbol{\eta}^w_m(s) - \boldsymbol{\omega}^\nu_m(s)\right]\Big), \\ &\text{where} \quad \boldsymbol{\alpha}^\nu(s) = (\mathbf{M}^\nu - \mathbf{I})\boldsymbol{\theta}(s), \qquad \boldsymbol{\omega}^\nu(s) = (\mathbf{M}^\nu - \mathbf{I})\boldsymbol{\eta}^w(s). \end{aligned} \tag{57}$$

{eq:kktsol}

The expressions in eq. (56) have the same form as for $s = 0$, which allowed us to argue that the model with maximal area must have the serial topology, [3].

The one missing ingredient is the scale invariance. Define the scale transformation:

$$\lambda * \mathbf{M}^\nu \equiv (1 - \lambda)\mathbf{I} + \lambda\mathbf{M}^\nu. \tag{58}$$

{eq:scale}

The Laplace transform has the following scaling property

$$\hat{A}\left(\lambda s; \lambda * \mathbf{M}^\nu\right) = \hat{A}\left(s; \mathbf{M}^\nu\right). \tag{59}$$

{eq:laplaces

At $s = 0$, it is scale invariant. This means that we can take a pair of matrices that violates the "diagonal" constraints (second inequality of (10)) and use the scale transform (58) to construct a pair of matrices that satisfy the "diagonal" constraints and have the same area. This allowed us to ignore the diagonal constraints. We can't do that for $s > 0$.

The corresponding symmetry here is scaling the shifted process from eq. (36), which amounts to

$$\lambda \circledast_s \mathbf{M} \equiv (1 - \lambda)\mathbf{I} + \lambda\mathbf{M}^\nu + \frac{(\lambda - 1)s}{r}(\mathbf{e}\boldsymbol{\pi} - \mathbf{I}). \tag{60}$$

{eq:shiftsca

Then we have

$$\hat{A}\left(s; \mathbf{M}^\nu\right) = \hat{A}\left(0, \widehat{\mathbf{M}}^\nu(s)\right) = \hat{A}\left(0, \lambda * \widehat{\mathbf{M}}^\nu(s)\right) = \hat{A}\left(s, \lambda \circledast_s \mathbf{M}^\nu\right). \tag{61}$$

{eq:lshiftsc

The problem here is that the lower bound, $\mathbf{M}^\nu_{ij} \geq 0$, is not invariant under this transformation, unlike the $s = 0$ case. Using this transformation to fix the upper bound can break the lower bound.

However, we do have numerical evidence that the model that maximises the Laplace transform does have the serial topology. In Figure 2 we see the result of numerical maximisation of (54). We see that restricting to the serial topology does not make the performance any worse. In fact, it slightly improves the performance, but this can be ascribed to the numerical optimisation problem being easier. Assuming that the optimiser did find the global maximum, this provides evidence for our conjecture that the model that maximises the Laplace transform must have the serial topology.

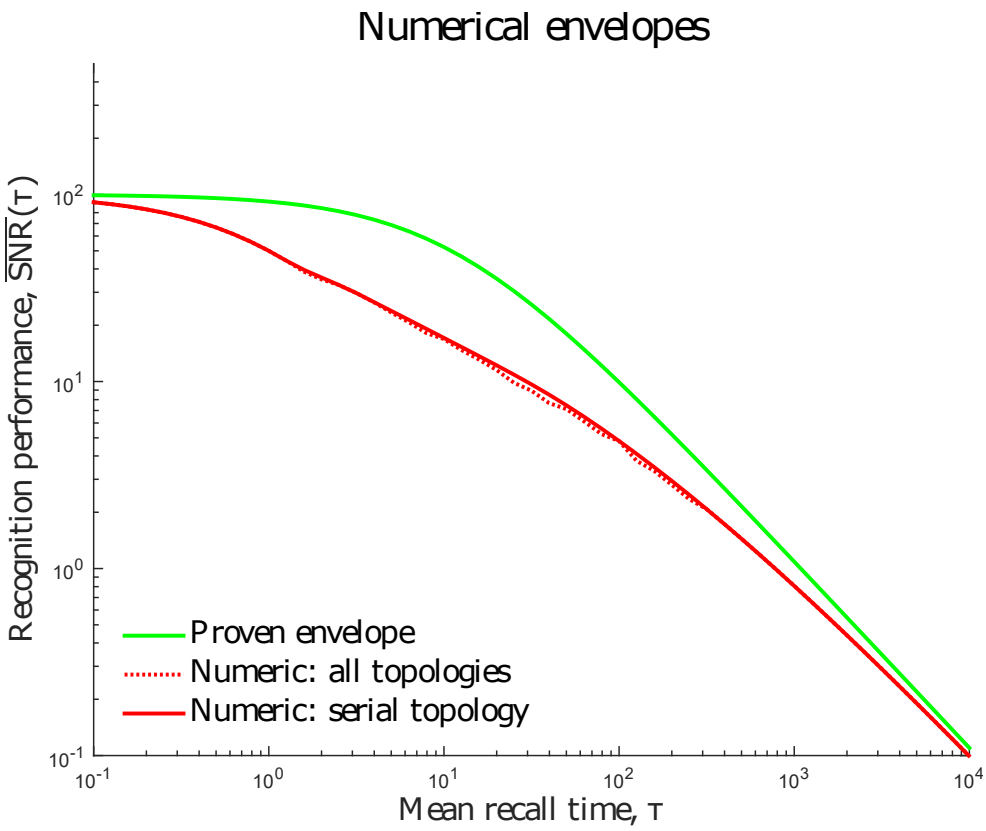

Figure 2: Numerical envelope for the signal-to-noise ratio, (4). Equations (54) to (56) are solved numericallly with $f^{\mathrm{pot/dep}} = \frac{1}{2}$, the number of synapses $N = 10^4$ and the number of states $M = 12$. The dashed red line indicates maximisation of the memory curve over the space of all matrices satisfying the constraints (9). The solid red line indicates maximisation over models with the serial topology, i.e. potentiation only moves up one state and depression only moves down one state. The solid grren line indicates the proven envelope, (53).

fig:envnum

4.1 Shifted problem

sec:lshift

The relation of the Laplace transform to the shifted process, eqs. (36) and (37) suggests an alternative approach: treat the off-diagonal components of $\widehat{\mathbf{M}}^{\nu}$ as the degrees of freedom and maximise the area instead. The constraints on $\widehat{\mathbf{M}}^{\nu}_{ij}$ are different from eq. (10):

$$\widehat{\mathbf{M}}^{\nu}_{ij} \geq \frac{s\boldsymbol{\pi}_j}{r}, \quad \forall\, \nu, i, j : i \neq j, \qquad \sum_{j:j\neq i} \widehat{\mathbf{M}}^{\nu}_{ij} \leq 1 + \frac{s}{r}(1 - \boldsymbol{\pi}_i). \quad \forall\, \nu, i. \tag{62}$$

{eq:lshiftco

The diagonal elements are still given by $\widehat{\mathbf{M}}^{\nu}_{ii} = 1 - \sum_{j:j\neq i} \widehat{\mathbf{M}}^{\nu}_{ij}$, but now they are allowed to be negative.

The Lagrangian we should consider is

$$\mathcal{L} = \sum_{ij} \boldsymbol{\pi}_i \mathbf{K}_{ij}(\boldsymbol{\eta}_i^w - \boldsymbol{\eta}_j^w) + \sum_{\nu ij} f^\nu \mu_{ij}^\nu \left(\mathbf{M}_{ij}^\nu - \frac{s}{r}(\boldsymbol{\pi}_j - \delta_{ij})\right) \quad + \lambda \left(\Delta \mathbf{p} - \sum_i \boldsymbol{\pi}_i \mathbf{w}_i\right), \tag{63}$$

{eq:lshiftla

where all of the first-passage times are evaluated at $s = 0$.

The derivatives, for $m \neq n$ are

$$\begin{aligned} \frac{1}{f^\nu} \frac{\partial \mathcal{L}}{\partial \widetilde{\mathbf{M}}_{mn}^\nu} = & \boldsymbol{\pi}_m \left(r\left[\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\right] + \left(r\mathbf{c}_m + (-1)^\nu + r\lambda\right)\left[\boldsymbol{\eta}_m^w - \boldsymbol{\eta}_n^w\right]\right) \\ & + \mu_{mn}^\nu - \mu_{mm}^\nu - \sum_{ij} \mu_{ij}^\nu s f^\nu \boldsymbol{\pi}_m (\overline{\mathbf{T}}_{mj} - \overline{\mathbf{T}}_{nj}) \boldsymbol{\pi}_j = 0. \end{aligned} \tag{64}$$

{eq:lshiftde

We can absorb the new terms by defining

$$\widehat{\boldsymbol{\theta}}^\nu(s) = \boldsymbol{\theta}(0) - \frac{s f^\nu}{r} \overline{\mathbf{T}} \boldsymbol{\Pi} (\boldsymbol{\mu}^\nu)^{\mathrm{T}} \mathbf{e}, \qquad \widehat{\boldsymbol{\alpha}}^\nu(s) = (\mathbf{M}^\nu + (s-1)\mathbf{I} - s\mathbf{e}\boldsymbol{\pi})\widehat{\boldsymbol{\theta}}^\nu(s),$$
$$\widehat{\boldsymbol{\omega}}^\nu(s) = (\mathbf{M}^\nu + (s-1)\mathbf{I} - s\mathbf{e}\boldsymbol{\pi})\boldsymbol{\eta}^w(0).$$

We note that

$$\mu_{ij}^\nu(\mathbf{M}_{ij}^\nu - s\boldsymbol{\pi}_j) = 0, \qquad \text{and} \qquad \sum_{j:j\neq i} \mu_{ii}^\nu(\mathbf{M}_{ij}^\nu - s\boldsymbol{\pi}_j) = \mu_{ii}^\nu(1 - \mathbf{M}_{ii}^\nu - s(1 - \boldsymbol{\pi}_i)) = \mu_{ii}^\nu,$$

then we can "solve" for the KKT multipliers:

$$\begin{aligned} \mu_{mm}^\nu &= -\boldsymbol{\pi}_m \left(r\widehat{\boldsymbol{\alpha}}_m^\nu(s) + \left[r\mathbf{c}_m + (-1)^\nu + r\lambda\right]\widehat{\boldsymbol{\omega}}_m^\nu(s)\right), \\ \mu_{mn}^\nu &= \boldsymbol{\pi}_m \left(r\left[\widehat{\boldsymbol{\theta}}_n(s) - \widehat{\boldsymbol{\theta}}_m(s) - \widehat{\boldsymbol{\alpha}}_m^\nu(s)\right] + \left[r\mathbf{c}_m + (-1)^\nu + r\lambda\right]\left[\boldsymbol{\eta}_n^w + \boldsymbol{\eta}_m^w - \widehat{\boldsymbol{\omega}}_m^\nu(s)\right]\right), \end{aligned} \tag{65}$$

{eq:lshiftso

although there are KKT multipliers hiding in $\widehat{\boldsymbol{\theta}}^\nu(s)$ and $\widehat{\boldsymbol{\alpha}}^\nu(s)$.

While this problem involves the scale invariant $\hat{A}(0)$, neither the upper nor lower bounds in eq. (62) are invariant. We then run into the same problems in proving our serial-topology hypothesis.

5 Nearly uniform serial models

sec:serial

Looking at them models that were found by numerical maximisation in Figure 3 (see the video for the full set of models), suggests we look at the particular set of models, shown in Figure 4.

At one particular time, the model is (a) the uniform serial model with M states. At later times, the model has (b) "sticky" end states with low exit transition probabilities, tending to the area maximising model as $\tau \to \infty$. At earlier times, the model is (c) progressively shortened by reducing the transition probability to the end states. Once this reaches zero,

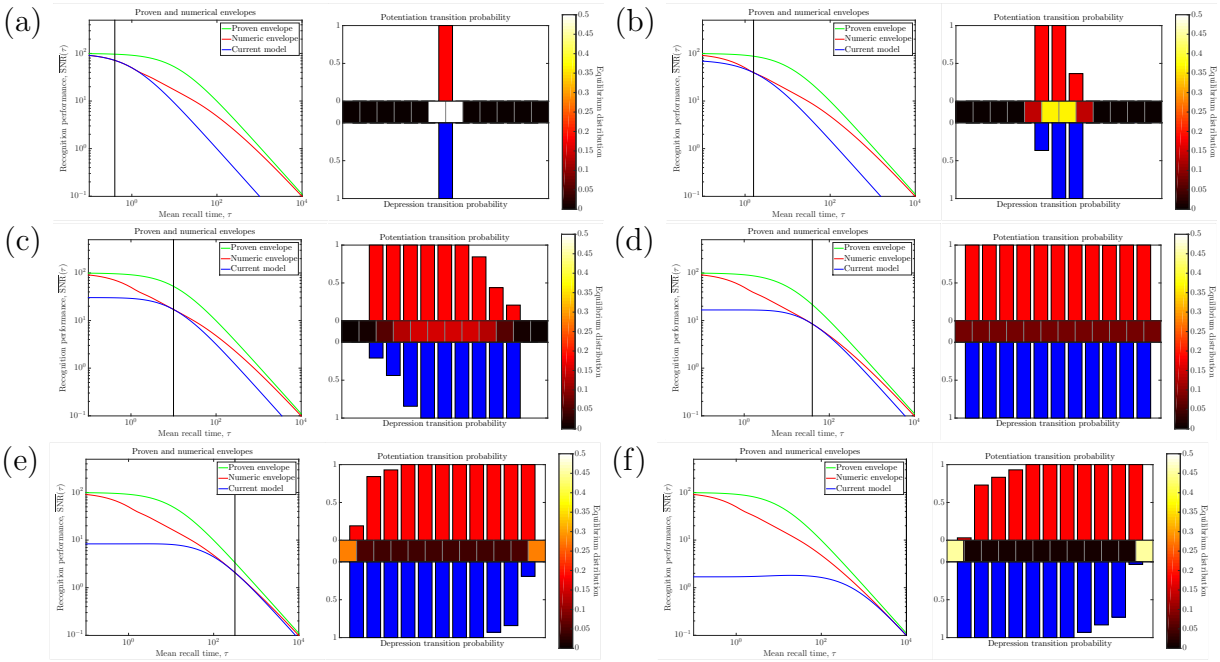

Figure 3: The models that constitute the numeric envelope in Figure 2. The plot to the left of each frame indicates the memory curve of the model shown to the right in blue, with the numeric envelope for models with the serial topology in red and the proven envelope, (53), in green. The bar graph in the upper right of each frame indicates the transition probabilities from the state to the left of each bar to the state on its right, under potentiation. The bar graph in the lower right of each frame indicates the transition probabilities from the state to the right of each bar to the state on its left, under depression. The heat map in the centre right of each frame indicates the equilibrium probability distribution for each state. See the video for the full set of models.

fig:envvid

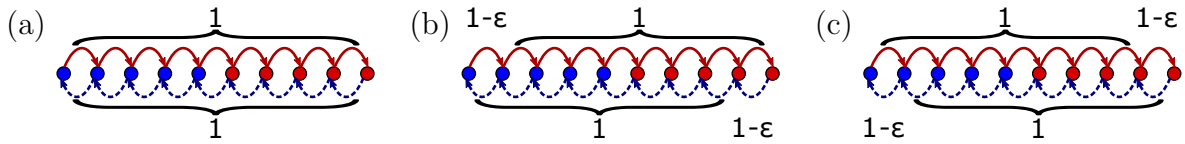

Figure 4: Heuristics for models that maximise memory curve. All unlabelled arrowshave transition probabilities equal to 1. At one particular time, the model is (a) the uniform serial model with M states. At later times, the model has (b) "sticky" end states with low exit transition probabilities, tending to the area maximising model. At earlier times, the model is (c) progressively shortened by reducing the transition probability to the end states. Once this reaches zero, we discard the old end state and work on the new end state.

we discard the old end state and shorten further by reducing the transition probability to the new end state (as one can see from Figure 3 and the video, this is not quite true, but it is close enough to true to get a reasonable approximation to the envelope). Once the length has been reduced to $M = 2$, the model is not shortened any further and the two-state model with deterministic transitions maximises the memory curve for all earlier times.

These models' Laplace transforms can be computed analytically, as we will see below, leading to a heuristic envelope. All of these calculations will be performed at $f^{\text{pot/dep}} = \frac{1}{2}$, so the denominator will play no role. We will work in units such that $r = 1$ temporarily, only restoring r at the end. We study models with m strong and m weak states for a total of $M = 2m$.

5.1 Uniform serial model

Here we will look at the model in Figure 4a. Looking at (35), the only thing we need is differences in $\boldsymbol{\eta}_i^w(s)$. In fact, given that the only nonzero elements of $\mathbf{K}$ are

$$\mathbf{K}_{ii+1} = \frac{1}{2} \quad \text{for} \quad i < 2m, \qquad \mathbf{K}_{ii-1} = -\frac{1}{2} \quad \text{for} \quad i > 1, \tag{66}$$

and $\boldsymbol{\pi}_i = \frac{1}{2m}$, we have

$$\hat{A}(s) = \frac{\boldsymbol{\eta}_1^w(s) - \boldsymbol{\eta}_{2m}^w(s)}{2m}. \tag{67}$$

Note that

$$\boldsymbol{\eta}^w(s) = \text{const.} - \mathbf{Z}(s)\mathbf{w}. \tag{68}$$

As we are only interested in differences of the $\boldsymbol{\eta}_i^w(s)$, we can drop the constant. Then

$$-(s\mathbf{I} + \mathbf{e}\boldsymbol{\pi} - \mathbf{W}^{\text{F}})\boldsymbol{\eta}^w(s) = \mathbf{w}. \tag{69}$$

We can then write

$$\boldsymbol{\eta}^w(s) = -\frac{(\boldsymbol{\pi}\boldsymbol{\eta}^w(s))\mathbf{e} + \mathbf{w}}{s} + \delta\eta(s), \qquad (\mathbf{W}^{\text{F}} - s\mathbf{I})\delta\eta(s) = \frac{\mathbf{W}^{\text{F}}\mathbf{w}}{s}. \tag{70}$$

As we are only interested in differences, we can drop any terms proportional to $\mathbf{e}$ in $\boldsymbol{\eta}^w(s)$.

The i'th row of this equation reads

$$\frac{\delta\eta_{i+1}+\delta\eta_{i-1}}{2}-(s+1)\delta\eta_i=0, \tag{71}$$

{eq:detarow}

with the following exceptions

$$\begin{aligned}\frac{\delta\eta_2}{2}-\left(s+\frac{1}{2}\right)\delta\eta_1&=0,\\ \frac{\delta\eta_{m+1}+\delta\eta_{m-1}}{2}-(s+1)\delta\eta_m&=\frac{1}{s},\\ \frac{\delta\eta_{m+2}+\delta\eta_m}{2}-(s+1)\delta\eta_{m+1}&=-\frac{1}{s},\\ \frac{\delta\eta_{2m-1}}{2}-\left(s+\frac{1}{2}\right)\delta\eta_{2m}&=0.\end{aligned} \tag{72}$$

{eq:detabndu

These equations are invariant under $\delta\eta_{2m+1-i}\to-\delta\eta_i$, so these two must be equal. Then we have

$$\begin{aligned}\frac{\delta\eta_{i+1}+\delta\eta_{i-1}}{2}-(s+1)\delta\eta_i&=0,\quad\text{for}\quad 1<i<m,\\ \frac{\delta\eta_2}{2}-\left(s+\frac{1}{2}\right)\delta\eta_1&=0,\\ \frac{\delta\eta_{m-1}}{2}-\left(s+\frac{3}{2}\right)\delta\eta_m&=\frac{1}{s}.\end{aligned} \tag{73}$$

{eq:detahalf

The first of these equations is solved by

$$\delta\eta_i=B\mathrm{e}^{\beta(i-1)}+C\mathrm{e}^{\beta(1-i)},\qquad\text{where}\qquad s=S(\beta)\equiv 2\sinh^2\left(\frac{\beta}{2}\right)=\cosh(m\beta)-1, \tag{74}$$

{eq:detagens

with B and C determined by the boundary conditions provided by the last two equations of (73):

$$B=-\frac{1}{s(1+\mathrm{e}^{-\beta})\cosh(m\beta)},\qquad C=-\frac{1}{s(\mathrm{e}^{\beta}+1)\cosh(m\beta)}. \tag{75}$$

{eq:unicoeff

This can be used to compute the Laplace transform

$$A(s)=\frac{\boldsymbol{\eta}_1^w(s)}{m}=\frac{1}{m}\left(\frac{1}{s}+B+C\right)=\frac{2\sinh^2\left(\frac{m\beta}{2}\right)}{ms\cosh(m\beta)}=\frac{2S(m\beta)}{Ms\left(S(m\beta)+1\right)}. \tag{76}$$

{eq:unilapla

In reality, M can only take even integer values. However, one can find a heuristic approximation to the envelope by maximising wrt. M. We find that

$$M_*=\pm\frac{2y_*}{\beta},\quad s=2\sinh^2\frac{y_*}{M_*},\qquad \hat{A}_*(s)=\left(\frac{2\sinh^2\frac{y_*}{2}}{y_*\cosh y_*}\right)\frac{|\beta|}{s},$$
$$\text{where}\quad y_*=\tanh\frac{y_*}{2}\cosh y_*\approx 1.50553. \tag{77}$$

{eq:unienv}

This is only valid for

$$2 \leq M_* \leq M, \qquad \Longrightarrow \qquad 2\sinh^2 \frac{y_*}{M} \leq s \leq 2\sinh^2 \frac{y_*}{2}. \tag{78}$$ {eq:univalid

When $M \gg 2$, the validity region is approximately

$$\frac{4.53327}{M^2} \lesssim s \lesssim 1.36423. \tag{79}$$ {eq:univalid

When $s \sim \mathcal{O}(M^{-2}) \ll 1$, we have $\beta \approx \pm\sqrt{2s}$ and the heuristic envelope is approximately

$$\hat{A}_*(s) \approx \frac{0.54203}{\sqrt{s}}. \tag{80}$$ {eq:unienvap

We will lok at a more accurate interpolation between different M in the next section.

For $s \gtrsim 1.36423$, we set $M = 2$ to find

$$\hat{A}(s) = \frac{1}{1+s}. \tag{81}$$ {eq:binaryen

This does not match up with (80) at the boundary of the two regions, as this is outside the regime of validity of that approximation. However, it does match up with the full formula, (77) at the boundary.

al_shorten

5.2 Shortened serial model

Here we will look at the model in Figure 4c. We proceed in the same manner as in §5.1, with the following differences.

The following elements of $\mathbf{K}$ and $\boldsymbol{\pi}$ differ from (66):

$$\begin{aligned} \mathbf{K}_{2m-1,2m} = -\mathbf{K}_{2,1} = \frac{(1-\varepsilon)}{2}, \qquad \boldsymbol{\pi}_1 = \boldsymbol{\pi}_M &= \frac{(1-\varepsilon)}{2(m-\varepsilon)}, \\ \boldsymbol{\pi}_i &= \frac{1}{2(m-\varepsilon)} \quad \text{for} \quad 1 < i < M2m. \end{aligned} \tag{82}$$ {eq:shortene

This means that the Laplace transform is given by

$$\hat{A}(s) = \frac{(1-\varepsilon)(\boldsymbol{\eta}_1^w(s) - \boldsymbol{\eta}_{2m}^w(s)) + \varepsilon(\boldsymbol{\eta}_2^w(s) - \boldsymbol{\eta}_{2m-1}^w(s))}{2(m-\varepsilon)} = \frac{2(1-\varepsilon)\boldsymbol{\eta}_1^w(s) + 2\varepsilon\,\boldsymbol{\eta}_2^w(s)}{2(m-\varepsilon)}. \tag{83}$$ {eq:shortena

The $\boldsymbol{\eta}_i^w(s)$ take the same general form (74) except for $i = 1$, and with different coefficients due to the different boundary conditions

$$\begin{aligned} \frac{\delta\eta_2}{2} - \left(s + \frac{1}{2}\right)\delta\eta_1 &= 0, \\ \frac{\delta\eta_3 + (1-\varepsilon)\delta\eta_1}{2} - \left(s + \frac{2-\varepsilon}{2}\right)\delta\eta_2 &= 0, \\ \frac{\delta\eta_{m-1}}{2} - \left(s + \frac{3}{2}\right)\delta\eta_m &= \frac{1}{s}. \end{aligned} \tag{84}$$ {eq:detabnds

These are solved by

$$\begin{aligned}
\delta\eta_i &= B\mathrm{e}^{(i-2)\beta} + C\mathrm{e}^{(2-i)\beta} \qquad \text{for} \quad 2 \le i \le m, \\
\delta\eta_1 &= \frac{1}{[(1-\varepsilon)\cosh(m\beta) + \varepsilon(2s+1)\cosh((m-1)\beta)]s}, \\
B &= -\frac{(2s+1)\mathrm{e}^{\beta/2} + (1-\varepsilon)\tanh(\beta/2)}{s[(1-\varepsilon)\cosh(m\beta) + \varepsilon(2s+1)\cosh((m-1)\beta)]}, \\
C &= -\frac{((2s+1)\mathrm{e}^{-\beta/2} - (1-\varepsilon)\tanh(\beta/2)}{s[(1-\varepsilon)\cosh(m\beta) + \varepsilon(2s+1)\cosh((m-1)\beta)]},
\end{aligned} \tag{85}$$

{eq:shortens

Which gives the following Laplace transform

$$\hat{A}(s) = \frac{(1-\varepsilon)S(m\beta) + \varepsilon(2s+1)S((m-1)\beta)}{s[m-\varepsilon][(1-\varepsilon)[S(m\beta)+1] + \varepsilon(2s+1)[S((m-1)\beta)+1]]}. \tag{86}$$

{eq:shortenl

This interpolates between (76) at $\varepsilon = 0$ and the same expression for $m-1$ at $\varepsilon = 1$.

To find the optimal ε at any given s, it helps if we define

$$\begin{aligned}
\Delta f_m(\beta) &= S(m\beta) - \zeta S((m-1)\beta), \quad & f_m(\beta) &= \frac{S(m\beta)}{\Delta f_m(\beta)}, \\
\Delta g_m(\beta) &= \Delta f_m(\beta) + 1 - \zeta, & g_m(\beta) &= \frac{S(m\beta)+1}{\Delta g_m(\beta)}, \\
\text{where} \quad \zeta &= 2s+1, \qquad \Longrightarrow & \hat{A}(s) &= \frac{\Delta f_m(f_m - \varepsilon)}{s\Delta g_m(m-\varepsilon)(g_m - \varepsilon)}.
\end{aligned}$$

The results are plotted in fig. 5.

We can maximise this wrt. ε.

$$(\varepsilon - f_m)^2 = (m - f_m)(g_m - f_m), \qquad \hat{A}(s) = \frac{\Delta f_m}{s\Delta g_m\left[\sqrt{|m-f_m|} - \sqrt{|g_m - f_m|}\right]^2}. \tag{87}$$

{eq:shortopt

We can find the boundary of the region of applicability by asking at what value of s is the optimal value of $\varepsilon = 0$ or 1?

$$\begin{aligned}
\left.\frac{\partial\hat{A}(s)}{\partial\varepsilon}\right|_{\varepsilon=0} &\propto m(2s+1)\left[S(m\beta) - S((m-1)\beta)\right] - S(m\beta)[S(m\beta)+1] = 0, \\
\left.\frac{\partial\hat{A}(s)}{\partial\varepsilon}\right|_{\varepsilon=1} &\propto (m-1)\left[S(m\beta) - S((m-1)\beta)\right] - (2s+1)S((m-1)\beta)[S((m-1)\beta)+1] = 0, \\
&\text{where} \quad s = S(\beta) = 2\sinh^2(\beta/2),
\end{aligned} \tag{88}$$

{eq:shortenv

with the $\varepsilon = 0$ case providing the lower boundary and $\varepsilon = 1$ providing the upper boundary. For lower s, $\varepsilon = 0$ is the optimum. For higher s, $\varepsilon = 1$ is the optimum, which is eqivalent to $\varepsilon = 0$ with $M-2$ states. This is depicted by the dotted lines in fig. 5b.

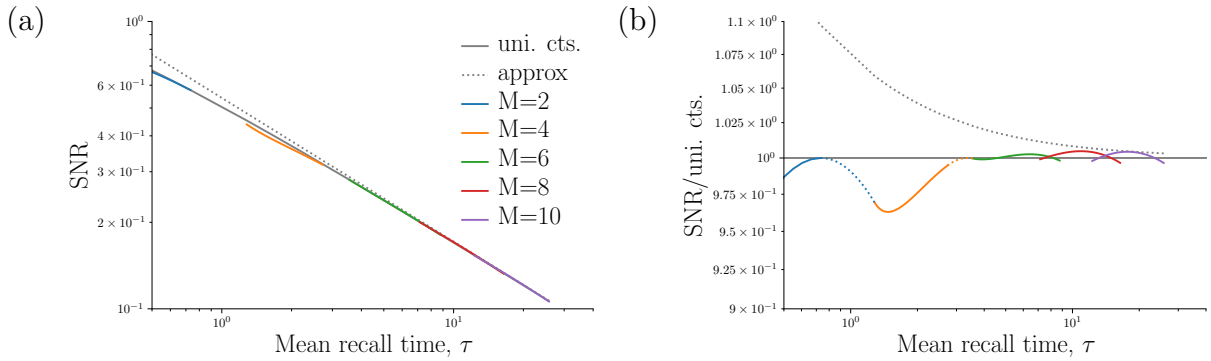

Figure 5: Signal-to-noise ratio for optimal shortened models, eq. (82) optimised over ε. (a): Solid grey line is eq. (77), the uniform model with M treated as a continuous parameter and optimised. Dotted grey line is the approximate expression from eq. (80). The other curves are from eq. (87), optimised shortened models for various values of M. (b): The ratios of the quantities plotted in (a) to eq. (77). The dotted blue and orange lines are the extensions of the $M = 2, 4$ lines with ε held at 0, as discussed below eq. (88). Noting the scale of the y-axis, we see that eqs. (77) and (80) do provide good approximations, especially for larger M.

$M = 2$ is a special case:

$$\hat{A}(s) = \frac{(\alpha - 1)^2}{[2(\alpha^2 - \alpha + 1) - (1 - \varepsilon)(\alpha - 1)^2]s},$$

so $\varepsilon = 0$ is always the maximum.

If these were really the models that maximised $A(s)$, the region of validity for M would not overlap with that for $M - 2$. In fact there are overlaps, as seen in fig. 5b for $M = 6$ and higher. This is related to the fact that Figure 3 and the video shows that the optimal models do not shorten by one state at a time, but several outgoing transition probabilities are reduced at once.

5.3 Sticky serial model

Here we will look at the model in Figure 4b. We proceed in the same manner as in §5.1, with the following differences.

The following elements of $\mathbf{K}$ and $\boldsymbol{\pi}$ are differ from (66):

$$\mathbf{K}_{1,2} = -\mathbf{K}_{2m,2m-1} = \frac{1-\varepsilon}{2}, \qquad \boldsymbol{\pi}_1 = \boldsymbol{\pi}_{2m} = \frac{1}{2(m-(m-1)\varepsilon)}, \qquad \boldsymbol{\pi}_i = \frac{1-\varepsilon}{2(m-(m-1)\varepsilon)} \quad \text{for} \quad 1 < i < 2m. \tag{89}$$

This means that the Laplace transform is given by

$$\hat{A}(s) = \frac{(1-\varepsilon)(\boldsymbol{\eta}_1^w(s) - \boldsymbol{\eta}_M^w(s))}{2(m-(m-1)\varepsilon)} = \frac{2(1-\varepsilon)\,\boldsymbol{\eta}_1^w(s)}{2(m-(m-1)\varepsilon)}. \tag{90}$$

{eq:stickyar

The $\boldsymbol{\eta}_i^w(s)$ take the same general form (74), but with different coefficients due to the different boundary conditions

$$\begin{aligned} \frac{\varepsilon\delta\eta_2}{2} - \left(s + \frac{1-\varepsilon}{2}\right)\delta\eta_1 &= 0, \\ \frac{\delta\eta_{m-1}}{2} - \left(s + \frac{3}{2}\right)\delta\eta_m &= \frac{1}{s}. \end{aligned} \tag{91}$$

{eq:detabnds

These are solved by

$$\begin{aligned} B &= -\frac{\mathrm{e}^\beta - \varepsilon}{s[\mathrm{e}^\beta + 1][\cosh(m\beta) - \varepsilon\cosh((m-1)\beta)]}, \\ C &= -\frac{1 - \varepsilon\mathrm{e}^\beta}{s[\mathrm{e}^\beta + 1][\cosh(m\beta) - \varepsilon\cosh((m-1)\beta)]}, \end{aligned} \tag{92}$$

{eq:stickyco

which leads to the Laplace transform

$$\hat{A}(s) = \frac{(1-\varepsilon)}{[m-(m-1)\varepsilon]s}\left[\frac{S(m\beta) - \varepsilon S((m-1)\beta)}{S(m\beta) - \varepsilon S((m-1)\beta) + 1 - \varepsilon}\right]. \tag{93}$$

{eq:stickyla

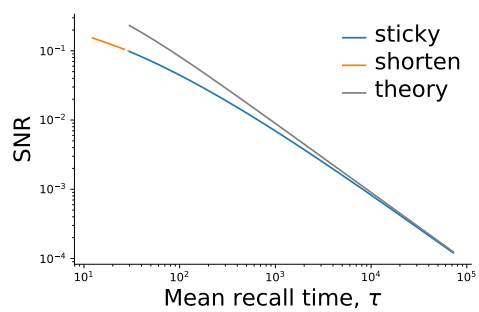

Figure 6: Signal-to-noise ratio for optimal sticky models, eq. (89) optimised over ε for $M = 10$, in blue. The orange line is the SNR for the optimised shortened models, eq. (82) with $M = 10$. The grey line is our theoretical envelope from eq. (53).

fig:sticky

We can maximise this wrt. ε. This results in a complicated quadratic equation for ε with an even worse solution that we will not reproduce here. We define

$$\begin{aligned} \Delta m &= (m-1), & \mu &= \frac{m}{\Delta m}, \\ \Delta f_m(\beta) &= S((m-1)\beta), & f_m(\beta) &= \frac{S(m\beta)}{\Delta f_m(\beta)}, \\ \Delta g_m(\beta) &= \Delta f_m(\beta) + 1, & g_m(\beta) &= \frac{S(m\beta)+1}{\Delta g_m(\beta)}, \end{aligned} \quad \Longrightarrow \quad \hat{A}(s) = \frac{\Delta f_m(1-\varepsilon)(f_m - \varepsilon)}{s\Delta m\Delta g_m(\mu-\varepsilon)(g_m-\varepsilon)}.$$

The minimum is at the solution of

$$[(1+f_m-g_m-\mu)\varepsilon-f_m+\mu g_m]^2=(\mu-1)(g_m-1)(\mu-f_m)(g_m-f_m).$$

The result is plotted in fig. 6.

We can find the boundary of the region of applicability by asking: at what value of s is the optimal value of $\varepsilon=0$?

$$\left.\frac{\partial\hat{A}(s)}{\partial\varepsilon}\right|_{\varepsilon=0}\propto m\left[S(m\beta)-S((m-1)\beta)\right]-S(m\beta)\left[S(m\beta)+1\right]=0,\qquad \text{where}\quad s=S(\beta)=2\sinh^2(\beta/2). \tag{94}$$

{eq:stickyva

The "sticky" models are optimal at values of s smaller than this. This is not quite the same as the lower boundary of the region eq. (88), or its approximate version eq. (78), so there is a gap between them where the uniform model with M states is optimal. The width of this gap is insignificant for larger M, as seen in fig. 6.

For small $s\ll M^{-2}$, it is given by

$$\varepsilon=1-\frac{2(M-1)\sqrt{s}}{\sqrt{M(M-2)}}\left(1+\sqrt{\frac{(M-2)^3s}{M}}+\mathcal{O}(M^2s)\right), \tag{95}$$

{eq:stickyep

and the Laplace transform is

$$\hat{A}(s)=(M-1)\left(1-\sqrt{M(M-2)s}+\mathcal{O}(M^2s)\right). \tag{96}$$

{eq:stickyen

euristicenv

5.4 Heuristic envelope

We can obtain a good approximation to the envelope by combining the results §5.1 and §5.3, ignoring the complications of §5.2. We will also ignore the small gap between the regions of validity, (78) and (94).

We will phrase it in terms of $\overline{\mathrm{SNR}}(\tau)$, reintroducing r:

$$\overline{\mathrm{SNR}}(\tau)=\frac{\sqrt{N}\hat{A}\left(\frac{1}{r\tau}\right)}{r\tau},\qquad \beta(\tau)=\sinh^{-1}\sqrt{1/2r\tau}. \tag{97}$$

{eq:envunits

Then, combining (81), (77) and (93), using the notation $\Delta_{M,\varepsilon}[T(M,\beta)]=T(M,\beta)-\varepsilon T(M-2,\beta)$., we find

$$\overline{\mathrm{SNR}}(\tau)=\begin{cases}\dfrac{\sqrt{N}}{1+r\tau}, & r\tau\le\frac{1}{2\sinh^2\frac{y_*}{2}},\\[2ex] \sqrt{N}\left(\dfrac{4\sinh^2\frac{y_*}{2}}{y_*\cosh y_*}\right)\beta(\tau), & \frac{1}{2\sinh^2\frac{y_*}{2}}\le r\tau\le\frac{1}{2\sinh^2\frac{y_*}{M}},\\[2ex] \max\limits_{\varepsilon\in[0,1]}\left\{\dfrac{2\sqrt{N}(1-\varepsilon)\,\Delta_{M,\varepsilon}\left[S\left(\frac{M\beta}{2}\right)\right]}{\Delta_{M,\varepsilon}[M]\,\Delta_{M,\varepsilon}\left[S\left(\frac{M\beta}{2}\right)+1\right]}\right\}, & \frac{1}{2\sinh^2\frac{y_*}{M}}\le r\tau.\end{cases} \tag{98}$$

{eq:henv}

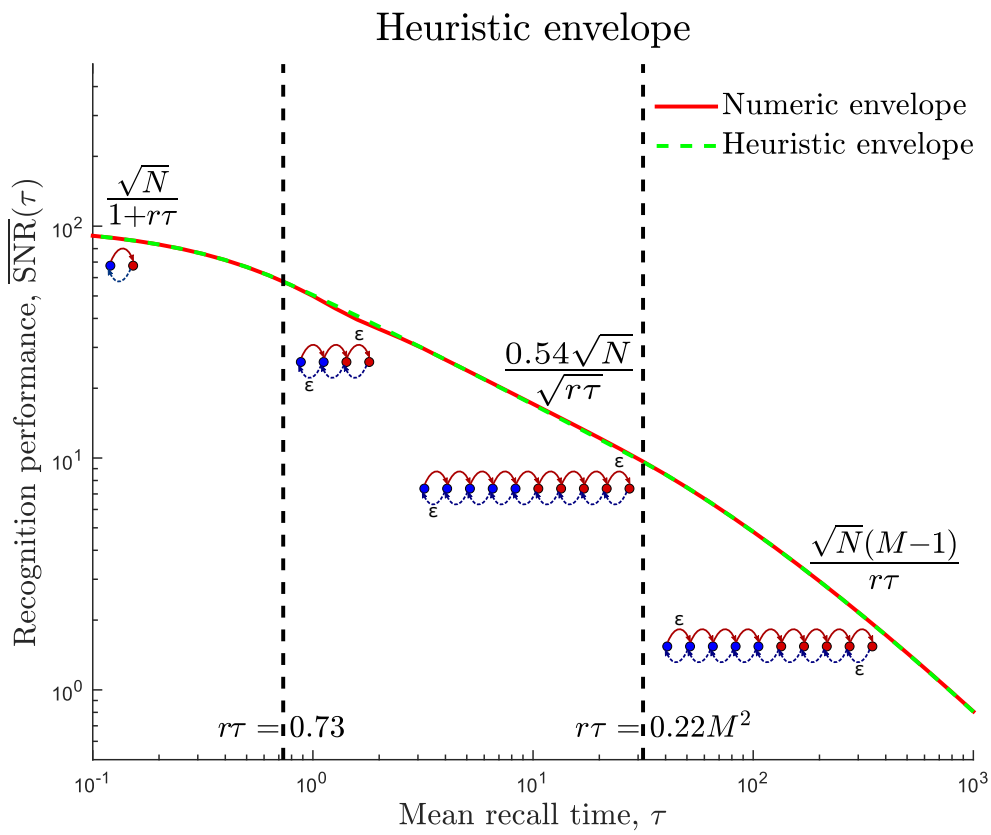

Figure 7: Heuristic envelope for the signal-to-noise ratio, eq. (98). Analytic expressions from eq. (99) are only valid to the right of each region.

euristicenv

For $M \gg 1$, and $r\tau$ much greater than the lower bound of each region, we can approximate this as

$$\overline{\mathrm{SNR}}(\tau) \approx \begin{cases} \dfrac{\sqrt{N}}{1+r\tau}, & r\tau \leq 0.73, \\ \dfrac{0.54\sqrt{N}}{\sqrt{r\tau}}, & 0.73 \leq r\tau \leq 0.22M^2, \\ \dfrac{\sqrt{N}(M-1)}{r\tau}, & 0.22M^2 \leq r\tau. \end{cases} \tag{99}$$

{eq:henvappr

This is plotted in Figure 7.

6 Conclusions

onclusions

We do not expect to find precisely these models inside real synapses. First, evolution was faced with stronger constraints than us when designing these synapses as it had to build them out of real molecules. Second, the set of priorities faced by evolution would have been longer than just performance in recognition memory at a single timescale.

Nevertheless, the qualitative message of our findings should be more robust to reality:

- *Short* timescale → *intermediate* timescales: transition topology goes from short and wide → long and thin.
- *Intermediate* timescale → *long* timescales: transition probabilities go from strong and deterministic → weak and stochastic.

This is what we expect to find in the different brain regions that store memories for different timescales.

The transition between the two regimes is set by the maximum number of states, M (we can't set the number of states to be exactly M, as bigger models can always mimic smaller models). So, if we want to lengthen the timescale at which a synapse is optimal, we start by adding extra states, using the topology mechanism. When we run out of states, we switch to the stochasticity mechanism and make the end states sticky. The stochasticity mechanism is only used when we no longer have the option of using the topology mechanism. It is always better to add more states, if allowed, than to make transitions stochastic.

We can get some insight into why this is by considering a serial model with M states and all transition topologies equal to q.

$$\begin{aligned} \mathrm{SNR}(0) &\propto q, & \max_a \tau_a &\propto \frac{1}{q}, \\ \mathrm{SNR}(0) &\propto \frac{1}{M}, & \max_a \tau_a &\propto M^2. \end{aligned}$$

The initial SNR scales as described because it is related to the equilibrium probability flux between the weak and strong states, and the equilibrium distribution is uniform over M states. The lifetime scales as described because rq sets the timescale for transitions, and for diffusion distance (M) scales as the square root of time.

If we want to lengthen the lifetime with a fixed hit on initial SNR, we get a larger increase by increasing M than by decreasing q. If we want to lengthen the lifetime by a fixed amount, we take a smaller hit on initial SNR by increasing M than by decreasing q. Thus, we always get more bang for our buck using the topology mechanism than the stochasticity mechanism.

References

98retrieval [1] F. T. Sommer and P. Dayan, "Bayesian retrieval in associative memories with storage errors.," *IEEE Trans. Neural Netw.* **9** (Jan., 1998) 705–13.

eny1981fund [2] J. G. Kemeny, "Generalization of a fundamental matrix," *Linear Algebra and its Applications* **38** (1981) no. 0, 193 – 206.

2013synapse [3] S. Lahiri and S. Ganguli, "A memory frontier for complex synapses," in *Adv. Neural Inf. Process. Syst. 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., pp. 1034–1042. NIPS, 2013.