THE RESTRICTED ISOMETRY PROPERTY FOR ECHO STATE NETWORKS WITH APPLICATIONS TO SEQUENCE MEMORY CAPACITY

Han Lun Yap, Adam S. Charles, and Christopher J. Rozell

ABSTRACT

The ability of networked systems (including artificial or biological neuronal networks) to perform complex data processing tasks relies in part on their ability to encode signals from the recent past in the current network state. Here we use Compressed Sensing tools to study the ability of a particular network architecture (Echo State Networks) to stably store long input sequences. In particular, we show that such networks satisfy the Restricted Isometry Property when the input sequences are compressible in certain bases and when the number of nodes scale linearly with the sparsity of the input sequence and logarithmically with its dimension. Thus, the memory capacity of these networks depends on the input sequence statistics, and can (sometimes greatly) exceed the number of nodes in the network. Furthermore, input sequences can be robustly recovered from the instantaneous network state using a tractable optimization program (also implementable in a network architecture).

Index Terms— Compressed Sensing, Echo State Networks, Sequence Memory

1. INTRODUCTION

The ability of networked systems (including networks of artificial or biological neurons) to perform complex data processing tasks relies in part on their ability to encode signals from the recent past in the current network state. For example, due to the long-scale temporal dependencies present in many interesting types of time-series data, using a network to predict future values of a time-series relies in part on how well the current network state preserves information about the sequence of inputs from the (sometimes distant) past [1]. Due to these reasons, recent research has investigated the Short Term Memory (STM) capacity of neural network models [2–5]. In contrast with early work on memory models in neural networks that relied on the notion of attractors of dynamical systems [6], STM is related to the preservation of input sequences in the nodes of the network while the input is streaming into the network.

Prior work related to STM imposes independent and identically distributed (i.i.d.) Gaussian statistics on the input sequence [2, 3]. Results with such a signal model have shown rather unsatisfactorily that the theoretical STM capacity is capped by the number of network neurons (i.e., the number of time samples of input history that can be recovered is equal to the number of nodes in the network). More recent work on STM imposes a sparsity structure on the input sequences. Such a structure is not uncommon for natural images and other signals [7]. With such a signal model, the authors in [5] used a statistical mechanics calculation on an annealed approximation of the network dynamics to demonstrate the potential for STM capacity to exceed the number of neurons.

Sparsity has been exploited extensively in the signal processing community, as demonstrated by the rapidly expanding literature on Compressed Sensing (CS) [8]. CS studies the conditions and algorithms for recovering a compressible signal from an underdetermined system of linear equations. The now canonical result in CS says that if a measurement system satisfies the Restricted Isometry Property (RIP), then a compressible signal can be robustly recovered from its measurements via ℓ_1-minimization, even if the length of the signal greatly exceeds the number of measurements. Geometrically, the RIP says that *any* sparse signal is uniquely and *stably* represented in the measurement space.

In this paper, we apply CS tools to study STM capacity in a particular class of neural networks called Echo State Networks (ESNs). Specifically, we show that such networks satisfy the RIP when the input sequences are compressible in certain bases and when the number of nodes scales linearly with the sparsity of the input sequence and logarithmically with its dimension. This implies that, not only are input sequences uniquely and *stably* determined from the network nodes, they can be recovered by using a tractable optimization program (that can also be implemented in a network architecture) [9]. As a direct consequence, the STM capacity of such networks can indeed greatly exceed the number of nodes in the network.

2. ECHO STATE NETWORKS AND SHORT TERM MEMORY

An ESN is a recurrent neural network whose set of M nodes are randomly connected without any prior training [1, 10]. The network state $\boldsymbol{x}[n] \in \mathbb{R}^M$ evolves with the input $s[n] \in \mathbb{R}$ at time n as

$$\boldsymbol{x}[n] = f\left(\boldsymbol{W}\boldsymbol{x}[n-1] + \boldsymbol{z}s[n]\right), \tag{1}$$

The authors are from the School of Electrical and Computer Engineering at Georgia Institute of Technology. This work was partially supported by NSF grant CCF-0830456 and DSO National Laboratories, Singapore.

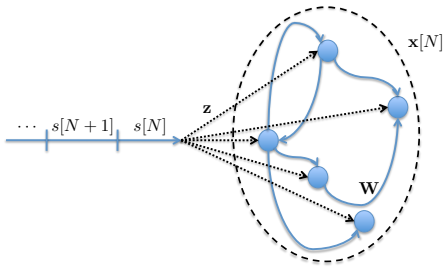

Fig. 1. A pictorial description of the ESN showing the input at time N, $s[N]$, being fed by the feed-forward vector $\boldsymbol{z}$ into the reservoir of nodes $\boldsymbol{x}[N]$ with connectivity pattern $\boldsymbol{W}$.

where $\boldsymbol{W} \in \mathbb{R}^{M\times M}$ is the network connectivity pattern and $\boldsymbol{z} \in \mathbb{R}^M$ is the input feed-forward vector. The function $f : \mathbb{R}^M \to \mathbb{R}^M$ is called the activation function and is usually taken to be a sigmoid function applied component-wise. Figure 1 shows a pictorial representation of an ESN. ESNs have drawn considerable interests since their introduction due to performance in dynamical systems modeling [10], time-series prediction tasks [1], and of course its general structural similarity to biological brains [2].

In [2,3], the authors used a linearized version of the ESN, together with a Gaussian statistics on the input sequences, to show that the theoretical recoverable STM length is capped at the number of neurons in the network. In a slightly different direction, the authors in [4] removed the Gaussian model on the inputs and used instead Fisher information to quantify the output SNR of the input sequences as a proxy to the memory capacity of an ESN. Our work is more closely related to the work done by the authors in [5] which exploits a sparsity structure (in the canonical basis) of the input stimuli and studies the STM capacity of an annealed system. Here, we study the exact formulation of the ESN and expand the class of input sequences to include sequences that are compressible in any arbitrary basis.

3. THE RESTRICTED ISOMETRY PROPERTY

Signal processing research has shown that many useful signals, including many natural stimuli [7], are inherently compressible in some basis $\boldsymbol{\Psi} \in \mathbb{C}^{N\times N}$ (i.e., the energy of the signal is concentrated in very few of its coefficients) [11]. CS is the theoretical and algorithmic study of the recoverability of compressible signals $\boldsymbol{s}$ from under-determined systems of linear equations, which can be written as $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$. In particular, it has been shown that if the measurement matrix $\boldsymbol{A}$ satisfies the RIP of order K and conditioning δ (RIP-(K,δ) [12], i.e. for all K sparse signals

$$(1-\delta)\,||\boldsymbol{s}||_2^2 \le ||\boldsymbol{A}\boldsymbol{s}||_2^2 \le (1+\delta)\,||\boldsymbol{s}||_2^2,$$

holds, then solving a constrained ℓ_1 optimization recovers any K-sparse input signal $\boldsymbol{s}$. In fact, the RIP ensures recovery of *all* K-sparse signals by the *same* measurement matrix $\boldsymbol{A}$, and guarantees robust recovery in the presence of noise.

Theorem 3.1. *Assume a matrix $\boldsymbol{A}$ satisfies RIP-(2K,δ) with $\delta < 0.4627$. Let $\boldsymbol{s} \in \mathbb{C}^N$ be any vector and suppose we acquire the noisy measurements $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon}$ with $||\boldsymbol{\epsilon}||_2 \le \eta$. Let $\widehat{\boldsymbol{s}}$ be the unique solution of:*

$$\min_{\boldsymbol{s}} ||\boldsymbol{s}||_1 \text{ subject to } ||\boldsymbol{A}\boldsymbol{s} - \boldsymbol{x}||_2 \le \eta. \tag{2}$$

Then

$$||\boldsymbol{s} - \widehat{\boldsymbol{s}}||_2 \le \alpha\eta + \beta\frac{||\boldsymbol{s} - \boldsymbol{s}_K||_1}{\sqrt{K}}, \tag{3}$$

where $\boldsymbol{s}_K$ is the best K-term approximation of $\boldsymbol{s}$, and α, β are some constants that depend only on δ.

Thus, for an input vector $\boldsymbol{s}$ and measurement operator $\boldsymbol{A}$ satisfying the RIP, solving the ℓ_1-minimization program (2) guarantees an output $\widehat{\boldsymbol{s}}$ whose distance from $\boldsymbol{s}$ is bounded both by the measurement noise level and by the distance from $\widehat{\boldsymbol{s}}$ to its best K-term approximation.

The power of the RIP comes from the fact that whenever a measurement matrix $\boldsymbol{A}$ satisfies the RIP of order $2K$, distances between the images of any 2 K-sparse signals are maintained in the measurement space, i.e. $||\boldsymbol{A}\boldsymbol{s}_1 - \boldsymbol{A}\boldsymbol{s}_2||_2 \approx ||\boldsymbol{s}_1 - \boldsymbol{s}_2||_2$. This distance-preservation guarantee, or *stable embedding*, allows different sparse input signals to be distinguishable by the ℓ_1-minimization recovery program. In fact, this stable embedding of sparse signals allows many signal-processing algorithms (e.g., signal detection and classification) to work directly in the measurement space instead of necessitating a prior recovery step [13].

Many random matrix constructions satisfy the RIP with high probability. For example, when $\boldsymbol{A}$ is an $M \times N$ matrix whose rows are chosen uniformly at random from a DFT matrix, it is known that $\boldsymbol{A}$ satisfies RIP-(K,δ) with high probability when $M \ge CK\log^4(N)$ [12]. Notice that when $K \ll N$, the number of measurements required M will be much less than N. In the ESN setting, such a result will imply that the number of nodes M can be much less than the recoverable input sequences' length N.

4. RIP FOR ECHO STATE NETWORKS

In this section, we show that ESNs, when written as an equivalent measurement matrix on input sequences, can satisfy the RIP. We restrict ourselves to a class of ESN proposed in [2,3] whereby the activation function f in (1) is identity:

$$\boldsymbol{x}[n] = \boldsymbol{W}\boldsymbol{x}[n-1] + \boldsymbol{z}s[n]. \tag{4}$$

Following [5] and iterating (4), we see that the network state at time N can be written as

$$\begin{aligned}\boldsymbol{x}[N] &= \boldsymbol{z}s[N] + \boldsymbol{W}\boldsymbol{z}s[N-1] + \cdots + \boldsymbol{W}^{N-1}\boldsymbol{z}s[1] \\ &= \left[\boldsymbol{z} \mid \cdots \mid \boldsymbol{W}^{N-1}\boldsymbol{z}\right]\begin{bmatrix} s[N] \\ \vdots \\ s[1] \end{bmatrix} =: \boldsymbol{A}\boldsymbol{s}.\end{aligned}$$

Thus, we have shown that the network connectivity and the feed-forward vector can be used to create an effective measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ on the input sequence $\boldsymbol{s}$ comprising of the past N time steps.

To advance the analysis of $\boldsymbol{A}$, we can use the eigenvalue decomposition of the connectivity matrix $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ to rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{U}\left[\tilde{\boldsymbol{z}} \mid \boldsymbol{D}\tilde{\boldsymbol{z}} \mid \boldsymbol{D}^2\tilde{\boldsymbol{z}} \mid \cdots \mid \boldsymbol{D}^{N-1}\tilde{\boldsymbol{z}}\right] \tag{5}$$

where $\tilde{\boldsymbol{z}} = \boldsymbol{U}^{-1}\boldsymbol{z}$. This setup can further be reorganized as

$$\boldsymbol{A} = \boldsymbol{U}\widetilde{\boldsymbol{Z}}\left[\boldsymbol{d}^0 \mid \boldsymbol{d} \mid \boldsymbol{d}^2 \mid \cdots \mid \boldsymbol{d}^{N-1}\right] = \boldsymbol{U}\widetilde{\boldsymbol{Z}}\boldsymbol{F} \tag{6}$$

where $\boldsymbol{d} = \text{diag}\,(\boldsymbol{D})$ is the column vector consisting of the eigenvalues of $\boldsymbol{W}$, $\widetilde{\boldsymbol{Z}} = \text{diag}(\tilde{\boldsymbol{z}})$ and the exponentiation of the vector $\boldsymbol{d}$ is defined as the element-wise exponentiation. Additionally, we will analyze the properties of $\boldsymbol{A}$ under some generally-accepted assumptions on the network. Deviations from these assumptions will be studied in a future work. First, we assume, as in [3, 5], that the connectivity matrix $\boldsymbol{W}$ is a real, random orthonormal matrix (i.e., $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{W}^T\boldsymbol{W} = I$). This implies that $\boldsymbol{U}$ is also orthonormal and when M is large enough, the eigenvalues are distributed uniformly on the unit complex circle.[1] As such, the matrix $\boldsymbol{F}$ becomes a subsampled discrete-time Fourier transform (DTFT) matrix. Second, we assume that we have control over the choice of the feed-forward vector $\boldsymbol{z}$. By letting $\boldsymbol{z} := \boldsymbol{U}\boldsymbol{1}$, where $\boldsymbol{1} := [1, \cdots, 1]^T$, $\widetilde{\boldsymbol{Z}}$ becomes an identity matrix.

Under these assumptions, we see that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{F}$. Because $\|\boldsymbol{U}\boldsymbol{F}\boldsymbol{s}\|_2^2 = \|\boldsymbol{F}\boldsymbol{s}\|_2^2$ for any $\boldsymbol{s}$, we can draw on previous RIP results on subsampled DTFT matrices [12] to arrive at the following theorem.

Theorem 4.1. *Let $\boldsymbol{W}$ be a random orthogonal $M\times M$ matrix with eigenvalues distributed uniformly on the complex unit circle as described and $\boldsymbol{z} := \boldsymbol{U}\boldsymbol{1}$ be the length-M feedforward vector. Then with probability at least $1 - O(N^{-1})$, for* any *input sequence $\boldsymbol{s} \in \mathbb{R}^N$ that is compressible in a basis $\boldsymbol{\Psi}$, $\boldsymbol{A}\boldsymbol{\Psi}$ satisfies RIP-$(2K, \delta)$ whenever*

$$M \geq CK\delta^{-2}\mu^2\left(\boldsymbol{\Psi}\right)\log^4(N).$$

The quantity $\mu\left(\boldsymbol{\Psi}\right)$ is the incoherence *of the basis $\boldsymbol{\Psi}$ with the subsampled DTFT $\boldsymbol{F}$ defined as*

$$\mu\left(\boldsymbol{\Psi}\right) := \max_{n=1,\cdots,N}\sup_{t\in[0,2\pi]}\left|\sum_{n'=0}^{N-1}\boldsymbol{\Psi}_{n',n}e^{-jtn'}\right|,$$

where $\boldsymbol{\Psi}_{n',n}$ is the (n', n)-th entry of the matrix $\boldsymbol{\Psi}$.

A few remarks are in order. First, we can use Theorem 3.1 to ensure recoverability of input sequences from the node values of the network. In particular, given noisy readings of the node values $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon}$ with error vector $\|\boldsymbol{\epsilon}\|_2 \leq \eta$, solving the ℓ_1-optimization (2) recovers any input sequence $\boldsymbol{s}$ compressible in the basis $\boldsymbol{\Psi}$ up to an ℓ_2 error:

$$||\widehat{\boldsymbol{s}} - \boldsymbol{s}||_2 \leq \alpha\eta + \beta\|\boldsymbol{\Psi}^H\boldsymbol{s} - (\boldsymbol{\Psi}^H\boldsymbol{s})_K\|_1/\sqrt{K},$$

where α and β are constants, and $(\boldsymbol{\Psi}^H\boldsymbol{s})_K$ is the best K terms approximation to $\boldsymbol{\Psi}^H\boldsymbol{s}$. If $\boldsymbol{s}$ is K-sparse in $\boldsymbol{\Psi}$ and there is no noise ($\eta = 0$), the above theorem says that we can perfectly recover *any* sparse input sequences $\boldsymbol{s}$ from $\boldsymbol{x}$ whenever the number of nodes M is proportional to the sparsity K and the coherence $\mu^2\left(\boldsymbol{\Psi}\right)$ (and poly-logarithmic in N).

Second, this recoverability guarantee allows us to look at the STM capacity of such networks. Suppose that the sparsity of input sequences to the ESN scales linearly with the sequence length N, i.e., $K = \lceil \rho N \rceil$ where $\rho \in (0,\ 1)$ is the *sparsity density*, and suppose that the node values are not corrupted by noise. Then, the STM capacity will be the largest N such that $M \geq C\rho N\delta^{-2}\mu^2\left(\boldsymbol{\Psi}\right)\log^4(N)$. Thus, if the sparsity density is low (i.e., $\rho \ll 1$) and the sparsity basis $\boldsymbol{\Psi}$ has small incoherence (i.e., $\mu^2(\boldsymbol{\Psi}) \approx 1$), then the STM capacity of the network can greatly exceed the number of nodes M.

Third, the RIP of $\boldsymbol{A}$ means that input sequences of length-N are *stably embedded* in the state of the network nodes. Thus, not only are different input sequences of length-N distinguishable, their distances are also preserved in the node space (i.e., sequences that are similar will have similar node values while sequences that are very different will have very different node values). The ESN architecture has often been characterized through the Echo State Property (ESP). Concisely, the ESP ensures that under certain compactness conditions every network state is uniquely determined by some left-infinite input sequence [10], i.e., $\boldsymbol{x}[n]$ is uniquely determined by $\cdots, s[n-2], s[n-1], s[n]$. In essence, the ESP implies that there is a one-to-one correspondence between the input time series and the current network state, further implying that a function can predict future inputs from the current state as well as if it had the entire previous input sequence. Since distances between sparse inputs are preserved in the node space, the RIP is a stronger guarantee of information preservation than the ESP (which may not preserve distances between inputs). This will provide some measure of stability to any algorithms operating on the network state (e.g. time series prediction).

5. SIMULATIONS OF STM CAPACITY

To show the validity of our theory on STM capacity, we create a plot demonstrating the total error in recovering an input sequence of length N with sparsity density ρ from M nodes in Figure 2. We use a plotting style similar to the phase transition diagrams of [14] where the relative mean-squared error (rMSE) of the reconstruction is shown for each pair of variables (in this case N and M). The wedge between the dashed line (M = N) and the solid line (recovery error = 0.1%) in each plot shows that the STM capacity can easily exceed the

[1] Since the connectivity matrix is real, its eigenvalues and eigenvectors come in complex conjugate pairs.

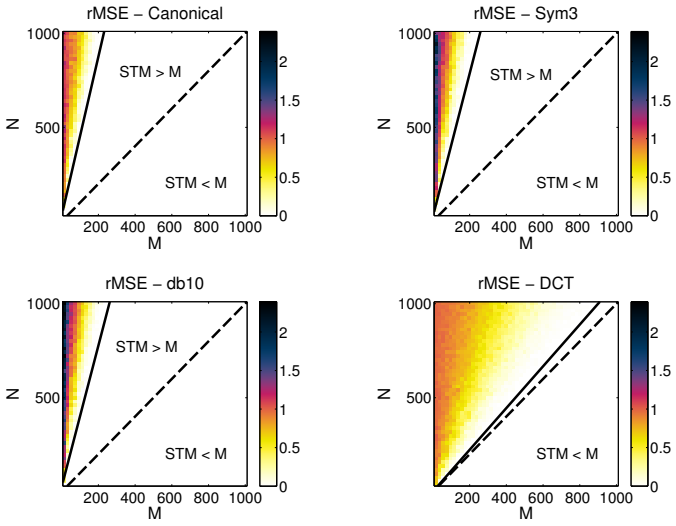

Fig. 2. rMSE for input sequences of length N from the M number of neurons where the input sequences are ρN-sparse in a basis $\mathbf{\Psi}$ with $\rho = 0.05$. The wedge between the dashed line ($M = N$) and the solid line (recovery error = 0.1%) in each plot shows where the STM capacity exceeds the number of network nodes. This wedge is large for sequences sparse in the canonical, Symlets and Daubechies-10 wavelet basis (up to 4 level decompositions) as these bases are incoherent with the subsampled DFT. For the discrete cosine transform (DCT) basis which is coherent with the subsampled DFT, recovery above $N = M$ suffers significantly.

number of nodes in the system for input sequences that are ρN-sparse in different bases.

6. CONCLUSIONS

In this paper, we quantified the STM capacity of an ESN by showing that, when written as a measurement matrix on input sequences, the network satisfies the RIP whenever the number of nodes scales linearly with the sparsity of the input sequence and logarithmically with the sequence dimension (with a factor depending on the input sparsity basis). These results lead us to conclude both that 1) the STM capacity of these networked architectures can be much larger than the size of the network (a barrier to previous STM analysis), and 2) the tools of CS can be powerfully applied beyond signal recovery to the study of some properties in dynamical systems.

These preliminary results lead to many interesting open questions. In particular, the current derivation makes several simplifying assumptions on the network structure. First, we imposed an orthogonal structure on the network connectivity matrix $\boldsymbol{W}$ and second, we chose a particular feed-forward vector $\boldsymbol{z} := \boldsymbol{U}\mathbf{1}$. In practice, ESNs, including networks that model neural systems, can have network structures that are different from those considered in this paper. For example, networks can have "small world" connectivity and the weights of the feed-forward vector can be randomly chosen. Effects of deviating from these assumptions on the RIP will be studied in a forthcoming paper. Additionally, we only considered ESN networks fed with finite length signals. In ongoing work, we extend results to networks subjected to infinite length inputs.

7. REFERENCES

[1] H. Jaeger and H. Haas, "Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication." *Sci.*, vol. 304, no. 5667, pp. 78–80, Apr. 2004.

[2] H. Jaeger, "Short term memory in echo state networks," *Technical Report GMD Report 152, Forschungszentrum Informationstechnik GmbH*, 2002.

[3] O. L. White, D. D. Lee, and H. Sompolinsky, "Short-term memory in orthogonal neural networks," *Physical Review Lett.*, vol. 92, no. 14, p. 148102, 2004.

[4] S. Ganguli, D. Huh, and H. Sompolinsky, "Memory traces in dynamical systems," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 48, pp. 18 970–5, Dec. 2008.

[5] S. Ganguli and H. Sompolinsky, "Short-term memory in neuronal networks through dynamical compressed sensing," in *Proc. Conf. Advances Neural Inform. Process. Syst.*, 2010.

[6] H. S. Seung, "How the brain keeps the eyes still," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 23, pp. 13 339–13 344, Nov. 1996.

[7] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[8] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. Mathematicians*, vol. 3, 2006, pp. 1433–1452.

[9] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neural Computation*.

[10] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," *Technical Report GMD Report 148, German National Research Center for Information Technology*, 2001.

[11] S. Mallat, *A wavelet tour of signal processing: the sparse way*, 3rd ed. Academic Press, 2008.

[12] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical Found. and Numerical Methods for Sparse Recovery*, vol. 9, pp. 1–92, 2010.

[13] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 445–460, Apr. 2010.

[14] D. L. Donoho and J. Tanner, "Observed Universality of Phase Transitions in High-Dimensional Geometry, with Implications for Modern Data Analysis and Signal Processing," *Phase Transitions*, p. 47, Jun. 2009.