HHS Public Access

Author manuscript

Neural Comput. Author manuscript; available in PMC 2017 July 26.

Published in final edited form as:

Neural Comput. 2012 February ; 24(2): 332–390. doi:10.1162/NECO_a_00234.

Noise Tolerance of Attractor and Feedforward Memory Models

Sukbin Lim and

Center for Neuroscience, University of California, Davis, Davis, CA 95618, U.S.A

Mark S. Goldman

Center for Neuroscience; Department of Neurobiology, Physiology, and Behavior; and Department of Ophthalmology and Visual Sciences, University of California, Davis, Davis, CA 95618, U.S.A

Abstract

In short-term memory networks, transient stimuli are represented by patterns of neural activity that persist long after stimulus offset. Here, we compare the performance of two prominent classes of memory networks, feedback-based attractor networks and feedforward networks, in conveying information about the amplitude of a briefly presented stimulus in the presence of gaussian noise. Using Fisher information as a metric of memory performance, we find that the optimal form of network architecture depends strongly on assumptions about the forms of nonlinearities in the network. For purely linear networks, we find that feedforward networks outperform attractor networks because noise is continually removed from feedforward networks when signals exit the network; as a result, feedforward networks can amplify signals they receive faster than noise accumulates over time. By contrast, attractor networks must operate in a signal-attenuating regime to avoid the buildup of noise. However, if the amplification of signals is limited by a finite dynamic range of neuronal responses or if noise is reset at the time of signal arrival, as suggested by recent experiments, we find that attractor networks can out-perform feedforward ones. Under a simple model in which neurons have a finite dynamic range, we find that the optimal attractor networks are forgetful if there is no mechanism for noise reduction with signal arrival but nonforgetful (perfect integrators) in the presence of a strong reset mechanism. Furthermore, we find that the maximal Fisher information for the feedforward and attractor networks exhibits power law decay as a function of time and scales linearly with the number of neurons. These results highlight prominent factors that lead to trade-offs in the memory performance of networks with different architectures and constraints, and suggest conditions under which attractor or feedforward networks may be best suited to storing information about previous stimuli.

1 Introduction

Short-term memory is thought to be maintained by patterns of neural activity that are initiated by a memorized stimulus and persist long after its offset. Because memory periods are relatively long compared to biophysical time constants of individual neurons, it has been suggested that network interactions can extend the time over which neural activities are sustained (Brody, Romo, & Kepecs, 2003; Durstewitz, Seamans, & Sejnowski, 2000; Major

An online supplement is available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00234.

& Tank, 2004; Wang, 2001). However, the form of such interactions is currently unknown in most systems, and experimental and theoretical work has suggested a range of different network architectures that could subserve short-term memory.

A critical factor for robustly maintaining the memory of a stimulus is being able to resist the effects of noise that can accumulate over time. This is a particularly acute problem for the representation of analog values in memory. In many memory-storing paradigms during which neurophysiological recordings have been obtained (for example, see Aksay, Baker, Seung, & Tank, 2000; Goldman-Rakic, 1995; Robinson, 1989; Romo, Brody, Hernandez, & Lemus, 1999; Sharp, Blair, & Cho, 2001; Taube & Bassett, 2003), neurons have been shown to exhibit what appear to be continuously varying response levels that change in a graded manner with the stored stimulus value. With such analog representations, any noise-induced change in neural activity has the potential to affect the encoding of the stimulus. Thus, such networks are faced with apparently conflicting demands. On the one hand, the networks must be able to maintain the value of a signal in memory for long durations. On the other hand, the mechanism for performing this maintenance must keep the signal from being contaminated by excessive buildup of noise.

The most common models for how activity evoked by a transient stimulus is maintained over time are the so-called attractor networks. In attractor networks, individual neurons do not intrinsically maintain activity over long timescales and thus cannot in isolation store a memory. Instead, activity is maintained by positive feedback whereby neurons that are connected by excitatory or disinhibitory positive feedback loops maintain one another's activity following the offset of the external drive provided by the stimulus. In such models, the network structure determines which patterns of activity can be sustained by positive feedback, and typically only a small, specially designed set of patterns can be maintained. These maintained patterns of activity are called attractors of the network dynamics, because perturbing the dynamics away from such patterns leads to a rapid return to the attractor. A number of models of analog memory storage have utilized attractor dynamics (for review, see Brody et al., 2003; Durstewitz et al., 2000; Major & Tank, 2004; Wang, 2001), and recent analyses of neocortical data provide suggestive evidence for such attractors in tasks involving a working memory component (Ganguli, Bisley et al., 2008).

Recently both theoretical models (Ganguli, Huh, & Sompolinsky, 2008; Goldman, 2009; Mauk & Buonomano, 2004; Rabinovich, Huerta, & Laurent, 2008; Savin & Triesch, 2009; White, Lee, & Sompolinsky, 2004) and experimental observations (MacDonald, Lepage, Eden, & Eichenbaum, 2011; Pastalkova, Itskov, Amarasingham, & Buzsaki, 2008) have suggested instead how purely feedforward networks can store the memory of a stimulus in their transient dynamics. Experimentally, a feedforward progression of neuronal activity has been reported in hippocampal neurons during memory delay periods (MacDonald et al., 2011; Pastalkova et al., 2008), and theoretical work suggests mechanistically how an analog signal can be represented over time by activity that slowly propagates through a feedforward chain of neurons or, in recurrent networks, through a sequence of distinct and nonoverlapping patterns of network activity (Ganguli, Huh, et al., 2008; Goldman, 2009; White et al., 2004).

Here, we compare the performance of attractor and feedforward models in the presence of noise. Our work builds on the information-theoretic frameworks for quantifying memory performance of White et al. (2004) and Ganguli, Huh et al. (2008), who considered the performance of linear neural networks with discrete dynamics (i.e., defined with difference equations so that time is measured in discrete units that facilitate analytic calculation). We measure memory performance by calculating the Fisher information that is maintained about a transient stimulus at a time T into the future. Unlike previous work in neuronal systems (but as in the fluid mechanics example of Ganguli, Huh et al., 2008), the networks we study are defined by differential equations that consider the more realistic situation of continuous time dynamics. However, to facilitate analytic calculations, we also, when appropriate, compare to networks constructed with discrete dynamics.

The structure of this letter is as follows. First, in analogy to previous studies of linear networks with discrete dynamics, we analytically calculate the memory-storing performance of linear, continuous-time networks and determine the properties that optimize the Fisher information storage capacity of both attractor and feedforward networks. We then consider the effects of two nonlinearities suggested by neuronal recording data. First, we consider the effects on memory performance of reset mechanisms that, for example, remove noise from the system near the time of stimulus arrival (Churchland et al., 2010; Rajan, Abbott, & Sompolinsky, 2010; Weber & Daroff, 1972) or keep the network from entering the memory-storing state until the time of stimulus onset (Amit & Brunel, 1997; Durstewitz et al., 2000; Wang, 2001). Second, we consider the effect of limiting neurons to having a finite range of firing rates with which they can encode amemorized stimulus.

2 Material and Methods

In this letter, we compare the performance of attractor and feedforward network models in maintaining the memory of a brief, analog-valued stimulus for a fixed or known delay period T in the presence of noise. Here, we define the dynamics of each network model, as well as the Fisher information used to quantify the memory performance.

2.1 Linear Network Models

The structure of the networks considered in this letter is illustrated in Figure 1A. The goal of the network models is to maintain information about the scalar-valued amplitude s of a briefly presented stimulus occurring at time $t = 0$. In the majority of this work, we employ a firing rate model with continuous linear dynamics described by

$$\tau \frac{d\vec{r}}{dt} = -\vec{r} + \overleftrightarrow{W}\vec{r} + s\vec{v}\delta(t) + \sigma\vec{\xi}(t), \quad (2.1)$$

where $\vec{r}$ is a vector containing the firing rates of the N neurons in the network, each of which has intrinsic time constant τ. Inputs to the neurons include recurrent feedback from other neurons $W\leftrightarrow\vec{r}$, the pulse-like input $s\vec{v}\delta(t)$ whose strength s is to be remembered, and gaussian white noise $\sigma\vec{\xi}(t)$ of mean 0 and amplitude σ that is presented at all times (see

Figure 1A). Here, the elements of the connectivity matrix W_{ij} represent the strength of the synaptic connection from the jth to the ith neuron, and the elements of the input vector $\vec{v}$ indicate the relative weights of the inputs to each neuron. The input is presented as a transient pulse at time 0, modeled by a delta function $\delta(t)$.

Equation 2.1 can be solved analytically to give the form of the neuronal activities at time t when the stimulus strength is s and noise starts entering the system at time t_0:

$$\overrightarrow{r}(s,t)=\frac{s}{\tau}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau]\overrightarrow{v}+\frac{\sigma}{\tau}\int_{t_0}^{t}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau]\overrightarrow{\xi}(t')\,dt'. \quad (2.2)$$

The first and second terms describe the evolution of neural activities in response to the deterministic pulse-like stimulus and continually presented noisy input, respectively. The resulting mean neural activity and covariance matrix of neural variability are given as

$$\mathrm{mean}(\overrightarrow{r}(s,t))=\frac{s}{\tau}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau]\overrightarrow{v}, \quad (2.3)$$

$$\mathrm{Cov}(\overrightarrow{r}(s,t))=\frac{\sigma^2}{\tau^2}\overleftrightarrow{C}=\frac{\sigma^2}{\tau^2}\int_{t_0}^{t}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau]\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})^{\mathrm{T}}(t-t')/\tau]dt' \quad (2.4)$$

where the superscript T here denotes a matrix transpose and should not be confused with the memory period duration T. Note that the mean neural activity scales linearly with s and the magnitude of $\vec{v}$. We set $||\vec{v}|| = 1$ except in section 3.3, where we consider the effects of imposing a finite dynamic range on neuronal responses. The covariance matrix scales linearly with σ^2, and we denote the integral factor in equation 2.4 by $C\leftrightarrow$. In linear networks with no reset, we set t_0 to $-\infty$ to account for noise building up continuously at all times before the stimulus onset. When considering networks in which the appearance of the stimulus resets the noise, t_0 is set to 0, so that only noise presented after the stimulus onset affects memory performance.

To facilitate analytic calculations in cases where evaluation of the Fisher information is difficult, we also consider in Figures 6, 9, and 10 a network with discrete dynamics defined as in Ganguli, Huh et al. (2008). To derive these discrete dynamics, we start from the continuous differential equation:

$$\tau\frac{d\overrightarrow{r}}{dt}=(-\overleftrightarrow{I}+\overleftrightarrow{W})\overrightarrow{r}+s\overrightarrow{v}\delta(t)+\sigma\overrightarrow{\xi}(t),$$
$$\Rightarrow \frac{d\overrightarrow{r}}{dt'}=(-\overleftrightarrow{I}+\overleftrightarrow{W})\overrightarrow{r}+\frac{s}{\tau}\overrightarrow{v}\delta(t')+\frac{\sigma}{\sqrt{\tau}}\overrightarrow{\xi}(t'), \quad \text{where } t'=\frac{t}{\tau}.$$

Note that the additional factor $\sqrt{\tau}$ multiplying $\vec{\xi}(t')$ reflects that the variance, rather than standard deviation, of white noise grows linearly with time:

$\text{mean}\ (\xi_i(t_1)\xi_i(t_2)) = \text{mean}\ \left(\xi_i(\tau t_1')\xi_i(\tau t_2')\right) = \delta(\tau(t_1'-t_2')) = \tau^{-1}\delta(t_1'-t_2')$. By discretizing time with time step $\Delta t'$ and replacing the derivative of $\vec{r}(t')$ with a finite difference, we obtain the discrete dynamics approximating the continuous dynamics as follows:

$$\begin{aligned}\frac{\vec{r}(t'+\Delta t')-\vec{r}(t')}{\Delta t'} &\approx (-\overleftrightarrow{I}+\overleftrightarrow{W})\vec{r}(t')+\frac{s}{\tau}\vec{v}\,\delta(t')+\frac{\sigma}{\sqrt{\tau}}\vec{\xi}(t') \\ \Rightarrow\ \vec{r}(n+1) &\sim \overleftrightarrow{W}\vec{r}(n)+\frac{s}{\tau}\vec{v}\,\delta(n)+\frac{\sigma}{\sqrt{\tau}}\vec{\xi}(n).\end{aligned} \tag{2.5}$$

Here, the discrete dynamics approximation can be seen to be equivalent to updating the continuous dynamics equation with a time step equal to the intrinsic neuronal time constant τ, that is, with $\Delta t' = 1$. $W\leftrightarrow$, $\vec{v}$, and σ are the same as in equation 2.1, and $\delta(n)$ and $\vec{\xi}$ are the delta function and gaussian white noise in discrete time, respectively. The mean and covariance matrix for the above equation can be obtained as

$$\text{mean}(\vec{r}(s,n+1)) = \frac{s}{\tau}\overleftrightarrow{W}^{n}\vec{v}, \tag{2.6}$$

$$\text{Cov}(\vec{r}(s,n+1)) = \frac{\sigma^2}{\tau}\sum_{i=n_0}^{n}\overleftrightarrow{W}^{(n-i)}\left(\overleftrightarrow{W}^{(n-i)}\right)^{\mathrm{T}}. \tag{2.7}$$

In equation 2.7, n_0 replaces t_0 in equation 2.4 and the power of τ in the denominator is reduced by one relative to that in equation 2.4 because the differential dt' in equation 2.4 is set equal to τ in discrete dynamics and therefore cancels one factor of τ in the denominator.

The evolution of the network activity under linear dynamics can be computed by decomposing the activity into linearly independent modes. Here, we consider two such decompositions and use them to characterize the dynamics of the attractor and feedforward network models in the absence of noise.

In attractor networks, positive feedback sustains the activity evoked by the transient stimulus, for example, due to mutual excitatory connections between neurons that form a positive feedback loop (see Figure 1B). To identify such positive feedback, the eigenvector decomposition is commonly used to decompose the coupled networks into noninteracting modes of activity that can be considered independently. In the eigenvector decomposition, the pattern of neural activity $\vec{r}$ at any given time is defined in terms of the eigenvectors $\vec{q}_i$ and corresponding eigenvalues λ_i of the connectivity matrix $W\leftrightarrow$, which satisfy the equation $W\leftrightarrow\vec{q}_i = \lambda_i\vec{q}_i$ for $i = 1$ to N. Geometrically, the eigenvector decomposition corresponds to a change of basis into a new coordinate system whose axes are defined by the eigenvectors $\vec{q}_i$. In this new basis, the connectivity matrix $W\leftrightarrow$ is represented by a diagonal matrix $D\leftrightarrow$ having eigenvalues as diagonal entries such that $Q\leftrightarrow^{-1} W\leftrightarrow\ Q\leftrightarrow = D\leftrightarrow$, where the column vectors of $Q\leftrightarrow$ are the eigenvectors. When the eigenvectors are orthogonal to each other, $W\leftrightarrow$ is known as a normal matrix. In this case, the activity in each mode is equal to the

Cartesian projection of the network activity onto that mode, and there is no overlap among the activities in the different modes.

Activity in any eigenmode exhibits exponential growth or decay with a time constant $\tau_{eff}^{i}=\tau/|1-\mathrm{Re}(\lambda_i)|$. If $\lambda_i = 1$, activity is sustained without decay, and the mode can integrate any input perfectly. If Re $(\lambda_i) < 1$, activity decays with a time constant that decreases as λ_i decreases, and for Re $(\lambda_i) > 1$, activity grows exponentially. Attractor networks are defined by having a small number of modes (the attractor modes) with λ_i's much larger than the other eigenvalues. For such networks, activity in all except the attractor modes decays exponentially quickly to zero so that after a transient period, the only remaining activity is along these modes. The resulting subspace spanned by these modes is then called an attractor of the network dynamics. We illustrate a simple attractor network consisting of two symmetric excitatory neurons in Figure 1B. In such a network, the noninteracting modes correspond to the sum and difference of the activities and are called the common and difference modes (see Figure 1C). In the common mode, which is proportional to the average activity in the network, activity evoked by a transient input is maintained by mutual excitation of the neurons. By contrast, because the symmetric mutual excitation tends to make the neurons fire at equal rates, the difference mode is sharply attenuated by the network interactions, leading to rapid decay of any initial activity in this mode. Thus, after a transient period, only the activity along the common mode remains and the common mode is called an attractor of the network dynamics. Generally if there exist multiple modes with strong positive feedback, the signal can be stored in any of these modes, and the network is called a d-dimensional attractor network, where d denotes the number of modes with large λ_i's (see Figure 1D). In the special case when d equals one or two, the attractor is called a line or plane attractor, respectively.

Feedforward networks use a different mechanism for storing a signal. Rather than maintaining a stable pattern of activity through positive feedback, as in the attractor networks, the signal is carried by different neurons at different times. For example, in feedforward networks composed of neurons connected as a chain, the activity can be maintained as long as activity continues to propagate along a chain in which the activity in the previous neuron is passed onto the next neuron and filtered at each stage (see Figures 1E and 1G). The feedforward networks cannot be decomposed into a full set of N eigenmodes because, by the definition of an eigenmode, the activity that starts in an eigenmode remains in that mode (see Figure 1F, left). By contrast, the fundamental characteristic of the feedforward networks is that the activities of all neurons except the final one are passed onto the next neurons instead of being sustained.

The Schur decomposition is more suitable for describing feedforward networks (Ganguli, Huh et al., 2008; Goldman, 2009; Murphy & Miller, 2009). Rather than diagonalizing the matrix $W\leftrightarrow$, the Schur decomposition changes to a basis in which $W\leftrightarrow$ is triangular, that is, it decomposes any connectivity matrix into orthogonal modes that can have both feedforward and self-connections, but no feedback connections from later-stage neurons to earlier neurons. Formally, the Schur decomposition transforms the matrix $W\leftrightarrow$ into a lower triangular matrix $M\leftrightarrow$ such that $Q\leftrightarrow^{-1} W\leftrightarrow\ Q\leftrightarrow = M\leftrightarrow$, where the columns of $Q\leftrightarrow$ are the

orthogonal modes, called Schur modes, and the values of $M\leftrightarrow$ along the diagonal are the eigenvalues of $W\leftrightarrow$ (equivalently, $M\leftrightarrow$ can be made into an upper triangular matrix; Horn & Johnson, 1985). As in the eigenvector decomposition for normal networks, the diagonal entries of $M\leftrightarrow$ give the feedback of the Schur modes onto themselves (for normal $W\leftrightarrow$, the Schur and eigenvector decompositions are identical). If $W\leftrightarrow$ is nonnormal, then the Schur decomposition will contain nonzero lower triangular entries that correspond to feedforward connections between the Schur modes. In this case, activity may be transiently amplified as it propagates through the feedforward connections between modes, even when all the eigenmodes are decaying (i.e., when all $\lambda_i < 1$; Trefethen & Embree, 2005).

Here, we consider two types of feedforward networks. First, we consider literally feedforward networks for which the connectivity matrix $W\leftrightarrow$ itself is lower triangular with zeros along the diagonal, so that all connections are feedforward. Thus, the Schur mode patterns of activity correspond to individual neurons (see Figure 1F, right). Especially, we consider a simple chain-like structure whose connectivity matrix between neurons in the literally feedforward networks is of the form $W_{ij} = a > 0$ for all i and j such that $i = j + 1$ and zero otherwise. For networks with many neurons arranged in a chain, the propagation of activity can continue for a duration proportional to the chain length, with each neuron's activity peaking at different times. With this diversity of temporal profiles of neural activities, the network can generate persistent activity with a simple readout that linearly sums the activities of the different neurons with appropriate weights (see Figure 1G; Goldman, 2009).

Second, we consider recurrent matrices $W\leftrightarrow$ whose Schur decomposition $M\leftrightarrow$ has a feedforward(lower triangular, with zeros along the diagonal) structure; we call these functionally feedforward networks because the activity patterns defined by the Schurmodes, rather than the neuronal activity itself, propagate in a feedforward manner (Ganguli, Huh et al., 2008; Goldman, 2009; Murphy & Miller, 2009). As in the case of the literally feedforward networks, we consider simple functionally feedforward chains of the form $M_{ij} = a > 0$ for all i and j such that $i = j + 1$ and zero otherwise. An example of a functionally feedforward chain is shown in Figures 1Hand 1I, in which a two-neuron network with one excitatory and one inhibitory neuron is decomposed into common and difference modes by the Schur decomposition. The modes make a feedforward chain such that the activity of the difference mode drives the activity in the common mode (see Figure 1I). More neurons can form a longer functionally feedforward chain, allowing progression of activity patterns that persist for longer periods of time (see Figure 1J).

2.2 Fisher Information Measure for Memory Performance

To achieve good memory performance, a network must maintain a memory of the stimulus while resisting the excessive accumulation of noise. The ability to achieve this can be quantified as the ratio between the amount of signal and noise stored in the system at a given time following stimulus offset. Here we use a closely related measure, the Fisher information I_F, which quantifies the amount of information carried about a signal by the distribution of neural activities and which, for linear networks and gaussian white noise, is

shown below to represent a ratio between the factor by which the network amplifies signals and the amount of noise accumulated by the network (Ganguli, Huh et al., 2008).

To get an intuition for this measure, we show in Figure 2A how to compute the Fisher information for an example of the activity of a single neuron (or a single eigenmode of an attractor network) in the presence of noise. The neuron (or mode) must distinguish between different pulse-like stimuli of amplitudes s and $s + \delta s$ that it receives at time 0. Making this discrimination more difficult, noise is presented to the neuron (or mode) continually in time (see Figure 2A). We model the transient stimulus as a delta function $\delta(t)$ so that the stimulus causes a jump in the mean neural activity at time 0, with size proportional to the stimulus strength s or $s + \delta s$ (see Figures 2B and 2C, thick lines). Due to the noise, each presentation of the stimulus leads to a different trajectory so that there is trial-to-trial variability in the response (the black and gray noisy trajectories in Figure 2B).

The memory of the stimulus is carried by the distribution of the firing activities of the neurons. In order to perform well in maintaining the distinction among stimuli, the distributions for different stimuli must remain well separated: the more the noise makes the two distributions overlap, the greater will be the corruption of the stored memory. In linear networks, the mean activities of the neurons (gray circle and black asterisk in Figure 2C) carry the information about the presented stimulus, and the signal gain is measured as the difference in the mean activities $\delta\langle r\rangle$ divided (i.e., normalized) by the separation δs of the signals to be discriminated (see Figure 2D). The noise in the neural activities is given by the spread in the firing rate distribution. The Fisher information I_F conveyed by the network is defined as the ratio of the square of the signal gain to the noise variance at time T. Thus, either a wider separation between the means (high signal gain) or narrower distributions about these means (small noise variance) lead to higher Fisher information.

Formally, the Fisher information is defined as

$$I_F(s,t)=\left\langle \left(\frac{\partial}{\partial s}\log\,(p(\vec{r}(t)|s))\right)^2\right\rangle_{\vec{r}(t)}=\left\langle -\frac{\partial^2}{\partial s^2}\log\,(p(\vec{r}(t)|s))\right\rangle_{\vec{r}(t)}, \quad (2.8)$$

where $\langle\,\rangle_{\vec{r}(t)}$ denotes taking the average over all $\vec{r}(t)$ for a given s.

For linear networks with gaussian white noise, the distribution of neural activity remains gaussian for all times and therefore can be described completely by the mean and noise covariance matrix of the firing rate distribution (see equations 2.3 and 2.4). Using that the logarithm of a gaussian distribution is proportional to the squared deviation from the mean divided by the covariance matrix, that is, $\log p(\vec{r}|s) = -c_1\,[\vec{r} - \text{mean}(\vec{r})]^T\,[Cov(\vec{r})]^{-1}\,[\vec{r} - \text{mean}(\vec{r})] + c_2$, where c_i are constants, I_F is in this case given by

$$I_F(t) = \vec{v}^T \exp\left[\left(-\overleftrightarrow{I} + \overleftrightarrow{W}\right)^T t/\tau\right] \frac{1}{\tau} \left[\frac{\sigma^2}{\tau^2} \overleftrightarrow{C}\right]^{-1} \frac{1}{\tau} \exp\left[\left(-\overleftrightarrow{I} + \overleftrightarrow{W}\right) t/\tau\right] \vec{v}$$
$$= \frac{1}{\sigma^2} \vec{v}^T \exp\left[\left(-\overleftrightarrow{I} + \overleftrightarrow{W}\right)^T t/\tau\right] \overleftrightarrow{C}^{-1} \exp\left[\left(-\overleftrightarrow{I} + \overleftrightarrow{W}\right) t/\tau\right] \vec{v}. \quad (2.9)$$

Here, $\frac{1}{\tau}\exp\left[\left(-\overleftrightarrow{I} + \overleftrightarrow{W}\right) t/\tau\right] \vec{v}$ is the derivative of the mean with respect to s, called the signal gain. Thus, equation 2.9 shows that I_F is of the form of a (matrix) ratio between the signal gain squared and the noise covariance. Note that $\sigma^2 I_F(t)$ is independent of the stimulus strength s and injected noise level σ^2 and depends only on the properties of the network connectivity. Therefore, in the following, we calculate $\tilde{I}_F = \sigma^2 I_F(t)$ instead of I_F and often refer to $\tilde{I}_F$ as the Fisher information for brevity (see Figure 2D). Note that this quantity has the same units as σ^2 since I_F is unitless (and assuming that s is unitless).

In equation 2.9, the readout of the network activities is not specified. In particular, in a linear system with gaussian noise, it can be shown that $\tilde{I}_F$ is greater than or equal to the (signal gain)-to-(noise gain) ratio in any linear readout of the network (see section A.1 in the appendix). The equality holds when the linear readout is in the direction $C\overleftrightarrow{}^{-1} \exp[(-I\overleftrightarrow{} + W\overleftrightarrow{})t/\tau]\vec{v}$ (see the optimal linear estimator in population decoding, as in Salinas & Abbott, 1994; Sompolinsky, Yoon, Kang, & Shamir, 2001). Note that in the feedforward networks, the optimal linear readout will generally vary over time, reflecting that information about the signal propagates from earlier to later stages in the feedforward chain.

3 Results

Here we compare how attractor and attractorless (literally feedforward or functionally feedforward) models perform in storing the amplitude of a brief stimulus. The memory performance is measured by the Fisher information, a measure of how much the network amplifies the signal corresponding to the stimulus compared to how much it amplifies ambient noise (see section 2.2). In section 3.1, we consider purely linear networks that allow us to isolate how the structures of attractor and feedforward networks influence memory performance in the absence of nonlinear influences. Then we consider the effect of two biologically observed nonlinearities. In section 3.2, we consider a condition that we term a reset nonlinearity under which either noise does not begin to accumulate strongly until the memory period commences (Amit & Brunel, 1997; Durstewitz et al., 2000; Wang, 2001) or in which noise that is present before the stimulus arrival is “reset” by the appearance of the stimulus (Churchland et al., 2010; Rajan et al., 2010; Weber & Daroff, 1972). In section 3.3, we consider the effects of restricting neurons to having a finite range of firing rates with which they encode a stimulus.

3.1 Linear Networks

In this section, we compare the memory performance of attractor and feedforward network models with continuous linear dynamics. This allows us to focus on how signal and noise are propagated through the network as a function of the structure of the network connectivity without the complicating influence of nonlinearities. We first compute the Fisher information $\tilde{I}_F = \sigma^2 I_F$ in line attractor networks and then extend our results to higher-

dimensional attractor networks. Then we compute $\tilde{I}_F$ for feedforward networks and compare their performance to that of the attractor networks.

3.1.1 One-Dimensional Attractor Networks—We first consider the line attractor networks, which are defined by having only a single stationary or slowly decaying (or possibly growing) pattern of activity that defines the attracting mode (see Figures 1B and 1C). When stimulated by a brief stimulus, both signal and accumulated noise in line attractor networks quickly converge to this attractor. As a result, for times after the transient responses of the nonattractor modes decay away, all information conveyed by the line attractor networks is contained in the attractor mode, and we can closely approximate the Fisher information $\tilde{I}_F$ by the (signal gain)-to-(noise gain) ratio in this mode (see Figure 2).

The memory-storing performance of the line attractor models reflects a balance between two factors. First, the network must be able to sustain the signal for the full duration of the memory period. As shown in Figure 1C, this is accomplished in attractor networks by having sufficiently large positive feedback in the attracting mode. Second, the network must not accumulate too much noise over time. Since noise is assumed to be presented at all times, including prior to stimulus onset, this implies that inputs to the network should not be sustained indefinitely or noise will accumulate without bound. Thus, there is a trade-off in attractor networks between sustaining signals for sufficiently long to maintain signal strength and having enough decay of signals that noise does not accumulate excessively.

To quantify this trade-off between sustaining the signal and accumulating noise, we examine the memory performance of the attractor network in terms of the amount of positive feedback a in the attracting mode, where a is the eigenvalue associated with the attracting eigenmode and the time constant of decay (or growth, for Re(a)>1) of activity in this mode is given by $\tau_{eff} = \tau/|1 - \mathrm{Re}(a)|$. When the feedback is too weak (see Figure 3A), the signal decays quickly to zero and any memory of the initial stimulus amplitude s is forgotten. Thus, $\tilde{I}_F$ is close to zero in this case (see Figure 3D). Increasing the recurrent feedback leads to slower decay of signals corresponding to the memorized stimulus, and when the feedback is tuned to be large enough to offset the intrinsic leak of the neurons ($a \approx 1$), the mean responses to different amplitude stimuli stay well separated (black and gray thick traces in Figure 3C). However, because noise along the attractor mode is subject to the same dynamics as signals along this mode, noise also accumulates without decay. Because noise is present at all times before the stimulus arrives, this leads to an extremely large variance in the responses (see Figure 3C; note the wide spread of trajectories even before time 0). For networks that are either nonforgetful ($a = 1$) or amplifying ($a > 1$), this noise becomes infinite in magnitude so that the Fisher information $\tilde{I}_F$ is zero (see Figure 3D). Thus, there is an optimal amount of feedback in linear attractor networks, and a corresponding optimal time constant of decay of network activity, that balances having a long time constant so that the signal does not decay and having a short time constant so that noise does not accumulate too much (see Figures 3B and 3D).

This example shows that the attractor network performance is benefited by having an imperfect memory-holding mechanism. To find the optimal forgetting time constant of network activity decay, we analytically calculated $\tilde{I}_F$ for the line attractor networks (see

sections A.2.1 and A.2.2). We find that $\tilde{I}_F$ achieves its maximum when the decay time constant of network activity $\tau_{eff,opt} = 2T$, where T is the duration over which the signal is to be stored. Thus, the memory duration T sets the scale for the optimal network decay time. When activity decays much faster than the memory duration T, the signal decays away before the end of the memory period. When activity decays much more slowly than T or grows exponentially, noise accumulation overwhelms the signal.

3.1.2 Higher-Dimensional Attractor Networks—We next extend the result for line attractor networks to higher-dimensional attractor networks having many modes with slow decay of activity. In line attractor networks, the input is stored along the one-dimensional attractor. In higher-dimensional attractor networks, signal and noise can accumulate in any direction spanned by the multiple attractor modes. We show that these extra dimensions do not affect $\tilde{I}_F$ of networks with optimally arranged inputs and readout. However, for imperfectly arranged inputs or outputs, we find that the memory performance of the line attractor networks is sensitive to the input direction but insensitive to the readout direction. In contrast, higher-dimensional attractor networks are more sensitive to the readout but less sensitive to the input direction.

For illustration, in Figure 4 we compare the line attractor networks to plane attractor networks defined by having two slowly decaying modes of activity. To convey geometrical intuition, we plot only the modes $\vec{q}_1$ and $\vec{q}_2$ with the two largest eigenvalues and assume all the eigenmodes are orthogonal. For simplicity, we assume that each attracting mode has the same eigenvalue, so that all directions in the attracting plane have equal decay times and the signal can be stored equally well in any direction on the plane. Likewise, noise accumulates in the same manner in any direction on the plane. Then, when gaussian white noise is presented equally to all neurons and thus to all orthogonal modes (see Figure 4A), the resulting noise at any time is also equivalent in all directions of the plane (see Figure 4C). By contrast, in the line attractor networks, noise along directions other than the line attractor is filtered out so that noise along the attracting mode has a larger variance than that along the other modes (see Figure 4B).

To maximize the strength of the signal carried by attractor networks, the inputs should be arranged so that none of the input is lost due to being sent into decaying modes whose amplitudes quickly fall to zero. In line attractor networks, this corresponds to aligning the input direction $\vec{v}$ along the direction of the attracting eigenmode $\vec{q}_1$ (see Figure 4E). When there is more than one attracting mode, as in the plane attractor networks, the optimal input direction $\vec{v}$ can be along any linear combination of these attracting modes (see Figure 4F). The Fisher information $\tilde{I}_F$ is proportional to the square of the projection of the signal onto the attracting modes. Thus, the Fisher information is identical and equal to its maximal value for both the line and higher-dimensional attractors as long as the input is aligned along the subspace defined by the attracting eigenmodes (see Figure 4D, $\theta = 0$). If $\vec{v}$ is not aligned along the attracting modes, a portion of the signal is lost to the decaying modes and $\tilde{I}_F$ decreases from this maximum value. In the line attractor, there exists only a single attracting mode storing the signal, so $\tilde{I}_F$ decreases as $\vec{v}$ deviates from $\vec{q}_1$ (see Figure 4D, solid curve). On the other hand, $\tilde{I}_F$ stays the same in the plane attractor networks for any $\vec{v}$ in the attracting plane (see Figure 4D, dashed line). Note that $\tilde{I}_F$ in the plane attractor networks

would decrease as in the line attractor networks if $\vec{v}$ were to deviate from the attracting plane (not shown). However, because the dimension of the plane attractor is higher than that of the line attractor, the alignment of the input vector is less restrained in the plane attractor (or, more generally, higher-dimensional attractor) networks.

Next we consider the arrangement of the readout for the maximal memory performance. As discussed in section 2.2, the Fisher information measure $\tilde{I}_F$ implicitly assumes an optimal readout because $\tilde{I}_F$ is equal to the (signal gain)-to-(noise gain) ratio along the optimal linear readout direction. However, for nonoptimal readout, the memory performance may be less than $\tilde{I}_F$ and the sensitivity to the direction of the readout may differ between the line and plane attractor networks (see section A.2.4).

In the line attractor networks, mistuning of the readout does not have much effect on the (signal gain)-to-(noise gain) ratio because the signal and noise accumulate along the one-dimensional attracting mode and their ratio is maintained for the projection onto any readout direction (see Figure 4H). Thus, the memory performance of line attractor networks remains near the maximal $\tilde{I}_F$ even when the readout direction is well away from the attractor mode (see Figure 4G, solid line). Only when the readout direction becomes close to orthogonal to the attractor direction, so that the signal becomes smaller than or comparable to the small but finite noise accumulated in the nonattracting modes, does the memory performance fall off by a significant amount. By contrast, in plane attractor networks, the memory performance is far more sensitive to the direction of the readout (see Figure 4G, dashed line). Because noise develops along all attractor dimensions but the signal lies only along the direction $\vec{v}$ defined by the input, projections that are not along the input direction pick up additional noise and lower the (signal gain)-to-(noise gain) ratio (see Figure 4I). Hence, optimal performance requires amore precise readout mechanism in higher-dimensional attractor networks.

In summary, both the line attractor and higher-dimensional attractor networks were shown to have the same maximal memory performance, as characterized by the Fisher information. For the line attractor networks, memory performance was highly insensitive to the readout direction but more sensitive to the direction of the input. Conversely, for higher-dimensional attractor networks, the memory performance was highly sensitive to the readout direction but less sensitive to the direction of the inputs. These results suggest that line attractor networks might be more useful if the stored memory needed to be used by multiple networks that each projected activity along a different direction. By contrast, higher-dimensional attractors might be more useful in storing memories that can arrive from multiple input networks that each sends in different input patterns encoding the stored variable.

3.1.3 Feedforward Networks—Next, we compute the memory performance of linear feedforward networks and focus on networks with a chain-like structure that were proposed recently as a neural substrate for short-term memory storage (Ganguli, Huh et al., 2008; Goldman, 2009; White et al., 2004). A critical difference between feedforward and attractor networks is that, unlike in attractor networks, activity in feedforward networks eventually exits out the end of the network. Thus, the memory of any input is lost after some finite time in feedforward networks. Although this finite time of signal propagation might at first seem to be disadvantageous, finite memory duration can be advantageous because it prevents

noise from building up in the network (Ganguli, Huh et al., 2008). These relative advantages and disadvantages are quantified below, where we compute the Fisher information conveyed by linear feedforward networks and compare their performance to that of attractor networks.

Here, we consider simple feedforward chains having uniform strength a of the feedforward connections and compute $\tilde{I}_F$ as a function of a. When the strength of the connections is weak, the activity decays before it reaches the last stage and $\tilde{I}_F$ is close to zero (see Figures 5A and 5C). On the other hand, if the feedforward connections are stronger, the signal decays more slowly (for $a < 1$) or can grow exponentially (for $a > 1$). Noise entering the system at any given time similarly gets amplified as it passes down the chain. However, unlike in the attractor networks, accumulation of noise in the feedforward networks is limited because noise exits the system when it reaches the end of the chain. Moreover, if signals are amplified along the feedforward chain, then inputs entering the first stage of the network get amplified more than inputs entering later stages. Thus, by arranging to have the signal enter the network at the first stage, the network can make the signal at time T arbitrarily larger than the noise entering at later times by using strong connection strengths a that allow the signal to be amplified faster than noise enters the system (see Figure 5B). This implies that $\tilde{I}_F$ can increase indefinitely with increasing a, so that linear feedforward networks could in principle convey signals to arbitrary precision (see Figure 5C; for how this result changes when neurons have a finite dynamic range, see section 3.3). This result is consistent with that of Ganguli, Huh et al. (2008) who also showed a monotonic increase of $\tilde{I}_F$ with a in models with discrete dynamics.

Comparison with the attractor networks reveals two important features of the feedforward networks that reflect the advantages and disadvantages of having finite memory duration (see Figure 5D). First, because the feedforward networks can transiently amplify signals over the memory period but still remove noise due to the eventual exiting of signals from the chain, these networks can greatly outperform the attractor networks. Second, for smaller values of a, the attractor networks outperform the feedforward networks. This latter result reflects the smearing out of signals by the continuous dynamics (see section 3.2 and Figure 6) and differs from that found when comparing feedforward and attractor networks with discrete dynamics (Ganguli, Huh et al., 2008), in which the feedforward networks outperformed the attractor networks for all settings of a.

In summary, with continuous buildup of noise, short-term memory networks need to forget to prevent the excessive accumulation of noise. In feedforward networks, this forgetting mechanism is inherent in the finite length of the feedforward chain, and the networks can amplify signals transiently without noise building up in an unbounded manner. In contrast, in attractor networks, the duration of signal and noise accumulation is not limited, and a perfect memory holding mechanism is inferior to a forgetful one in which signal decay and noise accumulation are optimally balanced. Comparing the purely linear attractor and feedforward network models, we find that the feedforward networks can outperform the attractor models due to their ability to transiently amplify signals without building up excessive noise. In the following sections, we show how these results may change in the presence of select, biologically motivated nonlinearities.

3.2 Networks with a Reset Mechanism

In the previous section, we found that the feedforward networks stored more information than the attractor networks because they could amplify the signal without infinite buildup of noise. In contrast, the attractor networks needed to be forgetful in order to prevent infinite noise buildup. However, what if the accumulation of noise before the memory period is limited, or there exists a mechanism to reset the network state near the onset of the signal? Recent experimental studies in several cortical regions showed that variability in neural activity is reduced with stimulus onset (Churchland et al., 2010), and theoretical work suggests this may be a general feature of many nonlinear recurrent networks (Rajan et al., 2010). Alternatively, networks may not switch into a memory-storing state that accumulates noise until close to the start of the memory period; for example, such a switch may occur due to a change in network state triggered by attention or neuromodulation (Amit & Brunel, 1997; Durstewitz et al., 2000; Wang, 2001). Finally, if a network receives feedback about its deviation from a desired level and is able to correct these errors, then infinite buildup of errors is also prohibited. For example, in the oculomotor system, drift in the networks that control eye position triggers corrective saccades that can correct errors caused by accumulation of noise or systematic drift of network activity (Weber & Daroff, 1972). Motivated by these examples, we here consider the effect of allowing a network to reset its activities with the arrival of a signal and remove previously accumulated noise. Note that the level of spontaneous activity before the memory period can differ between these different reset mechanisms, being low even before stimulus arrival if there is a stable low-rate spontaneous state (Amit & Brunel, 1997; Durstewitz et al., 2000; Wang, 2001) or being higher during spontaneous activity and reduced only at stimulus onset (Rajan et al., 2010; Churchland et al., 2010); however, for any reset mechanism, the variability of network activity would be low at the beginning of the memory period. For simplicity, we implement this “reset nonlinearity” by setting the noise to zero at the time of signal arrival, so that noise accumulates only during the memory period of duration T.

First, we consider the attractor networks. Before the signal arrives, noise accumulates, and this accumulation can grow without bounds along any nondecaying ($\alpha \geq 1$) modes of the network (see Figures 6A and 6B). However, at the time of signal arrival, the reset mechanism quenches the neural activities to zero. Therefore, only noise presented after $t = 0$ degrades the memory performance, and unlike in the attractor networks without reset, perfectly integrating or exponentially growing modes can convey information about the signal. In fact, $\tilde{I}_F$ monotonically increases with increasing α (see Figure 6C and section A. 2.2), showing that memory performance is enhanced by amplifying signals in the network. This result can be understood by recalling that the signal is presented only at time $t = 0$, whereas noise is equally presented during the entire memory performance: by amplifying the input over time, more weight is given to inputs at earlier times, allowing the signal to be amplified faster than noise enters the system. In the limit of infinite signal amplification, the signal can be made arbitrarily larger than the noise, so that the Fisher information $\tilde{I}_F$ approaches infinity and signals can be discriminated with perfect precision.

In feedforward networks, the reset mechanism also enhances memory performance by removing noise accumulated prior to the stimulus onset (see Figures 6D and 6E) and thereby

increases $\tilde{I}_F$ relative to a network without reset (see Figure 6F). However, the increase in $\tilde{I}_F$ for the feedforward chain network is not nearly as large as that in the attractor networks. This reflects that, even without an externally imposed reset, the finite length of the feedforward chain already provides a mechanism for removing noise because noise exits the system when it reaches the end of the chain.

By comparing the performance of the attractor and feedforward networks, we find that the attractor networks perform better than the feedforward networks when there exists a reset mechanism (see Figure 6I). To understand what factors contribute to this result, we first consider the case of networks with discrete dynamics (Ganguli, Huh et al., 2008). In feedforward chains exhibiting discrete dynamics, all activity at one stage of the chain passes in the next time step to the following stage. When $a = 1$ (see Figure 6G), activity is passed from neuron to neuron in discrete time steps without loss of amplitude. Thus, there is no smearing out of activity across neurons as the activity progresses through the feedforward chain. In this case, the Fisher information for the feedforward and attractor networks is identical (see section A.4), reflecting a deep similarity between the feedforward and attractor networks: whereas in attractor networks, activity at each time step is sent from a given neuron (or mode) onto the same neuron (or mode), in feedforward networks, the activity is similarly propagated over time, but instead from one neuron to the next (see Figure 6G; for discussion of a more general mathematical formalism that formalizes the similarity between feedforward and attractor networks, based on pseudospectral analysis (Trefethen & Embree, 2005) see the supplement of Goldman, 2009).

The discrete dynamics example suggests that the key factor explaining the poorer performance of feedforward networks with continuous dynamics is the spreading of activity across neurons or modes that occurs in the continuous feedforward networks. This spreading has two effects. First, it reduces the amplitude (vector length) of the signal carried by the network by spreading activity across different neurons. To understand this, note, for example, that dividing a signal equally among two neurons, so that the activity can be described by a vector ($s/2$, $s/2$), reduces the vector amplitude of the signal by a factor of $\sqrt{2}$ compared to when the entire signal is carried by a single neuron, that is, (s, 0). This loss of signal is evident in Figure 6H, which shows how the signal gain decreases over time in the continuous feedforward networks (dashed curve) but is maintained at a constant level in the discrete feedforward networks (circles). Second, the spreading of activity causes activity to exit the network before the end of the memory period. This is evidenced by the dip in signal gain seen near the end of the memory period ($T = 2$) in the same figure. Due to both the spreading of the signal and the loss of signal out the end of the chain, $\tilde{I}_F$ for continuous-time feedforward networks becomes lower than that of the attractor networks (see Figure 6I).

In the example above, we showed that the attractor network outperforms the feedforward network with the same strength between the modes. However, a more direct biological constraint is to compare the attractor and feedforward networks when the maximum connection strength between neurons is held fixed. In this case, we again find that the optimal attractor network outperforms any (literally or functionally) feedforward network. The proof is given in section A.5. There, we show that if the maximal synaptic strengths between the neurons are bounded by $w_{\max}$ and the eigenmodes or Schur modes are

orthogonal to each other, the connectivity strength between the modes is bounded and given by $Nw_{\max}$. For the attractor networks, we find a network that reaches this bound. By contrast, the feedforward networks cannot achieve this bound for all connections between modes. Further, even if we assume there exists a feedforward network with all connections between modes set to $Nw_{\max}$, the previous comparison of memory performance for a given connectivity between the modes shows that the attractor network still outperforms the feedforward network when both networks have connectivity strength $Nw_{\max}$ between modes (see Figures 6H and 6I). Thus, the optimal attractor network outperforms any feedforward network when the synaptic connectivity between the neurons is bounded. Alternatively, we can also consider the constraint that the total postsynaptic weight is bounded. We note that at least for the case of excitatory (literally feedforward or excitatory attractor) networks, the maximal memory performance under this constraint corresponds to the previous result, in which there were fixed connections between modes (see Figures 6H and 6I). The optimal feedforward networks use this maximal connection strength $w_{postsynaptic,\max}$ between all neurons, and the optimal attractor networks have a maximum eigenvalue $w_{postsynaptic,\max}$. Thus, the optimal attractor networks outperform the optimal feedforward networks.

In summary, in this section we considered the effect on memory performance of reset mechanisms that remove accumulated noise at the time of signal arrival. As a result of this reset, the attractor networks could amplify signals without having a buildup of noise prior to signal arrival affect the memory performance. Moreover, for a given level of amplification between the neurons or modes, the attractor networks perform better than the feedforward networks since the activity in the feedforward networks spreads out along the chain and is lost when it exits the end the chain.

3.3 Bounds on the Neuronal Activity

In the previous sections, we found that the networks exhibiting the best memory performance depended on strong amplification of signals that led to large and potentially unbounded growth of network activity. However, unbounded amplification of activities is not possible since neurons have a limited dynamic range. This limited range assumes several forms. Biophysically, there are absolute limits on the maximal firing rates that neurons can achieve (typically in the hundreds of Hz) due to postspike refractoriness. Additionally, neurons have been suggested experimentally (Baddeley et al., 1997) to have constraints on the average firing rates they can assume over long time periods. During working memory periods, most neurons do not sustain average firing rates beyond several tens of Hz, even though trial-to-trial fluctuations may be much larger than this for brief periods of time.

Given these observations, in this section we consider the effects of imposing constraints on the range of firing rates with which neurons encode signals in memory. Throughout much of the discussion, we confine ourselves to limits on the mean firing rates attained over the course of the memory period. This constraint is motivated by the observation that memory neurons typically have much lower (trial-averaged) mean firing rates than are allowed by their moment-to-moment biophysical constraints, and for analytical tractability is implemented by adjusting the inputs of an otherwise linear network such that the activity never exceeds hard bounds on the mean rates. Then, in order to gain insights into the effects

of constraints on the absolute size of neuronal fluctuations, we consider what happens when we additionally apply a hard bound on the variance of firing rates about the mean.

3.3.1 Effects of Bounded Rates on the Form of External Inputs—Before considering the effects of a finite dynamic range of firing activity in specific networks, we investigate the constraints it places on the form of the input vector. Since the input vector drives the mean firing activity in linear networks (see equation 2.3), putting a limit on the mean firing rate correspondingly constrains the input vector. Note that this differs from our treatment of networks with unconstrained firing rates, for which we normalized $||\vec{v}||$ to 1 because the Fisher information for all networks simply scaled with the square of the input magnitude and could be made arbitrarily large by increasing $||\vec{v}||$; see equation 2.9 and compare to Ganguli, Huh et al., (2008), who assumed that $||\vec{v}||$ is still limited to 1 under a similar constraint on the dynamic range). To implement the constraint on mean firing rates, we assume that each neuron has its mean (absolute) activity bounded by a maximal value r_0 (where negative rates can be considered as the firing rate of an "anti-neuron" with opposite stimulus preference; Shadlen, Britten, Newsome, & Movshon, 1996). This is illustrated geometrically in Figure 7A, which shows that the hard limits on the mean firing rate define a hypercube (a square in 2D) in the space of possible firing rates. To stay within these limits, the magnitude of the input vector must be set such that the mean firing rate of any given neuron never exceeds its bound. As we show further, this leads to different maximal amplitudes of the input vector for different network architectures.

The constraint on the mean firing rate has immediate implications as well for the spatial pattern of inputs that are conveyed most faithfully by the network. Given the limitation on how much information any given neuron can convey with its limited dynamic range, the maximal information carried by a network is achieved when all neurons are used and each of these neurons uses its full dynamic range. When this idea is represented geometrically, information storage is maximized if the attracting or Schur modes of the networks lie along the directions pointing to the vertices of the hypercube that defines the maximal range of mean responses (see Figure 7B, open circles). With this arrangement, the strength (vector length) of signals conveyed by the networks is proportional to $\sqrt{N}$, illustrating the benefits of having more neurons in the network when each individual neuron has limited dynamic range. As shown in section A.6, this scaling leads to the Fisher information for the best attractor and feedforward networks scaling with the network size N (see Figure 8G).

3.3.2 Attractor Networks with Finite Dynamic Range—We first consider the performance of attractor networks with a limited range of mean firing rates and no reset nonlinearity. In this case, our results follow closely that found for the linear networks of section 3.1: if the decay time constant of the network is too small, the signal decays to zero before the end of the memory period (see Figure 8A, bottom trace, and Figure 8B, probability distribution of activity in this mode). By contrast, if the network does not exhibit decay or decays too slowly, then noise builds up to the point that the signal becomes overwhelmed by noise (see Figures 8A and 8B, top traces). To perform optimally, the network must balance signal decay and the accumulation of noise, and we find that the optimal time constant of network decay to achieve this balance is $\tau_{eff}=\frac{\tau}{1-Re(\alpha)}=2T$ (see

Figures 8A and 8B, middle traces; in Figure 8E, the dotted line shows memory performance as a function of a). We note that this result is identical to that found in section 3.1 for networks with no bounds on the mean firing rates. This identical result reflects that, due to the need to remove noise through decay of network activity, the activity of the network never needs to be constrained by the limited dynamic range. However, the limited dynamic range does bound the total information that can be conveyed by the network because it constrains the amount of input that the network can receive at the time of the stimulus.

When there is a reset nonlinearity at the time of signal arrival, the optimal strength of network feedback does change compared to that obtained without a limited dynamic range. Recall that, without limits on the dynamic range (see section 3.2), the optimal networks were found to have strong feedback ($a > 1$) so that they could amplify their signals faster than noise entered the system. With a finite dynamic range, unconstrained amplification of activity is no longer possible. Figure 8C illustrates mean trajectories of three modes with different recurrent feedback a, or, equivalently τ_{eff}, corresponding to decaying ($\tau^i_{eff} > 0$), perfectly integrating ($\tau^i_{eff} = \infty$), and exponentially growing modes ($\tau^i_{eff} < 0$), respectively. Compared to the decaying mode, the perfectly integrating mode performs better because it maintains the signal faithfully yet has only a finite buildup of noise due to the reset at time $t = 0$. For the amplifying mode that exhibits exponential growth (the increasing trajectory in Figure 8C), we set $||\vec{v}||$ to a value such that activity propagates linearly through the network until, at the end of the memory period, it just reaches the limit of the dynamic range. Thus, the maximal signal that can be carried by the network is identical to that obtained in the perfectly integrating mode. However, due to the amplification, noise accumulates faster than in the perfectly integrating mode, resulting in a larger variance in the neuronal firing rates (see Figure 8D). Thus, with a finite dynamic range and reset of activities with arrival of the signal, we find that perfectly sustaining the activity during the memory period is optimal (see Figure 8E; see section A.6 for the detailed calculation).

More generally, in both the networks with and without a reset nonlinearity, we find that $\tilde{I}_F$ for optimally tuned networks increases linearly with the number of neurons N (see Figure 8G) and decreases inversely with the memory duration T (see Figure 8H). Thus, it scales with N/T. The former result reflects that more neurons allow more signal to be carried by the network, as discussed in the preceding section. This latter result reflects the accumulation of the continually presented noise, which results in a linear increase in noise variance over the memory period (see section A.6).

Note that the constraint on the mean firing rate alone may allow infinite accumulation of noise, which is not biologically plausible. Thus, in addition to constraining the mean activity, we further consider bounds on the variance of neural activity. For analytical tractability, we place a simple bound on the maximal variance of activity without affecting the underlying dynamics: if the variance of activity obtained from the dynamics exceeds the bound, the variance is saturated and the variance is set to the bound (see section A.6).

To see the effect of the bound on variance, we note that there are two possible cases. In the first case, the Fisher information is maximized in a regime where the noise variance bound is

saturated. Since the noise variance is saturated, this corresponds to a regime in which the signal-to-noise, at best, is very low and the information transmission is exceptionally small. Thus, without some additional nonlinear mechanism for noise reduction, the system would fail to transmit much information. Furthermore, it is not even clear that Fisher information provides a good metric in such cases where only very coarse discrimination of signals may be performed (see Butts & Goldman, 2006). We therefore do not consider this case further.

In the second case, corresponding to a higher signal-to-noise regime, the maximal Fisher information is obtained when the noise variance is not saturated. In this case, as shown in section A.6, we find that the optimal value of the network feedback a is not different from that obtained without a bound on the noise variance (compare Figures 8E and 8F). Furthermore, we derive conditions on σ and N such that this higher signal-to-noise regime is attained. In a similar manner, the optimal memory performance of the feedforward networks in the high signal-to-noise regime is not affected by the bound on the variability. Therefore, for the feedforward networks discussed below, we consider only the effect of the bound on the mean activity.

3.3.3 Feedforward Networks with Finite Dynamic Range—In the sections 3.1 and 3.2, we found that the optimal feedforward networks used transient amplification of signals to increase the (signal gain)-to–(noise gain) ratios that are represented by $\tilde{I}_F$. However, as we noted for the attractor networks with a reset, unbounded signal amplification is no longer possible when there is a finite dynamic range, and firing rates will saturate unless the inputs entering the network are reduced. The consequences of this limited dynamic range for the feedforward networks are delineated below.

We consider first the case of networks with discrete dynamics (see Figure 6G) for which analytical calculation of the optimal Fisher information is tractable. Similar to the above results for the attractor networks with a reset, we find that the optimal memory performance is obtained under two conditions. First, the input vector $\vec{v}$ should be made as strong as possible for each neuron so that each neuron in the network uses the full extent of its mean dynamic range. This immediately implies that the optimal feedforward networks must have a functionally, rather than literally, feedforward architecture (because in a literally feedforward architecture, by definition the first stage does not contain all neurons).

Second, the networks should use a value $a = 1$ that corresponds to perfect maintenance of the signal as it propagates down the chain of modes (see Figure 9A). If network activity decays more quickly than this ($a<1$), part of the signal will be lost. If network activity grows more quickly ($a>1$), then in order to use its full dynamic range of mean activity and not saturate, the network will need to have initial activity that is less than maximal and will need to amplify this activity over time, leading to an amplification of noise as well. This is precisely analogous to the case for the attractor network with a reset, and indeed the discrete feedforward and attractor networks with a reset have identical $\tilde{I}_F$ (see Figure 10C). Without a reset, the feedforward networks with discrete dynamics can outperform the attractor networks because they do not need to forget in order to remove noise and, in fact, the performance of the feedforward networks is identical with or without a reset (see Figures 10A and 10C).

We note that the two conditions above imply that for the feedforward networks to maintain neuronal activity at a level that uses the full dynamic range of all neurons, the activities of the neurons at all times need to be directed along the vertices of the hypercube that defines the allowed range of mean firing rates (see Figures 7B and 9B). It is not immediately obvious that this condition can be met for the feedforward networks, because it implies that there must be N orthogonal modes of the network that each lie along a different vertex of the hypercube. In section A.6, we show that networks can be constructed that obey this criterion, at least in the case that the number of neurons N is equal to a power of 2. When N is restricted to this case, we show that the Fisher information conveyed by the network scales as N/T, similar to the case of the attractor networks with a reset. Building on this case, we show in section A.6 that for general N, the maximal Fisher information is still of the order of N/T when the number of neurons is at least twice the number of feedforward stages.

Literally feedforward networks perform more poorly than the optimal, functionally feedforward networks already described. In the literally feedforward networks, different sets of neurons are used to convey information in each stage. In particular, since input is applied to neurons only in the first stage and is carried only by a subset of neurons at any time, the feedforward networks cannot convey as much information as functionally feedforward networks that use all neurons at every stage. Interestingly, when we considered storage of a single-dimensional stimulus in a literally feedforward network with number of stages equal to the duration of the memory period (as in Ganguli, Huh et al., 2008), we found that the memory performance of the optimal networks scaled only as N/T^2 rather than as N/T (see section A.6). Furthermore, although the previous study focused on networks having a fan-out structure with more neurons at later stages, we found that the optimal network architecture in this case contained equal numbers of neurons at all times (see section 4 for further commentary). This uniform structure provides an optimal balance between the fan-out architecture, which allows larger signal amplification between stages, and the fan-in architecture, which reduces noise (and particularly the amount of noise that is common among neurons) by pooling stages with more neurons into stages with fewer neurons (see Figures 9C to 9E). Figure 9D shows how $\tilde{I}_F$ depends on the rate of fan-out, which is defined as the ratio of the number of neurons in the successive stages: when the fan-out rate is less than 1 (greater than 1), it is a fan-in (fan-out) structure. As seen in this figure, $\tilde{I}_F$ is maximized when the fan-out rate is 1, that is, for a uniform structure. We note that this result holds regardless of whether $||\vec{v}|| = 1$, as in Ganguli, Huh et al. (2008) (calculation not shown).

For feedforward networks with continuous dynamics, the Fisher information $\tilde{I}_F$ cannot be expressed in a simple analytical form, making it difficult to find the structure that optimizes memory performance. To obtain a lower bound on the maximal $\tilde{I}_F$ and gain an intuition for how the results obtained in discrete dynamics might change when the dynamics are continuous, we therefore calculated $\tilde{I}_F$ for networks with the structure found to be optimal under discrete dynamics. Numerical simulation in this case shows that the feedforward networks perform worse than the attractor networks, either with or without reset (see Figures 10B and 10D). Furthermore, with a reset, we show in section A.6 that the attractor network saturates the bound on information transmission achievable by any network with a finite dynamic range, whereas no feedforward network can achieve this bound. Thus, at least for

networks with a reset, the attractor networks strictly outperform the feedforward networks. Without a reset, the worse performance of the feedforward networks could in principle be due to the nonoptimal architecture taken from the optimal discrete network. However, we think this is unlikely because the reduced memory performance for the feedforward networks with continuous dynamics is analogous to the similar result found in section 3.2 (Figures 6H and 6I and accompanying text), which could be explained by the combination of spreading of signals across the modes of the network and signal loss through the end of the chain.

4 Discussion

We have compared the memory performance of two prominent classes of short-term memory networks in storing the amplitude of a briefly presented stimulus in the presence of gaussian white noise. In one class of networks, memory was sustained by positive feedback that was mediated by recurrent connections and resulted in the formation of low-dimensional attractors (Robinson, 1989; Seung, 1996). In the other class, memory was sustained by passing activity through either long feedforward chains of neurons or through a chain of orthogonal activity patterns (Schurmodes) in a recurrent network (Ganguli, Huh et al., 2008; Goldman, 2009; White et al., 2004). In each case, memory performance was quantified with the Fisher information, which, for the linear network dynamics considered here, represents a ratio of the amount that the network amplifies the signals versus the noise received.

Our primary results were as follows. For the attractor networks, including those with a limited range of firing rates, we found that the best-performing networks were forgetful if noise is allowed to build up without constraint before the stimulus arrives. This forgetfulness reflected a fundamental trade-off between requiring a long time constant of decay of network activity to maintain signals throughout the memory period and needing some decay of network activity in order to remove noise that enters the system before the stimulus arrives. However, if there exists a mechanism to remove noise from the system near the time of stimulus arrival or if networks enter the memory-storing state only close to the time of the stimulus onset, then we found that the optimal networks with a limited dynamic range perfectly maintain their signals throughout the memory period.

Comparison of the memory performance between line attractor and higher-dimensional attractor networks showed that the optimal memory performance with or without reset is independent of the dimension of the attracting modes. However, optimal memory performance in the line attractor networks requires an optimal alignment of the input vector, whereas optimal memory performance in the higher-dimensional attractors requires an optimal alignment of the readout vector. These results suggest that line attractor networks might be more useful if the stored memory needs to be used by multiple networks that each project activity along a different direction. By contrast, higher-dimensional attractors might be more useful in storing memories that arrive from multiple networks that each encodes the stimulus along a different direction.

For the feedforward networks, the optimal network architectures did not depend strongly on the presence of a resetting mechanism because the feedforward networks naturally remove

noise from the system when it exits from the end of the feedforward chain. Due to this inherent noise-removal mechanism, the feedforward networks could transiently amplify signals without excessive noise buildup (Ganguli, Huh et al., 2008) and, for linear networks with no reset or bounds on activity, perform better than the attractor networks. However, when the firing rates were bounded, the ability of the feedforward network to amplify inputs was limited, and the optimal feedforward networks propagated activity without amplification or decay ($a = 1$).

Comparing the networks, we found that the Fisher information for both the optimal attractor and feedforward networks increased linearly with the number of neurons N, reflecting that additional neurons allow more signal to be carried by the network. Additionally, the optimal networks in both cases exhibited a power law decay in memory performance. For the attractor networks and for feedforward networks in a discrete approximation, this decay was inversely proportional to time and reflected the linear increase in noise variance over time. Interestingly, we note that such a linear increase in variance has been observed experimentally in spatial working memory tasks (Ploner, Gaymard, Rivaud, Agid, & Pierrot-Deseilligny, 1998; White, Sparks, & Stanford, 1994). Feedforward networks with continuous dynamics performed less well than those with discrete dynamics, reflecting two factors: the signals in continuous networks spread out over time, leading to a reduction in the signal gain; and due to this spreading, signals exit from the end of the chain before the end of the memory period. Together these factors lead to worse performance of the feedforward networks relative to the attractor networks when there is a noise reset, and quite likely (although we could only compute a lower bound approximation on the feedforward networks) even in the absence of such a reset.

4.1 Comparison to Previous Work

Many previous studies have proposed perfectly tuned attractor networks as a substrate for holding short-term memories in the absence of noise (for reviews, see Brody et al., 2003; Goldman, Compte, & Wang, 2009; Wang, 2001). Here, we have explicitly considered the effects of noise on both attractor and nonattractor (feedforward or functionally feedforward) networks. For networks with underlying linear dynamics and both a reset nonlinearity and a finite dynamic range on neuronal responses, our results are consistent with the optimality of perfectly tuned attractor networks. Similarly, we note that the perfectly tuned attractor network (integrator) was found in a recent study to be the optimal architecture for storing the running total of a continuously presented input in which noise likewise started with the arrival of the signal (Brown et al., 2005).

Perfect integrator networks face a fine-tuning problem of network connectivity in that the feedback connections must precisely offset intrinsic neuronal decay processes in order to sustain activity at a constant rate in the absence of external input. Several mechanisms have been suggested to lessen the strictness of this tuning requirement. These include the use of long intrinsic (Marder, Abbott, Turrigiano, Liu, & Golowasch, 1996) or synaptic (Hempel, Hartman, Wang, Turrigiano, & Nelson, 2000; Wang et al., 2006; Mongillo, Barak, & Tsodyks, 2008) time constants. In addition, bistability is a nonlinear mechanism for maintaining the robustness of memory storage (Camperi & Wang, 1998; Koulakov,

Raghavachari, Kepecs, & Lisman, 2002; Goldman, Levine, Major, Tank, & Seung, 2003) and homeostatic learning rules have been suggested to be able to keep short-term memory-storing circuits tuned (Goldman, 2009; Renart, Song, & Wang, 2003). Further investigation is needed to analyze the robustness of different network architectures to synaptic weight changes.

In the absence of a reset nonlinearity, we find that noise buildup before the time of the stimulus presentation makes the perfect attractor network nonoptimal; instead, at least in the high signal-to-noise regime, we find that the optimal attractor networks must be forgetful in order to reduce noise accumulation. This result is similar to that of White et al. (2004), who considered the storage of temporal sequences in memory networks with discrete dynamics and who noted that forgetting was necessary in order to prevent the buildup of noise that arrived at all times before the stimulus onset.

For the feedforward networks, previous work that examined networks with a finite dynamic range of neural activity focused on network architectures with a fan-out structure (Ganguli, Huh et al., 2008; Ganguli & Latham, 2009). This previous work showed that under a finite dynamic range constraint, a fan-out network can achieve the same scaling as the optimal network; however, this study did not check whether other structures may achieve this bound or whether there exists a structure having better memory performance. By contrast, at least for storage of a one-dimensional stimulus, we explicitly calculated that the optimal network architectures for the (discrete) feedforward networks had a uniform structure and that the fan-out structure was suboptimal. A key difference between our study and that of Ganguli, Huh et al. (2008) is that they primarily focused their discussion on memory for sequences, whereas here we explicitly focus on memory for a single-dimensional input. Higher-dimensional signals cannot be stored in attractor networks if the dimension of the attractor is lower than the dimension of the signal. Therefore, if the stimulus to be remembered is higher dimensional, such as remembering an entire sequence of inputs, this may favor a high-dimensional or feedforward network in which time is explicitly represented by patterns of activity that are sequentially activated as signals propagate through the network (Ganguli, Huh et al., 2008; Goldman, 2009; White et al., 2004). Ganguli, Huh et al. (2008) showed that the duration T for which a network could reliably convey information about a temporal sequence increased only in proportion to $\sqrt{N}$. This contrasts with our result for storing a single-dimensional stimulus, in which memory increases proportional to the network size N. The reason for this difference is that for storage of a single-dimensional stimulus, our optimal networks (both attractor and functionally feedforward) could use their entire finite dynamic range to store this one dimension. By contrast, when the stimulus dimension scales with time, as in sequence memory, the network must divide its dynamic range among all stored dimensions. This leads to memory performance, which scales as N/T_2, rather than N/T, so that the duration T for which a network reliably can convey information about a temporal sequence increases only in proportion to $\sqrt{N}$. Consistent with this observation, the memory performance of our optimal literally feedforward networks (which use approximately 1/T of the entire network's range at any given time) scaled only as N/T_2. Furthermore, literally feedforward networks might have an advantage over functionally feedforward networks or generic high-dimensional attractors because the literally

feedforward networks keep the elements of a sequence arriving at different times cleanly segregated.

4.2 Temporal Information in Memory Networks

In this work, we focused on mechanisms for storing the amplitude of a stimulus when the memory period is a fixed (or known) duration, so that there is no need for encoding the time since the pulse occurred. However, if the duration of the memory period is variable and unknown, then joint information about the amplitude of an input pulse and the time at which the pulse occurs needs to be encoded. A one-dimensional attractor network is not suitable to extract joint information about the amplitude and time of input since a one-dimensional network cannot represent such a two-dimensional quantity. Rather, at least a two-dimensional network is required. Feedforward networks seem advantageous for processing time and storing signals since different sets of neurons or modes are used at different times. However, it is unclear whether time and amplitude are dependently encoded by the network activity, as in feedforward networks, versus encoded in independent modes of activity (either with time encoded in a completely separate network from amplitude, or with independent modes of activity that represent time and that represent amplitude, as suggested by the recent work of Machens, Romo, and Brody, 2010). For example, it has been suggested in the circuits underlying bird song that time is represented through a feedforward chain of bistable units that are more robust to temporal encoding than graded networks (but with loss of any representation of amplitude information; see Long, Jin, & Fee, 2010). Alternatively, high-dimensional attractor networks have been suggested to encode both time and amplitude (Machens et al., 2010; Singh & Eliasmith, 2006). Further work, both experimental and theoretical, is needed to address the joint processing of amplitude and temporal information.

4.3 Effect of Correlated or Signal-Dependent Noise

In this study, we assumed for simplicity that the external noise received by each neuron was equal in amplitude and uncorrelated across neurons. However, similar to studies in sensory systems that have shown strong effects on neural coding in the presence of correlated noise (Abbott & Dayan, 1999; Averbeck, Latham, & Pouget, 2006; Latham, Deneve, & Pouget, 2003; Sompolinsky et al., 2001; Zohary, Shadlen, & Newsome, 1994), we found that the correlation structure of noise received by the network may dramatically affect the optimal architecture of memory networks. As illustrated in Figure 11A, line attractor networks may be advantageous when noise is correlated: if the input direction can be chosen independent of the profile of the injected noise, then the attractor and the input direction can be oriented orthogonal to the directions of high noise and along directions with low noise. In contrast, activity in feedforward networks is passed through many different orthogonal patterns of activity (see Figure 11B), so that it may be difficult to take advantage of correlated noise that has a particularly non-noisy direction. If instead the input direction and the profile of injected noise are dependent, a more careful examination is required to determine the architectures of the best-performing attractor and feedforward networks.

We considered only additive noise in this study. When noise is instead multiplicative or signal dependent, different optimal architectures may be necessary. Although this question deserves much further study, multiplicative noise is more disruptive to higher firing rates

than additive noise and therefore might lead to better memory performance for networks with relatively faster decay of signals, with smaller amplitudes of input that drive neurons to less high firing rates, or with other differences in network architecture that decrease the use of high firing rates.

4.4 Nonlinear Dynamics

In this letter, we have modeled the finite range of neuronal activities in an analytically tractable manner by imposing a finite dynamic range on the mean firing rates and their variances and arranging the network inputs so that the trajectories of neuronal firing never exceed this range. More realistically, neurons have hard or soft limits on their observed firing rates that are best modeled with explicitly nonlinear network models. Having the underlying dynamics of the network be nonlinear rather than imposing the finite dynamic range as a simple constraint on a linear network may influence the memory performance in various ways. For example, explicit inclusion of nonlinear dynamics may reduce the buildup of noise relative to a network with linear dynamics, and even without an explicit reset mechanism, the optimal architecture of the attractor networks may become less forgetful. Furthermore, recent theoretical work shows that randomly connected nonlinear networks with sigmoidal neuronal input-output functions exhibit a sharp reduction of neural variability with the arrival of a stimulus (Rajan et al., 2010), suggesting a mechanistic explanation for the reset mechanism considered in our study. More dramatic, the presence of strong nonlinearity can lead to bistable responses, which may be useful in robustly maintaining memories in the presence of noise (Toyoizumi, 2010) or lessening the need to fine-tune synaptic connection weights (Camperi & Wang, 1998; Goldman et al., 2003; Koulakov et al., 2002). Further work is needed to explore the possibilities offered by nonlinear networks and to develop analysis methodologies that allow a rigorous understanding of networks in which the conveniences offered by linear analysis no longer apply.

Acknowledgments

This research was supported by a Sloan Foundation Research Fellowship, NIH grant R01 MH069726, and a UC Davis Ophthalmology Research to Prevent Blindness grant. We thank T. Toyoizumi, A. T. Sornborger, and E. Aksay for valuable discussions and D. Fisher and J. Ditterich for valuable discussions and feedback on the manuscript. This research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

References

Abbott LF, Dayan P. The effect of correlated variability on the accuracy of a population code. Neural Comput. 1999; 11(1):91–101. [PubMed: 9950724]

Aksay E, Baker R, Seung HS, Tank DW. Anatomy and discharge properties of pre-motor neurons in the goldfish medulla that have eye-position signals during fixations. J Neurophysiol. 2000; 84(2): 1035–1049. [PubMed: 10938326]

Amit DJ, Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb Cortex. 1997; 7(3):237–252. [PubMed: 9143444]

Averbeck BB, Latham PE, Pouget A. Neural correlations, population coding and computation. Nat Rev Neurosci. 2006; 7(5):358–366. [PubMed: 16760916]

Baddeley R, Abbott LF, Booth MC, Sengpiel F, Freeman T, Wakeman EA. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proc Biol Sci. 1997; 264(1389): 1775–1783. [PubMed: 9447735]

Brody CD, Romo R, Kepecs A. Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations. Curr Opin Neurobiol. 2003; 13(2):204–211. [PubMed: 12744975]

Brown E, Gao J, Holmes P, Bogacz R, Gilzenrat M, Cohen JD. Simple neural networks that optimize decisions. International Journal of Bifurcation and Chaos. 2005; 15(3):803–826.

Butts DA, Goldman MS. Tuning curves, neuronal variability, and sensory coding. PLoS Biol. 2006; 4(4):e92. [PubMed: 16529529]

Camperi M, Wang XJ. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. J Comput Neurosci. 1998; 5(4):383–405. [PubMed: 9877021]

Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, Corrado GS. Stimulus onset quenches neural variability: A widespread cortical phenomenon. Nat Neurosci. 2010; 13(3):369–378. [PubMed: 20173745]

Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. Nat Neurosci. 2000; 3:1184–1191. [PubMed: 11127836]

Ganguli S, Bisley JW, Roitman JD, Shadlen MN, Goldberg ME, Miller KD. One-dimensional dynamics of attention and decision making in LIP. Neuron. 2008; 58(1):15–25. [PubMed: 18400159]

Ganguli S, Huh D, Sompolinsky H. Memory traces in dynamical systems. Proc Natl Acad Sci USA. 2008; 105(48):18970–18975. [PubMed: 19020074]

Ganguli S, Latham P. Feedforward to the past: The relation between neuronal connectivity, amplification, and short-term memory. Neuron. 2009; 61(4):499–501. [PubMed: 19249270]

Goldman MS. Memory without feedback in a neural network. Neuron. 2009; 61(4):621–634. [PubMed: 19249281]

Goldman, MS., Compte, A., Wang, XJ. Neural integrator models. In: Squire, LR., editor. Encyclopedia of Neuroscience. Vol. 6. Oxford: Academic Press; 2009. p. 165-178.

Goldman MS, Levine JH, Major G, Tank DW, Seung HS. Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. Cereb Cortex. 2003; 13(11):1185–1195. [PubMed: 14576210]

Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995; 14(3):477–485. [PubMed: 7695894]

Hempel CM, Hartman KH, Wang XJ, Turrigiano GG, Nelson SB. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. J Neurophysiol. 2000; 83(5):3031–3041. [PubMed: 10805698]

Horn, RA., Johnson, CR. Matrix analysis. Cambridge: Cambridge University Press; 1985.

Koulakov AA, Raghavachari S, Kepecs A, Lisman JE. Model for a robust neural integrator. Nat Neurosci. 2002; 5(8):775–782. [PubMed: 12134153]

Latham PE, Deneve S, Pouget A. Optimal computation with attractor networks. J Physiol Paris. 2003; 97(4–6):683–694. [PubMed: 15242674]

Long MA, Jin DZ, Fee MS. Support for a synaptic chain model of neuronal sequence generation. Nature. 2010; 468(7322):394–399. [PubMed: 20972420]

MacDonald C, Lepage K, Eden U, Eichenbaum H. Hippocampal “time cells” bridge the gap in memory for discontiguous events. Neuron. 2011; 71(4):737–749. [PubMed: 21867888]

Machens CK, Romo R, Brody CD. Functional, but not anatomical, separation of ”what” and ”when” in prefrontal cortex. J Neurosci. 2010; 30(1):350–360. [PubMed: 20053916]

Major G, Tank D. Persistent neural activity: Prevalence and mechanisms. Curr Opin Neurobiol. 2004; 14(6):675–684. [PubMed: 15582368]

Marder E, Abbott LF, Turrigiano GG, Liu Z, Golowasch J. Memory from the dynamics of intrinsic membrane constants. Proc Natl Acad Sci USA. 1996; 93:13481–13486. [PubMed: 8942960]

Mauk MD, Buonomano DV. The neural basis of temporal processing. Annu Rev Neurosci. 2004; 27:307–340. [PubMed: 15217335]

Mongillo G, Barak O, Tsodyks M. Synaptic theory of working memory. Science. 2008; 319(5869): 1543–1546. [PubMed: 18339943]

Murphy BK, Miller KD. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. Neuron. 2009; 61(4):635–648. [PubMed: 19249282]

Pastalkova E, Itskov V, Amarasingham A, Buzsaki G. Internally generated cell assembly sequences in the rat hippocampus. Science. 2008; 321(5894):1322–1327. [PubMed: 18772431]

Ploner CJ, Gaymard B, Rivaud S, Agid Y, Pierrot-Deseilligny C. Temporal limits of spatial working memory in humans. Eur J Neurosci. 1998; 10(2):794–797. [PubMed: 9749746]

Rabinovich M, Huerta R, Laurent G. Transient dynamics for neural processing. Science. 2008; 321(5885):48–50. [PubMed: 18599763]

Rajan K, Abbott LF, Sompolinsky H. Stimulus-dependent suppression of chaos in recurrent neural networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2010; 82(1 Pt 1):011903. [PubMed: 20866644]

Renart A, Song P, Wang XJ. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron. 2003; 38(3):473–485. [PubMed: 12741993]

Robinson DA. Integrating with neurons. Annu Rev Neurosci. 1989; 12:33–45. [PubMed: 2648952]

Romo R, Brody CD, Hernandez A, Lemus L. Neuronal correlates of parametric working memory in the prefrontal cortex. Nature. 1999; 399(6735):470–473. [PubMed: 10365959]

Salinas E, Abbott LF. Vector reconstruction from firing rates. J Comput Neurosci. 1994; 1(1–2):89–107. [PubMed: 8792227]

Savin C, Triesch J. Developing a working memory with reward-modulated STDP. CoSyNe Abs. 2009:T29.

Seung HS. How the brain keeps the eyes still. Proc Natl Acad Sci USA. 1996; 93(23):13339–13344. [PubMed: 8917592]

Shadlen MN, Britten KH, Newsome WT, Movshon JA. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J Neurosci. 1996; 16(4):1486–1510. [PubMed: 8778300]

Sharp PE, Blair HT, Cho J. The anatomical and computational basis of the rat head-direction cell signal. Trends Neurosci. 2001; 24(5):289–294. [PubMed: 11311382]

Singh R, Eliasmith C. Higher-dimensional neurons explain the tuning and dynamics of working memory cells. J Neurosci. 2006; 26(14):3667–3678. [PubMed: 16597721]

Sompolinsky H, Yoon H, Kang K, Shamir M. Population coding in neuronal systems with correlated noise. Phys Rev E Stat Nonlin Soft Matter Phys. 2001; 64(5 Pt 1):051904. [PubMed: 11735965]

Taube JS, Bassett JP. Persistent neural activity in head direction cells. Cereb Cortex. 2003; 13(11): 1162–1172. [PubMed: 14576208]

Toyoizumi T. An extensive memory lifetime is achieved by coupled nonlinear neurons. Soc Neurosci Abs. 2010:197.14.

Trefethen, LN., Embree, M. Spectra and pseudospectra: The behavior of non-normal matrices and operators. Princeton, NJ: Princeton University Press; 2005.

Wang XJ. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 2001; 24(8):455–463. [PubMed: 11476885]

Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat Neurosci. 2006; 9(4):534–542. [PubMed: 16547512]

Weber RB, Daroff RB. Corrective movements following refixation saccades: Type and control system analysis. Vision Res. 1972; 12(3):467–475. [PubMed: 5021911]

White JM, Sparks DL, Stanford TR. Saccades to remembered target locations: An analysis of systematic and variable errors. Vision Res. 1994; 34(1):79–92. [PubMed: 8116271]

White OL, Lee DD, Sompolinsky H. Short-term memory in orthogonal neural networks. Phys Rev Lett. 2004; 92(14):148102. [PubMed: 15089576]

Zohary E, Shadlen MN, Newsome WT. Correlated neuronal discharge rate and its implications for psychophysical performance. Nature. 1994; 370(6485):140–143. [PubMed: 8022482]

Appendix

A.1 Relation Between Fisher Information and (Signal Gain)-to-Noise Ratio in Linear Systems

Here, we prove that the Fisher information I_F is greater than or equal to the (signal gain)-to-noise ratio for a network with linear dynamics and linear readout of the network activity.

Denoting the signal gain vector $\frac{1}{\tau}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau]\overrightarrow{v}$ as $\frac{1}{\tau}\overrightarrow{g}$ and noise covariance matrix as $\frac{\sigma^2}{\tau^2}\overleftrightarrow{C}$, I_F in equation 2.9 can be expressed as

$$I_F(t)=\frac{1}{\sigma^2}\overrightarrow{g}^T\overleftrightarrow{C}^{-1}\overrightarrow{g}. \quad \text{(A.1)}$$

If the network activity is read out linearly by projecting along a direction $\vec{k}$, then the mean activity and the noise variance along $\vec{k}$ become

$$\text{mean}\,(\overrightarrow{k}^T\overrightarrow{r})=\frac{s}{\tau}\overrightarrow{k}^T\overrightarrow{g}, \quad \text{(A.2)}$$

$$\text{var}(\overrightarrow{k}^T\overrightarrow{r})=\frac{\sigma^2}{\tau^2}\overrightarrow{k}^T\overleftrightarrow{C}\overrightarrow{k}. \quad \text{(A.3)}$$

The quantity of activity in the readout that is analogous to I_F is the (signal gain)-to-noise ratio (SNR) defined as the square of the signal gain divided by the noise variance in the direction $\vec{k}$:

$$SNR(\overrightarrow{k}^T\overrightarrow{r})=\frac{1}{\sigma^2}\frac{(\overrightarrow{k}^T\overrightarrow{g})^2}{\overrightarrow{k}^T\overleftrightarrow{C}\overrightarrow{k}}. \quad \text{(A.4)}$$

The relation $I_F \geq$ SNR can be proven as follows. If $C\leftrightarrow$ is a diagonal matrix such that $C\leftrightarrow = \text{diag}(c_1, c_2, \ldots, c_n)$ with $c_i > 0$, then from equations A.1 and A.4,

$$\begin{aligned} I_F(t) &= \tfrac{1}{\sigma^2}\overrightarrow{g}^T\overleftrightarrow{C}^{-1}\overrightarrow{g}=\tfrac{1}{\sigma^2}\sum_i\tfrac{g_i^2}{c_i}, \\ SNR(t) &= \tfrac{1}{\sigma^2}\tfrac{(\overrightarrow{k}^T\overrightarrow{g})^2}{\overrightarrow{k}^T\overleftrightarrow{C}\overrightarrow{k}}=\tfrac{1}{\sigma^2}\tfrac{(\sum_i k_i g_i)^2}{\sum_i k_i^2 c_i}, \\ \tfrac{1}{\sigma^2}\sum_i\tfrac{g_i^2}{c_i} &\geq \tfrac{1}{\sigma^2}\tfrac{(\sum_i k_i g_i)^2}{\sum_i k_i^2 c_i} \iff \left(\sum_i\frac{g_i^2}{c_i}\right)\left(\sum_i k_i^2 c_i\right)\geq\left(\sum_i k_i g_i\right)^2. \end{aligned}$$

The above relation is the Cauchy-Schwarz inequality, and the equality holds when $k_i \propto g_i/c_i$. For a nondiagonal matrix $C\leftrightarrow$, we can get the same result by changing to a coordinate system in which $C\leftrightarrow$ is transformed to a diagonal matrix and the condition for the equality becomes that $\vec{k}$ is along $C\leftrightarrow^{-1}\vec{g}$.

An alternative proof of the relation between I_F and the SNR (not shown) can be obtained using the Cramer-Rao bound relationship between I_F and the maximum likelihood estimator of the stimulus.

A.2 Analytic Expression for the Fisher Information in Attractor Networks

A.2.1 Calculation of Fisher Information for Line Attractors

For the line attractor models, we assume the connectivity matrix is eigenvalue-decomposable such that

$$\overleftrightarrow{W} = \overleftrightarrow{Q} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \overleftrightarrow{Q}^{-1},$$

where the columns of $Q\leftrightarrow$ are the right eigenvectors of $W\leftrightarrow$, denoted by $\overrightarrow{q}_i^r$ such that $\overleftrightarrow{W}\overrightarrow{q}_i^r = \lambda_i \overrightarrow{q}_i^r$. The rows of $Q\leftrightarrow^{-1}$ are the left eigenvectors of $W\leftrightarrow$, denoted by $\overrightarrow{q}_i^l$ such that $(\overrightarrow{q}_j^l)^T \overleftrightarrow{W} = (\overrightarrow{q}_j^l)^T \lambda_j$. We assume that the left eigenvectors have unit length. If $W\leftrightarrow$ is not normal, meaning that the left (or right) eigenvectors corresponding to different eigenvalues are not necessarily orthogonal to one another, it is also the case that the right and left eigenvectors corresponding to the same eigenvalue are not necessarily parallel to each other. In this nonnormal case, the length of the right eigenvector is equal to the inverse of the cosine of the angle between the corresponding right and left eigenvectors.

The eigenvalue decomposition of the matrix $(-I\leftrightarrow + W\leftrightarrow)/\tau$ is given as

$$\begin{aligned}(-\overleftrightarrow{I} + \overleftrightarrow{W})/\tau &= \overleftrightarrow{Q} \begin{bmatrix} -1/\tau_{eff}^{(1)} & 0 & \cdots & 0 \\ 0 & -1/\tau_{eff}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1/\tau_{eff}^{(n)} \end{bmatrix} \overleftrightarrow{Q}^{-1} \\ &= \overleftrightarrow{Q}\,\overleftrightarrow{D}\,\overleftrightarrow{Q}^{-1} \quad \text{with } \tau_{eff}^{(i)} = \frac{\tau}{1-\lambda_i},\end{aligned}$$

so that

$$\frac{1}{\tau}e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau}\overrightarrow{v}=\frac{1}{\tau}\overleftrightarrow{Q}\,e^{\overleftrightarrow{D}t}\overleftrightarrow{Q}^{-1}\overrightarrow{v}=\sum_i\frac{\mathrm{e}^{-t/\tau_{eff}^{(i)}}(\overrightarrow{q}_i^{\,l}\cdot\overrightarrow{v})}{\tau}\overrightarrow{q}_i^{\,r}. \quad \text{(A.5)}$$

If the λ_i's are arranged in descending order by their real parts and the real part of the first eigenvalue is much larger than the real parts of the remaining eigenvalues as in the line attractors, then for $i > 1$, $\mathrm{e}^{-t/\tau_{eff}^{(1)}}$ decays much more slowly than $\mathrm{e}^{-t/\tau_{eff}^{(i)}}$ so that $\tau_{eff}^{(1)} \gg \tau_{eff}^{(i)}$. Furthermore, if the decay in the other modes relative to the activity along the line attractor is fast enough to overcome the nonorthogonality of the eigenvectors implicit in the length of $\overrightarrow{q}_i^{\,r}$, then the network activity in equation A.5 can be expressed approximately in terms of λ_1 and the corresponding left and right eigenvectors as

$$\frac{1}{\tau}e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau}\overrightarrow{v}\simeq\frac{\mathrm{e}^{-t/\tau_{eff}^{(1)}}(\overrightarrow{q}_1^{\,l}\cdot\overrightarrow{v})}{\tau}\overrightarrow{q}_1^{\,r}. \quad \text{(A.6)}$$

In a similar manner, noise also accumulates primarily along $\vec{q}_1^{\,r}$, and the Fisher information I_F is approximately the (signal gain)-to-noise ratio along $\vec{q}_1^{\,r}$. The detailed calculation is

$$\begin{aligned}I_F&=\overrightarrow{v}^T e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})^T t/\tau}\tfrac{1}{\tau}\left(\tfrac{\sigma^2}{\tau^2}\textstyle\int_{t_0}^t e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau}e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})^T(t-t')/\tau}dt'\right)^{-1}\times\tfrac{1}{\tau}e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})^{t/\tau}}\overrightarrow{v}\\&=\tfrac{1}{\sigma^2}\overrightarrow{v}^T(\overleftrightarrow{Q}^{-1})^T e^{\overleftrightarrow{D}t}\overleftrightarrow{Q}^T\left(\overleftrightarrow{Q}\textstyle\int_{t_0}^t e^{\overleftrightarrow{D}(t-t')}\overleftrightarrow{Q}^{-1}(\overleftrightarrow{Q}^{-1})^T e^{\overleftrightarrow{D}(t-t')}dt'\overleftrightarrow{Q}^T\right)^{-1}\times\overleftrightarrow{Q}\,e^{\overleftrightarrow{D}t}\overleftrightarrow{Q}^{-1}\overrightarrow{v}\\&=\tfrac{1}{\sigma^2}(\overleftrightarrow{Q}^{-1}\overrightarrow{v})^T e^{\overleftrightarrow{D}t}\left(\textstyle\int_{t_0}^t e^{\overleftrightarrow{D}(t-t')}\overleftrightarrow{Q}^{-1}(\overleftrightarrow{Q}^{-1})^T e^{\overleftrightarrow{D}(t-t')}dt'\right)^{-1}e^{\overleftrightarrow{D}t}(\overleftrightarrow{Q}^{-1}\overrightarrow{v}).\end{aligned}$$

In the last expression for I_F, the signal gain vector and the noise covariance are expressed in the coordinates of the right eigenvectors and computed as follows:

$$\begin{aligned}&e^{\overleftrightarrow{D}t}(\overleftrightarrow{Q}^{-1}\overrightarrow{v})\simeq\mathrm{e}^{-t/\tau_{eff}^{(1)}}(\overrightarrow{q}_1^{\,l}\cdot\overrightarrow{v})[1,0,\ldots,0]^T\\&\textstyle\int_{t_0}^t e^{\overleftrightarrow{D}(t-t')}\overleftrightarrow{Q}^{-1}(\overleftrightarrow{Q}^{-1})^T e^{\overleftrightarrow{D}(t-t')}dt'\\&=\textstyle\int_{t_0}^t\mathrm{diag}([e^{-(t-t')/\tau_{eff}^{(i)}}]_{i=1,\ldots,n})[\overrightarrow{q}_i^{\,l}\cdot\overrightarrow{q}_j^{\,l}]_{i,j=1,\ldots,n}\mathrm{diag}([e^{-(t-t')/\tau_{eff}^{(j)}}]_{j=1,\ldots,n})\,dt'.\end{aligned} \quad \text{(A.7)}$$

In equation A.7, $[\vec{q}_i^{\,l}\cdot\vec{q}_j^{\,l}]_{i,j=1,\ldots,n}$ is the matrix whose elements are the inner products of the left eigenvectors, and the (i, j)th element of the matrix $\int_{t_0}^t e^{\overleftrightarrow{D}(t-t')}\overleftrightarrow{Q}^{-1}(\overleftrightarrow{Q}^{-1})^T e^{\overleftrightarrow{D}(t-t')}dt'$ is computed as $(1/\tau_{eff}^{(i)}+1/\tau_{eff}^{(j)})^{-1}(1-e^{-(t-t_0)(1/\tau_{eff}^{(i)}+1/\tau_{eff}^{(j)})})\overrightarrow{q}_i^{\,l}\cdot\overrightarrow{q}_j^{\,l}$. If $Q\leftrightarrow$ is an orthogonal matrix, then all the off-diagonal terms become zero since $\vec{q}_i^{\,l}\cdot\vec{q}_j^{\,l}=0$ for $i\neq j$ and the (1, 1)th element of the inverse matrix is the reciprocal of $\frac{\tau_{eff}^{(1)}}{2}(1-e^{-2(t-t_0)/\tau_{eff}^{(1)}})$ where we

have used that $\|\vec{q}_1^{\,l}\| = 1$. If they are not orthogonal but $\tau_{eff}^{(1)}$ is large enough to overcome the nonorthogonal factor $\vec{q}_i^{\,l} \cdot \vec{q}_j^{\,l}$ in equation A.7, then only the (1, 1)th element is of order $\tau_{eff}^{(1)}$, whereas all the other elements are of order $\tau_{eff}^{(i)}$ for some $i > 1$. Thus, the (1,1)th element of the inverse matrix is still close to $1/\{\frac{\tau_{eff}^{(1)}}{2}(1-e^{-2(t-t_0)/\tau_{eff}^{1}})\}$.

Since the signal is predominantly along $[1, 0, \ldots, 0]^T$ in the coordinates of the right eigenvectors, the Fisher information I_F becomes the product of the square of the signal and the (1,1)th element of the inverse of the noise covariance matrix:

$$I_F \simeq \frac{\left(\vec{q}_1^{\,l} \cdot \vec{v}\right)^2 \exp\left(-2t/\tau_{eff}^{(1)}\right)}{\frac{\sigma^2 \tau_{eff}^{(1)}}{2}\left(1-\exp\left(-2(t-t_0)/\tau_{eff}^{(1)}\right)\right)}. \quad \text{(A.8)}$$

This approximation breaks down if the nonorthogonal factors $\vec{q}_i^{\,l} \cdot \vec{q}_j^{\,l}$ become large, for instance, in feedforward networks that are not eigen-decomposable.

Above, we showed general conditions under which the Fisher information of the line attractor networks is approximated by the signal-to-noise ratio along the attractor. For ease of computation, in the calculations below for the attractor networks, we consider only the case of (normal) networks in which all modes of the attractor networks are orthogonal, $Q\leftrightarrow Q\leftrightarrow^T = Q\leftrightarrow^T Q\leftrightarrow = I\leftrightarrow$. In this case, the left and right eigenvectors corresponding to a given eigenvalue are identical, and we denote the ith eigenvector by $\vec{q}_i$.

A.2.2 Optimal Decay Time Constant for Line Attractors with or Without Reset

Here we show the expression for the Fisher information I_F in line attractors with or without reset and obtain the optimal I_F in each case. First, we consider the case of an attractor network with no reset. In this case, noise builds up at all times before the signal arrives, so that $t_0 = -\infty$, and I_F is obtained from equation A.8 as

$$I_F(t) \simeq \frac{(\vec{q}_1 \cdot \vec{v})^2 \exp\left(-2t/\tau_{eff}^{(1)}\right)}{\sigma^2 \tau_{eff}^{(1)}/2}. \quad \text{(A.9)}$$

Differentiating equation A.9 with respect to $\tau_{eff}^{(1)}$, I_F attains a maximum value at time T equal to $(\vec{q}_1 \cdot \vec{v})^2/(e\sigma^2 T)$ when $\tau_{eff}^{(1)} = 2T$.

Alternatively, if we assume the existence of a reset mechanism and set the start of noise accumulation t_0 as 0, then equation A.8 becomes

$$I_F(t) \simeq \frac{(\vec{q}_1 \cdot \vec{v})^2 \exp\left(-2t/\tau_{eff}^{(1)}\right)}{\frac{\sigma^2 \tau_{eff}^{(1)}}{2}\left(1-\exp\left(-2t/\tau_{eff}^{(1)}\right)\right)} = \frac{2(\vec{q}_1 \cdot \vec{v})^2(1-\lambda_1)\exp(-2(1-\lambda_1)t/\tau)}{\sigma^2\tau(1-\exp(-2(1-\lambda_1)t/\tau))}. \quad \text{(A.10)}$$

Instead of taking a differential with respect to $\tau_{eff}^{(1)}$, which is undefined at $\lambda_1 = 1$, taking the differential with respect to λ_1 shows that I_F monotonically increases with λ_1. That is, it is an increasing function for both the signal-decaying regime, $\lambda_1 < 1$, and the nondecaying regime, $\lambda_1 \geq 1$.

A.2.3 Fisher Information for Plane Attractors

In plane attractor networks, there are two prominent modes with strong recurrent feedback, $\vec{q}_1$ and $\vec{q}_2$, that together define a plane. We assume for simplicity that the recurrent feedback amounts and thus the effective time constants are the same in $\vec{q}_1$ and $\vec{q}_2$ and denote them as λ and τ_{eff}. Since any neural activity in the modes other than the plane attractor decays to zero after a transient time, the remaining neural activity lies along the projection of the input vector onto the plane, $\vec{v}_{proj}$. Then for any mode $\vec{q}_x$ on the plane, the projection of neural activity onto this mode, denoted as x, evolves as

$$\tau \frac{dx}{dt} = -x + \lambda x + s(\vec{v}_{proj} \cdot \vec{q}_x)\delta(t) + \sigma\xi(t),$$
$$x(t) = \frac{s(\vec{v}_{proj} \cdot \vec{q}_x)}{\tau}\exp(-t/\tau_{eff}) + \frac{\sigma}{\tau}\int_{t_0}^{t}\exp(-(t-t')/\tau_{eff})\xi(t')dt', \quad \text{with } \tau_{eff} = \frac{\tau}{1-\lambda}.$$

I_F gives the maximal (signal gain)-to-noise ratio among these modes, which occurs when $\vec{q}_x$ is along $\vec{v}_{proj}$. Then the form of I_F in the plane attractor networks becomes

$$I_F \simeq \frac{\|\vec{v}_{proj}\|^2 \exp(-2t/\tau_{eff})}{\frac{\sigma^2\tau_{eff}}{2}(1-\exp(-2(t-t_0)/\tau_{eff}))}, \quad \text{(A.11)}$$

which is similar to I_F for the line attractor (with $\|\vec{v}_{proj}\|^2$ in place of $(\vec{q}_1{}^T \cdot \vec{v})^2$ in equation A.8).

A.2.4 SNR for Line and Plane Attractor Networks with a Linear Readout

Here we calculate the memory performance of line and plane attractor networks if neural activity is linearly read out by projecting along a direction $\vec{k}$.

First, we consider the line attractor networks. If the ratio of the signal in the nonattractor modes to that in the attractor mode decays to zero as in equation A.6, then the signal will be predominantly along $\vec{q}_1$. Then its projection on $\vec{k}$ is closely approximated as

$\frac{1}{\tau}(\vec{q}_1 \cdot \vec{k})(\vec{q}_1 \cdot \vec{v})\exp(-t/\tau_{eff}^{(1)})$. Also, noise accumulates only along $\vec{q}_1$ so that the signal-to-noise ratio along $\vec{k}$ becomes

$$SNR_1(t) \simeq \frac{(\vec{q}_1 \cdot \vec{k})^2 (\vec{q}_1 \cdot \vec{v})^2 \exp\left(-2t/\tau_{eff}^{(1)}\right)}{\sigma^2 (\vec{q}_1 \cdot \vec{k})^2 \int_{t_0}^{t} \exp\left(-2(t-t')/\tau_{eff}^{(1)}\right) dt'} = \frac{(\vec{q}_1 \cdot \vec{v})^2 \exp\left(-2t/\tau_{eff}^{(1)}\right)}{\sigma^2 \int_{t_0}^{t} \exp\left(-2(t-t')/\tau_{eff}^{(1)}\right) dt'}. \quad \text{(A.12)}$$

Second, we consider the plane attractor networks. In this case, the signal is stored in the direction of $\vec{v}_{proj}$, and the amplitude of the projection along the readout becomes $\vec{k} \cdot \vec{v}_{proj}$ ($=\vec{k}_{proj} \cdot \vec{v}_{proj}$) where $\vec{k}_{proj}$ is the projection of $\vec{k}$ onto the plane. Noise accumulates along all directions in the plane, and the projection of noise along $\vec{k}$ is proportional to $\|\vec{k}_{proj}\|$. Thus, the (signal gain)-to-noise ratio for the plane attractor network is

$$SNR_2(t) \simeq \frac{\|\vec{k}_{proj} \cdot \vec{v}_{proj}\|^2 \exp(-2t/\tau_{eff})}{\sigma^2 \|\vec{k}_{proj}\|^2 \int_{t_0}^{t} \exp(-2(t-t')/\tau_{eff})\, dt'}. \quad \text{(A.13)}$$

In line attractors, the (signal gain)-to-noise ratio is independent of the choice of $\vec{k}$ as long as it is not close to orthogonal to the attracting mode, since both the signal and noise accumulate almost exclusively along the one-dimensional attracting mode. However, in the plane attractor networks, noise develops isotropically in the plane, and in order not to collect noise in the noninput direction, the readout should match the input direction exactly such that $\vec{k}_{proj} \| \vec{v}_{proj}$ (see Figure 4).

A.3 Calculation of Noise Covariance Matrix of Feedforward Networks

Here, we describe how we obtain the noise covariance matrix of the feedforward networks for the computation of the Fisher information with or without reset. First, we consider the case without reset. Without reset, t_0 in equation 2.4 is set to $-\infty$ and $C\leftrightarrow$ can be written as

$$\overleftrightarrow{C} = \int_{-\infty}^{t} \exp[(-\overleftrightarrow{I} + \overleftrightarrow{W})(t-t')/\tau] \exp[(-\overleftrightarrow{I} + \overleftrightarrow{W})^T (t-t')/\tau]\, dt' = \int_{0}^{\infty} \exp[(-\overleftrightarrow{I} + \overleftrightarrow{W})t'/\tau] \exp[(-\overleftrightarrow{I} + \overleftrightarrow{W})^T t'/\tau]\, dt' \text{ with the substitution, } t-t' \to t'.$$

As noted in the supplement of Ganguli, Huhet al. (2008), a recursive relation for $C\leftrightarrow$ can be derived by differentiating the integrand above and using the fundamental theorem of calculus:

$$(-\overleftrightarrow{I}+\overleftrightarrow{W})\overleftrightarrow{C}+\overleftrightarrow{C}(-\overleftrightarrow{I}+\overleftrightarrow{W})^{T}+\tau\overleftrightarrow{I}=0. \quad \text{(A.14)}$$

This recursion relationship is in the form of a continuous Lyapunov equation and can be solved in Matlab using the lyap function.

For the case with a reset, in which t_0 is set to 0, no simple recursive relation can be obtained. In this case, we obtain the analytical form of the noise covariance matrix directly from the explicit expression of neural variability at time t in equation 2.2. The expression for the neural variability at time t is given as follows:

$$\begin{aligned}
&\tfrac{\sigma}{\tau}\textstyle\int_0^t \exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau]\,\overrightarrow{\xi}(t')\,dt' \quad \text{for } \overleftrightarrow{W}=\begin{pmatrix} 0 & \cdots & & 0 \\ \alpha & 0 & & \vdots \\ 0 & \ddots & \ddots & \\ 0 & 0 & \alpha & 0 \end{pmatrix} \\
&=\tfrac{\sigma}{\tau}\sqrt{\tau}\textstyle\int_0^{t/\tau}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t']\,\overrightarrow{\xi}(t')\,dt' \quad \text{with}\tfrac{t-t'}{\tau}\to t' \\
&=\sigma\sqrt{\tau}\textstyle\int_0^{t/\tau}e^{-t'}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha t' & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \frac{(\alpha t')^{n-1}}{(n-1)!} & \cdots & \alpha t' & 1 \end{pmatrix}\begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} dt' \\
&=\sigma\sqrt{\tau}\begin{pmatrix} \int_0^{t/\tau}e^{-t'}\xi_1 dt' \\ \int_0^{t/\tau}\alpha t' e^{-t}\xi_1 dt' + \int_0^{t/\tau}e^{-t'}\xi_2 dt' \\ \vdots \\ \int_0^{t/\tau}\frac{(\alpha t')^{n-1}}{(n-1)!}e^{-t'}\xi_1 dt' + \cdots + \int_0^{t/\tau}e^{-t'}\xi_n dt' \end{pmatrix}.
\end{aligned}$$

In this expression, the variability in the ith neuron contains the filtered noise $\int_0^{t/\tau}\frac{(\alpha t')^{i-k}}{(i-k)!}e^{-t'}\xi_k dt'$ generated from the kth neuron for $k \le i$. The (i, j)th component of the noise covariance matrix is the sum of correlation due to noise generated by all k neurons with $k \le \min(i, j)$ and is

$$\begin{aligned}
&\sigma^2\tau\sum_{k=1}^{\min(i,j)}\textstyle\int_0^{t/\tau}e^{-2t'}\frac{(\alpha t')^{i+j-2k}}{(i-k)!(j-k)!}dt' \\
=&\sigma^2\tau\sum_{k=1}^{\min(i,j)}\frac{(\alpha/2)^{i+j-2k}}{(i-k)!(j-k)!}\gamma(i+j-2k+1, 2t/\tau),
\end{aligned}$$

where $\gamma(i + j - 2k + 1, 2t/\tau)$ is the lower incomplete gamma function.

A.4 Analysis of Fisher Information of Attractor and Feedforward Networks with Reset

In section A.2, we derived analytical expressions for I_F for attractor networks as a function of the strength of network feedback (or, equivalently, network time constant). In this section, we provide an analysis of I_F as a function of the signal gain achieved by the network (Ganguli, Huh et al., 2008). Specifically, here we provide analytical bounds for the Fisher information I_F and show which types of attractor and feedforward networks achieve these bounds.

First, to gain intuition and for comparison with previous work (Ganguli, Huh et al., 2008), we consider an approximation of continuous dynamics by discrete dynamics (see section 2.1). With discrete dynamics, I_F has an upper bound that can be expressed solely as a function of the magnitude of the signal gain vectors at time step m, $W\overleftrightarrow{}^m\vec{v}$:

$$I_F(\vec{r}(l)) \leq \frac{1}{\sigma^2\tau\sum_{m=1}^{l}\frac{1}{\|\overleftrightarrow{W}^m\vec{v}\|^2}}. \quad \text{(A.15)}$$

For networks without a reset, Ganguli, Huh et al. (2008) showed that only feedforward networks satisfy the equality. However, with a reset, the condition for the equality is relaxed to (Ganguli, Huh et al., 2008)

$$\vec{v}^T\overleftrightarrow{W}^{mT}\overleftrightarrow{W}^{m-i}\left[\overleftrightarrow{I}-\frac{\overleftrightarrow{W}^i\vec{v}\vec{v}^T\overleftrightarrow{W}^{iT}}{\|\overleftrightarrow{W}^i\vec{v}\|^2}\right]=0 \quad \text{for all } m \geq i. \quad \text{(A.16)}$$

In this condition, the operator $\overleftrightarrow{I}-\frac{\overleftrightarrow{W}^i\vec{v}\vec{v}^T\overleftrightarrow{W}^{iT}}{\|\overleftrightarrow{W}^i\vec{v}\|^2}$ projects out the activity in the direction of the signal gain vector $W\overleftrightarrow{}^i\vec{v}$ at the ith step, so that the resulting activity is orthogonal to $W\overleftrightarrow{}^i\vec{v}$. Then equation A.16 gives that the evolution of any activity orthogonal to $W\overleftrightarrow{}^i\vec{v}$ after $m-i$ steps, $\overleftrightarrow{W}^{m-i}[\overleftrightarrow{I}-\frac{\overleftrightarrow{W}^i\vec{v}\vec{v}^T\overleftrightarrow{W}^{iT}}{\|\overleftrightarrow{W}^i\vec{v}\|^2}]$, remains orthogonal to the signal $W\overleftrightarrow{}^{m-i}(W\overleftrightarrow{}^i\vec{v}) = W\overleftrightarrow{}^m\vec{v}$ during the evolution of the dynamics.

To find the networks satisfying the above condition, we consider the $W\overleftrightarrow{}$-cyclic subspace generated by $\vec{v}$ and defined as Z = span($\{\vec{v}, W\overleftrightarrow{}\vec{v}, W\overleftrightarrow{}^2\vec{v}, \ldots\}$). First, if we denote the space orthogonal to Z as $Z^\perp$, then the above condition yields that the evolution by $W\overleftrightarrow{}$ does not mix the two spaces, so $W\overleftrightarrow{}$ can be decomposed into blocks of the form

$$\begin{pmatrix} & Z & Z^\perp \\ Z & \overleftrightarrow{W}Z & 0 \\ Z^\perp & 0 & \overleftrightarrow{W}Z^\perp \end{pmatrix}.$$

Moreover, the form of the upper-left submatrix of $W\overleftrightarrow{}$, that is, the transformation of Z to $W\overleftrightarrow{}$Z, is constrained by equation A.16. If we choose $\vec{v}$ as the first coordinate of Z and choose the remaining coordinates as vectors of Z orthogonal to $\vec{v}$, then equation A.16 implies that for any power n, all columns of $W\overleftrightarrow{}^{n}Z$ except the first column remain orthogonal to the first column $W\overleftrightarrow{}^{n}\vec{v}$.

Using the above considerations, it can be verified directly that the upper bound in equation A.16 is achieved by all orthogonal matrices, feedforward chains, and networks with a ring structure constructed by connecting the final element of a feedforward chain to the first (this list is not exclusive; other matrices can also satisfy the bound).

Attractor networks can also satisfy the upper bound exactly or very closely. If we assume that all modes of the attractor network are orthogonal and $\vec{v}$ is aligned to one of the modes, denoted as $\vec{q}_1$, then the attractor networks satisfy equation A.16 with dim(Z) = 1 since $(W\overleftrightarrow{}\vec{q}_i)^T W\overleftrightarrow{}\vec{q}_1 = \lambda_1\lambda_i\vec{q}_i^T\vec{q}_1 = 0$ for $i \neq 1$ due to the orthogonality. If the modes are not orthogonal but there exists a mode with strong recurrent feedback compared to the other modes, we can treat the activity in the other modes as negligibly small so that the network performs similarly to a network with zero eigenvalues in the modes other than the attractor mode. Thus, the low-dimensional attractor networks also satisfy the upper bound closely.

Next, we consider networks with continuous dynamics. The bound for I_F with continuous dynamics is obtained in the same way as for discrete dynamics:

$$I_F(r(t)) \leq \frac{1}{\sigma^2 \int_0^t \frac{1}{\|e^{(-\overleftrightarrow{I}+\overleftrightarrow{W})t'/\tau}\vec{v}\|^2} dt'}, \quad \text{(A.17)}$$

where the equality condition is given as

$$\vec{v}^T \exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})^T t/\tau] \exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau] \left[\overleftrightarrow{I} - \frac{\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t'/\tau]\vec{v}\vec{v}^T \exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})^T t'/\tau]}{\|\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t'/\tau]\vec{v}\|^2}\right] = 0. \quad \text{(A.18)}$$

As in the discrete networks, this equation holds for attractor networks with orthogonal eigen modes and for line attractor networks as long as the input vector is set along one of the attractor modes. For feedforward networks, equation A.18 does not strictly hold. However, we have checked that the feedforward networks come close to achieving the bound given in equation A.17. This is shown in Figure 12, where the numerically calculated $\tilde{I}_F$ (panel B, solid line) is compared to the bound of equation A.17 (panel B, dotted line) calculated numerically from the signal gain vector (panel A).

A.5 Relation Between the Bounds on the Connectivity Strength Between Neurons and the Connectivity Strength Between Modes

Here, we calculate bounds on the connectivity strength between modes when synaptic strengths are bounded by a maximal strength $w_{\max}$ and the modes are orthogonal to each other. First, we consider the attractor networks. If $\vec{q}_i$ and λ_i denote the ith eigenvector and eigenvalue of $W\leftrightarrow$, respectively, then the connectivity matrix $W\leftrightarrow$ is decomposed into a diagonal matrix $D\leftrightarrow$ such that $Q\leftrightarrow^{-1} W\leftrightarrow Q\leftrightarrow = D\leftrightarrow$, where the diagonal matrix $D\leftrightarrow$ has the eigenvalues as the diagonal entries and the column vectors of $Q\leftrightarrow$ are the eigenvectors. If $Q\leftrightarrow$ is an orthogonal matrix, the Frobenius norms $\|\cdot\|_F$ of the two matrices, $W\leftrightarrow$ and $D\leftrightarrow$, are the same, which can be proven by using that the trace of the matrix is preserved under an orthogonal change of coordinates:

$$\begin{aligned}\|\overleftrightarrow{D}\|_F &\equiv \sqrt{\sum_i \lambda_i^2} = \sqrt{\mathrm{Tr}(\overleftrightarrow{D}\,\overleftrightarrow{D}^T)} = \sqrt{\mathrm{Tr}(\overleftrightarrow{W}\,\overleftrightarrow{W}^T)} \\ &= \sqrt{\sum_{i,j} w_{i,j}^2} \equiv \|\overleftrightarrow{W}\|_F. \end{aligned} \tag{A.19}$$

If $w_{\max}$ denotes the maximal synaptic strength, that is, if each element of $W\leftrightarrow$ is bounded above by $w_{\max}$, then the Frobenius norm of $W\leftrightarrow$ is bounded above by $Nw_{\max}$. From equation A.19, the Frobenius norm of $D\leftrightarrow$ has the same bound, and thus each eigenvalue λ_i is at most $Nw_{\max}$. Furthermore, there exists an attractor network that reaches this bound: the matrix with all elements equal to $w_{\max}$ has a maximal eigenvalue equal to $Nw_{\max}$.

Similarly, it can be shown that the bound on the synaptic connectivity between neurons leads to bounds on the strengths of the feedforward connectivity between the Schur modes. For the feedforward networks, $\vec{q}_i$ and $M\leftrightarrow$ denote the Schur modes and the Schur decomposition of $W\leftrightarrow$, satisfying $Q\leftrightarrow^{-1} W\leftrightarrow Q\leftrightarrow = M\leftrightarrow$. Since $Q\leftrightarrow$ is an orthogonal matrix for any Schur decomposition, the Frobenius norm of $M\leftrightarrow$ is equal to that of $W\leftrightarrow$, which is at most $Nw_{\max}$ as in equation A.19. The Frobenius norm of a lower triangular matrix $M\leftrightarrow$ is $\sqrt{\sum_{i\geq j} m_{ij}^2}$. Furthermore, not all m_{ij} can be $Nw_{\max}$ at the same time. As discussed in section 3.2, these bounds lead to the result that the attractor networks with a reset outperform the feedforward networks for a given bound on synaptic strengths.

A.6 Optimal Network Structures When Neuronal Activity is Bounded

Here we obtain the optimal arrangement of the attractor and feedforward networks and calculate their Fisher informations I_F, under constraints on the dynamic range of neural activity. First, we consider a constraint on the mean firing rates in which each neuron is constrained to have the absolute value of its firing rate bound by a maximal value r_0. Formally, such a bound on every element is expressed by the infinity norm of the vector of mean firing rates and denoted below by $\|\mathrm{mean}\,(\vec{r})\|_\infty \leq r_0$. If we assume that the stimulus strength to be remembered is in the range $[-s_0\ s_0]$, then the constraint on the signal gain is given as

$$\left\|\frac{s}{\tau}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau]\overrightarrow{v}\right\|_\infty \le r_0 \quad \text{for } s \text{ in } [-s_0, s_0]$$
$$\text{i. e. } \left\|\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau]\overrightarrow{v}\right\|_\infty \le \frac{\tau r_0}{s_0} \equiv c_0. \tag{A.20}$$

Geometrically, this constraint corresponds to limiting the signal gain vector to reside within a hypercube (see Figure 7). The magnitude of the signal gain vector is then bounded by the distance to the vertices of the hypercube as

$$\left\|\exp\left[(-\overleftrightarrow{I}+\overleftrightarrow{W})t/\tau\right]\overrightarrow{v}\right\|_2 \le c_0\sqrt{N}. \tag{A.21}$$

In addition to the constraint on mean activity, we considered in the main text constraints on the variance of activity as a heuristic for constraints on the absolute size of fluctuations allowed in neuronal firing rates. The bound on variance is applied as follows. When the variance is less than the bound, the network dynamics proceeds normally (but still with bound on the mean). When the variance reaches the bound, the variability of neural activity is assumed to saturate and is set to the bound. Mathematically we assume that the neural activity has a larger dynamic range than the mean activity, so that the bound on the maximum standard deviation of activity can be modeled as a constant multiple of the maximum mean activity r_0. Then the variance of activity is given as

$$\text{var}\ (r_i) = \min\left(c_1 r_0^2, \left[\frac{\sigma^2}{\tau^2}\int_{t_0}^{t}\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})(t-t')/\tau]\times\exp[(-\overleftrightarrow{I}+\overleftrightarrow{W})^{\mathrm{T}}(t-t')/\tau]dt'\right]_i\right), \tag{A.22}$$

where the second term is the amount of accumulated noise in the ith neuron when the dynamics are not constrained.

A.6.1 Attractor Networks

We first consider the line attractor networks without reset. In the following, we denote the attracting eigenmode simply by $\vec{q}$ and the time constant of this mode by τ_{eff}. If only the mean is constrained, we can obtain the maximal Fisher information from the expression for I_F in equation A.9 and the constraint on the mean activity, equation A.20, as

$$\text{Max}\, I_F(T) = \text{Max}\frac{(\overrightarrow{v}\cdot\overrightarrow{q})^2\exp(-2T/\tau_{eff})}{\sigma^2\tau_{eff}/2} \tag{A.23}$$

$$\text{with } \|(\vec{v}\cdot\vec{q})\exp(-t/\tau_{eff})\vec{q}\|_\infty \le c_0 \text{ for } 0 \le t \le T. \quad \text{(A.24)}$$

Without a reset, the network should have exponentially decaying dynamics (τ_{eff} positive) to prevent the infinite accumulation of noise. In this case, the magnitude of the signal gain vector decreases over time and is largest at time 0, so that equation A.24 can be replaced by a constraint on the initial gain, $\|(\vec{v}\cdot\vec{q})\vec{q}\|_\infty \le c_0$. I_F can be maximized by separately maximizing the term $(\vec{v}\cdot\vec{q})^2$ and the term $\exp(-2T/\tau_{eff})/(\sigma^2\tau_{eff}/2)$ in equation A.23. As noted, the magnitude of the vector $(\vec{v}\cdot\vec{q})\vec{q}$ attains its maximal value of $c_0\sqrt{N}$ when $\vec{q}$ points to one of the vertices. Moreover, in section A.2.2, it was found the second term in I_F achieves its maximum when τ_{eff} is equal to the optimal time constant of decay. Thus, altogether, the maximum of I_F becomes

$$I_F(T) = \frac{(\vec{v}\cdot\vec{q})^2\exp(-2T/\tau_{eff})}{\sigma^2\tau_{eff}/2} \le \frac{c_0^2 N}{\sigma^2 eT}. \quad \text{(A.25)}$$

Next, we consider what happens when the maximal variability is also bounded, as in equation A.22. In this case, the maximum Fisher information is obtained as

$$\begin{gathered}\text{Max } I_F(T) = \text{Max}\frac{(\vec{v}\cdot\vec{q})^2\exp(-2T/\tau_{eff})}{\text{var}(\vec{r}\cdot\vec{q})} \\ \text{with var}(r_i) \le \min(c_1 r_0^2, \sigma^2\tau_{eff}q_i^2/2) \\ \text{and } \|(\vec{v}\cdot\vec{q})\exp(-t/\tau_{eff})\vec{q}\|_\infty \le c_0 \quad \text{for } 0 \le t \le T,\end{gathered}$$

where var$(\vec{r}\cdot\vec{q})$ denotes the noise along $\vec{q}$. Note that the accumulated noise in each neuron is not independent since it is the projection of noise in the attractor onto each neuron and var$(\vec{r}\cdot\vec{q})$ is the sum of the noise variances for each neuron. In the optimally arranged networks, in which $\vec{q}$ points along a vertex (i.e., has all components equal in magnitude), the maximal variance is equal for all neurons so that the maximal value of var$(\vec{r}\cdot\vec{q})$ equals $Nc_1r_0^2$. The variance of noise from the dynamics exceeds the maximal variability when $\sigma^2\tau_{eff}q_i^2/2 = \sigma^2\tau_{eff}/(2N) \ge c_1r_0^2$ or, in terms of the feedback strength α when $\alpha \ge \alpha_0 = 1-\sigma^2\tau/(2Nc_1r_0^2)$. Then I_F becomes

$$\begin{gathered}I_F(T) = \frac{(\vec{v}\cdot\vec{q})^2\exp(-2T/\tau_{eff})}{\sigma^2\tau_{eff}/2} \quad \text{for } \alpha<\alpha_0 \quad \text{with } \|(\vec{v}\cdot\vec{q})\vec{q}\|_\infty \le c_0, \\ I_F(T) = \frac{(\vec{v}\cdot\vec{q})^2\exp(-2T/\tau_{eff})}{Nc_1r_0^2} \quad \text{for } \alpha \le \alpha_0 \quad \text{with } \|(\vec{v}\cdot\vec{q})\exp(-t/\tau_{eff})\vec{q}\|_\infty \le c_0 \quad \text{for } 0 \le t \le T.\end{gathered}$$

(A.26)

As discussed in the main text, we consider only the higher signal-to-noise regime in which the maximal Fisher information is obtained when the noise variance is not saturated. Here, we derive a simple estimate of when this regime is attained and show that the optimal value of the network feedback a is not different from that obtained without a bound on the noise variance. In this high signal-to-noise regime, a_0 is greater than the value $a_{opt} = 1 - \tau/(2T)$ (see Figure 8F, peak of dashed line) at which I_F was maximized with only a constraint on the mean activity. Then, for $a < a_0$, because the noise has not yet saturated, there is a maximum in I_F at $a = a_{opt}$ of value $c_0^2 N/(\sigma^2 eT)$ (see equation A.25). For $a \geq a_0$, the maximal I_F is given from equations A.26 and A.21 as $c_0^2 N/(N c_1 r_0^2) = c_0^2/(c_1 r_0^2)$. This value is attained when the numerator of equation A.26 (the signal gain) has reached its maximal value, which occurs for $a \geq 1$. Comparing the expressions for the variance-saturating and nonsaturating regimes, we see that the maximal I_F is achieved in the nonsaturating regime $a < a_0$ when $\frac{\sigma^2}{N} < \frac{c_1 r_0^2}{eT}$, that is, for small σ or large N. Furthermore, we note as claimed above that this maximum occurs at the same $a = a_{opt}$ that was optimal without considering the finite variance of activity.

Next, we consider the line attractor with a reset and bounded mean activity. The maximal I_F is calculated in a similar manner to the case without a reset. However, noise accumulates only during the memory period, so that exponential growth of activity is possible (τ_{eff} can be negative). I_F with the constraint on mean activity is given from equations A.10 and A.20 as

$$\mathrm{Max}\, I_F(T) = \mathrm{Max} \frac{(\vec{v} \cdot \vec{q})^2 \exp(-2T/\tau_{eff})}{\frac{\sigma^2 \tau_{eff}}{2}(1 - \exp(-2T/\tau_{eff}))} \quad \text{(A.27)}$$

$$\text{with } \| (\vec{v} \cdot \vec{q}) \exp(-t/\tau_{eff}) \vec{q} \|_\infty \leq c_0 \quad \text{for } 0 \leq t \leq T. \quad \text{(A.28)}$$

To find the optimal τ_{eff}, we consider separately the cases of positive and negative τ_{eff}. For positive τ_{eff} (decaying attractor mode), the signal gain is maximal at $t = 0$ as in the case without reset, and I_F is maximal for a perfect integrator ($\tau_{eff} = \infty$) with $\vec{q}$ pointing to one of the vertices so that $(\vec{v} \cdot \vec{q}) = c_0 \sqrt{N}$. For negative τ_{eff} (amplifying attractor mode), both the signal gain in the numerator and the noise variance in the denominator increase with τ_{eff}. However, the signal gain is limited by the constraint in equation A.28, so that the maximum of I_F occurs when the noise variance is minimized. This occurs at $\tau_{eff} = -\infty$, corresponding to the perfect integrator. Combining the results for positive and negative τ_{eff} provides that the perfect integrator is optimal and its maximal I_F is $c_0^2 N/(\sigma^2 T)$.

The optimal Fisher information for line attractors with a reset and bound on the variability, as well as the mean, can be computed analogously to the case without a reset. The variance of the noise reaches the bound when $\frac{\sigma^2 \tau_{eff} q_i^2}{2}(1 - \exp(-2T/\tau_{eff})) \geq c_1 r_0^2$ and, for sufficiently

large N or small σ, the maximum I_F over a still occurs at the same value $a = 1$, which was optimal without considering the finite variance of activity.

In summary, we have shown that the optimal I_F for the line attractor networks occurs for the perfect integrator. Moreover, the perfect integrator with a reset has the optimal memory performance of any continuous-dynamics networks with a bounded firing rate. In section A.4, we found the upper limit of I_F in terms of the magnitude of the signal gain vector (see equation A.17). The uniform bound on the mean firing rate sets the upper limit on the magnitude of the signal gain vector to $c_0\sqrt{N}$. Substituting this bound into the expression for the upper limit of I_F, we obtain that $I_F \le c_0^2 N/(\sigma^2 T)$, which (comparing to above) shows that the perfect integrator saturates the bound on memory performance.

A.6.2 Feedforward Networks

Here, we calculate the optimal structure of feedforward networks when the mean activity is constrained. For analytical tractability, we perform this calculation under a discrete dynamics approximation so that, as noted in section A.4, I_F for the feedforward networks achieves the equality in equation A.15.

We first consider literally feedforward networks. We assume the number of stages l is equal to T/τ so that activity reaches the final stage at time T. In the literally feedforward networks, the maximal signal amplification in each stage is restrained to lie in an N_m-dimensional hypercube, where N_m denotes the number of neurons in the mth stage. Then the bound of I_F becomes

$$I_F(\vec{r}(l)) = \frac{1}{\sigma^2\tau\sum_{m=1}^{l}\frac{1}{\|\mathbf{w}^m\vec{v}\|^2}} \le \frac{1}{\sigma^2\tau\sum_{m=1}^{l}\frac{1}{c_0^2 N_m}} \quad \text{with } N = \sum_{m=1}^{l} N_m, \tag{A.29}$$

where c_0 is defined in equation A.20. Using the inequality $\sum_{m=1}^{l}\frac{1}{N_m} \ge l^2/(\sum_{m=1}^{l} N_m)$ and noting that the equality holds when all N_m are equal with $N_m = N/l$, we find that I_F attains a maximal value

$$I_F(\vec{r}(l)) = \frac{1}{\sigma^2\tau l^2/(c_0^2 N)} = \frac{c_0^2\tau N}{\sigma^2 T^2}. \tag{A.30}$$

Finally, we consider functionally feedforward networks. In this case, the maximal signal gain $\|W^{\leftrightarrow m}\vec{v}\|$ for the feedforward networks is $c_0\sqrt{N}$, where N is the total number of neurons. Then the upper bound of I_F is

$$I_F(\overrightarrow{r}(l)) \leq \frac{c_0^2 N}{\sigma^2 \tau l} = \frac{c_0^2 N}{\sigma^2 T}. \quad \text{(A.31)}$$

The equality holds when the signal gain at every stage achieves its maximal bound, that is, when each mode of the feedforward network points to the vertices of the hypercube in the state space (see Figure 9B). It is not obvious that this condition can be attained given that the states of the functionally feedforward network are additionally required to be orthogonal to each other. Below, we show that at least for the case when N is a power of 2, we can construct N mutually orthogonal modes that point to the vertices of the hypercube and use this result to show more generally that the maximal Fisher information of functionally feedforward networks is of order N/T if N is at least twice the number of feedforward stages. We next present the proof of the existence of N orthogonal modes pointing to the vertices of the N-hypercube when $N = 2^i$, where i is a natural number:

Proof—We perform the proof by induction. For $N = 2$, $(1, 1)^T$ and $(1,-1)^T$ satisfy the condition. Now assume that there exist 2^i orthogonal modes whose $N = 2^i$ elements are either −1 or 1. If we denote this set as $(u_1, u_2, \ldots, u_{2^i})$, then for $N = 2^{i+1}$, we can construct 2^{i+1} orthogonal modes from the orthogonal modes corresponding to $N = 2^i$ as follows:

$$\left\{ \left[\begin{pmatrix} u_1 \\ u_1 \end{pmatrix}, \begin{pmatrix} u_1 \\ -u_1 \end{pmatrix} \right], \left[\begin{pmatrix} u_2 \\ u_2 \end{pmatrix}, \begin{pmatrix} u_2 \\ -u_2 \end{pmatrix} \right], \ldots, \left[\begin{pmatrix} u_{2^i} \\ u_{2^i} \end{pmatrix} \begin{pmatrix} u_{2^i} \\ -u_{2^i} \end{pmatrix} \right] \right\}.$$

The pairs of modes $\left[\begin{pmatrix} u_j \\ u_j \end{pmatrix}, \begin{pmatrix} u_j \\ -u_j \end{pmatrix} \right]$ are orthogonal due to the negative sign. The different groups constructed from different u_j are orthogonal by definition of these being orthogonal modes from the case when $N = 2^i$.

The above construction can also be used to show that for any N, the maximal Fisher information of functionally feedforward networks is of order N/T if N is larger than twice the number of feedforward stages $l = T/\tau$. For general N, we can generate at least $N/2$ orthogonal modes by applying the above construction to the maximal power of 2 less than N. This creates an $N/2$-length feedforward network that uses more than half of the full dynamic range. As in the calculation leading to equation A.31, the Fisher information of such networks is of order $(N/2)/T$ when the number of neurons N is greater than twice l ($N/2 \geq l$). Thus, for general N, the maximal Fisher information is still of order N/T for $N \geq 2l$. By contrast, for a continuous (rather than discrete) feedforward network, the spreading out of activity over time implies that it is impossible for the network to maintain a signal gain vector pointing to a vertex at all times. Therefore, the continuous feedforward networks, unlike their discrete counterparts, strictly cannot attain the maximal bound on the Fisher information.

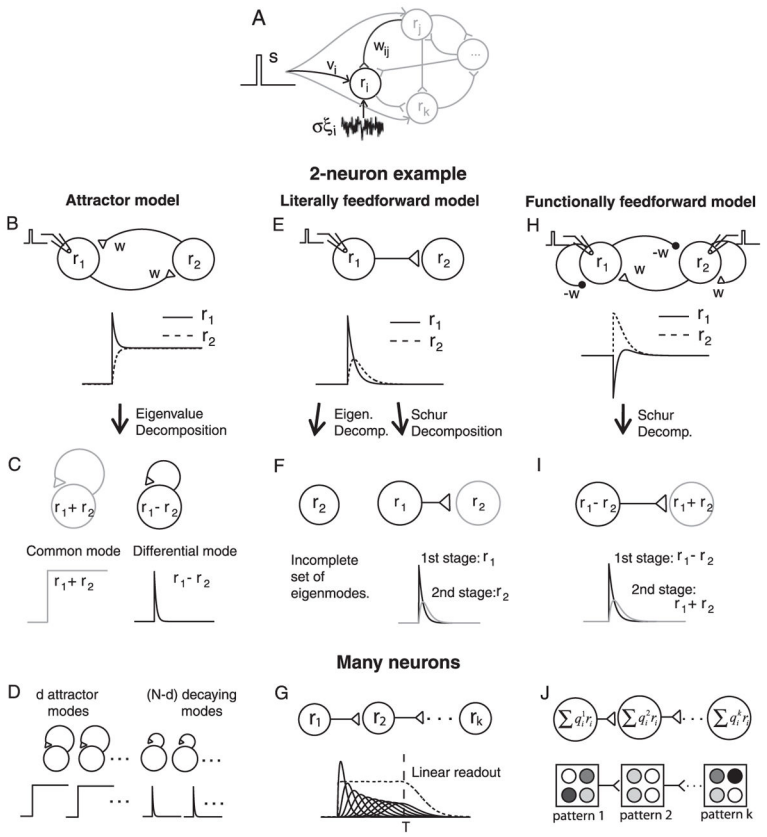

Figure 1.

Network models for short-term memory. (A) Schematic of network model that memorizes the amplitude of a transient stimulus. (B–D) Attractor networks. (B) A simple attractor network composed of two mutually excitatory neurons. (C) Eigenvalue decomposition of two-neuron attractor network. Network activity is decomposed into common and difference modes corresponding to the sum and difference of neural activities. The feedback strength of the mode onto itself represents the corresponding eigenvalue of each mode. The common mode is an attractor mode. (D) *d*-dimensional attractor network having *d* attractor modes. (E–G) Literally feedforward networks. (E) Two-neuron feedforward chain. The activity in the early neuron is passed on to the next neuron and is filtered. (F) Decomposition of feedforward networks. Feedforward networks cannot be decomposed into eigenvectors because there is only one eigenvector. Instead, they can be characterized by the Schur decomposition, which allows feedforward connections between orthogonal activity patterns. (G)Activity of a longer feedforward chain of neurons and a linear readout of this chain (dashed) giving persistent activity. (H–J) Functionally feedforward networks. (H) Recurrent network consisting of one excitatory and one inhibitory neuron. (I) Schur decomposition reveals the feedforward connection between the modes. In this network, the difference mode (black) is projected to the common mode (gray), and temporal profiles of these Schurmodes are the same as the neural activities in the literally feedforward network shown in *F*. (J) More neurons may implement a longer functionally feedforward chain. Note that superscripts here denote Schur modes, not powers.

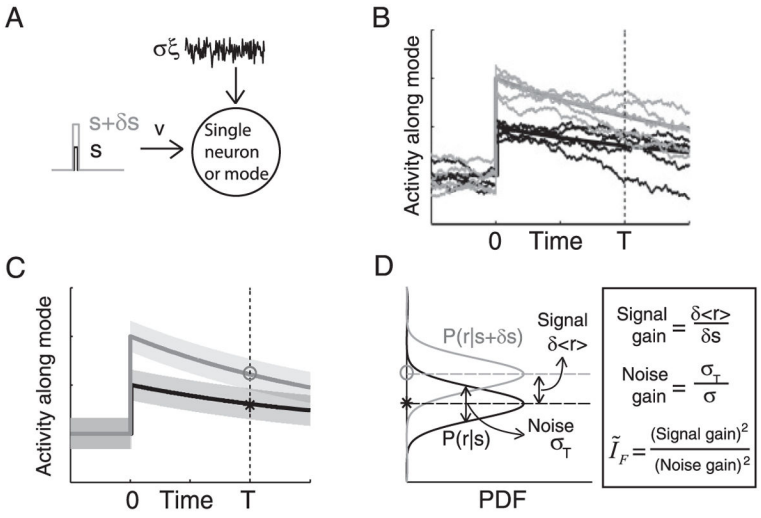

Figure 2.

Competition between signal and noise accumulation. (A) Single neuron or eigenmode in attractor networks subject to two different stimuli s (black) and $s+\delta s$ (gray), and noise. (B) Time course of noisy neural activities. The stimuli with different strengths generate different-size jumps. The mean trajectories given in the solid curves remain separated, whereas individual trajectories may overlap due to noise. (C) Mean neural activities with spread representing the variability across trials. The circle and the asterisk are the mean neural activities after the memory duration T. (D) Computation of Fisher information. Distributions of neural activities carry the information about the signal. The ratio between the square of the signal gain and noise gain is the normalized Fisher information, $\tilde{I}_F = \sigma^2 I_F$ (PDF: probability density function).

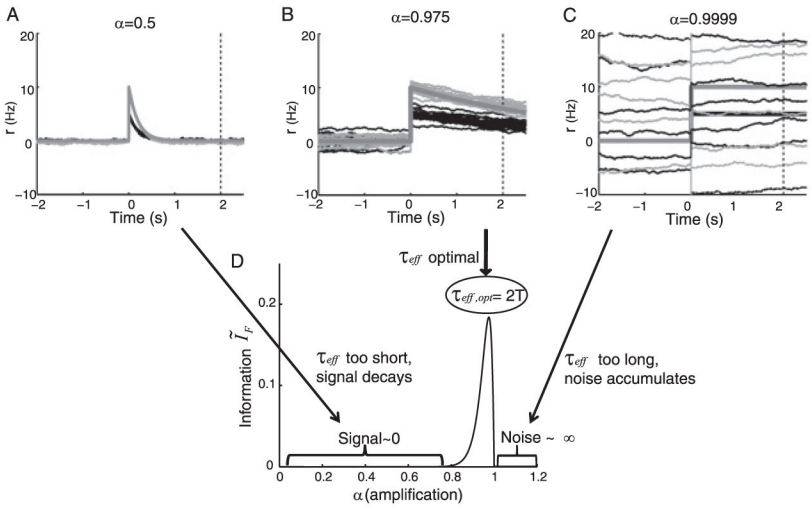

Figure 3.

Memory performance of attractor networks without reset. (A–C) Mean and single-trial firing rate trajectories for different strengths of recurrent feedback a. For small a, the signal decays quickly to zero (A). For large a, noise accumulates infinitely, and distributions of trajectories have infinite variance (C). The optimal I_F occurs when the signal decay and noise accumulation are appropriately balanced (B). T=2 s, τ = 0.1 s and σ = 2 Hz$^{1/2}$ are chosen for illustration. (D) Fisher information $\tilde{I}_F$ as a function of a. $\tilde{I}_F$ attains a maximum at the optimal time constant of decay, $\tau_{eff,opt} = 2T$.

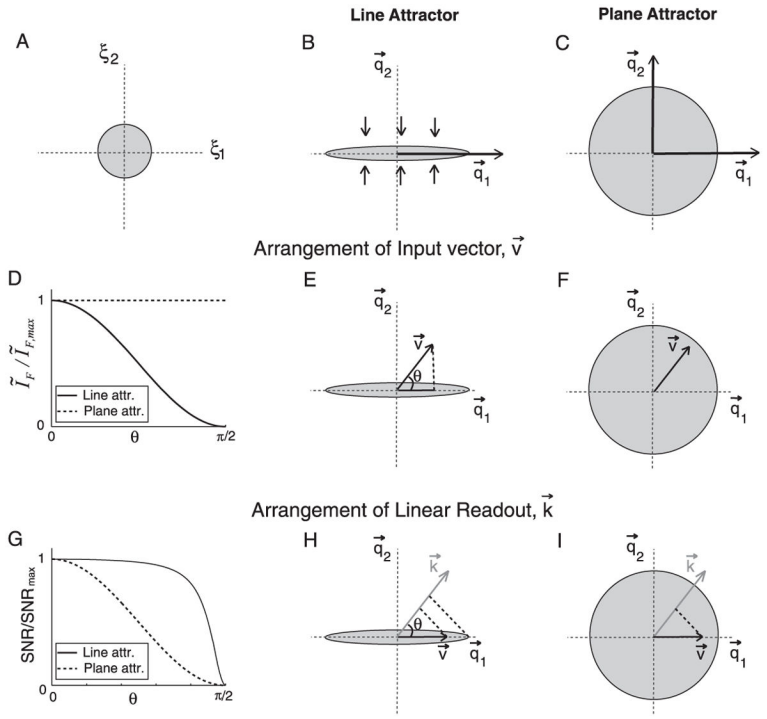

Figure 4.

Comparison between line and plane attractor networks. (A–C) Profile of noise covariance matrix in two modes. We assume that eigenmodes are orthogonal to each other and plot the two eigenmodes, $\vec{q}_1$ and $\vec{q}_2$, with the two largest eigenvalues (dimensions orthogonal to $\vec{q}_1$ and $\vec{q}_2$ are not shown). (A) Profile of covariance matrix of injected noise, which has equal strength in all directions. (B) Accumulated noise covariance matrix in line attractor networks. Noise except along the line attractor $\vec{q}_1$ decays rapidly so that only noise along $\vec{q}_1$ is prominent. (C) Accumulated noise covariance matrix in plane attractor networks. Noise amplitude is equal in every direction on the plane defined by $\vec{q}_1$ and $\vec{q}_2$. (D–F) Fisher information $\tilde{I}_F$ for different arrangements of the input vector $\vec{v}$. (D, E) In the line attractor, the memory performance is proportional to the amplitude of the projection of $\vec{v}$ onto $\vec{q}_1$, so that $\tilde{I}_F$ monotonically decreases as the angle between $\vec{q}_1$ and $\vec{v}$ increases to $\pi/2$. (D, F) In the plane attractor network, the signal can be stored in any direction on the plane, and $\tilde{I}_F$ remains the same if $\vec{v}$ is on the plane attractor. (G–I) $\tilde{I}_F$ for different arrangements of the linear readout $\vec{k}$. (G, H) In the line attractor, only $\vec{q}_1$ stores the signal and noise, and the signal-to-noise ratio remains approximately constant for different $\vec{k}$ if noise in the other modes is small. (G, I) In the plane attractor, noise accumulates in all directions, and when the projection of $\vec{k}$ onto the plane is not along the projection of $\vec{v}$ onto the plane, noise in the non-readout directions lowers $\tilde{I}_F$.

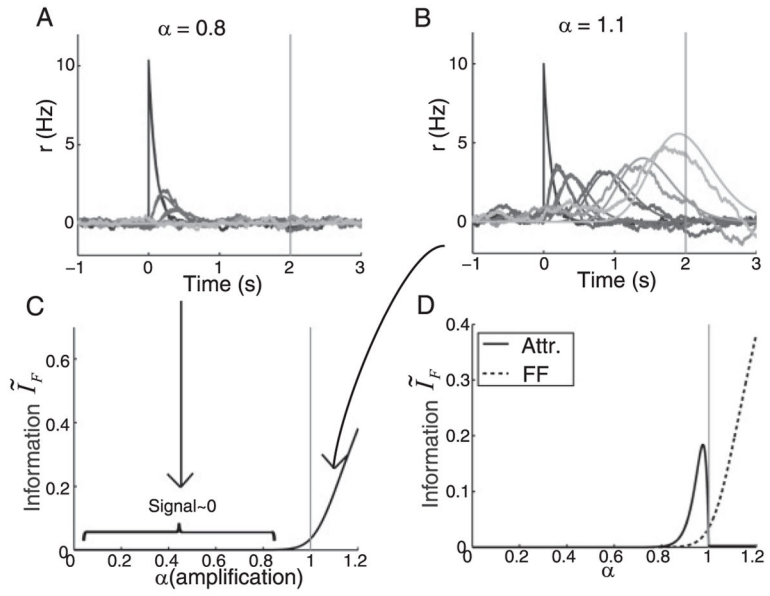

Figure 5.

Memory performance of feedforward networks without reset and comparison to attractor networks. (A, B) Trajectories without reset for different strengths of feedforward connectivity a. (A) For small a, the amplitude of activity of neurons or Schur modes decays quickly to zero as it propagates along the feedforward chain (smooth lines: mean activities; noisy lines: activities for one realization of noise). (B) For $a > 1$, activity is amplified until it reaches the end of the feedforward chain. For both a values, noisy trajectories are not highly different from the mean trajectories shown in the solid curves, reflecting lack of noise buildup. Network parameters are the same as in Figure 3, and the total number of stages is 20. Only activities at stages 1, 3, 5, 10, 15, and 20 (different gray scales; different colors in Supplemental Figure S1) are shown. (C) Fisher information $\tilde{I}_F$ increases monotonically with a in linear feedforward networks, reflecting that feedforward networks are able to amplify signals without excessive accumulation of noise. (D)Comparison of $\tilde{I}_F$ between attractor and feedforward networks. (See Supplemental Figure S1 for a color version of this figure.)

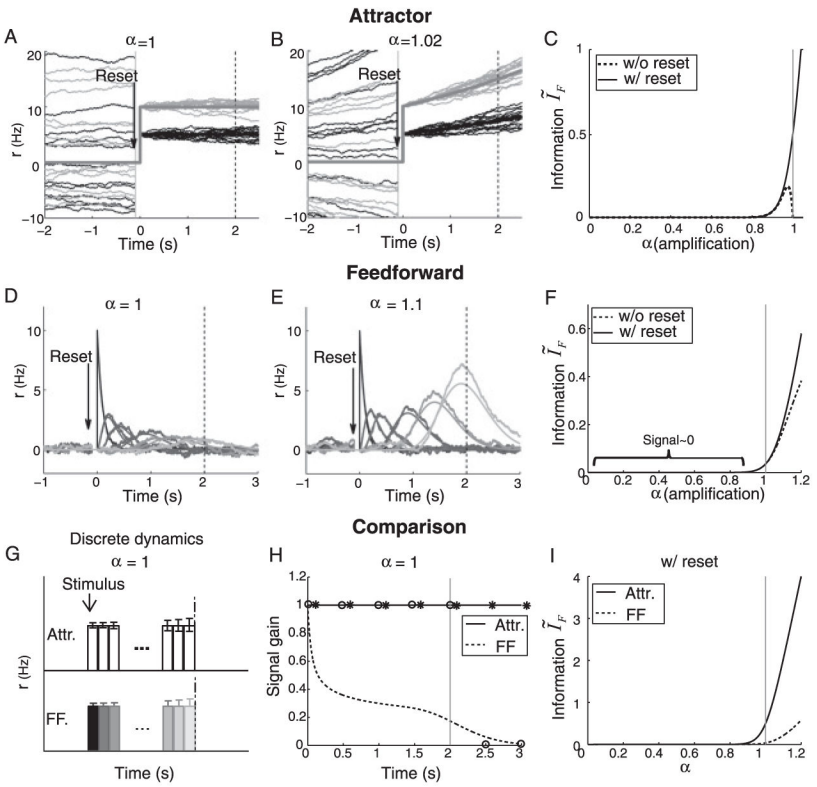

Figure 6.

Effect of a reset mechanism on the memory performance of attractor and feedforward networks. (A, B) Trajectories of attractor networks when the neural activity is reset to zero near the stimulus onset. For the perfect integrator (a = 1; A) or an amplifying mode (a = 1.02; B), trajectories for different stimuli are well separated. (C) Fisher information $\tilde{I}_F$ of attractor networks with or without a reset. With a reset, $\tilde{I}_F$ increases monotonically with the feedback strength a. (D, E) Trajectories of feedforward networks with a reset. With a reset, the variability is reduced for any a. (F) $\tilde{I}_F$ of feedforward networks with or without a reset. Including a reset increases $\tilde{I}_F$ but not as much as in attractor networks. (G) Comparison of activity of attractor and feedforward networks in discrete dynamics. Different grayscales (different colors in Supplemental Figure S2) represent activities in different neurons or modes. For a = 1, the signal amplitude is equal at all times for both networks, but activity remains in the same neuron or mode in the attractor networks, whereas the activity propagates along the chain in the feedforward networks. Activity in the feedforward networks shifts perfectly to the next stage without loss until exiting from the chain, in contrast to the spread of neural activity and corresponding signal loss in networks with continuous dynamics (panels D, and H). (H) Signal gain of attractor and feedforward networks with reset for a = 1. While the magnitude of the signal remains constant up to time T for the networks with discrete dynamics (circles and asterisks; red circles and blue asterisks in Supplemental Figure S2) and for continuous attractor networks (solid curve; blue curve in Supplemental Figure S2), it decreases toward zero in feedforward networks with continuous dynamics (dashed curve; dashed red curve in Figure S2). (I) $\tilde{I}_F$ of attractor and feedforward networks with reset. Due to lower signal gain (panel H), the feedforward

networks maintains less information $\tilde{I}_F$ than the attractor networks. (Parameters for this figure are the same as in Figure 3. See Supplemental Figure S2 for a color version of this figure.)

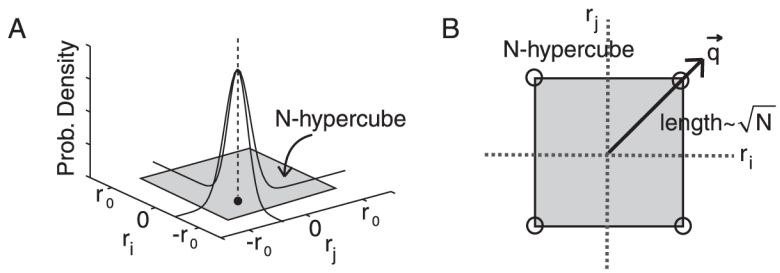

Figure 7.

Constraint on mean firing activity. (A) Schematic of the constraint on mean firing rates. The mean firing rates of each neuron are constrained to be less than or equal to r_0 in magnitude. This defines an N-dimensional hypercube in activity space (gray square for $N = 2$). However, the actual neural activity may lie outside this bound due to noise (example probability distribution of activities outlined with solid lines, with mean indicated by dashed line and circle). (B) Optimal arrangement of attracting or Schur modes. The magnitude of neural activity is maximal when each attractor or Schur mode, denoted as $\vec{q}$, points to one of the vertices (open circles) of the hypercube.

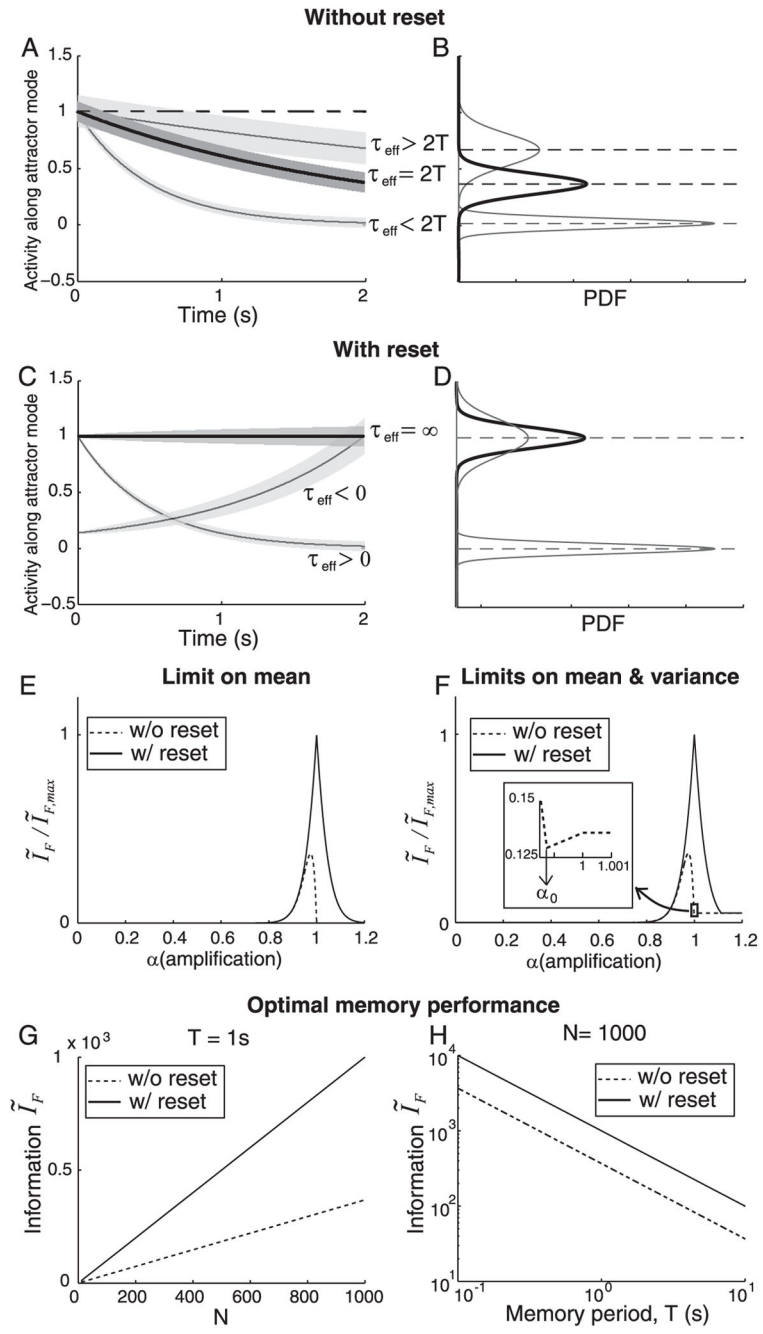

Figure 8.

Optimal architecture of line attractor networks with a finite dynamic range. (A, B) Optimal recurrent feedback a (or equivalently τ_{eff}) in attractor networks without a reset under the constraint on mean activities. (A) Time course of the mean and standard deviation of activity along the attractor for different amounts of recurrent feedback (gray, a = 0.8 and 0.99; black, a = 0.975; other parameters are the same as in Figure 3). (B) Distribution of the activity at $t = T$. With a fixed bound on each mean firing rate, the memory performance without reset is maximized at $\tau_{eff,opt} = 2T$ as in the case (see Figure 3) without a constraint

on the mean firing rate. (C,D)Optimal recurrent feedback in attractor networks with a reset under the constraint on mean activities. With a reset, amplifications $a > 1$ are allowed, but the initial trajectory must be adjusted so that the whole trajectory lies under the bound (gray, $a = 0.8$ and 1.1; black, $a = 1$). In this case, the perfect integrator performs best. (E) Fisher information $\tilde{I}_F$ as a function of a under the constraint on the mean firing rate. (F) $\tilde{I}_F$ under the constraints on the mean firing rate and maximal variability. The minimum feedback strength that saturates the bound on the variance of firing rates is denoted by a_0. For $a > a_0$, the noise variance is set to a constant, and changes in $\tilde{I}_F$ only reflect changes in the signal gain (inset). Adding the bound on the maximal variance of noise prevents $\tilde{I}_F$ from being zero for large a but does not change the location of the maximum $\tilde{I}_F$ for sufficiently large N and small σ (see section A.6). (G) Linear growth of the maximal $\tilde{I}_F$ with increasing N (solid, with reset; dashed, without reset). The maximal $\tilde{I}_F$ with or without reset differs by a constant multiplicative factor e. (H) Maximal $\tilde{I}_F$ decays with increasing T as a power law with exponent −1. For networks without reset, the shown $\tilde{I}_F$ for a given value of T reflects the performance of a network with decay time optimized for this memory duration.

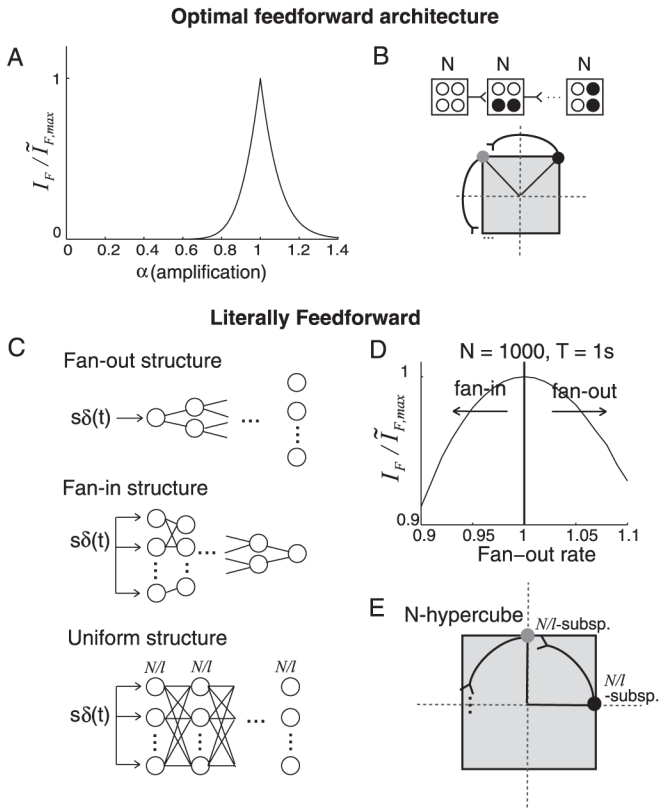

Figure 9.

Optimal architecture of feedforward networks with finite dynamic range. (A, B) Optimal architecture for feedforward networks with discrete dynamics. (A) Fisher information $\tilde{I}_F$ as a function of the feedforward strength a. $\tilde{I}_F$ for feedforward networks with discrete dynamics is the same for networks with and without a reset. As in the attractor networks with a reset, the optimal $\tilde{I}_F$ occurs for networks that maintain activity at a constant level. (B) For the optimal feedforward network, the Schur modes correspond to the vertices of the N-dimensional hypercube (bottom), which give the maximal magnitude of neural activity within the bound (top, schematic of Schur modes in which networks attain their maximum and minimum firing activity within the bound). (C–E) Optimal literally feedforward networks with number of stages equal to the duration of the memory period. (C) Schematics of different architectures of literally feedforward networks. (D) $\tilde{I}_F$ for different fan-out rates. A fan-out rate equal to 1 corresponds to a uniform structure having an equal number of neurons at each stage (C, bottom). $\tilde{I}_F$ is maximal for this uniform structure in discrete dynamics. (E) Uniform structure in the activity space. The literally feedforward network with an equal number of neurons in each stage can be visualized as propagating activity from the previous to the next N/l-dimensional hypercubes where N and l are numbers of neurons and stages, respectively.

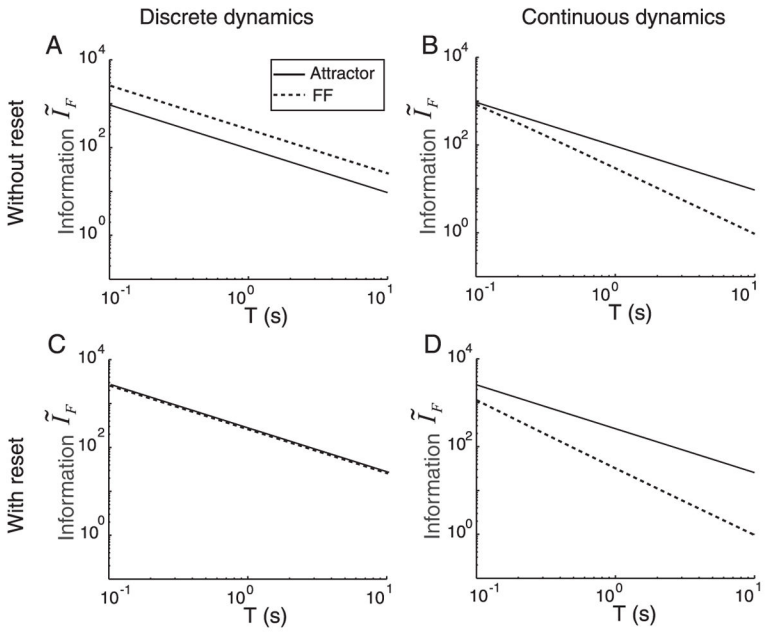

Figure 10.

Optimal memory performance of the attractor and feedforward networks with a finite dynamic range for different memory durations T. (A, C) Fisher information $\tilde{I}_F$ of attractor and feedforward networks with discrete dynamics as a function of T (A, without a reset; C, with a reset). The maximal $\tilde{I}_F$ in all cases decays as a power law with exponent −1. The total number of neurons $N = 256$. (B, D) $\tilde{I}_F$ in continuous dynamics without (B) or with reset (D). The maximal $\tilde{I}_F$ in attractor networks is compared to $\tilde{I}_F$ for functionally feedforward networks with architectures that optimize the memory performance in discrete dynamics. The attractor networks have an exponent −1 (solid line), and the feedforward networks approximately obey a power law but with an exponent less than −1 (dashed line). With or without reset, the attractor networks perform better.

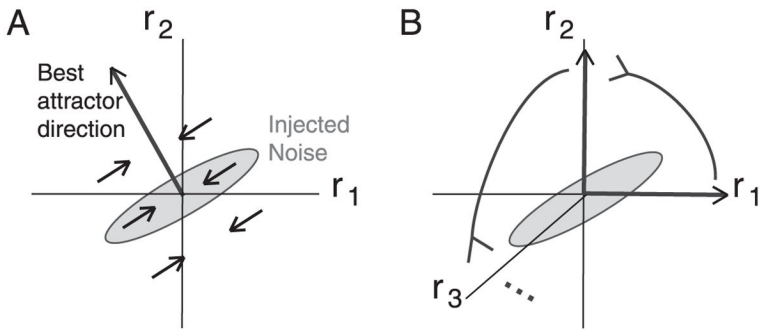

Figure 11.

Effect of correlated noise on network structures. (A) Optimal arrangement of attractor mode for attractor networks in the presence of correlated noise. In the presence of noise having nonuniform noise variance (gray ellipse), the attractor networks can reduce noise most and maximize memory performance by setting the attractor mode and the input vector to be orthogonal to the largest noisy direction. (B) Arrangement of the Schurmodes in feedforward networks. Because activity sweeps between many different directions defined by the Schur modes, excluding specific noisy directions is more difficult in the feedforward networks.

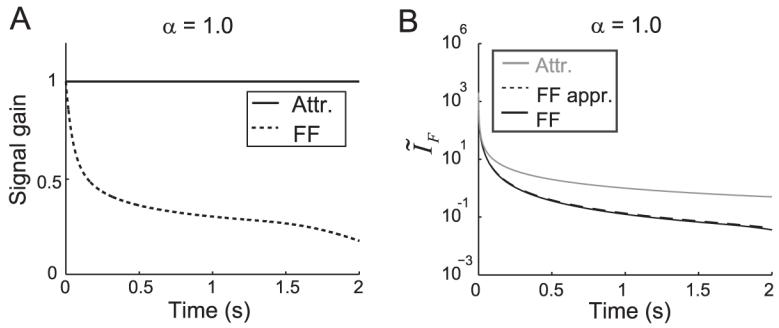

Figure 12.

Calculation of the Fisher information $\tilde{I}_F$ in continuous dynamics. (A) The magnitude of the signal gain vector in attractor and feedforward networks with continuous dynamics. (B) Comparison of $\tilde{I}_F$ for feedforward networks with the upper bound given in equation A.17. $\tilde{I}_F$ of attractor networks can be calculated analytically for continuous dynamics (gray curve) and satisfies the upper bound. $\tilde{I}_F$ of feedforward networks obtained numerically (black solid curve) is close to the upper bound obtained semianalytically from the magnitude of the signal gain vector given in panel A (black dashed curve). Note that the upper bounds for the attractor and feedforward networks are different, because these networks have different signal gain.