Memory traces in dynamical systems

Surya Ganguli[a,b,1], Dongsung Huh[c], and Haim Sompolinsky[d,e]

[a]Sloan-Swartz Center for Theoretical Neurobiology, University of California, San Francisco, CA 94143; [b]Center for Theoretical Neuroscience, Columbia University, New York, NY 10032; [c]Computational Neurobiology Program, University of California at San Diego, La Jolla, CA 92093; [d]Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem 91904, Israel; and [e]Center for Brain Science, Harvard University, Cambridge, MA 02138

Communicated by David W. McLaughlin, New York University, New York, NY, October 3, 2008 (received for review April 3, 2008)

To perform nontrivial, real-time computations on a sensory input stream, biological systems must retain a short-term memory trace of their recent inputs. It has been proposed that generic high-dimensional dynamical systems could retain a memory trace for past inputs in their current state. This raises important questions about the fundamental limits of such memory traces and the properties required of dynamical systems to achieve these limits. We address these issues by applying Fisher information theory to dynamical systems driven by time-dependent signals corrupted by noise. We introduce the Fisher Memory Curve (FMC) as a measure of the signal-to-noise ratio (SNR) embedded in the dynamical state relative to the input SNR. The integrated FMC indicates the total memory capacity. We apply this theory to linear neuronal networks and show that the capacity of networks with normal connectivity matrices is exactly 1 and that of any network of N neurons is, at most, N. A nonnormal network achieving this bound is subject to stringent design constraints: It must have a hidden feedforward architecture that superlinearly amplifies its input for a time of order N, and the input connectivity must optimally match this architecture. The memory capacity of networks subject to saturating nonlinearities is further limited, and cannot exceed $\sqrt{N}$. This limit can be realized by feedforward structures with divergent fan out that distributes the signal across neurons, thereby avoiding saturation. We illustrate the generality of the theory by showing that memory in fluid systems can be sustained by transient nonnormal amplification due to convective instability or the onset of turbulence.

Fisher information | fluid mechanics | network dynamics

Critical cognitive phenomena such as planning and decision-making rely on the ability of the brain to hold information in short-term memory. It is thought that the neural substrate for such memory can arise from persistent patterns of neural activity, or attractors, that are stabilized through reverberating positive feedback, either at the single-cell (1) or network (2, 3) level. However, such simple attractor mechanisms are incapable of remembering sequences of past inputs.

More recent proposals (4–6) have suggested that an arbitrary recurrent network could store information about recent input sequences in its transient dynamics, even if the network does not have information-bearing attractor states. Downstream readout networks can then be trained to instantaneously extract relevant functions of the past input stream to guide future actions. A useful analogy (4) is the surface of a liquid. Even though this surface has no attractors, save the trivial one in which it is flat, transient ripples on the surface can nevertheless encode information about past objects that were thrown in.

This proposal raises a host of important theoretical questions. Are there any fundamental limits on the lifetimes of such transient memory traces? How do these limits depend on the size of the network? If fundamental limits exist, what types of networks are required to achieve them? How does the memory depend on the network topology, and are special topologies required for good performance? To what extent do these traces degrade in the presence of noise? Previous analytical work has addressed some of these questions under restricted assumptions about input statistics and network architectures (7). To answer these questions in a more general setting, we use Fisher information to construct a measure of memory traces in networks and other dynamical systems. Traditionally, Fisher information has been applied in theoretical neuroscience to quantify the accuracy of population coding of static stimuli (see, e.g., ref. 8). Here, we extend this theory by combining Fisher information with dynamics.

The Fisher Memory Matrix in a Neuronal Network

We study a discrete time network dynamics given by

$$\mathbf{x}_i(n) = f([\mathbf{W}\mathbf{x}(n-1)]_i + \mathbf{v}_i s(n) + z_i(n)), \quad i = 1 \ldots N. \qquad [1]$$

Here a scalar, time-dependent signal $s(n)$ drives a recurrent network of N neurons (Fig. 1*B*). $\mathbf{x}(n) \in \mathcal{R}^N$ is the network state at time n, $f(\cdot)$ is a general sigmoidal function, $\mathbf{W}$ is an $N \times N$ recurrent connectivity matrix, and $\mathbf{v}$ is a vector of feedforward connections from the signal into the network. We keep $\mathbf{v}$ time independent to focus on how purely temporal information in the signal is distributed in the N spatial degrees of freedom of the network state $\mathbf{x}(n)$. The norm $\|\mathbf{v}\|$ sets the scale of the network input, and we will choose it to be 1. The term $\mathbf{z}(n) \in \mathcal{R}^N$ denotes a zero mean Gaussian white noise with covariance $\langle \mathbf{z}_i(k_1)\mathbf{z}_j(k_2)\rangle = \varepsilon\delta_{k_1,k_2}\delta_{i,j}$.

We build upon the theory of Fisher information to construct useful measures of the efficiency with which the network state $\mathbf{x}(n)$ encodes the history of the signal. Because of the noise in the system, a given past signal history $\{s(n-k)|k \geq 0\}$induces a conditional probability distribution $P(\mathbf{x}(n)|\mathbf{s})$ on the network's current state. Here, we think of this history $\{s(n-k)|k \geq 0\}$as a temporal vector $\mathbf{s}$ whose kth component $\mathbf{s}_k$ is $s(n-k)$. The Fisher memory matrix (FMM) between the present state $\mathbf{x}(n)$ and the past signal is then defined as

$$\mathbf{J}_{k,l}(\mathbf{s}) = \left\langle -\frac{\partial^2}{\partial_{\mathbf{s}_k}\partial_{\mathbf{s}_l}} \log P(\mathbf{x}(n)|\mathbf{s}) \right\rangle_{P(\mathbf{x}(n)|\mathbf{s})}. \qquad [2]$$

This matrix captures [see supporting information (SI) *Appendix*] how much the conditional distribution $P(\mathbf{x}(n)|\mathbf{s})$ changes when the signal history $\mathbf{s}$ changes (Fig. 1). Specifically, if one were to perturb the signal slightly from $\mathbf{s}$ to $\mathbf{s} + \delta\mathbf{s}$, the Kullback Leibler divergence between the 2 induced distributions $P(\mathbf{x}(n)|\mathbf{s})$ and $P(\mathbf{x}(n)|\mathbf{s} + \delta\mathbf{s})$ would be approximated by $(\frac{1}{2})\delta\mathbf{s}^T\mathbf{J}(\mathbf{s})\delta\mathbf{s}$ (*SI Appendix*). Thus the FMM (Eq. **2**) measures memory through the ability of the past signal to perturb the network's present state. In this work, we will focus on the diagonal elements of the FMM. Each diagonal element $J(k) \equiv \mathbf{J}_{k,k}$ is the Fisher information that $\mathbf{x}(n)$ retains about a pulse entering the network at k time steps in the past. Thus, the diagonal captures the decay of the memory

Author contributions: S.G., D.H., and H.S. performed research; and S.G. and H.S. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: surya@phy.ucsf.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804451105/DCSupplemental.

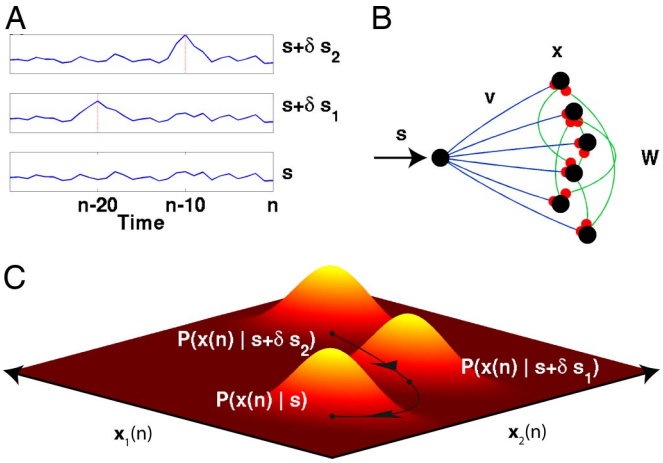

Fig. 1. Conversion of temporal to spatial information. (*A*) Three scalar signals: a base signal, $s(k)$, and 2 more signals obtained by perturbing s by the addition of an identical pulse centered at time $n - 10$ and $n - 20$. (*B*) Each of these signals is fed to a recurrent network $\mathbf{W}$ through a feedforward connectivity $\mathbf{v}$. (*C*) At time n, the temporal structure of each signal is encoded in the spatial distribution of the network state $\mathbf{x}(n)$, here shown in 2 dimensions. A stronger memory trace for the recent input perturbation $\delta\mathbf{s}_2$, relative to the remote perturbation $\delta\mathbf{s}_1$, is reflected by the larger difference between $P(\mathbf{x}(n)|\mathbf{s} + \delta\mathbf{s}_2)$ and $P(\mathbf{x}(n)|\mathbf{s})$ relative to that between $P(\mathbf{x}(n)|\mathbf{s} + \delta\mathbf{s}_1)$ and $P(\mathbf{x}(n)|\mathbf{s})$. As both perturbations recede into the past, both memory traces decay, and the 3 distributions become identical.

trace of a past input, and so we call $J(k)$ the Fisher memory curve (FMC).

For a general nonlinear system, the FMC depends on the signal itself and is hard to analyze. In this article, we focus on linear dynamics where the transfer function in Eq. **1** is defined by $f(x) = x$. Because the noise is Gaussian, the conditional distribution $P(\mathrm{x}(n)|\mathbf{s})$ is also Gaussian, with a mean that is linearly dependent on the signal $\delta\mathbf{x}(n)/\delta s(n-k) = \mathbf{W}^k\mathbf{v}$, and a noise covariance matrix $\mathbf{C}_n = \varepsilon \sum_{k=0}^{\infty} \mathbf{W}^k\mathbf{W}^{kT}$, which is independent of the signal. Hence, the FMC is independent of signal history and takes the form

$$J(k) = \mathbf{v}^T\, \mathbf{W}^{kT}\, \mathbf{C}_n^{-1}\, \mathbf{W}^k\, \mathbf{v}. \qquad [3]$$

We focus on two related features of Eq. **3**: the form of its dependence on the time lag k and the total area under the FMC, denoted by J_{tot}. An important parameter is the SNR in the input vector $\mathbf{v}s(n) + \mathbf{z}(n)$ at a single time n, which is $\frac{1}{\varepsilon}$. Because $J(k)$ depends on this input SNR only through the multiplicative factor $\frac{1}{\varepsilon}$, we will henceforth express $J(k)$ in units of $\frac{1}{\varepsilon}$. In these units, $J(k)$ is the fraction of the input SNR remaining in the system k time steps after an input pulse, and J_{tot} is the total SNR in the system state $\mathbf{x}(n)$ about the entire past signal history, relative to the SNR of the instantaneous input.

FMCs for Normal Networks

In the following, we uncover a fundamental dichotomy in the memory properties of two different classes of networks: normal and nonnormal. We first focus on the class of normal networks, defined as having a normal connectivity matrix $\mathbf{W}$. A matrix $\mathbf{W}$ is normal if it has an orthogonal basis of eigenvectors or equivalently commutes with its transpose. For normal networks, the relationship between the connectivity and the FMC simplifies considerably. Denoting the eigenvalues of $\mathbf{W}$ by λ_i, the FMC reduces to

$$J(k) = \sum_{i=1}^{N} v_i^2 |\lambda_i|^{2k} (1 - |\lambda_i|^2), \qquad [4]$$

where v_i is the projection of the input connectivity vector $\mathbf{v}$ on the ith eigenmode. Thus, for normal matrices, the orthogonal eigenvectors do not yield any essential contribution to memory performance.

First we note that summing Eq. **4** over k yields the important sum rule for normal networks,

$$J_{\mathrm{tot}} \equiv \sum_{k=0}^{\infty} J(k) = 1, \qquad [5]$$

which is independent of the network connectivity $\mathbf{W}$ and $\mathbf{v}$. This sum rule implies that normal networks cannot change the total SNR relative to that embedded in the instantaneous input but can only redistribute it across time. Whereas in the input vector, the SNR $\frac{1}{\varepsilon}$ is concentrated fully in the immediate signal $s(n)$, in the network state $\mathbf{x}(n)$, the dynamics has spread this information across time. This implies a tradeoff in memory performance for normal networks; if one attempts to optimize $\mathbf{W}$ or $\mathbf{v}$ to remember inputs occurring in the recent past, then one will necessarily take a performance loss in the ability to remember inputs occurring in the remote past and vice versa. The way different networks balance this tradeoff is reflected in the form of the time dependence of their FMC.

The reduction of the FMC to eigenvalues allows us to understand its asymptotics. For large k, the decay of the FMC in Eq. **4** is determined by the distribution of magnitudes of the largest eigenvalues. Dynamic stability requires that the largest eigenvalue magnitude, denoted by $\sqrt{\alpha}$, is less than 1. If α is a finite distance from 1, then the FMC for large k is dominated by this single eigenvalue, and it decays exponentially as $J(k) \propto \alpha^k$. However, if α is close to 1, then the multiplicity of long time scales associated with the large eigenvalues at the border of instability induces a power-law decay of the FMC. Specifically, if the density of eigenvalue magnitudes $\rho(r)$ near the edge of the spectrum behaves as $\rho(r) \propto (\sqrt{\alpha} - r)^{\nu}$, then for large k and α close to 1, the FMC decays algebraically (see *SI Appendix*) as

$$J(k) \propto k^{-(\nu+2)}, \quad k \gg 1. \qquad [6]$$

Note that because $\nu > -1$, the integral of Eq. **6** remains finite, consistent with the sum rule (Eq. **5**).

Examples of Normal Networks

An important class of normal matrices includes translation invariant lattices, i.e., circulant matrices. In the 1D case, $\mathbf{W}$ is of the form $\mathbf{W}_{ij} = \mathbf{d}_{(i-j)} \bmod N$, where $\mathbf{d}$ is any vector, and the eigenvectors of $\mathbf{W}$ are the Fourier modes. The signal enters at a single neuron so that $\mathbf{v}_k = \delta_{k,0}$ and couples to all of the modes with a uniform strength $1/\sqrt{N}$. An important special case is the delay ring (Fig. 2*A*), with $\mathbf{d}_k = \sqrt{\alpha}\delta_{k,1}$. Its FMC is $J(k) = \alpha^k(1 - \alpha)$. For any value of α, the FMC always displays exponential rather than power-law decay. This occurs because the eigenvalues of $\mathbf{W}$ all lie on a circle of radius $\sqrt{\alpha}$. Thus $\rho(r) = \delta(r - \sqrt{\alpha})$, and there is no continuous spectrum near the boundary of instability. Instead, there is only 1 time constant governed by α. Extensions of the delay ring are the ensemble of orthogonal networks (studied in ref. 7). These are normal networks in which $\mathbf{W}$ is a rotation matrix.

Another class of normal networks consists of networks with symmetric connectivity matrices. An example is a symmetric d-dimensional lattice (a 2D example is shown in Fig. 2*B*). Near the edge of the eigenvalue spectrum $\rho(r) \propto (\sqrt{\alpha} - r)^{(d-2)/2}$. Hence, for $\alpha \to 1$, FMC exhibits a power-law decay with

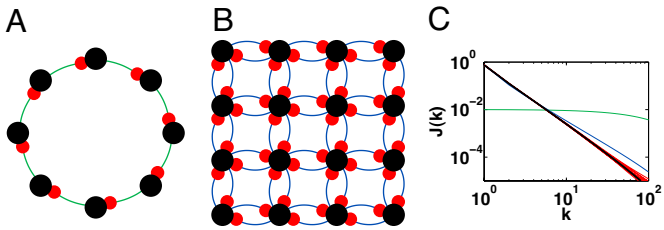

Fig. 2. Normal networks. (*A*) A delay ring. (*B*) A 2D symmetric invariant lattice (periodic boundary conditions not shown). (*C*) The FMC for the delay ring (green, $\alpha = 0.99$, $N = 1{,}000$), the 2D Lattice (blue, $\alpha = 0.99$, $N = 1{,}024$), and 10 random symmetric matrices (red, $\alpha = 0.99$, $N = 1{,}000$). The black trace is the analytic prediction of a mean field theory for the FMC of random symmetric matrices, derived in *SI Appendix*.

exponent $-(d + 2)/2$ (Fig. 2*C*). Finally, we consider large random symmetric networks defined by matrix elements $\mathbf{W}_{ij} = \mathbf{W}_{ji}$ that are chosen independently from a zero mean Gaussian distribution with variance $\alpha/4N$. The eigenvalues of $\mathbf{W}$ are distributed on the real axis r according to Wigner's semicircular law, which, near the edge, behaves as $\rho(r) \propto (\sqrt{\alpha} - r)^{1/2}$. Hence, Eq. **6** predicts a power-law decay of the FMC for $\alpha \to 1$, with exponent $-5/2$ as verified in Fig. 2*C*.

Preferred Input Patterns in Nonnormal Networks

For nonnormal networks, J_{tot} depends not only on the network connectivity $\mathbf{W}$ but also on the feedforward connectivity $\mathbf{v}$. To investigate the sensitivity to $\mathbf{v}$, we note from Eq. **3** that, in general, J_{tot} can be expressed as

$$J_{\text{tot}} = \mathbf{v}^T \mathbf{J}^s \mathbf{v}, \qquad [7]$$

where we have introduced the spatial FMM

$$\mathbf{J}^s_{ij} = \sum_{k=0}^{\infty} [\mathbf{W}^{kT} \mathbf{C}_n^{-1} \mathbf{W}^k]_{i,j}. \qquad [8]$$

This matrix, $\mathbf{J}^s$, and the temporal FMM $\mathbf{J}$ in Eq. **2**, can be unified into a general space–time framework (see *SI Appendix*). $\mathbf{J}^s$ measures the information in the network's spatial degrees of freedom $\mathbf{x}_i(n)$ about the entire signal history. The total information in all N degrees of freedom is $\text{Tr}\, \mathbf{J}^s = N$, independent of $\mathbf{W}$. Because $\mathbf{J}^s$ is positive definite with trace N, Eq. **7** yields a fundamental bound on the total area under the FMC of any network $\mathbf{W}$ and unit input vector $\mathbf{v}$:

$$J_{\text{tot}} \leq N. \qquad [9]$$

If $\mathbf{W}$ is normal, then $\mathbf{J}^s_{i,j} = \delta_{i,j}$, implying that all directions in space provide the same amount of total temporal information, and so J_{tot} is independent of the spatial structure $\mathbf{v}$ of the input, consistent with the sum rule (Eq. **5**). However, if $\mathbf{W}$ is nonnormal, $\mathbf{J}^s$ has nontrivial spatial structure, reflecting an inherent anisotropy in state space induced by the connectivity matrix $\mathbf{W}$. There will be preferred directions in state space, corresponding to the large principal components of $\mathbf{J}^s$, that contain a large amount of information about the total history, whereas other directions will perform relatively poorly. The choice of $\mathbf{v}$ that maximizes J_{tot} is the eigenvector of largest eigenvalue of $\mathbf{J}^s$.

The spatial anisotropy of nonnormal networks is demonstrated by evaluating the FMC for random asymmetric networks, where each matrix element $\mathbf{W}_{ij}$ is chosen independently from a zero mean Gaussian with variance α/N. If the feedforward connectivity $\mathbf{v}$ is chosen to be a random vector, the distribution of J_{tot} (Fig. 3*A*, blue) is centered around 1 as expected, because the trace of $\mathbf{J}^s$ equals N. However, if $\mathbf{v}$ is chosen as the maximal principal component of $\mathbf{J}^s$, the resultant J_{tot} is approximately 4 times as large (Fig. 3*A*, red).

Additional insight into the structure of the preferred $\mathbf{v}$ comes from the Schur decomposition of $\mathbf{W}$. Whereas every normal matrix is unitarily equivalent to a diagonal matrix (Fig. 3*B Upper*), every nonnormal matrix is unitarily equivalent to an upper triangular matrix (Fig. 3*B Lower*). On this basis, it may, in general, be preferable to distribute the signal near the beginning of the network to counterbalance the noise propagation along the network. We have tested this hypothesis by plotting the magnitude of the components of the optimal input vector for the random asymmetric networks in their Schur basis. As Fig. 3*C* shows, the optimal choice of feedforward weights $\mathbf{v}$ does indeed exploit the hidden feedforward structure by coupling the signal more strongly to its source than to its sink.

Transient Amplification and Extensive Memory

Comparing Eqs. **5** and **9** motivates defining networks with extensive memory as networks in which J_{tot} is proportional to N for large N. With this definition, normal networks do not have extensive memory. Furthermore, as indicated in Fig. 3*A*, despite the enhanced performance of generic asymmetric networks, their total memory remains $O(1)$, prompting the question whether in fact there exist nonnormal networks with extensive memory. Surprisingly, such networks do exist.

A particularly simple example is the delay line shown in Fig. 4*A Upper*). In this example, the only nonzero matrix elements are $\mathbf{W}_{i+1,i} = \sqrt{\alpha}$ for $i = 1 \ldots N - 1$. The FMC depends on how the signal enters the delay line. The optimal choice is to place the signal at the source, so that $\mathbf{v}_i = \delta_{i,1}$. Then the FMC takes the form $J(k) = \alpha^k(1 - \alpha)/(1 - \alpha^{k+1})$, $k = 0 \ldots N - 1$, and 0 otherwise. For values of $\alpha < 1$, J decays exponentially as α^k and

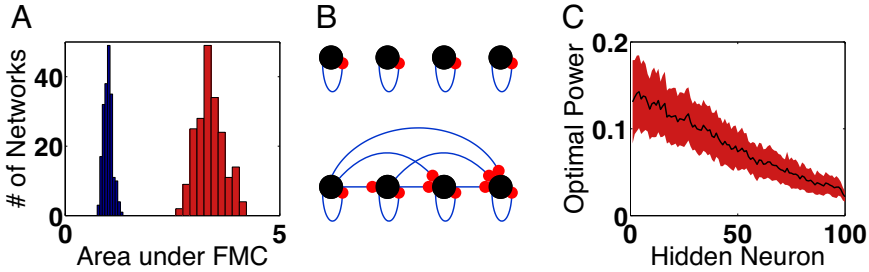

Fig. 3. Matching input connectivity to nonnormal architectures. (*A*) Histogram of J_{tot} for 200 random Gaussian matrices with $N = 100$, and $\alpha = 0.99$, for random (blue) and optimized (red) input weights $\mathbf{v}$. (*B*) (*Upper*) Every normal matrix can be diagonalized by a unitary matrix, and so its memory properties are equivalent to a set of N disconnected neurons each exerting positive feedback on itself. (*Lower*) A nonnormal matrix can only be converted to an upper triangular matrix through a unitary transformation and thus has a hidden feedforward architecture. (*C*) For the same 200 matrices in *B*, the mean (black) and standard deviation (red) of the magnitude of the components of the optimal input vector $\mathbf{v}$, in the Schur basis, ordered according to the hidden feedforward structure.

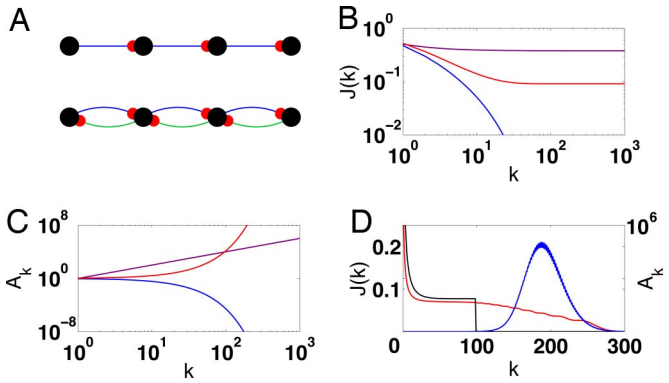

Fig. 4. Memory from transient amplification. (*A*) Delay line architecture (*Upper*) and a delay line with feedback (*Lower*). (*B*) The FMC for 3 delay lines of length $N = 10^3$ whose corresponding signal amplification profiles are shown in the same color in *C*. The red curves correspond to exponential amplification ($A_k = \alpha^k$, $\alpha = 1.1$), whereas the blue curves correspond to exponential decay ($A_k = \alpha^k$, $\alpha = 0.9$). The magenta curves correspond to power law amplification in an inhomogeneous delay line ($A_k = k^2$). (*D*) The signal amplification profile (blue) and corresponding FMC (red) for a delay line with feedback of length $n = 100$ in the transient amplification regime $\sqrt{\alpha}$ and $\sqrt{\beta} = 0.2$. The black curve shows the FMC of a delay line of length N with the same signal amplification profile up to time $N - 1$.

$J_{\text{tot}} \approx 1$ (Fig. 4*B*, blue). However, for $\alpha > 1, J$ saturates to a finite value, $1 - \frac{1}{\alpha}$ for large $k < N$, so that $J_{\text{tot}} \approx N\,(1 - \frac{1}{\alpha})$ (Fig. 4*B*, red).

The delay line with extensive memory is an example of a dynamical system with strong transient amplification. Network amplification can be characterized by the behavior of $A_k \equiv \|\mathbf{W}^k\mathbf{v}\|^2$ for $k \geq 0$. Whereas in normal systems, A_k is montonically decreasing for all $\mathbf{v}$, in nonnormal networks, A_k may initially increase before decaying to zero for large k (9). In the case of Fig. 4*C* (red), $A_k = \alpha^k$ increases exponentially, and this amplification lasts for a time of order N. It is important to note that, in such a system, not only the signal but also the noise is exponentially amplified as it propagates along the chain. Introducing the signal at the beginning of the chain guarantees that the signal and noise are amplified equally, resulting in the saturation of $J(k)$ (Fig. 4*B*).

It is not necessary to have purely feedforward connectivity to have large transient amplification and extensive memory. As an example, we consider a delay line with feedback (Fig. 4*A Lower*). In addition to the feedforward connections, $\sqrt{\alpha}$, there are feedback connections, $\mathbf{W}_{i,i+1} = \sqrt{\beta}$, for $i = 1 \ldots N - 1$. We consider a scenario in which $\sqrt{\alpha} < 1$, so that in the absence of the feedback, inputs would decay exponentially. If the feedback is in the range $1 - \sqrt{\alpha} < \sqrt{\beta} < 1/(4\sqrt{\alpha})$, then the system exhibits transient exponential amplification while maintaining global stability (9). Fig. 4*D* shows an example of the amplification and the extensive FMC in this regime. One advantage of the feedback is that the amplification as well as the tail of the FMC lasts longer than N, which cannot be achieved in a delay line of length N without feedback (Fig. 4*D*, black curve).

Transient exponential amplification, as in the above examples, is not a necessary condition for extensive memory. Consider, for example, a delay line with inhomogeneous weights, $\mathbf{W}_{i+1,i} = \sqrt{\alpha_i}$ for $i = 1 \ldots N - 1$, where $A_k = \Pi_{p=1}^{k}\alpha_p$ for $0 < \mathrm{k} < N$ and $A_0 = 1$. The FMC equals

$$J_{\text{delay}}(k) = \frac{1}{\sum_{m=0}^{k} A_m^{-1}}, \quad 0 \leq k \leq N - 1. \qquad [10]$$

From Eq. **10**, it is evident that J_{delay} saturates to a finite value at large k as long as A_k increases superlinearly in k, i.e., $A_k > O(k)$, as shown in Fig. 4 *B* and *C* (magenta). If this superlinear amplification lasts a time of order N, then J_{tot} will be extensive. The above results raise the question whether superlinear transient amplification lasting for a time of order N is a necessary prerequisite for extensive memory in general nonnormal networks with feedback. Interestingly, we have proven (see *SI Appendix*) that this is, indeed, a necessary condition for a general network. Specifically, we have shown that for any network with a given sequence of signal amplification $A_m, 0 \leq m \leq k$, the FMC up to time k cannot be larger than that of a delay line of length $k + 1$, with input vector $\mathbf{v}_i = \delta_{i,1}$, that possesses the same set of amplification factors, i.e., $\mathbf{W}_{i+1,i} = \sqrt{A_i/A_{i-1}}$ for $i = 1 \ldots k$. Thus for any network,

$$J(m) \leq J_{\text{delay}}(m) \quad \text{for } m = 0, \ldots, k, \qquad [11]$$

where $J_{\text{delay}}(\mathrm{m})$ is given by Eq. **10**. We have further shown (see *SI Appendix*) that, remarkably, in the space of networks with a given signal amplification profile, the only networks that saturate the bound Eq. **11** are those that are unitarily equivalent to the corresponding delay line, with the signal placed at the source. Thus the delay line is essentially the unique network that achieves the minimal possible noise amplification for a given amount of signal amplification. Therefore, the strong inequality (Eq. **11**) reveals that, in general architectures, noise undergoes stronger amplification than it otherwise would in the corresponding delay line. However, the length of the delay line necessary to realize amplification up to k time steps is exactly $k + 1$, so that for $k > N$, the number of neurons in the corresponding delay line is larger than that of the actual network with feedback, as noted in the example of Fig. 4*D*.

Consequences of Finite Dynamic Range

The networks discussed above achieve extensive memory performance through transient superlinear amplification that lasts for $O(N)$ time steps. However, such amplification may not be biophysically feasible for neurons that operate in a limited dynamic range, due, for example, to saturating nonlinearities. This raises the question, what are the limits of memory capacity for networks with saturating neurons? To address this question, we assume that the network architecture is such that all neurons have finite dynamic range, i.e., $\langle \mathbf{x}_i(n) \rangle^2 < R$ for $i = 1, \ldots, N$. We show (see *SI Appendix*) that in this case,

$$J(k) \leq \frac{1}{1 + \dfrac{k(k+1)}{2NR}} \quad \text{for all } k \geq 0. \qquad [12]$$

This bound implies that such a network cannot achieve an area under the FMC that is larger than $O(\sqrt{N})$ and, in particular, cannot achieve extensive memory.

Can a network of neurons with finite dynamic range achieve the $O(\sqrt{N})$ limit? To do so, a network must distribute the signal among many neurons so that as the distributed signal is amplified, the local input to any individual neuron does not grow. An example of such a network is the divergent fan-out architecture shown in Fig. 5*A*. It consists of L layers where the number of neurons N_k in layer k grows with k. The signal enters the first layer and, for simplicity, the connections between neurons in layer k to those in $k + 1$ are all equal to $\sqrt{\gamma_k}$. We show in *SI Appendix* that if N_k grows linearly in k, and $\sqrt{\gamma_k}$ decreases inversely with k, then as the overall signal propagates through the layers, it is amplified linearly, whereas single-neuron activities neither grow nor decay. Memory traces in such a network last a time proportional to the depth L, but because the number of neurons N is $O(L^2)$, in terms of neurons, the area under the FMC is $O(\sqrt{N})$, which is the limit.

A comparison between the performance of the fan-out architecture and a random Gaussian network of the same size, $N \approx$

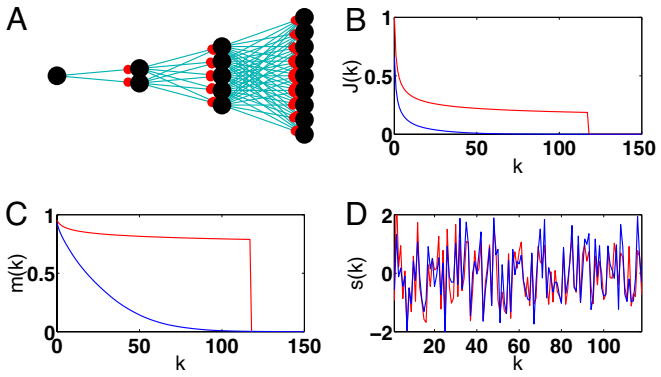

Fig. 5. Memory in fan-out vs. generic architectures. (*A*) A feedforward, fan-out architecture, or divergent chain. (*B*) The FMC of the divergent chain (red) and a random Gaussian network (blue) of comparable size. (*C*) The reconstruction performance of the divergent chain (red) and Gaussian network (blue). $m(k)$ is defined to be the average correlation between the actual past input $s(n-k)$ and an optimal estimate of this input, constructed from the current network state $\mathbf{x}(n)$ (7); see also *SI Appendix*. (*D*) An example of the actual input sequence (red) and its optimal linear reconstruction (blue) from the final state of a nonlinear divergent chain.

7,000, is shown in Fig. 5 *B* and *C*. The first network consists of $n = 7{,}021$ neurons organized in a divergent chain of length $L = 118$ with the number of neurons at each layer growing as $N_k = k$ and the connection strengths are $\sqrt{\gamma_k} = \frac{1}{k}$. The Gaussian network consists of 7,000 neurons with a Gaussian connectivity matrix; the square magnitude of its maximal eigenvalue is $\alpha = 0.95$. Fig. 5*B* shows the marked difference between the FMC of the two systems. The enhanced FMC of the chain translates into a better reconstruction of the signal (see *Discussion*). To demonstrate this relation, we have computed for the two systems the correlation coefficients between a white noise input with SNR $\frac{1}{\varepsilon} = 20$ and the estimated signal using an optimal readout of the network state. In the case of the divergent chain the optimal estimate of the signal $s(n-k)$ is the summed activity of the neurons at layer $k+1$. Fig. 5*C* shows the vastly improved signal reconstruction of the divergent chain.

Finally, to test the robustness of the fan-out architecture to saturation, we have simulated the dynamics of Eq. **1** with a saturating nonlinear transfer function $f(x) = \tanh(x)$ (see *SI Appendix*). As before, the input is white noise with SNR of 20. A sample of the signal and its reconstruction from the layers' activity is shown in Fig. 5*D*. The correlation coefficient of the 2 traces, roughly 0.8, is in accord with the theoretical prediction of the linear system, Fig. 5*C*. Thus, the fan-out architecture achieves impressive memory capacity by distributed amplification of the signal across neurons without a significant amplification of the input to individual neurons.

Nonnormal Amplification and Memory in Fluid Dynamics

To illustrate the generality of the connection between transient, nonnormal amplification and memory performance, we consider an example from fluid mechanics. Indeed, nonnormal dynamics is thought to play an important role in various fluid mechanical transitions, including the transition from certain laminar flows to turbulence (10). Here, we focus on a particular type of local instability known as a convective instability (11) that plays a role in describing fluid flow perturbations around wakes, mixing and boundary layers, and jets. For example, the fluid flow just behind the wake of an object, or in the vicinity of a mixing layer where two fluids at different velocities meet, is especially sensitive to perturbations, which transiently amplify but then decay away as they are convected away from the object or along the mixing layer as the two velocities equalize.

Following refs. 9 and 11, we model these situations phenomenologically through the time evolution of a 1D flow perturbation $u(x,t)$ obeying the linear evolution operator

$$\partial_t u = h^2 \partial_x^2 u - h\,\partial_x u + \left(\frac{1}{4} - x^2\right)u + v(x)s(t) + \eta. \qquad [13]$$

This describes rightward drift plus diffusion in the presence of a quadratic feedback potential (Fig. 6*A*) driven by a 1D signal $s(t)$ and zero mean, unit variance additive white Gaussian noise in time and space, $\eta(x,t)$. Perturbations in the region $|x| \leq \frac{1}{2}$ receive positive feedback and are exponentially amplified. However, the system is still globally stable, because these perturbations convect downstream and enter a region of exponential decay for $x > \frac{1}{2}$. The time spent by any perturbation in the amplification region is $O(1/h)$, and thus the total transient amplification will be $O(e^{1/h})$. Thus, for small h, significant nonnormal amplification occurs.

Although the sum rule (Eq. **5**) holds also for continuous time (see *SI Appendix*) and discrete space, there is no analogous bound (Eq. **9**) for continuous space because the number of degrees of freedom is infinite. Nevertheless, for a given system, J_{tot} is finite and is bounded by the amplification time, or equivalently by the effective number of amplified degrees of freedom, which, in our case, is $O(1/h)$.

The optimal way for the signal to enter the network, i.e., the first principal eigenvector of $\mathbf{J}^s$, is shown in Fig. 6*A*. This optimal input profile $v(x)$ is a wave packet poised to travel through the convective instability (Fig. 6*B*). Fig. 6*C* (red) shows that the optimal memory performance, or maximal eigenvalue of $\mathbf{J}^s$, scales linearly with $1/h$. Thus, consistent with the results above, the maximal area under the FMC is proportional to the time over which inputs are superlinearly amplified. For comparison, we have computed the value of the optimal J_{tot} in the case of a fluid dynamics that contains only diffusion and drift, the first 2 terms of the right-hand side of Eq. **13** but not the amplifying potential. In this case, J_{tot} is low for all values of h (Fig. 6*C*, blue).

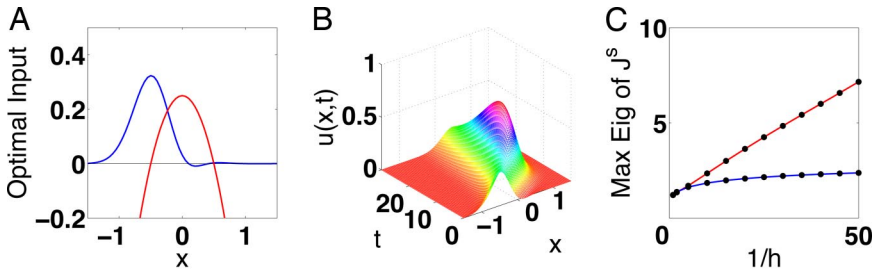

Fig. 6. Memory through convective instability. (*A*) The optimal input profile (blue) and quadratic potential (red). (*B*) The rightward time evolution and transient amplification of this input ($h = 0.1$). (*C*) Optimal memory in the presence (red) and absence (blue) of the convective instability.

Discussion

In this work, we focused on the diagonal part of the FMM (Eq. **2**. In networks that are unitarily equivalent to simple delay lines (e.g., Fig. 4*A Upper* and Fig. 5*A*), this matrix is diagonal. However, in general, the off-diagonal elements are not all zero. Their value reflects the interference between two signals injected into the system at two different times, and their analysis provides an interesting probe into the topology of (partially directed) loops in the system, which give rise to such interference (see *SI Appendix*).

It is interesting to note the relation between the FMC $J(k)$ and the more conventional memory function $m(k)$ defined through the correlation between an optimal estimate of the past signal $\hat{s}(n-k)$ based on the network state $\mathbf{x}(n)$ and the original signal $s(n-k)$ (see Fig. 5*C*). Even in the linear version of Eq. **1** studied here, $m(k)$ depends on the full FMM. Furthermore, it depends in a complex manner on the signal statistics (see *SI Appendix*), whereas Fisher information is local in signal space and in the present case is, in fact, independent of the signal except for an overall factor of the input SNR. Both features render signal reconstruction a much more complex measure to study analytically. Nevertheless, it is important to note that the FMC measures the SNR embedded in the network state, relative to the input SNR $\frac{1}{\varepsilon}$. Hence, for small-input SNR, high Fisher memory is crucial for accurate signal reconstruction. On the other hand, when $\frac{1}{\varepsilon}$ is sufficiently large, $m(k)$ may be close to 1 even for low Fisher information (see *SI Appendix*).

Our results indicate that generic recurrent neuronal networks are poorly suited for the storage of long-lived memory traces, contrary to previous proposals (4–6). In systems with substantial noise, only networks with strong and long-lasting signal amplification can potentially sustain such traces. However, signal amplification necessarily comes at the expense of noise amplification, which could corrupt memory traces. To avoid this, long-lived memory maintenance at high SNR further requires that the input connectivity pattern be matched to the architecture of the amplifying network. By analyzing the dynamical propagation of signal and noise through arbitrary recurrent networks, we have shown (see *SI Appendix*), remarkably, that for a given amount of signal amplification, no recurrent network can achieve less noise amplification (i.e., higher SNR) than a delay line possessing the same signal-amplification profile, with the input entering at its source. However, a recurrent network, unlike a delay line, can amplify signals four times larger than its network size (see Fig. 4*D*).

Although most of our analysis was limited to linear systems, we have shown that systems with a divergent fan out architecture (see Fig. 5*A*) can achieve signal amplification in a distributed manner and thereby exhibit long-lived memory traces that last a time $O(\sqrt{N})$, even in the presence of saturating nonlinearities. Indeed, this duration of memory trace is the maximum possible for any network operating within a limited dynamic range (see *SI Appendix*). We further note that it is not necessary for a network to manifestly have a connectivity as in Fig. 5*A* to achieve this limit. We have tested numerically the memory properties of networks with saturating nonlinearites whose connectivity arises from random orthogonal rotations of the divergent fan-out architecture. Such networks appear to have unstructured connectivity, and the underlying feedforward architecture is hidden. Nevertheless, these networks have memory traces that last a time $O(\sqrt{N})$ (S.G. and H.S., unpublished work).

Given the poor memory performance of generic networks, our work suggests that neuronal networks in the prefrontal cortex or hippocampus specialized for working memory tasks involving temporal sequences may posses hidden, divergent feedforward connectivities. Other potential systems for testing our theory are neuronal networks in the auditory cortex specialized for speech processing or networks in the avian brain specialized for song learning and recognition.

The principles we have discovered hold for general dynamical systems, as illustrated in the example from fluid dynamics. In light of the results of Fig. 6, it is not surprising that reconstruction of acoustic signals injected into the surface of water in a laminar state, attempted in ref. 12, fared poorly. Our theory suggests that performance could be substantially improved if, for example, the signal were injected behind the wake of a fluid flowing around an object, or in the vicinity of a mixing layer, or even into laminar flows at high Reynolds numbers just below the onset of turbulence.

In this work, we have applied the framework of Fisher information to memory traces embedded in the activity of neurons, usually identified as short-term memory. However, the same framework can be applied to study the storage of spatio-temporal sequences through synaptic plasticity, i.e., long-term memory (S.G. and H.S., unpublished work). More generally, memory of past events is a ubiquitous feature of biological systems, and they all face the problem of noise accumulation, decaying signals, and interference. In revealing fundamental limits on the lifetimes of memory traces in the presence of these various effects, and in uncovering general dynamical design principles required to achieve these limits, our theory provides a useful framework for studying the efficiency of dynamical processes underlying robust memory maintenance in biological systems.

ACKNOWLEDGMENTS. We have benefited from useful discussions with Kenneth D. Miller, Eran Mukamel, and Olivia White. This work was supported by the Israeli Science Foundation (H.S.) and the Swartz Foundation (S.G.). We also acknowledge the support of the Swartz Theoretical Neuroscience Program at Harvard University.

1. Lowenstein Y, Sompolinsky H (2003) Temporal integration by calcium dynamics in a model neuron. *Nat Neurosci* 6:961–967.
2. Seung HS (1996) How the brain keeps the eyes still. *Proc Natl Acad Sci USA* 93:13339–13344.
3. Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. *Science* 319:1543–1546.
4. Maass W, Natschlager T, Markram H (2002) Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput* 14:2531–2560.
5. Jaeger H (2001) GMD Report No. 148 (German National Research Center for Information Technology, Sankt Augustin, Germany).
6. Jaeger H, Haas H (2004) Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* 304:78–80.
7. White O, Lee D, Sompolinsky H (2004) Short-term memory in orthogonal neural networks. *Phys Rev Lett* 92:148102.
8. Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. *Proc Natl Acad Sci USA* 90:10749–10753.
9. Trefethen LN, Embree M (2005) *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators* (Princeton Univ Press, Princeton, NJ).
10. Trefethen LN, Trefethen AE, Reddy SC, Driscoll TA (1993) Hydrodynamic stability without eigenvalues. *Science* 261:578–584.
11. Cossu C, Chomaz JM (1997) Global measures of local convective instabilities. *Phys Rev Lett* 78:4387–4390.
12. Fernando C, Sojakka S (2003) Pattern recognition in a bucket: A real liquid brain. *Proc of ECAL* (Springer, New York).