Supporting Text: Fundamental limits on persistent activity in networks of noisy neurons

Yoram Burak[1,2,3] and Ila Fiete[4]

[1]*Edmond and Lily Safra Center for Brain Sciences,*
Hebrew University, Jerusalem 91904, Israel
[2]*Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel*
[3]*Center for Brain Science, Harvard University, Cambridge, MA 23108*
[4]*Center for Learning and Memory, University of Texas at Austin, Austin, TX 78712*

CONTENTS

I. STOCHASTIC DYNAMICS IN THE LARGE N LIMIT

To formally derive the dynamic equations in the large N limit, we consider a network with M sub-populations, each containing n neurons, so that the total number of neurons is $N = nM$. Within each sub-population, all neurons share the same incoming and outgoing weights, as well as the same bias. We scale the weights with n such that the total synaptic input to a neuron in each population is independent of n:

$$W_{ij} = \frac{1}{n} w_{IJ}, \tag{S1}$$

where the indices I and J represent the sub-populations that contain neurons i and j, respectively. The population weights w_{IJ}, represent the overall synaptic input to a neuron in sub-population J from all neurons in sub-population I. With this scaling, the noise-free rate dynamics are independent of n.

It is useful to write the dynamic equations in terms of g_i, since this quantity is identical for all neurons in the same sub-population, and is thus denoted below by g_I. From Eqs. (7) and (8),

$$\dot{g}_I + \frac{g_I}{\tau} = \eta_I \,, \tag{S2}$$

where

$$\eta_I = \frac{1}{n} \sum_J \sum_{j \in J} w_{IJ} \xi_j + \frac{1}{\tau} b_I \,, \tag{S3}$$

and ξ_j is the spike train generated by neuron j. The mean of η_I is

$$\langle \eta_I \rangle = \sum_J w_{IJ} r_J + \frac{1}{\tau} b_I \,, \tag{S4}$$

where b_I is the bias in sub-population I and r_J is the firing rate of neurons in sub-population J (identical for all neurons within the same sub-population). The variance of η_I is equal to

$$\mathrm{Var}(\eta_I) = \frac{1}{n} \sum_J w_{IJ}^2 r_J \,. \tag{S5}$$

The contribution to η_I from each sub-population is a sum over n independent random variables, which is then smoothed over a time scale τ in Eq. (S2). Therefore in the limit of large n, it is justified to treat η_I as if it is normally distributed. This is the formal basis for Eqs. (10)–(11): although individual synaptic variables are not normally distributed, Eqs. (10)–(11) (with normal distribution of ξ) are exactly equivalent to Eqs. (S2), (S4), and (S5) once the Gaussian approximation has been made.

In a network where the continuous attractor arises due to a symmetry, a natural way to define the large N limit is to distribute the neuron labels more and more densely as the number of neurons is increased. We briefly illustrate this for the ring network: we associate with each neuron an angle θ_i, where the angles are uniformly distributed in the interval $[0, 2\pi)$. We assume that the weights have the form

$$W_{ij} = \frac{1}{N} w(\theta_i - \theta_j) \,, \tag{S6}$$

where $w(\theta)$ is periodic. The rest of the derivation is very similar to the one outlined above: the formal justification for the Gaussian approximation is obtained by considering the dynamics for g_i.

II. THE COEFFICIENT OF DIFFUSION

Consider an expansion of the dynamic equations around a steady state $\bar{s}(\theta)$:

$$s_i(t) = \bar{s}_i(\theta) + \delta s_i(t). \tag{S7}$$

Then,

$$\dot{\delta s_i} = \sum_j K_{ij}(\theta)\delta s_j + \xi_i\,, \tag{S8}$$

where K is a kernel derived by linearizing Eq. **10** around $\bar{s}$:

$$K_{ij}(\theta) = -\frac{1}{\tau}\delta_{ij} + \phi'\left[\bar{g}_i(\theta)\right]W_{ij}\,, \tag{S9}$$

and

$$\langle\xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')\phi\left[\bar{g}_i(\theta)\right] \tag{S10}$$

To calculate the diffusivity, we evaluate the contribution of ξ to the mode of K with vanishing eigenvalue, which corresponds to motion within the line attractor,

$$u_i(\theta) = \frac{\mathrm{d}\bar{s}_i}{\mathrm{d}\theta} = \tau\phi'\left[\bar{g}_i(\theta)\right]\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}. \tag{S11}$$

In general, the matrix K is not symmetric (this is true even if the synaptic weights are symmetric, due to the nonlinear transfer function). Therefore, we need to project ξ on the *left* eigenvector v of K, which satisfies

$$\sum_i K_{ij}(\theta)v_i(\theta) = 0\,, \tag{S12}$$

and we chose the normalization of v such that

$$\sum_i v_i u_i = 1\,. \tag{S13}$$

Projecting Eq. (S8) onto v we find that over a time interval δt, the change in s within the attractor manifold is $\delta s_i = \zeta u_i$, where

$$\mathrm{var}\left(\zeta\right) = \delta t \cdot \sum_i \left[v_i(\theta)\right]^2 \phi\left[\bar{g}_i(\theta)\right]. \tag{S14}$$

Identifying $\mathrm{var}\left(\zeta\right)$ with $2\mathcal{D}$, we see that

$$2\mathcal{D}(\theta) = \sum_i \left[v_i(\theta)\right]^2 \phi\left[\bar{g}_i(\theta)\right]\,. \tag{S15}$$

If the connectivity matrix W is symmetric, it is straightforward to check that the transpose of K is the linear kernel specifying the dynamics of perturbations to the summed synaptic input g. Hence, in this case, we can infer immediately that

$$v_i(\theta) = \frac{1}{\alpha} \frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}, \tag{S16}$$

where the prefactor α is set by Eq. (S13):

$$\alpha = \tau \sum_i \phi'[\bar{g}_i(\theta)] \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2. \tag{S17}$$

Thus, for symmetric synaptic weights,

$$2\mathcal{D}(\theta) = \frac{\sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2 \phi[\bar{g}_i(\theta)]}{\left[\tau \sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2 \phi'[\bar{g}_i(\theta)]\right]^2}, \tag{S18}$$

which can be written as in Eq. **3** using the relation: $\bar{r}_i(\theta) = \phi[\bar{g}_i(\theta)]$.

Systematic drift

Above, we considered the variance of random drift along the attractor manifold. As noted under Results in the main manuscript, it may be possible for noise to additionally induce a systematic drift in the state along the attractor. Our focus is on variance because in the context of memory, variance is the essential quantity: an ideal estimator could subtract the effects of a systematic drift, but variance cannot be removed. Thus, it is variance that limits information.

Nevertheless, it is of interest to consider the conditions under which noise could induce a systematic drift in the network state, and to estimate the magnitude of such drift if it exists. Below, we briefly discuss these questions.

A noise-induced systematic drift can occur in principle even if the weights and biases are tuned to be drift-free under the deterministic dynamics: if the diffusion coefficient $\mathcal{D}$ depends on θ, and noise couples nonlinearly with the dynamics along the attractor manifold, detailed balance may be broken and produce biased motion along the attractor. This would be similar to the mechanism by which particles undergo biased flow in temperature gradients even in the absence of other applied forcing potentials [1–4].

The exact noise-induced drift, when it exists, cannot be derived using the approach taken in the earlier parts of this section; the reason why is articulated in detail in [4]. Instead, it is necessary

to start from the Master equation generated by the precise microscopic dynamics of Eqs. (7)–(9). Nevertheless, we may assess the magnitude of systematic drift when it exists: plausibly, the drift roughly lies in the range given by the difference between the Itō *vs.* Stratonovich interpretations of the Langevin equation for dynamics along the attractor:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \sqrt{2D(\theta)}\xi \,, \tag{S19}$$

where ξ is white Gaussian noise. While the Itō interpretation yields a zero drift term, the Stratonovich interpretation generates a systematic drift that, over a short interval δt, is equal to

$$\langle\delta\theta\rangle = \frac{1}{2}\frac{\partial D(\theta_0)}{\partial\theta}\delta t \,, \tag{S20}$$

when starting from a state θ_0 along the attractor. Given that D scales as $1/N$, the systematic drift term, if it exists, is of order $\delta t/N$. Thus, for small δt or large N, the leading contribution to the deviation of the network state from its initial value along the attractor comes from the diffusion term, which scales as $\sqrt{\delta t/N}$, compared to systematic drift, which scales as $\delta t/N$.

Even these small contributions from the drift may be eliminated if the network is tuned to display persistent activity states under stochastic dynamics, rather than being tuned in the deterministic limit. This may be achieved by adding small asymmetries to those neural weights (or neural biases) that would generate a precisely tuned continuous attractor in the determinsitic limit. Such corrections are expected to be at a higher order in $1/N$ relative to the deterministically determined weights, because of the relative size of the drift term, and will not (to leading order in $1/N$) affect our expressions for the diffusion coefficient. By contrast, the random component of drift cannot be eliminated by adjusting the weights.

For these reasons, in the rest of this work and throughout the main manuscript, we assume that the network is tuned so there is no drift under the stochastic dynamics.

Generalization to a multi-dimensional attractor

In a noisy network possessing a multi-dimensional attractor, the instantaneous attractor state is characterized by parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$, where M is the dimensionality of the attractor. In a short time interval Δt,

$$\langle\Delta\theta_\alpha\Delta\theta_\beta\rangle = 2\mathcal{D}_{\alpha\beta}\Delta t \,, \tag{S21}$$

where $\mathcal{D}$ is the diffusion tensor. To evaluate $\mathcal{D}$, we follow a similar procedure as in the one-dimensional case: we linearize the deterministic dynamics around a steady state $\bar{s}(\vec{\theta}_0)$. The operator that describes the linearized dynamics has an M-dimensional space of null eigenvectors. We chose a basis of right eigenvectors as follows,

$$u_i^\alpha(\vec{\theta}) = \tau \frac{\partial}{\partial \theta_\alpha} \phi \left[\bar{g}_i(\vec{\theta}) \right] . \tag{S22}$$

The eigenvector with label α corresponds to drift within the attractor manifold along the direction θ_α. We chose a basis of conjugate left eigenvectors, v_i^α, such that

$$\sum_{i=1}^N v_i^\alpha u_i^\beta = \delta_{\alpha\beta} . \tag{S23}$$

As in the one-dimensional case, we evaluate the diffusion tensor by projecting the change in s, in a short time interval δt, on the attractor manifold. This yields

$$2\mathcal{D}_{\alpha\beta} = \sum_i v_i^\alpha(\vec{\theta}) v_i^\beta(\vec{\theta}) \phi \left[\bar{g}_i(\vec{\theta}) \right] . \tag{S24}$$

If the synaptic weights are symmetric $(w_{ij} = w_{ji})$, the left eigenvectors can be written as

$$v_i^\alpha(\vec{\theta}) = \sum_\beta (A^{-1})_{\alpha\beta} \frac{\partial}{\partial \theta_\beta} \bar{g}_i(\vec{\theta}) , \tag{S25}$$

where, using, Eq. (S23),

$$A_{\alpha\beta} = \tau \sum_i \frac{\partial \bar{g}_i}{\partial \theta_\alpha} \phi' \left[\bar{g}_i(\vec{\theta}) \right] \frac{\partial \bar{g}_i}{\partial \theta_\beta} . \tag{S26}$$

Hence, in this case,

$$2\mathcal{D}_{\alpha\beta}(\vec{\theta}) = \sum_{\mu\nu} A_{\alpha\mu}^{-1} A_{\beta\nu}^{-1} \sum_i \phi \left[\bar{g}_i(\vec{\theta}) \right] \frac{\partial \bar{g}_i}{\partial \theta_\mu} \frac{\partial \bar{g}_i}{\partial \theta_\nu} . \tag{S27}$$

III. DIFFUSIVE DYNAMICS ALONG THE ATTRACTOR

Imagine that at time $t = 0$ the network state is set by an external sensory input to a point Θ on the attractor, and from that time onward the network state evolves without this sensory input. Because the statistics of the noise ξ depends only on the present state of the network, and not on its history, the dynamics of the instantaneous attractor state are those of a simple diffusion process, where the diffusion coefficient $\mathcal{D}(\theta)$ is obtained from the derivation in Sec. II. Thus, the statistics at short times also determine the statistics of the long-time dynamics. Here we briefly review how to derive the growth of noise-driven variance in the network state over time [5]. We

assume that the network is tuned so that there is no systematic (biased) drift along the attractor manifold.

In this case, the probability distribution of $\theta(t)$ across multiple trials of the experiment, $p(\theta, t)$, is calculated by solving the Fokker-Planck (diffusion) equation, with no drift term and with a diffusion term involving $D(\theta)$:

$$\frac{\partial p(\theta, t)}{\partial t} = \frac{\partial^2}{\partial \theta^2} \left[\mathcal{D}(\theta) p(\theta, t) \right] \tag{S28}$$

with the initial condition

$$p(\theta, 0) = \delta(\theta - \Theta), \tag{S29}$$

where the diffusion term is based on Eq. (S14). The variance of $\theta(t)$ can then be evaluated directly from $p(\theta, t)$.

Below, we evaluate the diffusive variance in several particular situations.

1. State-independent $\mathcal{D}$, unbound domain of diffusion.

In the simplest example, if $\mathcal{D}$ is not state dependent, and the domain for diffusion on which θ lies is unbounded, the solution of the diffusion equation is:

$$p(\theta, t) = \frac{1}{\sqrt{4\pi \mathcal{D} t}} \exp \left[-\frac{(\theta - \Theta)^2}{4\mathcal{D} t} \right] \tag{S30}$$

and the variance grows linearly with time: $\sigma^2(t) = 2\mathcal{D}t$.

2. State-independent $\mathcal{D}$, periodic θ

If θ is periodic, then clearly at long times $p(\theta, t)$ will be uniform, and the variance of $\theta(t)$ must saturate. The solution of the Fokker-Planck (diffusion) equation with periodic boundary conditions, $p(2\pi, t) = p(0, t)$ (i.e., θ lies on a unit circle), is the wrapped Gaussian distribution with parameter $\sigma^2(t) = 2\mathcal{D}t$:

$$p(\theta, t) = \frac{1}{\sqrt{2\pi\sigma(t)^2}} \sum_{n=-\infty}^{\infty} \exp \left[-\frac{(\theta - \Theta + 2\pi n)^2}{2\sigma(t)^2} \right] = \frac{1}{2\pi} \vartheta_3 \left(\frac{\theta - \Theta}{2}, e^{-\mathcal{D}t} \right) \tag{S31}$$

where the wrapped Gaussian density is evaluated only in the range $[-\pi, \pi)$, and is zero elsewhere. Here, $\vartheta_n(z, \tau)$ is the Jacobi Theta function. A simple approximation of this distribution is given by the von Mises distribution, $p(\theta, t) \propto \exp \{\cos [\kappa(\theta - \Theta)]\}$ with the parameter $\kappa^2 = 1/(2\mathcal{D}t)$.

The moments of $(\theta - \Theta)$ follow from this distribution. For example, define the variance on the circle by setting $\theta(t) - \Theta$ to be in the range $[-\pi, \pi)$. Then,

$$\text{Var}\left[\theta(t) - \Theta\right] = \left\langle (\theta(t) - \Theta)^2 \right\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \theta^2 \vartheta_3 \left(\frac{\theta}{2}, e^{-\mathcal{D}t}\right) d\theta. \tag{S32}$$

3. State-dependent diffusion, expansion in t

When the diffusion coefficient depends on the state within the attractor, the variance of $\theta(t)$ can be obtained from $\mathcal{D}(\theta)$ as outlined above, but in general it cannot be expressed using a closed analytical form. Here we consider the behavior of the variance for small t, using a perturbative expansion.

The Langevin equation corresponding to Eq. (S28) is

$$\frac{d\theta}{dt} = \xi(t) \tag{S33}$$

where

$$\langle \xi(t)\xi(t') \rangle = 2\mathcal{D}\left[\theta(t)\right] \delta(t - t') \tag{S34}$$

(interpreted as the Itō stochastic differential equation: $d\theta_t = \sqrt{2\mathcal{D}(\theta_t, t)}dW_t$), and where $\mathcal{D}$ is the coefficient of diffusion derived in the Sec. II. Thus,

$$\text{Var}\left(\Delta\theta\right) \equiv \left\langle [\theta(t) - \Theta]^2 \right\rangle = 2\left\langle \int_0^t \mathcal{D}\left[\theta(t')\right] dt' \right\rangle, \tag{S35}$$

and to leading order in t,

$$\text{Var}\left(\Delta\theta\right) = 2\mathcal{D}(\Theta)t, \tag{S36}$$

as derived in Sec. (II). If $\mathcal{D}$ is independent of θ (and for a bounded variable, as long as boundary effects can be ignored), this relationship holds also for large t. More generally, we can expand $\mathcal{D}(\theta)$ around $\theta = \Theta$ to evaluate corrections to this behavior due to the dependence of $\mathcal{D}$ on θ. Making use of the fact that $\langle \theta(t) \rangle = \Theta$, we find

$$\text{Var}\left(\Delta\theta\right) = 2\mathcal{D}(\Theta)t + \mathcal{D}(\Theta)\mathcal{D}''(\Theta)t^2 + \ldots \tag{S37}$$

4. Example: 2-neuron network with mutual excitation

As a concrete example where $\mathcal{D}$ depends on θ, we consider a 2-neuron (or 2-population) similar to the one in Fig. 2 D–F, but instead of mutual inhibition the connections between the neurons

are excitatory: $\phi(x) = x/\tau$, and

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} . \tag{S38}$$

To produce a continuous set of attractors, we provide positive bias to one of the neurons and opposite negative bias to the other neuron: $b = (b_0, -b_0)$. The attractor states of the noise-free dynamics are then

$$\begin{pmatrix} \theta + \frac{1}{2}b_0 \\ \theta - \frac{1}{2}b_0 \end{pmatrix} . \tag{S39}$$

In a Poisson-spiking network, the instantaneous attractor state θ undergoes random diffusion as described by Eq. (S33), where

$$D(\theta) = \frac{\theta}{2\tau} . \tag{S40}$$

Note that in contrast to the case of mutual inhibition, $\mathcal{D}$ depends on θ. However, the dependence of $\mathcal{D}$ on θ is linear. Therefore only the first term in Eq. (S37) survives, and the dependence of $\mathrm{Var}(\theta)$ on t is linear even for large t:

$$\left\langle [\theta(t) - \Theta]^2 \right\rangle = \frac{\Theta}{2\tau} t \tag{S41}$$

In other words, the variance increases linearly with a slope determined by the diffusivity at the initial state Θ. Fig. S1 demonstrates this result in a numerical simulation.

IV. INTERNAL FISHER INFORMATION

The internal Fisher information (FI) rate J is defined as the Fisher information [6] carried by spikes in a short interval δt about the instantaneous attractor state θ, divided by δt. Using the standard definition of FI,

$$J\delta t = -\left\langle \frac{\partial^2}{\partial \theta^2} \log p\left(\{n_i\} | \theta\right) \right\rangle , \tag{S42}$$

where n_i is the number of spikes produced by neuron i in the interval δt. To leading order in $1/N$, the neurons fire as Poisson processes with rates determined by the attractor state θ in the noise-free ($N \to \infty$) limit,

$$p\left(\{n_i\} | \theta\right) = \prod_i \frac{[\bar{r}_i(\theta)\delta t]^{n_i} \, e^{-\bar{r}_i(\theta)\delta t}}{n_i!} . \tag{S43}$$

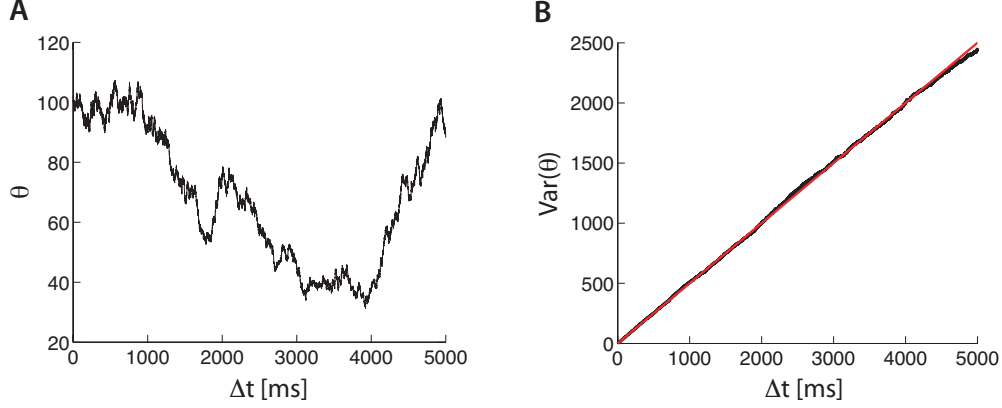FIG. S1. Simulation results from a 2-neuron network with mutual excitation and Poisson noise. In each trial we set the initial attractor state θ to $\Theta = 100$. **A** An example for the dynamics of θ from one trial. **B** Variance of $\theta(\Delta t)$ across 4000 random trials (black trace), compared with the theoretical prediction (red trace). Deviations between the analytical curve and the numerics at large Δt can be attributed to sampling effects from a simulation of finite duration.

The rest of the derivation of Eq. (3) is identical to the derivation of FI for a population of neurons encoding a sensory variable [7, 8].

In the multi-dimensional case, activity of the N neurons encode the instantaneous state within an M dimensional attractor parametrized by $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$. The Fisher information matrix is then

$$
J_{\alpha\beta}(\vec{\theta}) = -\left\langle \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta} \log p\left(\{n_i\} | \vec{\theta}\right) \right\rangle = \sum_{i=1}^{N} \frac{\phi'^2\left[\bar{g}_i(\vec{\theta})\right] \frac{\partial\bar{g}_i}{\partial\theta_\alpha} \frac{\partial\bar{g}_i}{\partial\theta_\beta}}{\phi\left[\bar{g}_i(\vec{\theta})\right]} .
\tag{S44}
$$

V.   INFORMATION-DIFFUSION INEQUALITY

We next prove that the Fisher information sets a lower bound on the diffusivity of the attractor state. We start with the one-dimensional case, then consider the more general case of a mult-dimensional attractor. Using Eqs. (S15) and (3),

$$
2\mathcal{D}J = \frac{1}{\tau^2}\sum_i v_i^2\phi_i \sum_j \frac{u_j^2}{\phi_j} = \frac{1}{\tau^2}\sum_{ij}(v_iu_i)(v_ju_j)\frac{v_i\phi_i}{u_i}\frac{u_j}{v_j\phi_j}
$$
$$
= \sum_{ij}(v_iu_i)(v_ju_j)\cdot\frac{1}{2}\left(\frac{x_j}{x_i} + \frac{x_i}{x_j}\right)
\tag{S45}
$$

where we used the short-hand notation:

$$
\phi_i = \phi\left[\bar{g}_i(\theta)\right]
\tag{S46}
$$

and

$$x_i = \frac{u_i}{v_i \phi_i}, \tag{S47}$$

and in the last step in Eq. (S45) we symmetrized the expression inside the sum over i and j with respect to the two indices.

We note that if x_i/x_j is positive, $(x_i/x_j + x_j/x_i)/2 \geq 1$. Furthermore, using the definition of x_i [Eq. (S47)], and the non-negativity of ϕ, $(v_i u_i)(v_j u_j)$ is positive as well. On the other hand, if x_i/x_j is negative, $(x_i/x_j + x_j/x_i)/2 \leq -1$ and $(v_i u_i)(v_j u_j)$ is negative. Thus, term by term,

$$(u_i v_i)(u_j v_j) \cdot \frac{1}{2} \left(\frac{x_j}{x_i} + \frac{x_i}{x_j} \right) \geq (u_i v_i)(u_j v_j) \tag{S48}$$

and we conclude that

$$2\mathcal{D}J \geq \frac{1}{\tau^2} \sum_{ij} (u_i v_i)(u_j v_j) = \frac{1}{\tau^2}. \tag{S49}$$

Saturation of the inequality

A sufficient and necessary condition for Eq. (S49) to be an equality is that x_i/x_j must be equal to unity for all i and j for which u_i and v_i do not vanish. In other words, x_i must be the same for all of these indices.

If the weight matrix is symmetric, we have an explicit expression for v_i, Eq. (S16). In this case,

$$x_i \propto \frac{\phi_i'}{\phi_i}, \tag{S50}$$

where we use the short-hand notation: $\phi_i' = \phi'[\bar{g}_i(\theta)]$. Therefore, ϕ'/ϕ must be constant within the full range of conductances present in the stationary solution. Assuming that values of g present in all the attractor states span a continuous range of values, this condition is met if and only if the neural transfer function is exponential in this range:

$$\phi(x) = \phi_0 e^x. \tag{S51}$$

Thus, the bound is saturated for any shape of the steady state attractor, if and only if the nonlinearity is exponential.

Generalization to multi-dimensional attractors

Intuitively, based on the Cramér-Rao inequality in the multi-dimensional case, we expect

$$2\mathcal{D} \geq \frac{1}{\tau^2} J^{-1}, \tag{S52}$$

where the inequality designates that the difference between the left hand side and the right hand side is a positive-semidefinite matrix. To prove the inequality, we define two $N \times M$ matrices:

$$U_{i\alpha} = \frac{u_i^\alpha}{\sqrt{\phi_i}} \quad ; \quad V_{i\alpha} = \sqrt{\phi_i} v_i^\alpha , \tag{S53}$$

where we adopted the short-hand notation: $\phi_i = \phi \left[\bar{g}_i \left(\vec{\theta} \right) \right]$ and from Eq. (S23),

$$V^\mathrm{T} U = U^\mathrm{T} V = I \tag{S54}$$

is an $M \times M$ identity matrix. In terms of U and V we can rewrite Eqs. (S24) and (S44) as:

$$2\mathcal{D} = V^\mathrm{T} V \quad ; \quad J = \frac{1}{\tau^2} U^\mathrm{T} U . \tag{S55}$$

Thus, we need to prove that $V^\mathrm{T} V \geq \left(U^\mathrm{T} U \right)^{-1}$ which is equivalent to proving:

$$U^\mathrm{T} U V^\mathrm{T} V U^\mathrm{T} U \geq U^\mathrm{T} U , \tag{S56}$$

since J (as well as D) is a positive definite matrix. The inequality follows immediately by: (i) noting that the quadratic form $A^\mathrm{T} A \geq 0$ for any matrix, (ii) choosing $A = \left(V U^\mathrm{T} - U V^\mathrm{T} \right) U$, and ($iii$) using Eq. (S54). It is straightforward to see that if the weight matrix is symmetric and the nonlinearity is exponential, $A = 0$ and the inequality (S52) is saturated.

VI.   READOUT

Suppose the memory network is initialized at a state Θ at time $t = 0$, but then its state begins to drift diffusively. We imagine that an observer receives a recall cue at time $T > 0$. The observer has access to all spikes emitted by neurons in the memory network after the cue, and is faced with the task of inferring Θ. Obviously, the best the observer could do is to estimate the instantaneous attractor state at time T. Therefore (assuming a one-dimensional linear variable) the variance of the estimator will have the structure:

$$\Delta \hat{\Theta}(T, \Delta T) = V_\mathrm{d}(T) + V_\mathrm{sp}(\Delta T) \tag{S57}$$

where $V_\mathrm{d}(T)$ is the variance of the attractor state over the memory storage period (*e.g.*, if $\mathcal{D}$ is independent of θ, $V_\mathrm{d}(T) = 2\mathcal{D}T$); and $V_\mathrm{sp}(\Delta T)$ is the variance of an estimator, of the instantaneous state at time T, based on observing spikes from T to $T + \Delta T$. The rest of this section is concerned with the evaluation of $V_\mathrm{sp}(\Delta T)$.

Ideal estimator

We begin by considering an ideal observer. We assume that spike rates are sufficiently high that the likelihood for the observed variable θ, given the spikes observed in a short time window Δt, can be approximated as a Gaussian variable with variance $1/(J\Delta t)$. Within the scope of this approximation, the Fisher information rate J and the diffusion coefficient $\mathcal{D}$ encompass all that we need to know about the coding properties of the network. We further assume below that $\mathcal{D}$ is a constant during the readout interval. This assumption is approximately correct (even for state-dependent $\mathcal{D}$), if the readout interval is short. The assumption of a short readout interval is self-consistent, due to our result at the end of this section, that readout performance saturates within the single-neuron time-constant of τ.

To simplify the derivation we model this scenario in discrete time. Later on, we take the limit where the discretization time step $\Delta t \to 0$. The dynamic equation for θ is written as

$$\theta_i = \theta_{i-1} + \xi_i \,, \tag{S58}$$

where

$$\langle \xi_i \xi_j \rangle = \sigma_1^2 \delta_{ij} = 2\mathcal{D}\Delta t \delta_{ij} \,, \tag{S59}$$

and we imagine that the estimator has access to noisy observations of θ,

$$y_i = \theta_i + \zeta_i \,, \tag{S60}$$

where

$$\langle \zeta_i \zeta_j \rangle = \sigma_2^2 \delta_{ij} = \frac{\delta_{ij}}{J\Delta t} \,. \tag{S61}$$

The likelihood of a trajectory $\vec{x} = (x_0, x_1, \ldots, x_n)$ based on observed spikes is

$$P(\vec{x}|\text{spikes}) \propto \int d\xi_1 \ldots d\xi_n \exp\left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^n \xi_i^2 - \frac{1}{2\sigma_2^2} \sum_{i=1}^n \left[y_i - x_0 - (A\vec{\xi})_i \right]^2 \right\} \,, \tag{S62}$$

where A is an n by n lower triangular matrix of the following form

$$A = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ & & & \ddots \end{pmatrix} \,. \tag{S63}$$

By marginalizing over all x_i for $i = 1 \ldots n$, we find that

$$p(x_0|\text{spikes}) \propto \exp\left\{ -\frac{1}{2} \vec{u}^{\mathrm{T}} \Xi \vec{u} \right\} , \tag{S64}$$

where

$$u_i = y_i - x_0 . \tag{S65}$$

The matrix

$$\Xi = \left[\sigma_2^2 I + \sigma_1^2 A A^T \right]^{-1} , \tag{S66}$$

and I is the identity matrix, from which we deduce that the optimal estimate is

$$\langle x_0 \rangle = \frac{\vec{y}^{\mathrm{T}} \Xi \vec{1}}{\vec{1}^{\mathrm{T}} \Xi \vec{1}} , \tag{S67}$$

where $\vec{1}$ is a vector whose entries are all equal to 1. The variance of the estimate is

$$\mathrm{Var}(x_0) = \frac{1}{\vec{1}^T \Xi \vec{1}} . \tag{S68}$$

In the limit $\Delta t \to 0$, Ξ becomes the operator

$$\Xi = \frac{1}{\sigma_2^2} \left\{ 1 - \left[1 + \epsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2} \right]^{-1} \right\} \equiv \frac{\Xi_0}{\sigma_2^2} , \tag{S69}$$

where

$$\epsilon = \frac{1}{2 \mathcal{D} J} . \tag{S70}$$

It is straightforward to derive that the entries of the vector $\Xi_0 \vec{1}$ then obey the differential equation

$$f(t) + \epsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(t) = 0 . \tag{S71}$$

The boundary conditions are found by considering the first and last rows of the matrix AA^T:

$$f(0) = 1 \quad ; \quad \left. \frac{\mathrm{d}f(t)}{\mathrm{d}t} \right|_{t=\Delta T} = 0 . \tag{S72}$$

Therefore,

$$f(t) = \frac{\cosh\left(\frac{\Delta T - t}{\sqrt{\epsilon}} \right)}{\cosh\left(\frac{\Delta T}{\sqrt{\epsilon}} \right)} , \tag{S73}$$

and $\vec{1}^{\mathrm{T}} \Xi \vec{1}$ becomes

$$J \int_0^T f(t) \mathrm{d}t = \sqrt{\frac{J}{2D}} \tanh\left(\sqrt{2DJ} \Delta T \right) . \tag{S74}$$

Hence,

$$V_{\text{sp}} = \sqrt{\frac{2\mathcal{D}}{J}}\coth\left(\sqrt{2\mathcal{D}J}\Delta T\right) . \tag{S75}$$

We thus see that the variance of the ideal estimator saturates at $\sqrt{2\mathcal{D}/J}$ after a characteristic time $\Delta T^* = 1/\sqrt{2\mathcal{D}J}$. Using the information-diffusion inequality, $\Delta T^* \leq \tau$.

We note that this ideal estimator is formally equivalent to the reverse Kalman filter [9] applied to the problem of inferring the state at the beginning of the readout period, given all subsequent spikes over the readout interval.

Naive estimator

We next consider a naive estimator that is tailored for the case where there is no diffusion. Under these conditions, the best estimate is simply an average over all observations:

$$\hat{\theta}_{\text{naive}} = \frac{1}{n}\sum_{i=1}^{n} y_i . \tag{S76}$$

Qualitatively, we expect the variance of this estimator to include two terms: a term coming from the noisiness of spikes, proportional to $1/(J\Delta T)$, and a term coming from diffusion of the instantaneous state, proportional to $\mathcal{D}\Delta T$. Therefore the balance between these two terms should lead to an optimal readout time $\Delta T \propto 1/\sqrt{\mathcal{D}J}$. The calculation below confirms this intuition, and provides the precise prefactors for the two terms.

Let us assume, without loss generality that $\theta(0) = 0$. The variance of the estimator is thus given by

$$V_{\text{sp,naive}} = \frac{1}{n^2}\left\langle\sum_{i=1}^{n} y_i^2\right\rangle = \frac{1}{n^2}\left\langle\sum_{i=1}^{n} \theta_i^2\right\rangle + \frac{1}{J\Delta T} , \tag{S77}$$

where we used the relation: $n\Delta t = \Delta T$. We next make use of the fact that

$$\langle\theta_i\theta_j\rangle = \begin{cases} \langle\theta_i^2\rangle = 2\mathcal{D}\cdot i\Delta t & , \quad i \leq j \\ \langle\theta_j^2\rangle = 2\mathcal{D}\cdot j\Delta t & , \quad i > j \end{cases} \tag{S78}$$

to obtain an expression in the limit $\Delta t \to 0$:

$$V_{\text{sp,naive}}(\Delta T) = \frac{2}{3}D\Delta T + \frac{1}{J\Delta T} . \tag{S79}$$

We see that the variance is minimized when ΔT is equal to

$$\Delta T^*_{\text{naive}} = \sqrt{\frac{3}{2\mathcal{D}J}} \leq \sqrt{3}\tau , \tag{S80}$$
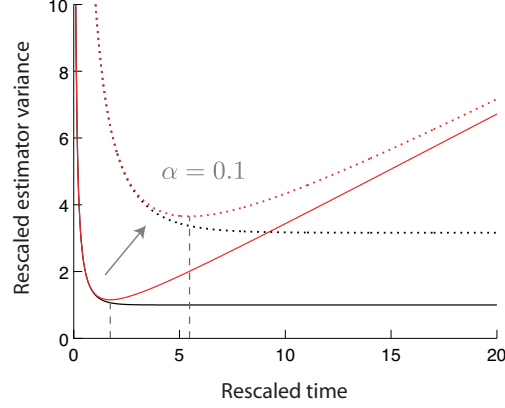
FIG. S2. **Time course of estimation error for incomplete readout.** Inference error V_s of estimators with full access to all spikes from the memory network (solid black line, ideal estimator and solid red line, naive estimator), compared with the inference error of similar estimators that have access to a fraction $\alpha = 0.1$ of the spikes (dotted black line, ideal estimator and dotted red line, naive estimator). Time T is rescaled by ΔT^* and variance V_s is rescaled by $\sqrt{2\mathcal{D}/J}$ for the case $\alpha = 1$. The ideal estimation interval ΔT of the naive estimator increases, in comparison to complete readout by a factor $1/\sqrt{\alpha} \simeq 3.16$, and the optimal variance increases by the same factor. Similiarly, the characteristic saturation time and saturated variance of the ideal estimator increase by the same factor. Note, however, that the variance of the naive estimator at large times is independent of α, since it is dominated by the ignored diffusion of the memory state.

and the estimator variance for $\Delta T = \Delta T^*_{\text{naive}}$ is

$$V_{\text{sp,naive}}(\Delta T^*_{\text{naive}}) = \frac{2}{\sqrt{3}} \sqrt{\frac{2\mathcal{D}}{J}} \,. \tag{S81}$$

Incomplete readout

We assumed so far that t gddhe estimators have access to all spikes emitted by neurons in the memory network. We can also imagine that only some of the neurons in the memory network project to the readout area, and therefore, only a fraction $\alpha < 1$ of spikes are available. In this case, J is replaced everywhere by αJ. The characteristic observation time increases by a factor $1/\sqrt{\alpha}$:

$$\Delta T^* = 1/\sqrt{2\alpha\mathcal{D}J} \tag{S82}$$

and the variance of the readout (for both estimators) degrades by the same factor. For the ideal estimator, and for $\Delta T \gtrsim \Delta T^*$,

$$\Delta\hat{\Theta} = V_d(T) + \sqrt{\frac{2\mathcal{D}}{\alpha J}} \geq V_d(T) + \frac{2\mathcal{D}\tau}{\sqrt{\alpha}} \,. \tag{S83}$$

Integration time constant

For continuous attractor networks that function as neural integrators, the neural noise reduces the accuracy of the integrator, but it does not shorten the integration time-constant: The integrated output of the network is statistically centered at the true integral of the past inputs, with the variance growing linearly with the duration of integration. Thus, the network does not tend towards a single fixed point, which would erase the effect of past inputs and is a signature of a shortened integration time-constant. In other words, although neural noise-driven diffusion limits the time-constant over which it is profitable for a readout network to accumulate information from the memory network (Sec. VI), it does not limit the time-constant of the network itself.

VII.   GENERALIZATIONS OF THE NEURAL MODEL

A model with two time scales

Neural processing may involve more than one time scale. To understand the interplay between different time scales in the information diffusion inequality, we consider a rate model that involves both a synaptic time constant and a membrane time constant. The model is described by the dynamic equation **10** of the main text, but instead of setting the firing rate to be a function of the instantaneous total synaptic current, as in Eq. **9**, the synaptic current is first averaged over a time scale τ_{m}:

$$\dot{r}_i + \frac{1}{\tau_m} r_i = \frac{1}{\tau_m} \phi(g_i) \tag{S84}$$

Note that the averaging occurs after the nonlinearity. We comment that temporal averaging can be introduced instead within the nonlinearity, and this generates the same results as below, in particular, Eq. (S91). For simplicity, we assume in this section that the weight matrix is symmetric.

The attractor states obeys Eq. **1** as before, and $\bar{r}_i = \tau^{-1} \bar{s}_i$. The linearized dynamic equations around a state $(\bar{s}, \bar{r})$ take the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \delta s \\ \delta r \end{pmatrix} = \begin{pmatrix} -\tau^{-1} \delta_{ij} & \delta_{ij} \\ \tau_m^{-1} \phi'(\bar{g}_i) W_{ij} & \tau_m^{-1} \delta_{ij} \end{pmatrix} \begin{pmatrix} \delta s \\ \delta r \end{pmatrix} . \tag{S85}$$

From the existence of the attractor manifold we can deduce immediately that the right eigenvector of this operator has the structure

$$u = \frac{\mathrm{d}}{\mathrm{d}\theta} \begin{pmatrix} \bar{s} \\ \tau^{-1} \bar{s} \end{pmatrix} , \tag{S86}$$

where the neural indices are suppressed to simplify the notation.

To find the left eigenvector, we consider a second set of dynamic equations for synaptic and rate variables s', r':

$$\dot{s}'_i + \tau_m^{-1} s'_i = r'_i,$$
$$\dot{r}'_i + \tau^{-1} r'_i = \tau_m^{-1} \phi \left[\sum_j W_{ij} s'_j + b_i \right].$$

(S87)

This is similar to the dynamics of s and r, but note that τ and τ_m have been interchanged on the left hand sides of the two equations. It is straightforward to see that every attractor state $(\bar{s}, \bar{r})$ of the variables s, r maps to an attractor state $(\bar{s}', \bar{r}') = \left(\bar{s}, \tau_m^{-1} \bar{s} \right)$ of the variables s', r'. Next, define:

$$x_i = \sum_j W_{ij} r'_j \quad ; \quad y_i = \sum_j W_{ij} s'_j.$$

(S88)

The matrix describing the linearlized dynamics of (x, y) is the transpose of the matrix describing the linearized dynamics of (s, r). Therefore the left eigenvector v for (s, r) is the same as the right eigenvector for (x, y). Using this observation we find

$$v_i = \frac{1}{\alpha} \frac{\mathrm{d}}{\mathrm{d}\theta} \left(\frac{\sum_j W_{ij} \bar{r}'_j}{\sum_j W_{ij} \bar{s}'_j} \right) = \frac{1}{\alpha} \frac{\mathrm{d}}{\mathrm{d}\theta} \left(\frac{\tau_m^{-1} \bar{g}_i}{\bar{g}_i} \right),$$

where α is set such that the vector product of v and u is equal to unity,

$$\alpha = \tau \left(\tau_m^{-1} + \tau^{-1} \right) \sum_i \phi'(\bar{g}_i) \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta} \right)^2.$$

(S89)

As in the case with one time constant, we project the noise on the left eigenvector and find that the change in (s, r) along the attractor manifold is given by ζu where

$$\langle \zeta^2 \rangle = \frac{1}{(\tau_m \alpha)^2} \sum_i \phi(\bar{g}_i) \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta} \right)^2,$$

(S90)

from which we obtain

$$2\mathcal{D}(\theta) = \frac{\sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta} \right)^2 \phi \left[\bar{g}_i(\theta) \right]}{\left[(\tau + \tau_m) \sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta} \right)^2 \phi' \left[\bar{g}_i(\theta) \right] \right]^2}.$$

(S91)

The expresion for the Fisher information in terms of $\bar{g}$ (or $\bar{r}$) remains the same as in the case without a membrane time constant. Therefore, the information-diffusion inequality for this model takes the form

$$2\mathcal{D} \geq \frac{J^{-1}}{(\tau + \tau_m)^2}.$$

(S92)

White Gaussian noise

Consider Eq. **10** where ξ is white, normal-distributed Gaussian noise with variance

$$\langle \xi(\theta, t)\xi(\theta', t')\rangle = \frac{2\pi\sigma^2}{N}\delta(t - t'). \tag{S93}$$

The parameter σ^2 sets the variance of the noise, and has dimensions of inverse time. Our primary goal in this calculation is not to to suggest a realistic neural model, but to investigate mathematically what happens if the Poisson noise is replaced by noise with some other statistics.

As before, we consider a short time window δt and evaluate the Fisher information. In contrast to the calculation for Poisson spiking neurons, we regard r as a continuous variable with Gaussian statistics. The Fisher information rate in r about the instantaneous state of the attractor is readily evaluated,

$$J = \sum_i \left(\frac{\mathrm{d}\bar{r}_i}{\mathrm{d}\theta}\right)^2 = \sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2 \phi'^2\left[\bar{g}_i(\theta)\right]. \tag{S94}$$

The coefficient of diffusion can be evaluated by following the steps outlined in Sec. II for Poisson spiking neurons. The only difference arises from the variance of the noise ξ:

$$2\mathcal{D} = \frac{\displaystyle\sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2}{\left[\displaystyle\sum_i \left(\frac{\mathrm{d}\bar{g}_i}{\mathrm{d}\theta}\right)^2 \phi'\left[\bar{g}_i(\theta)\right]\right]^2}. \tag{S95}$$

By comparing Eqs. (S94) and (S95) we see that $2\mathcal{D} = J^{-1}/\tau^2$ if ϕ is linear.

---

[1] Richard Feynman, Robert Leighton, and Matthew Sands, *The Feynman Lectures on Physics* (Addison Wesley, 1963).

[2] Rolf Landauer, "Inadequacy of entropy and entropy derivatives in characterizing the steady state," Phys. Rev. A **12**, 636–638 (1975).

[3] M. O. Magnasco and G. Stolovitzky, "Feynman's ratchet and pawl," J Stat Phys **93**, 615 (1998).

[4] N.G. VanKampen, "Diffusion in inhomogeneous media," Chemistry and Materials Science **68**, 135–138 (1987).

[5] N.G. VanKampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, 1981).

[6] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory* (Wiley & Sons, 1991).

[7] M A Paradiso, "A theory for the use of visual orientation information which exploits the columnar structure of striate cortex." Biol Cybern **58**, 35–49 (1988).

[8] H S Seung and H Sompolinsky, "Simple models for reading neuronal population codes." Proc Natl Acad Sci U S A **90**, 10749–10753 (1993).

[9] Ross H. Taplin, "The reverse kalman filter," Commun. Statist.– Theory methods **27**, 2547–2558 (1998).