Short-Term Memory in Orthogonal Neural Networks

Olivia L. White,[1] Daniel D. Lee,[2] and Haim Sompolinsky[1,3]
[1]*Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*
[3]*Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel*
(Received 11 November 2003; published 9 April 2004)

We study the ability of linear recurrent networks obeying discrete time dynamics to store long temporal sequences that are retrievable from the instantaneous state of the network. We calculate this temporal memory capacity for both distributed shift register and random orthogonal connectivity matrices. We show that the memory capacity of these networks scales with system size.

DOI: 10.1103/PhysRevLett.92.148102 PACS numbers: 87.18.Sn, 05.45.Tp, 84.35.+i, 89.75.Hc

The brain holds information in short-term memory for use in prospective action. It is thought that persistent firing patterns in cortical networks subserve such working memory and several mechanisms have been proposed [1]. In some, a stimulus, such as a spoken word or a picture, activates a pattern of neuronal activity that persists for several seconds because it corresponds to an isolated stable fixed point of the dynamics [2]. However, working memory often involves memorization of graded signals, such as stimulus location in 2D space or the eye's gaze, which are hard to associate with discrete attractors. Recent models propose that short-term memory is associated with low dimensional continuous manifolds of attractors. Both network [3] and single cell mechanisms [4] have been implicated in generating these manifolds.

More recently it was suggested that a generic recurrent network can store arbitrary temporal inputs in its transient responses, even though these responses do not correspond to attractors, and that working memory operates by reconstructing input history from the network's current state [5,6]. This proposal has been investigated numerically for some recurrent networks but its theoretical underpinnings are unexplored. For instance, how does a network's storage capacity for temporal memory scale with system size? What network architectures are suitable? How do noise and structural perturbations affect memory? Such issues have important implications for general dynamical systems. To what extent can the history of perturbations on a dissipative system be reconstructed from its current state? How does the transient memory of a dynamical system depend on its number of degrees of freedom and the amount of noise? In this Letter we develop theoretical understanding of the capacity of linear recurrent networks to store temporal signals.

Model.—We consider here the discrete time model proposed by Jaeger [6]. A time dependent scalar signal $s(n)$ is memorized by a linear recurrent network with N neurons obeying the discrete time dynamics

$$\mathbf{x}(n) = \mathbf{W}\mathbf{x}(n-1) + \mathbf{v}s(n) + \mathbf{z}(n). \tag{1}$$

$\mathbf{x}(n)$ gives the network state at time n, $\mathbf{v}$ is a unit norm constant vector of connections from the input source, $\mathbf{W}$ is the matrix of recurrent connections which need not be symmetric, and $\mathbf{z}(n)$ is a noise vector. To ensure dynamical stability, we require $\alpha < 1$, where α is the norm squared of $\mathbf{W}$'s largest eigenvalue. The goal is to extract the scalar history of the signal, $\{s(m)|m \le n\}$, from the current network state $\mathbf{x}(n)$. This is achieved using a layer of linear readout neurons, the state of which at time n is given by $\{y_k(n) = \mathbf{u}_\mathbf{k}^\mathrm{T}\mathbf{x}(n), k = 0, 1, 2, \ldots\}$; see Fig. 1. $\mathbf{u}_k$ is a constant vector of output connections from the recurrent network to the kth readout neuron. It is chosen to minimize mean square deviations $\langle|y_k(n) - s(n-k)|^2\rangle_n$ so that $y_k(n)$ is close to $s(n-k)$, where $\langle\cdots\rangle_n$ denotes a time average. The resultant optimal output weight is $\mathbf{u}_\mathbf{k} = \mathbf{C}^{-1}\mathbf{p}_\mathbf{k}$, where $\mathbf{C} = \langle\mathbf{x}(n)\mathbf{x}^\mathrm{T}(n)\rangle_n$ is the covariance matrix of $\mathbf{x}$, and $\mathbf{p}_\mathbf{k} = \langle s(n-k)\mathbf{x}(n)\rangle_n$. The ability to embed signals in the network may depend on their statistics. Here we characterize the signal ensemble by $\langle s(n)\rangle_n = 0$ and $\langle s(n)s(n+k)\rangle_n = \delta_{k,0}$. The noise vectors have zero mean $\langle\mathbf{z}(n)\rangle_n = 0$ and variance $\langle\mathbf{z}_i(n)\mathbf{z}_j(n+k)^\mathrm{T}\rangle_n = \epsilon\delta_{k,0}\delta_{i,j}$. With the above signal and noise statistics, $\mathbf{p}_\mathbf{k} = \mathbf{W}^k\mathbf{v}$, and

$$\mathbf{C} = \sum_{k=0}^{\infty}\mathbf{p}_\mathbf{k}\mathbf{p}_\mathbf{k}^\mathrm{T} + \epsilon\mathbf{C}_\mathbf{n}, \tag{2}$$

where the scaled noise covariance is $\mathbf{C}_\mathbf{n} = \sum_k \mathbf{W}^k\mathbf{W}^{k\mathrm{T}}$.

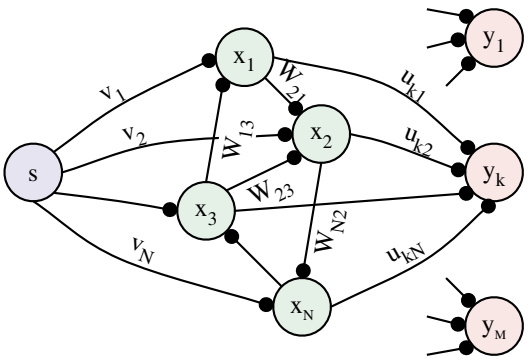

FIG. 1 (color online). Architecture of the short-term memory network. A single input unit $\mathbf{s}$ feeds into an N unit recurrent network. Each of an array of readout units reconstructs the input at a given past time.

Memory function.—We define the system's memory function as the overlap between the past input and its reconstructed value, $m(k) = \langle s(n-k)y_k(n)\rangle_n$. With the above statistics,

$$m(k) = \mathbf{p}_\mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{p}_\mathbf{k}. \tag{3}$$

$m(k) = 1$ corresponds to perfect reconstruction, $y_k(n) = s(n-k)$, whereas $m(k) = 0$ indicates no memory of $s(n-k)$. For an arbitrary, stable connection matrix $\mathbf{W}$,

$$\sum_{k=0}^{\infty} m(k) = \mathrm{Tr}\mathbf{C}^{-1}\sum_{k=0}^{\infty}\mathbf{p}_\mathbf{k}\mathbf{p}_\mathbf{k}^{\mathrm{T}} = N - \epsilon\mathrm{Tr}\mathbf{C}^{-1}\mathbf{C}_\mathbf{n}. \tag{4}$$

This sum rule provides a useful indication of the network's short-term memory. For zero noise, the area under the memory function is exactly N, implying that all N degrees of freedom are useful for storage. Storage capacity decreases with strength of noise. System performance also relates to the shape of $m(k)$. Since $0 \le m(k) \le 1$, it follows that the length of signal that can be exactly reproduced is also bounded. To characterize the length of time over which a signal can be retrieved with reasonable accuracy, we define the temporal capacity k_C as the minimum value of k such that $m(k) < \frac{1}{2}$. We focus particularly on the conditions under which the system's capacity is extensive, namely, $k_C \propto N$ as $N \to \infty$. For given noise ϵ, we define α^{opt} as the value of α at which capacity achieves its maximum, k_C^{opt}.

Distributed shift register (DSR) network.—A straightforward candidate for a short-term memory system is a delay line, or a *shift register* network, which corresponds to a one-dimensional network with $W_{ij} = \sqrt{\alpha}\delta_{i,j+1}$, and $v_i = \delta_{i,1}$. One drawback of this system is its extreme sensitivity to removal of a single neuron. A more robust, distributed architecture of the shift register operation is a fully connected network with

$$\mathbf{W} = \sqrt{\alpha}\sum_{k=1}^{N-1}\mathbf{v}^{(\mathbf{k}+1)}\mathbf{v}^{(\mathbf{k})\mathrm{T}}; \qquad \mathbf{v}^{(\mathbf{1})} \equiv \mathbf{v}, \tag{5}$$

where $\{\mathbf{v}^{(\mathbf{k})}\}$ is an arbitrary set of N orthonormal vectors. Note that $\mathbf{W}\mathbf{v}^{(\mathbf{k})} = \sqrt{\alpha}\mathbf{v}^{(\mathbf{k}+1)}$ for $k \le N-1$ and $\mathbf{W}\mathbf{v}^{(\mathbf{N})} = 0$, implying that $\mathbf{W}^N = 0$ [7]. In this network the covariance matrix is

$$\mathbf{C} = \sum_{k=1}^{N}[\alpha^{k-1} + \tilde{\epsilon}(1-\alpha^k)]\mathbf{v}^{(\mathbf{k})}\mathbf{v}^{(\mathbf{k})\mathrm{T}}, \tag{6}$$

where $\tilde{\epsilon} = \epsilon/(1-\alpha)$. The memory function is given by

$$m(k) = \frac{\alpha^k}{\alpha^k + \tilde{\epsilon}(1-\alpha^{k+1})}, \qquad k = 0, \ldots, N-1 \tag{7}$$

and $m(k) = 0$ for $k \ge N$. In zero noise, $\mathbf{x}(n) = \sum_{k=0}^{N-1}\alpha^{k/2}s(n-k)\mathbf{v}^{(\mathbf{k}+1)}$. Thus the network embeds each of the previous N signal values in a distinct orthogonal direction $\mathbf{v}^{(\mathbf{k})}$ for k up to $N-1$. An important question in both this and the following models is how the value of α affects the system performance. In the absence of noise the present model retrieves perfectly the most recent N inputs for all values of α, as implied by Eq. (7). However, the required readout weights $\mathbf{u}_\mathbf{k} = \alpha^{-k/2}\mathbf{v}^{(\mathbf{k}+1)}$ for the retrieval of these memories increase with decreasing α, limiting the choice of α to values close to 1.

Nonzero noise contributes to $\mathbf{x}(n)$, polluting the signal but leaving fixed the directions along which temporal signals are embedded. When $\epsilon > 0$, $m(k) < 1$ for all k. The capacity k_C is greater than 0 for all $\epsilon < 1$. It increases with decreasing ϵ and saturates to $k_C = N$ at zero noise. For fixed noise, increasing α increases signal-to-noise ratio and hence $m(k)$ increases; see Fig. 2. Thus $\alpha^{\mathrm{opt}} = 1$ for all values of noise $\epsilon > 0$.

Random orthogonal network.—We next ask whether broader classes of connection matrices can also store long temporal signals. A plausible extension of the above model is to a network with $\mathbf{W} = \sqrt{\alpha}\mathbf{O}$, where $\mathbf{O}$ is an $N \times N$ orthogonal matrix (i.e., $\mathbf{O}\mathbf{O}^{\mathrm{T}} = 1$) and $\alpha < 1$. Similar to the DSR model, $\mathbf{W}$ performs a rotation followed by a shrinking with factor $\sqrt{\alpha}$. However $\mathbf{W}^k\mathbf{v}$ and $\mathbf{W}^{k+1}\mathbf{v}$ are not necessarily orthogonal, in contrast to the DSR. Moreover while $\mathbf{W}^N = 0$ for the DSR, orthogonal $\mathbf{W}$ is full rank. Consequently, inputs from times earlier than N can interfere with current inputs. For any choice of $\mathbf{O}$ and $\mathbf{v}$ the covariance matrix is

$$\mathbf{C} = \sum_{k=0}^{\infty}\alpha^k\mathbf{O}^k\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{O}^{-k} + \tilde{\epsilon}\mathbf{I}, \qquad \tilde{\epsilon} \equiv \epsilon/(1-\alpha). \tag{8}$$

However the system behavior can depend upon the particular $\mathbf{O}$ and $\mathbf{v}$. We therefore consider $\mathbf{O}$ drawn from the

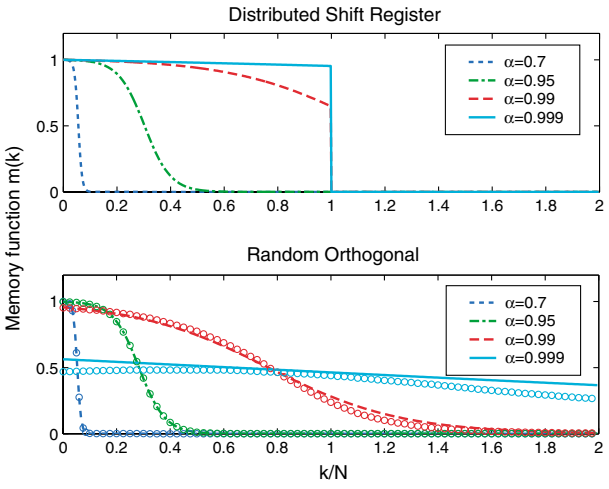

FIG. 2 (color). Memory capacities of DSR and orthogonal networks for $\epsilon = 10^{-4}$ and $N = 400$ at various values of α. For the orthogonal network, the circles show simulation results and the solid lines show predictions of the annealed approximation, which begins to break down for α very near 1. While the memory of the DSR improves with increasing α, the capacity of the orthogonal network peaks at $\alpha^{\mathrm{opt}} = 0.98$.

Gaussian orthogonal ensemble and input connections $\mathbf{v}$ from a Gaussian distribution with $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$. We evaluate $m(k) = \langle \mathbf{p}_{\mathbf{k}}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{p}_{\mathbf{k}} \rangle$ where the average is over these ensembles and captures the typical behavior for N large. Exact analytical evaluation of $m(k)$ is complex and requires accounting for statistical correlations between powers of random orthogonal matrices. In the following we solve the problem under the "annealed approximation" (AA), in which $\mathbf{W}$ and $\mathbf{v}$ are not quenched in time but drawn randomly at each time step. Under this approximation,

$$m(k) = \frac{\alpha^k q}{1 + \alpha^k q}, \tag{9}$$

where q satisfies

$$1 = N^{-1} \sum_{k=0}^{\infty} \frac{\alpha^k q}{1 + \alpha^k q} + \tilde{\epsilon} q. \tag{10}$$

To see this, first note that in the annealed scenario $\mathbf{p}_k = \mathbf{v}^{(\mathbf{k})}$, where $\{\mathbf{v}^{(\mathbf{k})}\}$ is an infinite set of independent random normalized vectors. Hence,

$$m(k) = \langle \alpha^k \mathbf{v}^{(\mathbf{k})\mathrm{T}} [\mathbf{I} + \mathbf{C}_0^{-1} \alpha^k \mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{k})\mathrm{T}}]^{-1} \mathbf{C}_0^{-1} \mathbf{v}^{(\mathbf{k})} \rangle, \tag{11}$$

where $\mathbf{C}_0 \equiv \mathbf{C} - \alpha^k \mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{k})\mathrm{T}}$ is independent of the random variable $\mathbf{v}^{(\mathbf{k})}$. Expanding in powers of α^k and averaging over $\mathbf{v}^{(\mathbf{k})}$ yields Eq. (9) for the memory function where

$$q = \langle \mathbf{v}^{(\mathbf{k})\mathrm{T}} \mathbf{C}_0^{-1} \mathbf{v}^{(\mathbf{k})} \rangle = \frac{1}{N} \langle \mathrm{Tr} \mathbf{C}_0^{-1} \rangle = \frac{1}{N} \langle \mathrm{Tr} \mathbf{C}^{-1} \rangle. \tag{12}$$

The last equality holds for large N because $\langle \mathrm{Tr} \mathbf{C}^{-1} \rangle - \langle \mathrm{Tr} \mathbf{C}_0^{-1} \rangle \sim \alpha^k \langle \mathbf{v}^{(\mathbf{k})\mathrm{T}} \mathbf{C}_0^{-2} \mathbf{v}^{(\mathbf{k})} \rangle$, which is of order 1 due to the normalization of $\mathbf{v}^{(\mathbf{k})}$. Equation (10) for q is obtained from the sum rule Eq. (4) by substituting $\mathbf{C}_{\mathbf{n}} = (1 - \alpha)^{-1} \mathbf{I}$. Though the annealed approximation neglects quenched correlations, it agrees surprisingly well with the numerical solution of the quenched system for all α values except for $\alpha \to 1$, seen in the examples of Fig. 2 for $\epsilon = 10^{-4}$.

To analyze the network's behavior, we first consider the limits $(1 - \alpha)/N$, $\epsilon/N \to 0$ for $N \to \infty$. In this case, the sum of $m(k)$ is finite so Eq. (10) gives $q = 1/\tilde{\epsilon}$, and thus Eq. (9) reduces to

$$m(k) = \frac{\alpha^k}{\alpha^k + \tilde{\epsilon}}, \qquad k = 0, 1, 2, \ldots \tag{13}$$

Capacity is nonzero for $\tilde{\epsilon} < 1$, in which case $k_C = \log(\tilde{\epsilon})/\log(\alpha)$. Here $\alpha^{\mathrm{opt}}(\epsilon)$ is less than unity and decreases with increasing noise because it results from a balance between signal suppression on one hand and amplification of error input ϵ by the factor $(1 - \alpha)^{-1}$ on the other. For small ϵ, maximizing k_C with respect to α yields $\alpha^{\mathrm{opt}} \approx 1 - \epsilon e$ and $k_C(\alpha^{\mathrm{opt}}) \approx 1/\epsilon e$. For $\epsilon \to 1$, $\alpha^{\mathrm{opt}} \to 0$ and $k_C \to 0$. Equation (13) is exact in the limit of large N if ϵ and α are kept fixed.

In order to yield extensive capacity, ϵ and $1 - \alpha$ must decrease inversely with N, so that $1 - \alpha = \rho/N$ and $\epsilon = \bar{\epsilon}/N$ with ρ and $\bar{\epsilon}$ finite. To see that in this case k_C scales with system size N, we write $q = \exp(\rho\mu)$. Capacity $k_C = \mu N$ is extensive for $\mu = O(1)$ and $k_C = 0$ for $\mu < 0$; see Eq. (9). The sum rule Eq. (10) determines the value of μ. In the present regime the sum can be approximated by an integral, yielding

$$\rho = \log[1 + \exp(\rho\mu)] + \bar{\epsilon} \exp(\rho\mu). \tag{14}$$

Solving for μ yields a nonmonotonic function of ρ (see Fig. 3) which attains its maximum at ρ^{opt}. For $\rho < \rho^{\mathrm{opt}}$, μ decreases with decreasing ρ and $\mu < 0$ for ρ smaller than the critical value $\rho_- = \log(2) + \bar{\epsilon}$. For large noise, i.e., $\bar{\epsilon} \gg 1$, ρ^{opt} increases with noise level as $\rho^{\mathrm{opt}} \approx e\bar{\epsilon}$, as in the finite capacity limit. However, at low noise levels ρ^{opt} does not approach 0 (as predicted by the finite capacity limit) but increases with decreasing $\bar{\epsilon}$ as $\rho^{\mathrm{opt}} \approx \frac{1}{2}\log(1/\bar{\epsilon})$. This is because for $\bar{\epsilon}$ sufficiently small and α sufficiently close to 1, strong long-time interference prevents faithful reconstruction even of recent inputs. Therefore, $\alpha^{\mathrm{opt}} < 1$ even in the $\bar{\epsilon} \to 0$ limit. Additionally, choosing ρ close to ρ^{opt} reduces retrieval error for values of $k < k_C$. This behavior is demonstrated in Fig. 2. As in the DSR model, the choice of ρ should be bounded not only by capacity limits but also by the magnitudes of the output weights which increase with ρ roughly as $\|\mathbf{u}_{\mathbf{k}}\| \approx \alpha^{-k/2} = \exp(\rho k/N)$ at zero noise.

The annealed approximation.—An interesting issue is the range of validity of the annealed approximation. As indicated above, finite capacity results should be exact in

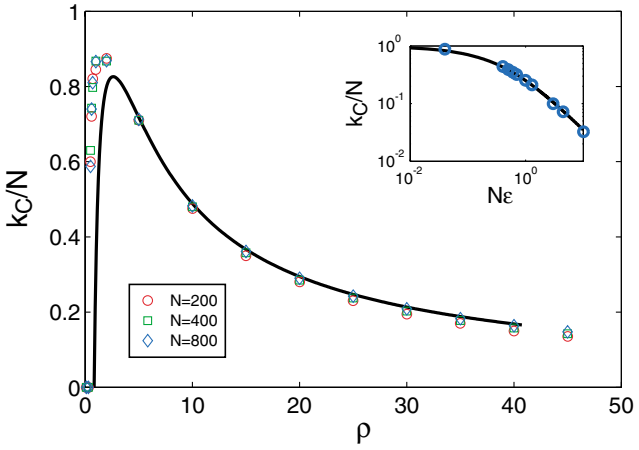

FIG. 3 (color). Capacity per neuron as a function of $\rho = N(1 - \alpha)$, at $\bar{\epsilon} = 0.04$, showing nonmonotonic dependence. Points from simulations on differently sized systems fall on essentially the same curve, confirming that in this regime capacity is extensive. The inset shows the dependence of k_C/N on scaled noise $\bar{\epsilon}$ for ρ^{opt} both in the AA and from simulation.

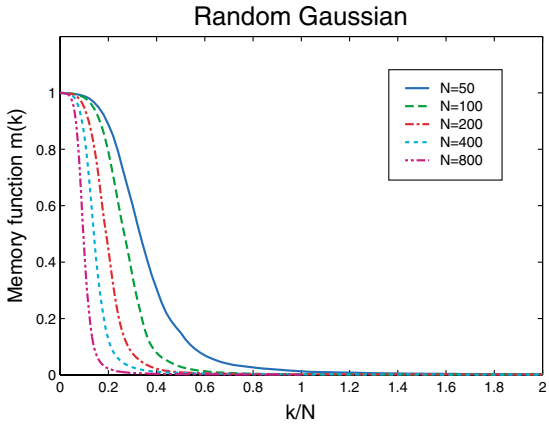

FIG. 4 (color). Numerical calculation of the memory function of Gaussian random matrices for $\bar{\epsilon} = 0.01$, $\alpha = 0.999$ and different system sizes. Results are averages over 50 realizations of the connectivity matrix. Exploring different values of α we find that up to the above mentioned value the memory improves with increasing α. Increasing α beyond this value results in irregular and highly variable $m(k)$.

the large N limit with fixed noise and suppression coefficient. In this limit, at any given time only a small number of directions in $\mathbf{x}$ space contribute to the current state and, since $\mathbf{O}$ is random, the correlations among them are negligible. On the other hand, when ϵ and $1 - \alpha$ are proportional to $1/N$, the number of unsuppressed modes is of order N and hence the combined effect of their correlations become important. Breakdown of the AA occurs when memory of early times begins to decrease due to strong long-time interference. Our simulations indicate that this occurs for $\rho \equiv N(1 - \alpha) \lesssim 10$, as seen for small ρ in Figs. 2 and 3. Derivation of a full quenched theory requires appropriate handling of the intricate correlations among high powers of random orthogonal matrices.

Robustness.—Our results show that systems with the DSR or orthogonal architectures are tolerant to stochastic noise in their network dynamics up to noise amplitudes significantly larger than $1/N$. An important issue is the sensitivity of the network to structural noise. We have tested numerically the robustness of the orthogonal recurrent network to neuron deletion. We find that this perturbation does not affect drastically the capacity of the system provided that the output weights are retrained after the neuron's removal, in contrast to the simple delay line. If the output weights are not retrained, however, capacity drops substantially. Therefore for this to be a viable model of working memory, the system would need to relearn weights $\mathbf{u}_k$ sufficiently quickly upon neuron loss [6]. Note also that these results imply that $\mathbf{W}$ need not be exactly orthogonal since neuron removal is also a perturbation away from orthogonality.

Random Gaussian matrices.—In this work we have assumed rather special network architectures. On the other hand, there are claims that any generic (stable) connection matrix $\mathbf{W}$ can robustly store long temporal signals [5,6]; if substantiated, this is indeed a powerful result. The theoretical study of more generic ensembles is difficult. However, our simulations of fully connected Gaussian random matrices indicate that their capacity is not extensive. If $\epsilon = 0$ and α is sufficiently small then it is likely that $m(k) = 1$ for $k \leq N$ and 0 for larger N, as is the case in the models studied here. This is because, for small α, interference from long past times is negligible and hence the sum rule Eq. (4) implies the above square form for $m(k)$. However, this capacity is unusable in large systems because it requires exponentially large output weights (reflecting the near singularity of the correlation matrix $\mathbf{C}$). Taking α close to 1, so that $\mathbf{C}$ is well conditioned, results in strong fluctuations in $m(k)$ and does not seem to yield extensive capacity. The presence of noise also regularizes the system but again does not contribute to give extensive capacity, as indicated in Fig. 4. We are currently developing a fuller understanding of the short-term memory properties of generic connectivity matrices.

H. S. is partially supported by the Israel Science Foundation (Center of Excellence Grant No. 8006/00), and D. D. L. by the U.S. Army Research Office. O. W. acknowledges support from the National Science Foundation via Grant No. DMR 02209243 and the National Institute of Health via Grant No. RO1 EY01473701. We acknowledge helpful discussions with Misha Tsodyks and are grateful to Daniel Fisher for many illuminating discussions on this work.

[1] X.-J. Wang, Trends Neurosci. **24**, 455 (2001).
[2] M. Usher and J. D. Cohen, in *Connectionist Models in Cognitive Neuroscience*, edited by D. Heinke, G.W. Humphryes, and A. Olson (Springer-Verlag, Berlin, 1999).
[3] H. S. Seung, Proc. Natl. Acad. Sci. U.S.A. **93**, 13 339 (1996).
[4] Y. Loewenstein and H. Sompolinsky, Nat. Neurosci. **6**, 961 (2003).
[5] W. Maass, T. Natschläger, and H. Markram, Neural Comput. **14**, 2531 (2002).
[6] H. Jaeger, German National Research Center for Information Technology, GMD Report No. 148, 2001.
[7] Since $\mathbf{W}^N = 0$, $|\mathbf{x}|$ converges for any finite system size N. However, in order for $|\mathbf{x}|$ to not grow exponentially with N we still require $\alpha < 1$.