Supplementary Material

Contents:

1. Review of Fisher Information and Basic Results.

We recall that the Kullback-Leibler (KL) divergence between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$, defined as

$$D_{\mathrm{KL}}(p||q) = \int d\mathbf{x}\, p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}, \tag{1}$$

is a measure of the statistical difference between the two distributions, and is 0 if and only if p and q are the same distribution. Now given a multidimensional family of distributions $p(\mathbf{x}|\mathbf{s})$, parameterized by the vector $\mathbf{s}$, the infinitesimal KL divergence between two nearby distributions given by $\mathbf{s}$ and $\mathbf{s}+\delta\mathbf{s}$ is in general approximated by the Fisher information matrix $\mathbf{J}_{k,l}(\mathbf{s})$ defined in the main paper, so that $D_{\mathrm{KL}}\big[p(\mathbf{x}|\mathbf{s})||P(\mathbf{x}|\mathbf{s}+\delta\mathbf{s})\big] \approx \frac{1}{2}\,\delta\mathbf{s}^T\mathbf{J}(\mathbf{s})\delta\mathbf{s}$.

In the special case when the family $p(\mathbf{x}|\mathbf{s})$ is a family of gaussian distributions whose mean $\mu(\mathbf{s})$ depends explicitly on $\mathbf{s}$, and whose covariance $\mathbf{C}$ is independent of $\mathbf{s}$, both the KL-divergence and Fisher information matrix take simple forms. In this case we have

$$D_{\mathrm{KL}}\big[p(\mathbf{x}|\mathbf{s}^1)||P(\mathbf{x}|\mathbf{s}^2)\big] = \frac{1}{2}\big[\mu(\mathbf{s}^1)-\mu(\mathbf{s}^2)\big]^T\mathbf{C}^{-1}\big[\mu(\mathbf{s}^1)-\mu(\mathbf{s}^2)\big], \tag{2}$$

and

$$\mathbf{J}_{k,l}(\mathbf{s}) = \frac{\partial\mu(\mathbf{s})^T}{\partial\mathbf{s}_k}\,\mathbf{C}^{-1}\,\frac{\partial\mu(\mathbf{s})}{\partial\mathbf{s}_l}. \tag{3}$$

Futhermore, when the mean μ of the family only depends linearly on the parameter $\mathbf{s}$, it is easy to see that the Fisher information (3) becomes independent of $\mathbf{s}$, and the quadratic approximation to the KL-divergence through the Fisher information becomes exact.

Now specializing to the linear network considered in the main paper, using (3) it is straightforward to compute the full Fisher information matrix. The network state at time n has the solution $\mathbf{x}(n) = \sum_{k=0}^{\infty}\mathbf{W}^k\mathbf{v}\mathbf{s}_k + \sum_{k=0}^{\infty}\mathbf{W}^k\mathbf{z}(n-k)$. Because the noise is gaussian, the

conditional distribution $P(\mathbf{x}(n)|\mathbf{s})$ is also gaussian, with mean $\mu(\mathbf{s}) = \sum_{k=0}^{\infty} \mathbf{W}^k \mathbf{v} s_k$ and noise covariance matrix $\mathbf{C}_n = \epsilon \sum_{k=0}^{\infty} \mathbf{W}^k \mathbf{W}^{kT}$. As above, the mean is only linearly dependent on the signal, and the noise covariance is independent of the signal, so $\mathbf{J}$ is independent of the signal history and takes the form

$$\mathbf{J}_{k,l} = \mathbf{v}^T \mathbf{W}^{kT} \mathbf{C}_n^{-1} \mathbf{W}^l \mathbf{v}. \tag{4}$$

We call this matrix the Fisher Memory Matrix (FMM). Its individual matrix elements have simple interpretations. First consider a simpler scenario in which a single input pulse s_k enters the network at time $n-k$. Then the network state at time n has the solution $\mathbf{x}(n) = \mathbf{W}^k \mathbf{v} s_k + \sum_{m=0}^{\infty} \mathbf{W}^m \mathbf{z}(n-m)$. The Fisher information that $\mathbf{x}(n)$ retains about this pulse is a single number, and is easily shown to be identical to the diagonal element $\mathbf{J}_{k,k}$ of the FMM in (4). Thus the digonal of the FMM is a memory curve, called the Fisher Memory Curve (FMC) in the main text, and it captures the decay of the memory trace of a single pulse.

Now consider a scenario in which two pulses s_k and s_l enter the network at times $n-k$ and $n-l$. A short computation shows that the Fisher information between $\mathbf{x}(n)$ and both pulses is a 2×2 symmetric matrix whose diagonals are $\mathbf{J}_{k,k}$ and $\mathbf{J}_{l,l}$ and whose off diagonals are both $\mathbf{J}_{k,l}$. Thus the off diagonal elements of the FMM in (4) capture the interference between two pulses entering the system at k and l timesteps in the past. As we shall see below, such interference can arise physically through the collision of signals entering the network at the two different times.

We note that in general, the Fisher information that an output variable $\mathbf{x}$ contains about an input s can be thought of as a signal to noise ratio. This interpretation is especially precise in the situation where $\mathbf{x}$ is linearly proportional to s plus some additive gaussian noise. For example, consider the one dimensional case $x = g * s + z$, where g is a scalar input-output gain and z is a gaussian random variable with variance ϵ. Then the Fisher information that x has about s is simply $\frac{g^2}{\epsilon}$, the ratio of the squared signal gain to the noise variance. Similarly, at any given time, the Fisher information that the vector input $\mathbf{v}s + \mathbf{z}$ to the network has about the signal s can be computed to be $\frac{\mathbf{v} \cdot \mathbf{v}}{\epsilon}$, which is again a squared signal gain to noise ratio. Finally, k timesteps, after a pulse has entered the network, it is embedded in the network state in the direction $\mathbf{W}^k \mathbf{v}$ (the signal gain). $\mathbf{J}_{k,k}$ is essentially this squared signal gain divided by the noise covariance $\mathbf{C}_n$. When measured in units of the input SNR $\frac{1}{\epsilon}$ (recall we normalize $\mathbf{v}$ so that $\mathbf{v} \cdot \mathbf{v} = 1$) $\mathbf{J}_{k,k}$ then represents the fraction of the input SNR remaining in the network state $\mathbf{x}$ about an input pulse entering the network k timesteps in the past.

2. Spacetime unification of spatial and temporal Fisher information.

Here we show that the spatial Fisher information matrix $\mathbf{J}^s$, defined in Eqn. 8 in the main text, and the temporal FMM, defined in the main text and computed above in (4), can both be understood within a unified spacetime framework. In this framework, the network receives a more general spatiotemporal input in which each neuron i receives an independent signal $s_i(n-k)$. This generalizes the case considered in the main paper in which $s_i(n-k) = \mathbf{v}_i s(n-k)$. The Fisher information that $\mathbf{x}(n)$ retains about this spatiotemporal signal is then a matrix with indices in spacetime,

$$\mathbf{J}^{st}_{(k,i),(l,j)} = [\mathbf{W}^{kT} \mathbf{C}_n^{-1} \mathbf{W}^l]_{i,j}. \tag{5}$$

The structure of $\mathbf{J}^{st}$ is exceedingly simple. First, it obeys the spacetime sum rule,

$$\mathrm{Tr}\, \mathbf{J}^{st} = \frac{N}{\epsilon}, \tag{6}$$

where the trace is over the diagonal spacetime elements. Thus the total spacetime SNR of the network is simply the number of degrees of freedom over the noise variance, *independent* of the recurrent connectivity $\mathbf{W}$. Furthermore, this total SNR is distributed equally among N spacetime eigenvectors of $\mathbf{J}^{st}$, each with eigenvalue $1/\epsilon$. The (unnormalized) form of these eigenvectors is $U_{k,i} = [\mathbf{W}^{kT}\mathbf{v}]_i$ where $\mathbf{v}$ is an arbitrary non-zero N-dimensional vector, as can be easily verified. The rest of the eigenvalues are zero. Thus, $\epsilon\mathbf{J}^{st}$ is a projection operator that projects the $N \times T$ (where T is the duration of the signal) dimensional space of all possible input trajectories onto those that are realizable by the system dynamics. This simple structure arises due to the linearity of the dynamics and the fact that both the gaussian noise and the signal enter the network in the same way, and hence are modified by the network dynamics in the same way regardless of the structure of $\mathbf{W}$.

Both the FMM $\mathbf{J}$ and the spatial Fisher information $\mathbf{J}^s$ arise naturally from the spacetime Fisher information $\mathbf{J}^{st}$. The FMM (4) is a projection of the spacetime Fisher information (5) onto the fixed spatial structure $\mathbf{v}$ of the temporal input. $\mathbf{J}^s$, arises as a partial temporal trace over $\mathbf{J}^{st}$, and thus measures the information in the network's spatial degrees of freedom $\mathbf{x}_i(n)$ about the entire signal history. As a consequence of the spacetime sum rule (6), the total information in all N degrees of freedom is $\mathrm{Tr}\,\mathbf{J}^s = \frac{N}{\epsilon}$, independent of $\mathbf{W}$. But as noted in the main text, $\mathbf{J}^s$ serves as an interesting order parameter for nonnormality. It is proportional to the identity matrix for every normal matrix, whereas for nonnormal matrices the departure of $\mathbf{J}^s$ from the identity captures the anisotropy inherent in the hidden feedforward connectivity of a nonnormal matrix.

3. Analytical Results for Fisher Memory Matrices.

Delay Ring

We first consider the delay ring of length N. Here $\mathbf{W}_{ij} = \mathbf{d}_{(i-j) \mod N}$ where $\mathbf{d}$ is a vector with components $\mathbf{d}_k = \sqrt{\alpha}\,\delta_{k,1}$. When the input $\mathbf{v}$ is localized to a single neuron, the FMM is

$$\mathbf{J}_{k_1 k_2} = \begin{cases} \frac{1}{\epsilon}\alpha^{\frac{k_1+k_2}{2}}(1-\alpha) & |k_1 - k_2| = mN, m = 0, 1, \dots \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

In addition to the diagonal FMC discussed in the main text, the full FMM also has a series of off diagonal bands indicating interference between signals entering the network at time differences that are integer multiples of N. Physically this interference arises because a past signal can percolate m times around the ring and collide with an incoming signal. The $m'th$ off diagonal registers this effect. It is smaller than the diagonal by a factor of α^{mN} which is the square of the attenuation experienced by a signal traversing the ring m times.

Inhomogenous Delay Line

A delay line is a feedforward chain whose only nonzero matrix elements are $\mathbf{W}_{i+1,i} = \sqrt{\alpha_i}$ for $i = 1 \dots N-1$. The noise covariance $\mathbf{C}_n$ is diagonal, with $(\mathbf{C}_n)_{ii} = \sum_{m=1}^{i-1} \prod_{p=m}^{i-1} \alpha_p$. 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

$$\mathbf{J}_{k,k} = \frac{\prod_{m=1}^{k} \alpha_m}{\epsilon \sum_{m=1}^{k+1} \left(\prod_{p=m}^{k} \alpha_p \right)}, \quad k = 0 \dots N-1, \tag{8}$$

and 0 otherwise. This result can be directly understood as a signal to noise ratio. Information about the signal k timesteps in the past is stored exclusively in neuron $k+1$. The numerator

is the squared signal amplification as it propagates from the source to this neuron. Each product term in the denominator sum represents the amplification of upstream noise that enters neuron m at $k+1-m$ timesteps in the past and collides with the signal at neuron $k+1$.

By dividing the top and bottom of (8) by the numerator $\prod_{m=1}^{k} \alpha_m$, and using $A_m \equiv ||\mathbf{W}^m \mathbf{v}||^2 = \prod_{p=1}^{m} \alpha_p$, one can see that the FMC takes the form stated in Eqn. 10 of the main paper:

$$\mathbf{J}_{k,k} = \frac{1}{\epsilon \sum_{m=0}^{k} A_m^{-1}}, \quad k = 0 \dots N-1. \tag{9}$$

Finally we note that the FMM for the delay line is diagonal so the FMC determines the FMM. More generally, this is true for any layered feedforward structure which has feedforward connections only between successive layers, and in which the signal enters within a single layer. Physically, this lack of interference occurs because it is impossible for two signals entering the network at two different times to collide with each other.

Random Symmetric Matrices

We next consider the ensemble of random symmetric matrices in which the elements $\mathbf{W}_{ij}$ are chosen independently from a zero mean gaussian distribution with variance $\frac{\alpha}{4N}$, subject to the constraint $\mathbf{W}_{ij} = \mathbf{W}_{ji}$. The eigenvalues of $\mathbf{W}$ are distributed on the real axis r according to Wigner's semicircular law

$$\rho_{\alpha}^{\text{Wig}}(r) = \begin{cases} \frac{2}{\pi\alpha}\sqrt{\alpha - r^2} & |r| < \sqrt{\alpha} \\ 0 & |r| > \sqrt{\alpha}. \end{cases} \tag{10}$$

When $\mathbf{v}$ couples to all modes of $\mathbf{W}$ with equal strength, the FMM reduces to the moments of this distribution as $N \to \infty$:

$$\mathbf{J}_{k_1 k_2} = \langle r^{k_1+k_2} \rangle_{\rho_{\alpha}^{\text{Wig}}} - \langle r^{k_1+k_2+2} \rangle_{\rho_{\alpha}^{\text{Wig}}}. \tag{11}$$

These moments are given by

$$\langle r^k \rangle_{\rho_{\alpha}^{\text{Wig}}} = \begin{cases} \frac{2}{k+2}\binom{k}{k/2}(\frac{\alpha}{4})^{\frac{k}{2}} & k \text{ even} \\ 0 & k \text{ odd}. \end{cases} \tag{12}$$

Thus a generic symmetric matrix displays interference between signals entering at even, but not odd, time separations. Physically this interference arises from loops of length 2 in the random symmetric ensemble. Explictly, the FMC is

$$\mathbf{J}_{k,k} = \frac{1}{k+1}\binom{2k}{k}(\frac{\alpha}{4})^k - \frac{1}{k+2}\binom{2k+2}{k+1}(\frac{\alpha}{4})^{k+1}. \tag{13}$$

This curve is the mean field theory for the FMC of random symmetric matrices, plotted in black in Fig. 2C of the main paper. As seen in Fig. 2C, it agrees well with numerical simulations of the FMC for random symmetric matrices for $N = 1000$.

Random Orthogonal Matrices

Let $\mathbf{W} = \sqrt{\alpha}\mathbf{O}$, where $\mathbf{O}$ is a random orthogonal matrix. The eigenvalues of $\mathbf{O}$ are uniformly distributed in the complex plane on a circle of radius $\sqrt{\alpha}$. Thus they are of the

form $\sqrt{\alpha}e^{i\phi}$ with a uniform density over the phase ϕ. The Fisher memory matrix is then

$$\mathbf{J}_{k_1 k_2} = \frac{1}{\epsilon}\int_0^{2\pi} \frac{d\phi}{2\pi}\, \alpha^{\frac{k_1+k_2}{2}} e^{i(k_2-k_1)\phi}(1-\alpha). \tag{14}$$

$$= \frac{1}{\epsilon}\,\alpha^{\frac{k_1+k_2}{2}}\,(1-\alpha)\,\delta_{k_1,k_2}. \tag{15}$$

Thus the FMM is diagonal with the memory curve decaying exponentially as α^k for all $\alpha < 1$. We note that for large N, with $1-\alpha$ finite, the FMM of a random orthogonal matrix on average is identical to the FMM of the delay ring.

4. Asymptotics of the Fisher Memory Curve for Normal Matrices.

When the input $\mathbf{v}$ couples to each eigenmode i of a normal connectivity matrix $\mathbf{W}$ with a uniform strength $v_i = \frac{1}{\sqrt{N}}$, the FMC only depends on the distribution $\rho(r)$ of eigenvalue magnitudes of $\mathbf{W}$. We will now derive the asymptotics of the FMC under this assumption on the input coupling $\mathbf{v}$, indicating afterwards how this assumption can be relaxed. We have,

$$\mathbf{J}_{k,k} = \frac{1}{\epsilon}\int_0^{\sqrt{\alpha}} dr\, \rho(r) r^{2k}(1-r^2) = \frac{1}{\epsilon}\int_0^{\sqrt{\alpha}} dr\, e^{\ln(\rho(r))+2k\ln(r)+\ln(1-r^2)}, \tag{16}$$

where $\sqrt{\alpha}$ is the magnitude of the largest eigenvalue of $\mathbf{W}$. If $\sqrt{\alpha}$ remains far from 1, then for large k this integral is dominated by the largest value of r, and so the asymptotic FMC decays exponentially: $J_{kk} \propto \alpha^k$ for large k. If $\sqrt{\alpha}$ is close to 1, and $\rho(r)$ has a continuous density near 1, this integral is dominated by the behavior of $\rho(r)$ near 1. In general this behavior is characterized by an exponent ν so that $\lim_{r\to\sqrt{\alpha}} \rho(r) \propto (\sqrt{\alpha}-r)^\nu$. Integrability of $\rho(r)$ requires that $\nu > -1$. To extract the k dependence in J_{kk} in the case where $\sqrt{\alpha}$ is close to 1, we perform the change of variables $r = \sqrt{\alpha} - \frac{x}{2k}$ in (16) to obtain

$$\mathbf{J}_{k,k} = \frac{1}{\epsilon}\frac{1}{2k}\int_0^{2k\sqrt{\alpha}} dx\, e^{\ln(\rho(\sqrt{\alpha}-\frac{x}{2k}))+2k\ln(\sqrt{\alpha}-\frac{x}{2k})+\ln(1-(\sqrt{\alpha}-\frac{x}{2k})^2)}. \tag{17}$$

For large k, this integral is dominated by small values of $\frac{x}{2k}$. By substituting the asymptotic behavior of $\rho(r)$ into (17), taking the limit $\sqrt{\alpha} \to 1$, and expanding in the small parameter $\frac{x}{2k}$, we obtain

$$\mathbf{J}_{k,k} \approx \frac{1}{\epsilon}\frac{1}{2k}\int_0^{2k\sqrt{\alpha}} dx\, e^{-x+(\nu+1)\ln\frac{x}{2k}+\ln 2} \approx \frac{1}{\epsilon}\frac{1}{(2k)^{\nu+2}}\int_0^{\infty} dx\, 2x^{\nu+1}e^{-x}. \tag{18}$$

Thus asymptotically, J_{kk} decays as a power law: $\mathbf{J}_{k,k} \propto \frac{1}{k^{\nu+2}}$.

It is now clear that the original assumption that $\mathbf{v}$ couple with equal strength to all eigenmodes of $\mathbf{W}$ is not strictly necessary for this derivation. As long as $\mathbf{v}$ couples to the slowest modes roughly evenly, then the power law $\mathbf{J}_{k,k} \propto \frac{1}{k^{\nu+2}}$ will remain unaltered. More precisely, if $v(r)$ denotes the total strength with which $\mathbf{v}$ couples to all eigenmodes of $\mathbf{W}$ with eigenvalue magnitude r, as long as the product obeys $\lim_{r\to\sqrt{\alpha}} v(r)\rho(r) \propto (\sqrt{\alpha}-r)^\nu$, the asymptotics of of the FMC are the same.

5. Fisher Information, Signal Statistics, and Alternative Memory Measures.

Both the ability to use the current state $\mathbf{x}(n)$ to reconstruct the past signal history $\mathbf{s}$, and the mutual information between $\mathbf{x}(n)$ and $\mathbf{s}$ constitute alternative measures of memory performance. However, unlike the FMM, these measures both depend on the signal statistics. Here

we assume a zero mean gaussian signal with covariance matrix $\mathbf{S}_{k_1 k_2} = \langle s(n-k_1)s(n-k_2)\rangle$ and outline the relationship between the FMM $\mathbf{J}$, the signal statistics $\mathbf{S}$, and both these alternative measures.

First, the FMM provides a lower bound on signal reconstruction performance. Consider a network that has been placed into a particular state $\mathbf{x}(n)$ given a signal history $\mathbf{s}$. Suppose that readout neurons offer a corresponding signal reconstruction $\hat{\mathbf{s}}$ based on $\mathbf{x}(n)$. We allow for the possiblity of a bias $\mathbf{b}(\mathbf{s}) = \langle \hat{\mathbf{s}}|\mathbf{s}\rangle - \mathbf{s}$ in the reconstruction. Then the Cramer-Rao theorem for biased estimators places a lower bound on the reconstruction uncertainty, or covariance of $\hat{\mathbf{s}}$, through the FMM:

$$\mathrm{Cov}(\hat{\mathbf{s}}|\mathbf{s}) \geq (\mathbf{I}+\mathbf{B})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{B})^T, \tag{19}$$

where $\mathbf{B}$ is a bias matrix with elements $\mathbf{B}_{k_1 k_2} = \partial \mathbf{b}_{k_1}/\partial_{\mathbf{s}_{k_2}}$.

Because of the underlying gaussianity, the Cramer-Rao bound is indeed saturated by the optimal linear estimator of the input history. This estimator can be realized by an array of T readout neurons whose output is $\hat{\mathbf{s}} = \mathbf{U}\mathbf{x}(n)$, where $\mathbf{U}$ is a $T \times N$ matrix of readout weights. Optimizing the mean square error of the estimate (averaging over both noise and signal statistics) one obtains $\mathbf{U}^{opt} = \mathbf{S}\mathbf{P}^T\mathbf{C}^{-1}$ where $\mathbf{P}$ is an $N \times T$ matrix with elements $\mathbf{P}_{ik} = (\mathbf{W}^k v)_i$ and $\mathbf{C}$ is the total covariance of $\mathbf{x}(n)$. $\mathbf{C}$ decomposes into signal and noise covariances, $\mathbf{C} = \mathbf{C}_s + \mathbf{C}_n$, where $\mathbf{C}_n$ is defined above, and $\mathbf{C}_s = \mathbf{P}\mathbf{S}\mathbf{P}^T$. The estimator covariance and bias matrices can be computed in terms of the FMM:

$$\mathrm{Cov}(\hat{\mathbf{s}}|\mathbf{s}) = \mathbf{S}^{1/2}\tilde{\mathbf{J}}(\mathbf{I}+\tilde{\mathbf{J}})^{-2}\mathbf{S}^{1/2}, \tag{20}$$

$$\mathbf{B} = \mathbf{S}^{1/2}\tilde{\mathbf{J}}(\mathbf{I}+\tilde{\mathbf{J}})^{-1}\mathbf{S}^{-\frac{1}{2}} - \mathbf{I}, \tag{21}$$

where we have introduced $\tilde{\mathbf{J}} = \mathbf{S}^{1/2}\mathbf{J}\mathbf{S}^{1/2}$. Using (20) and (21) one can check that (19) is saturated.

Also, a useful quantity that is related to the error of the estimator is the matrix of estimator-signal correlation, $\mathbf{M} = \langle \hat{\mathbf{s}}\mathbf{s}^T\rangle$ which for the optimal estimator reduces to $\mathbf{M} = \mathbf{S}\mathbf{P}^T\mathbf{C}^{-1}\mathbf{P}\mathbf{S}$. This performance measure can also be written in terms of the FMM and signal statistics as

$$\mathbf{M} = \mathbf{S}^{1/2}\tilde{\mathbf{J}}(\mathbf{I}+\tilde{\mathbf{J}})^{-1}\mathbf{S}^{1/2}. \tag{22}$$

The digaonal of this matrix was the memory curve, studied in [1] primarily for random orthogonal matrices.

Alternatively, the mutual information between the present state and the past signal, given by $I[\mathbf{x}(n)\,;\,\mathbf{s}] = H[\mathbf{x}(n)] - H[\mathbf{x}(n)|\mathbf{s}]$, where H denotes ensemble entropy, also has simple relationship with $\mathbf{J}$ and $\mathbf{S}$:

$$I[\mathbf{x}(n)\,;\,\mathbf{s}] = \frac{1}{2}\log\,\det(\mathbf{I}+\tilde{\mathbf{J}}). \tag{23}$$

Overall relations (19), (22) and (23) dissect more complex measures of memory performance into simpler contributions from the FMC, interference represented by off diagonal elements of the FMM, and the signal statistics. For example, from (23), one can show that for uncorrelated signal statistics, so that $\tilde{\mathbf{J}} = \mathbf{J}$, interference due to off-diagonal elements in $\mathbf{J}$ will degrade memory performance. This can be proven using the Hadamard inequality $\det\mathbf{A} \leq \prod_{k=0}^{T}\mathbf{A}_{k,k}$, which holds for any matrix $\mathbf{A}$. Applying it to the matrix $\mathbf{A} = \mathbf{I}+\mathbf{J}$ we see that for a given network that has no interference, or a diagonal FMM $\mathbf{J}$, modifying it to introduce interference by turning on off-diagonal FMM elements while keeping the FMC

fixed, will reduce the mutual information (23) in the system. However, if the input signal is correlated, it may be beneficial to introduce interference matched to the signal correlations to improve memory performance.

The relation between the reconstruction memory and the Fisher memory (22) simplifies considerably when the signal is white, so that $\mathbf{S} = \mathbf{I}$, and there is no network interference, so that the FMM is diagonal. In this case, the memory function $m(k)$ is related to the FMC $J(k)$ through $m(k) = \frac{J(k)}{1+J(k)}$. A useful measure of the time delay over which inputs can be reconstructed is the time k_c at which the memory function $m(k)$ falls to $\frac{1}{2}$, i.e. k_c is the solution to $m(k_c) = \frac{1}{2}$. This occurs when the SNR $J(k)$ drops to 1.

To see the effects of strong transient nonnormal amplification on the reconstruction memory $m(k)$ it is useful to compare the memory properties of random orthogonal matrices (or equivalently, for large N, the delay ring), and the delay line with amplification. In the former case from (15), k_c is determined by the condition $\alpha^{k_c}(1-\alpha) = \epsilon$. Since the FMC on the left hand side decays exponentially, k_c can be extensive only when both the decay constant α is close to 1, so that $1 - \alpha = \frac{\rho}{N}$, and when the noise is inversely proportional to N, so that $\epsilon = \frac{\epsilon_o}{N}$, for ρ and ϵ_o both $O(1)$ [1]. On the other hand, from the expression for the homogenous delay line with gain $\sqrt{\alpha}$ in the main paper, the condition for k_c is $\alpha^{k_c}\frac{1-\alpha}{1-\alpha^{k_c+1}} = \epsilon$. When $\alpha > 1$, for large $k < N$, the left hand side asymptotes to a finite value $1 - \frac{1}{\alpha}$. Thus as long as the $\alpha > \frac{1}{1-\epsilon}$, $m(k)$ will remain above $\frac{1}{2}$ for any finite value of $\epsilon < 1$. Thus due to transient nonnormal amplification, the reconstruction memory can be extensive, even for finite $\epsilon = O(1)$, whereas in the normal case, or for any case in which the FMC decays exponentially, the noise must be small ($\epsilon < O(1/N)$) in order for reconstruction memory to be extensive.

6. A Dynamical Bound on the Fisher Memory Curve.

Here we prove the upper bound

$$\mathbf{J}_{k,k} \leq \frac{1}{\epsilon \sum_{m=0}^{k} \frac{1}{||\mathbf{W}^m \mathbf{v}||^2}} \qquad \forall\, k \geq 0. \tag{24}$$

Consider a single input pulse s_0 entering the network at time 0. The state of the system at a time $k \geq 0$ is then

$$\mathbf{x}(k) = \mathbf{W}^k \mathbf{v} s_0 + \sum_{m=0}^{k} \mathbf{W}^{k-m} \mathbf{z}(m) + \sum_{m=-\infty}^{-1} \mathbf{W}^{k-m} \mathbf{z}(m). \tag{25}$$

$\mathbf{J}_{k,k}$ is the Fisher information that $\mathbf{x}(k)$ retains about s_0. In (25) we have explicitly divided the noise into two parts: 1) noise that enters the network either at the same time as, or after, the signal (times $m = 0, \ldots, k$) and 2) noise that enters the network before the signal (times $m \leq -1$). Now consider an "improved" network state $\bar{\mathbf{x}}(k)$ in which some of the noise entering the network in (25) is removed. Specifically, we will remove all noise entering the network before the signal ($\mathbf{z}(m) \to 0$, $\forall\, m \leq -1$). Now consider noise entering the network at time m after the signal, with $0 \leq m \leq k$. At this time the signal is embedded in the direction $\mathbf{W}^m \mathbf{v}$ in network space. We will further remove noise that is orthogonal to the direction in which the signal is embedded. This can accomplished by applying the projection operator $\frac{\mathbf{W}^m \mathbf{v}\mathbf{v}^T \mathbf{W}^{mT}}{||\mathbf{W}^m \mathbf{v}||^2}$ to $\mathbf{z}(m)$ before it enters the network. Both these noise removal operations

lead to the improved state

$$\bar{\mathbf{x}}(k) = \mathbf{W}^k \mathbf{v} s_0 + \mathbf{W}^k \mathbf{v} \sum_{m=0}^{k} \frac{\mathbf{v}^T \mathbf{W}^{mT} \mathbf{z}(m)}{||\mathbf{W}^m \mathbf{v}||^2}. \tag{26}$$

By construction, the actual dynamical state $\mathbf{x}(k)$ is related to the improved state $\bar{\mathbf{x}}(k)$ by the addition of gaussian noise that is uncorrelated with the signal s_0, and with the noise that is already present in $\bar{\mathbf{x}}(k)$. In general, the addition of such noise can never increase the Fisher information. Thus if we denote by $\bar{\mathbf{J}}_{k,k}$, the Fisher information that $\bar{\mathbf{x}}(k)$ retains about the input pulse s_0, then $\bar{\mathbf{J}}_{k,k}$ constitutes an upper bound on $\mathbf{J}_{k,k}$, i.e. $\mathbf{J}_{k,k} \leq \bar{\mathbf{J}}_{k,k}$.

We now compute the upper bound $\bar{\mathbf{J}}_{k,k}$. $\bar{\mathbf{x}}(k)$ is gaussian distributed with conditional mean $\langle \bar{\mathbf{x}}(k) | \mathbf{s}_0 \rangle = \mathbf{W}^k \mathbf{v} \mathbf{s}_0$ and noise covariance matrix $\bar{\mathbf{C}} = \epsilon \, \mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT} \sum_{m=0}^{k} \frac{1}{||\mathbf{W}^m \mathbf{v}||^2}$. We then have $\bar{\mathbf{J}}_{k,k} = \frac{1}{\epsilon} \mathbf{v}^T \mathbf{W}^{kT} \bar{\mathbf{C}}^{-1} \mathbf{W}^k \mathbf{v}$. Although $\bar{\mathbf{C}}$ is a rank 1 matrix, to compute $\bar{\mathbf{J}}_{k,k}$, we need only compute $\bar{\mathbf{C}}^{-1}$ within the 1 dimensional subspace spanned by $\mathbf{W}^k \mathbf{v}$. Within this subspace $\bar{\mathbf{C}}$ is invertible, and using the relation $\mathbf{v}^T \mathbf{W}^{kT} (\mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT})^{-1} \mathbf{W}^k \mathbf{v} = 1$, we obtain $\bar{\mathbf{J}}_{k,k}$ to be equal to the righthand side of (24), thus proving (24).

7. Uniqueness of the Delay Line.

Here we show that the delay line is essentially the only network that saturates the upper bound (24). More precisely, for any connectivity $\mathbf{W}$ and $\mathbf{v}$ which saturates (24), we show that there exists an $L \leq N$ dimensional orthonormal basis of network space such that

(1) When restricted to this basis, the matrix elements of $\mathbf{W}$ are identical to the connectivity matrix of a delay line of length L, i.e. $\mathbf{W}_{ij} = \sqrt{\alpha_i} \, \delta_{i,i-1}$, for $i = 2, \ldots, L$ and $\mathbf{W}_{1j} = 0, \forall \, j$.

(2) $\mathbf{v}$ feeds into its source ($\mathbf{v}_i = \delta_{i,1}$).

(3) For any vector $\mathbf{u}$ orthogonal to the L dimensional subspace spanned by this basis, and for any vector $\mathbf{x}$ in this subspace, both $\mathbf{u}^T \mathbf{W} \mathbf{x}$ and $\mathbf{x}^T \mathbf{W} \mathbf{u}$ are 0. Thus no other states feed into or out of this delay line.

The idea behind the proof is to examine the conditions under which the re-addition of the removed noise to $\bar{\mathbf{x}}(k)$ in (26) to get back $\mathbf{x}(k)$ in (25) does not reduce the Fisher information $\bar{\mathbf{J}}_{k,k}$, so that $\bar{\mathbf{J}}_{k,k} = \mathbf{J}_{k,k}$.

First we note that any noise added to the network state $\bar{\mathbf{x}}(k)$, that lives in the direction in which the signal s_0 is embedded, necessarily decreases the Fisher information. Consider a network state $\tilde{\mathbf{x}}(k)$ related to $\bar{\mathbf{x}}(k)$ in (26) by $\tilde{\mathbf{x}}(k) = \bar{\mathbf{x}}(k) + \frac{\mathbf{W}^k \mathbf{v}}{||\mathbf{W}^k \mathbf{v}||} \eta$, where η is a gaussian noise with variance $\langle (\delta \eta)^2 \rangle$. This extra, rank 1 noise lives only in the direction $\mathbf{W}^k \mathbf{v}$ in which the signal s_0 is embedded in $\bar{\mathbf{x}}(k)$. Denoting $\tilde{\mathbf{J}}_{k,k}$ the Fisher information that the "corrupted" $\tilde{\mathbf{x}}(k)$ retains about s_0, we show that $\tilde{\mathbf{J}}_{k,k} < \bar{\mathbf{J}}_{k,k}$. The noise covariance of $\tilde{\mathbf{x}}(k)$ is $\tilde{\mathbf{C}} = \bar{\mathbf{C}} + \frac{\mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT}}{||\mathbf{W}^k \mathbf{v}||^2} \langle (\delta \eta)^2 \rangle$. Its inverse is

$$\tilde{\mathbf{C}}^{-1} = \bar{\mathbf{C}}^{-1} - \frac{\bar{\mathbf{C}}^{-1} \mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT} \bar{\mathbf{C}}^{-1}}{\frac{||\mathbf{W}^k \mathbf{v}||^2}{\langle (\delta \eta)^2 \rangle} + \mathbf{v}^T \mathbf{W}^{kT} \bar{\mathbf{C}}^{-1} \mathbf{W}^k \mathbf{v}}, \tag{27}$$

where again $\bar{\mathbf{C}}$ is the noise covariance of $\bar{\mathbf{x}}(k)$. Then

$$\tilde{\mathbf{J}}_{k,k} = \mathbf{v}^T \mathbf{W}^{kT} \tilde{\mathbf{C}}^{-1} \mathbf{W}^k \mathbf{v} = \bar{\mathbf{J}}_{k,k} - \frac{\bar{\mathbf{J}}_{k,k}^2}{\frac{||\mathbf{W}^k \mathbf{v}||^2}{\langle (\delta\eta)^2 \rangle} + \bar{\mathbf{J}}_{k,k}}, \tag{28}$$

which clearly, is strictly less than $\bar{\mathbf{J}}_{k,k}$.

Now we introduce a particular network state $\tilde{\mathbf{x}}(k)$ whose Fisher information about s_0 is intermediate between that of $\mathbf{x}(k)$ in (25) and that of $\bar{\mathbf{x}}(k)$ in (26). We do this by reintroducing the noise that was removed back into $\bar{\mathbf{x}}(k)$, allowing it to propagate up to time k, but then keeping only noise that lives in the the direction $\mathbf{W}^k \mathbf{v}$ in which the signal s_0 is embedded at time k, using the projection operator $\frac{\mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT}}{||\mathbf{W}^k \mathbf{v}||^2}$. This procedure yields

$$\tilde{\mathbf{x}}(k) = \bar{\mathbf{x}}(k) + \frac{\mathbf{W}^k \mathbf{v} \mathbf{v}^T \mathbf{W}^{kT}}{||\mathbf{W}^k \mathbf{v}||^2} \left\{ \sum_{m=0}^{k} \mathbf{W}^{k-m} \left[\mathbf{I} - \frac{\mathbf{W}^m \mathbf{v} \mathbf{v}^T \mathbf{W}^{mT}}{||\mathbf{W}^m \mathbf{v}||^2} \right] \mathbf{z}(m) + \sum_{m=-\infty}^{-1} \mathbf{W}^{k-m} \mathbf{z}(m) \right\}. \tag{29}$$

By construction, we have $\mathbf{J}_{k,k} \leq \tilde{\mathbf{J}}_{k,k} \leq \bar{\mathbf{J}}_{k,k}$. Now if $\mathbf{W}$ and $\mathbf{v}$ are such that the bound (24) is saturated (i.e. $\mathbf{J}_{k,k} = \bar{\mathbf{J}}_{k,k}$), we must then have $\tilde{\mathbf{J}}_{k,k} = \bar{\mathbf{J}}_{k,k}$. Now $\tilde{\mathbf{x}}(k)$ has the form $\bar{\mathbf{x}}(k) + \frac{\mathbf{W}^k \mathbf{v}}{||\mathbf{W}^k \mathbf{v}||} \eta$, where η is a random variable which can be read off from (29). Then we have seen above that we will have $\tilde{\mathbf{J}}_{k,k} = \bar{\mathbf{J}}_{k,k}$ only when this variable is identically zero. Requiring this condition then yields a set of necessary conditions that $\mathbf{W}$ and $\mathbf{v}$ must obey to saturate (24):

$$\mathbf{v}^T \mathbf{W}^{kT} \mathbf{W}^{k-m} \left[\mathbf{I} - \frac{\mathbf{W}^m \mathbf{v} \mathbf{v}^T \mathbf{W}^m}{||\mathbf{W}^m \mathbf{v}||^2} \right] = 0, \quad m = 0, \ldots, k, \quad \forall\, k \geq 0. \tag{30}$$

$$\mathbf{v}^T \mathbf{W}^{kT} \mathbf{W}^{k-m} = 0, \quad \forall\, m \leq -1,\, k \geq 0. \tag{31}$$

These conditions are intuitive. (30) states that any network state $\mathbf{x}$ which is orthogonal to the signal $\mathbf{W}^m \mathbf{v}$ at time m, remains orthogonal to the signal for all future time, as both $\mathbf{x}$ and the signal dynamically propagate through the network $\mathbf{W}$. (31) requires that for any state $\mathbf{x}$ the network may be in at a time $m \leq -1$ before the signal enters, the evolution of this state will always remain orthogonal to the instantaneous signal embedding direction $\mathbf{W}^k \mathbf{v}$ for all future time $k \geq 0$.

We now examine the mathematical consequences of these constraints on $\mathbf{W}$ and $\mathbf{v}$. First note that by multiplying both sides of (31) on the right by $\mathbf{v}$ one can conclude that (31) implies that $\mathbf{v}^T \mathbf{W}^{kT} \mathbf{W}^n \mathbf{v} = ||\mathbf{W}^k \mathbf{v}||^2 \, \delta_{k,n} \forall\, k, n \geq 0$. Thus the sequence of L vectors $\mathbf{v}, \mathbf{W}\mathbf{v}, \ldots, \mathbf{W}^{L-1}\mathbf{v}$ form an orthogonal basis for network space. We leave open the possibility that $\mathbf{W}^L \mathbf{v} = 0$ for some $L < N$, so that these vectors do not span all of network space.

Then the set of vectors $\mathbf{e}_i = \frac{\mathbf{W}^{i-1}\mathbf{v}}{||\mathbf{W}^{i-1}\mathbf{v}||}, i = 1, \ldots L$ forms an orthonormal basis for an L dimensional subspace of network space. We compute the matrix elements of $\mathbf{W}$ and $\mathbf{v}$ in this basis. It is straightforward to see that these elements are given by $\mathbf{W}_{ij} = \frac{||\mathbf{W}^{i-1}\mathbf{v}||}{||\mathbf{W}^{i-2}\mathbf{v}||} \delta_{j-1,i-2}$, for $i = 2, \ldots, N$, $\mathbf{W}_{1j} = 0$, $\forall j$, and $\mathbf{v}_i = \delta_{i,1}$. Thus, restricted to this basis, $\mathbf{W}$ is a delay line of length L and $\mathbf{v}$ feeds into its source, proving claims (1) and (2) above.

We now examine the interaction of this delay line with states orthogonal to the L dimensional subspace above. Let $\mathbf{u}$ be such a state. We begin by showing that $\mathbf{v}^T \mathbf{W}^{lT} \mathbf{W}^n \mathbf{u} = 0$

for $l = 0, \ldots, L-1$ and $\forall\, m \geq 0$. For any l, and $n \geq l+1$, this condition follows from applying (31) to $\mathbf{u}$ with $k = l$ and $m = l-n$. For any l and $n \leq l$, this condition follows from applying (30) to $\mathbf{u}$ with $k = l$ and $m = l-n$, which yields $\mathbf{v}^T \mathbf{W}^{lT} \mathbf{W}^n \left[\mathbf{I} - \frac{\mathbf{W}^{l-n} \mathbf{v} \mathbf{v}^T \mathbf{W}^{l-n}}{||\mathbf{W}^{l-n}\mathbf{v}||^2} \right] \mathbf{u} = 0, \forall\, \mathbf{u}$. However, under the special assumption that $\mathbf{u}$ is orthogonal to $\mathbf{W}^l \mathbf{v}, \forall\, l = 0, \ldots, L-1$, we can drop the projection operator, and conclude that $\mathbf{v}^T \mathbf{W}^{lT} \mathbf{W}^n \mathbf{u} = 0$, as claimed. Now we simply note that for $n = 1$, we have $\mathbf{v}^T \mathbf{W}^{lT} \mathbf{W} \mathbf{u} = 0, \forall\, l = 0, \ldots, L-1$, which means no state orthogonal to the delay line feeds into the delay line. Also the statement that no state along the delay line evolves into a state orthogonal to the delay line follows trivially from the fact that $\mathbf{u}^T \mathbf{W} \mathbf{W}^l \mathbf{v} = \mathbf{u}^T \mathbf{W}^{l+1} \mathbf{v} = 0$. This proves claim (3) above, completing the proof that the delay line is the unique network that saturates (24).

8. Transient Amplification, Extensive Memory and Finite Dynamic Range.

Here we examine the consequences of the dynamical upper bound (24). We first make precise the statement that superlinear amplification for a time of order N is required for extensive memory. We then show that networks whose neurons operate within a limited dynamic range cannot achieve extensive memory; the area under the FMC for such networks is at most proportional to $\sqrt{N}$.

Both results depend upon the following theorem,

$$\left(\sum_{m=1}^{k} \frac{1}{A_m} \right) \left(\sum_{m'=1}^{k} A_{m'} \right) \geq \frac{k(k+1)}{2}, \tag{32}$$

where A_m is any positive real number. This can be proved as follows. First note that the left hand side of (32) does not depend on the order of A_m for $m = 1, \ldots, k$, so we are free to reorder this set of numbers in decreasing order, so that $A_{m'} \geq A_m$ for $m' \leq m$. Then the left hand side of (32) becomes

$$\sum_{m=1}^{k} \sum_{m'=1}^{k} \frac{A_{m'}}{A_m} \quad \geq \quad \sum_{m=1}^{k} \sum_{m'=1}^{m} \frac{A_{m'}}{A_m} \quad \geq \quad \sum_{m=1}^{k} \sum_{m'=1}^{m} 1 \quad = \quad \frac{k(k+1)}{2}, \tag{33}$$

thereby proving (32). The first inequality arises because we drop all terms in which $m' > m$ and the second inequality is a consequence of the reordering.

Now we use (24) and (32) to make precise the relationship between extensive memory and superlinear amplification. We have,

$$\frac{1}{\epsilon \mathbf{J}_{k,k}} \quad \geq \quad 1 + \sum_{m=1}^{k} \frac{1}{||\mathbf{W}^m \mathbf{v}||^2} \quad \geq \quad 1 + \frac{k(k+1)}{2 \sum_{m=1}^{k} ||\mathbf{W}^m \mathbf{v}||^2}. \tag{34}$$

The first inequality is equivalent to (24) while the second is an application of (32) with $A_m = ||\mathbf{W}^m \mathbf{v}||^2$. Thus

$$\mathbf{J}_{k,k} \leq \frac{1/\epsilon}{1 + \frac{k(k+1)}{2T_k}}, \tag{35}$$

where we have defined $T_k \equiv \sum_{m=1}^{k} ||\mathbf{W}^m \mathbf{v}||^2$ to be the area under the signal amplification profile up to time k. Now for any network to possess extensive memory, the FMC must remain above a finite value (that is independent of N) up to time N. In order to achieve this for large N, (35) reveals that the area under the signal amplification profile must at least

grow quadratically in time up to time N. This means the amplification profile itself must grow superlinearly. Any slower growth would result in a decay of the FMC according to (35), thereby precluding the possibility of extensive memory.

Now we consider the case of finite dynamic range. This means that the activity of each neuron i is constrained to lie between $-\sqrt{R}$ and $\sqrt{R}$. Thus the norm of the network state $\mathbf{x}^T\mathbf{x}$ cannot exceed NR. On the other hand, the average norm of the network state while the network is receiving both noise and signal is

$$\langle \mathbf{x}^T\mathbf{x} \rangle = \sum_{m=0}^{\infty} ||\mathbf{W}^m\mathbf{v}||^2 + \epsilon \sum_{m=0}^{\infty} \mathrm{Tr}\, \mathbf{W}^{Tm}\mathbf{W}^m. \tag{36}$$

The constraint that $\langle \mathbf{x}^T\mathbf{x} \rangle \leq NR$ then automatically limits the area under the signal amplification profile so that $T_k \leq NR \; \forall \; k$. Substituting this result into (35) yields the following bound on the FMC for any network operating within a finite dynamic range R:

$$\mathbf{J}_{k,k} \leq \frac{1/\epsilon}{1 + \frac{k(k+1)}{2NR}}. \tag{37}$$

This reproduces Eqn. 12 in the main text. The area under this bound is $O(\frac{\sqrt{NR}}{\epsilon})$. Thus finite dynamic range precludes the possiblity of extensive memory.

9. The Divergent Fan Out Network.

Here we analyze the divergent fan out network shown in Fig. 5A of the main paper. This network consists of L layers labelled by $l = 1, \ldots, L$ with N_l neurons in layer l. The signal enters the first layer which has a single neuron ($N_1 = 1$). We consider for simplicity all to all connectivity from each layer l to $l+1$ of uniform strength $\sqrt{\gamma_l}$.

We first analyze the dynamical amplification provided by this network. If all neurons of layer l have activity x_l, the propagated signal at each node of the next layer is $g_l x_l$ where the local gain g_l is $g_l = N_l\sqrt{\gamma_l}$. As long as $g_l \leq 1$, single neuron activities will remain within their finite dynamic range. On the other hand, the total amplification of the signal as it arrives in layer $k+1$, namely $||\mathbf{W}^k\mathbf{v}||^2 \equiv A_k$ can still be large. The feedforward input $\mathbf{v}$ is a vector that has only one nonzero entry that is 1 in the component corresponding to the single neuron in the first layer. Then $\mathbf{W}^k\mathbf{v}$ is a vector whose N_k nonzero entries all take the value $\prod_{p=1}^{k} g_p$ in the components corresponding to the N_k neurons in layer k. Thus the squared norm of this network state is $A_k = N_{k+1}\prod_{p=1}^{k} g_p^2$ for $0 \leq k \leq L-1$. One can choose $\sqrt{\gamma_l}$ so that $g_l = 1$ for each l. In this regime, single neuron activities neither grow nor decay, but network signal amplification is still achieved by spreading the activity out across neurons as the signal propagates down the layers.

Such signal amplification can lead to significant Fisher memory for this network. Indeed the FMC for this network saturates the dynamical upper bound (24), because it is equivalent under a unitary change of basis to a delay line of length L. Under this equivalence, the activity of the l'th neuron in the effective delay line corresponds to a projection of the activity at the l'th layer of the tree. More precisely, consider a set of L orthonormal basis vectors $\mathbf{e}_l$, $l = 1, \ldots, L$ of network space for the divergent fan out network. $\mathbf{e}_l$ is defined to have precisely N_l nonzero entries which all take the value $\frac{1}{\sqrt{N_l}}$ in the components corresponding to the N_l neurons in layer l. Now if $\mathbf{W}$ is the connectivity matrix of the divergent fanout architecture, then the matrix elements of $\mathbf{W}$ between these basis vectors are $\mathbf{e}_l^T W \mathbf{e}_m = \delta_{m,l+1} g_l \sqrt{\frac{N_{l+1}}{N_l}}$.

Although these basis vectors do not of course span the full network space, they do span the space that is reachable by the signal, i.e. the span of $\{\mathbf{W}^l\mathbf{v}\}$ for $l = 0, \ldots, L-1$. Furthermore, $\mathbf{W}$ has no nonzero matrix elements between any vector in this space, and any vector orthogonal to this space. In essence the rest of the network degrees of freedom in the divergent fanout network do not interact with the signal as it propagates down the layers. Thus to compute the memory properties of the network, we can simply restrict to the subspace reachable by the signal. We have seen that in this subspace, the connectivity is simply a delay line of length L with an effective gain $g_l\sqrt{\frac{N_{l+1}}{N_l}}$ from "effective" neuron l to $l+1$.

Because of the equivalence to the delay line, we then know immediately the FMC of the divergent architecture, namely the right hand side of (24) with $||\mathbf{W}^k\mathbf{v}||^2 = N_{k+1}$ (where as above, we have chosen γ_l so that the local gain $g_l = 1$). The FMC J_{kk} is nonzero for $k = 0, \ldots, L-1$ (the signal arrives at layer L at time $L-1$ and can propagate no further). As long as the network amplification is superlinear, then the FMC will asymptote to a finite quantity in this range. Thus if we choose the number of neurons in the divergent architecture to grow as a power law $N_l = O(l^s)$, with $s \geq 1$, we achieve superlinear amplification and the area under the FMC will be extensive in the *depth* L of the network. However the number neurons N in the network will grow as $O(L^{s+1})$. Thus in terms of the number of neurons N, the area under the FMC will scale as $O(N^{\frac{1}{s+1}})$. The optimal scaling is achieved with $s = 1$, in which case the area under the FMC is $O(\sqrt{N})$. Thus a divergent fan out network with the number of neurons in layer l growing linearly in l is an example of a network that achieves the limit of $O(\sqrt{N})$ memory capacity for networks whose neurons operate within a finite dynamic range.

10. Details of the Nonlinear Dynamics in the Divergent Chain

The dynamical system from which the signal reconstruction was obtained in Fig. 5D of the main paper is described by Equation (1) in the main paper where the sigmoidal nonlinearity $f(x)$ was specifically taken to be $f(x) = R\tanh(\frac{x}{R})$. Thus the input-output relation of these neurons is approximately linear for inputs x in the dynamic range $-R < x < R$. We used a value of $R = 5$, or 5 times the signal strength.

11. Fisher Information in Continuous Time.

It is straightforward to generalize the above theory to contiuous time. Here we give the basics. We consider the time evolution

$$\tau\frac{d\mathbf{x}}{dt} = \mathbf{W}\mathbf{x} + \mathbf{v}s(t) + \mathbf{z}(t), \tag{38}$$

where $\mathbf{z}(t)$ is now a continous time white gaussian process with covariance $\langle \mathbf{z}_i(t_1)\mathbf{z}_j(t_2)\rangle = \epsilon\delta_{ij}\delta(t_1 - t_1)$. The solution is

$$\mathbf{x}(t) = \int_{-\infty}^{t} dt'\, e^{\mathbf{W}\frac{(t-t')}{\tau}}\mathbf{v}s(t') + \int_{-\infty}^{t} dt'\, e^{\mathbf{W}\frac{(t-t')}{\tau}}\mathbf{z}(t'). \tag{39}$$

The FMM is

$$\mathbf{J}(t_1, t_2) = \mathbf{v}^T e^{\mathbf{W}^T\frac{t_1}{\tau}}\mathbf{C}_t^{-1}e^{\mathbf{W}\frac{t_2}{\tau}}\mathbf{v}, \tag{40}$$

where the noise covariance $\mathbf{C}_t$ is

$$\mathbf{C}_t = \epsilon\int_{-\infty}^{t} dt'\, e^{\mathbf{W}\frac{t'}{\tau}}e^{\mathbf{W}^T\frac{t'}{\tau}}. \tag{41}$$

The spatial Fisher information $\mathbf{J}^s$ is

$$\mathbf{J}^s = \int_0^\infty dt\, e^{\mathbf{W}^T \frac{t}{\tau}} \mathbf{C}_t^{-1} e^{\mathbf{W} \frac{t}{\tau}}, \tag{42}$$

and for normal $\mathbf{W}$, $\mathbf{J}^s$ remains the $\frac{1}{\epsilon}$ times the identity, and hence the area under the FMC for any normalized $\mathbf{v}$ is $\frac{1}{\epsilon}$. Tr$\mathbf{J}^s$ for any matrix remains $\frac{N}{\epsilon}$, and so for nonnormal matrices, this remains a fundamental limit on the area under the FMC in continuous time as well.

12. Numerical Computation of Fisher Memory Matrices.

The spatial and temporal Fisher memory matrices can be computed through the method of lyapunov equations, which were used to obtain many of the results in Fig. 3 through 6 in the main paper. First, the noise covariance $\mathbf{C}_n = \sum_{k=0}^\infty \mathbf{W}^k \mathbf{W}^{kT}$ (here we work in units of $\frac{1}{\epsilon}$) obeys the discrete lyapunov equation

$$\mathbf{W}\mathbf{C}_n\mathbf{W}^T + \mathbf{I} = \mathbf{C}_n. \tag{43}$$

Similarly, the spatial FMM obeys the equation

$$\mathbf{W}^T\mathbf{J}^s\mathbf{W} + \mathbf{C}_n^{-1} = \mathbf{J}_s. \tag{44}$$

We used the MATLAB command dlyap to solve these equations.

Analogous results hold in continuous time. Both $\mathbf{C}_t$ in (41) and $\mathbf{J}^s$ in (42) obey continuous time lyapunov equations:

$$\mathbf{W}\mathbf{C}_t + \mathbf{C}_t\mathbf{W}^T + \mathbf{I} = 0, \tag{45}$$

and

$$\mathbf{W}^T\mathbf{J}_s + \mathbf{J}_s\mathbf{W} + \mathbf{C}_t^{-1} = 0. \tag{46}$$

We used the MATLAB command lyap to solve these equations to obtain the results in Fig. 6 of the main paper. The optimal input profile in Fig. 6A is simply the principal eigenvector of $\mathbf{J}^s$ where $\mathbf{W}$ is a simple finite difference approximation to the differential operator on the right hand side of Eq. 11 in the main paper.

References

[1] White O, Lee D, Sompolinsky, H (2004) Short-term memory in orthogonal neural networks. *Phys Rev Lett* 92:148102.