Compression of multiple input streams into recursive neural networks

Adam S. Charles*, Dong Yin† and Christopher J. Rozell‡
*Princeton Neuroscience Institute, Princeton University
†School of Electrical Engineering and Computer Sciences, University of California, Berkeley
‡School of Electrical and Computer Engineering, Georgia Institute of Technology
*adamsc@princeton.edu†dongyin@berkeley.edu‡crozell@gatech.edu

Recursive neural networks (RNNs) are becoming an increasingly important part of the machine learning toolbox [1], [2], [3] in applications such as video [4], audio [5], EEG data [6]. These networks have been used as either stand-alone tools for training classifiers, or as in layers in conventional deep neural networks to expand their use to time-varying data [7], [5], [6]. Mathematically, RNNs are a state of M nodes $\boldsymbol{x} \in \mathbb{R}^M$ that evolve as

$$\boldsymbol{x}_n = f(\boldsymbol{W}\boldsymbol{x}_{n-1} + \boldsymbol{Z}\boldsymbol{s}_n + \boldsymbol{\epsilon}),$$

where $\boldsymbol{W} \in \mathbb{R}^{M\times NM}$ is the recurrent connectivity matrix, $\boldsymbol{s}_n \in \mathbb{R}^L$ is the input into the network at time n, $\boldsymbol{Z} \in \mathbb{R}^{M\times L}$ is the feed-forward connectivity matrix, $\boldsymbol{\epsilon} \in \mathbb{R}^M$ is a network error vector, and $f(\cdot) : \mathbb{R}^M \to \mathbb{R}^M$ is a potential point-wise non-linearity. The short-term memory (STM) of RNNs is loosely defined as the number of past inputs $\boldsymbol{s}_n$ whose information is present in the network state. As the abilities of RNNs to process temporal data are often attributed to RNNs accumulating information from input data over time into the network nodes (the STM), here we analyze the ability of RNNs to store inputs in the network state. In particular, since many real-world signals have low-dimensional representations, we study the STM of RNNs when the inputs are either sparse in a basis (e.g. audio or video signals), or where the input vectors are correlated such that they form low-rank martix (e.g. two-photon microscopy)

STM of randomly connected networks have been analyzed with a number of methods, including nonlinear networks [8], [9], [10], [11], [12] and linear networks [1], [13], [14], [15], [16], [17] with both discrete-time and continuous-time dynamics. These methods can be broadly be classified as either correlation-based methods [14], [15] or uniqueness methods [1], [18], [13], [17]. Traditional analyses of STM or RNNs focus on the case where $\boldsymbol{W}$ is a random orthogonal matrix, $\boldsymbol{Z}$ is random and the network is linear (i.e. $f(\boldsymbol{x}) = \boldsymbol{x}$). These analysis typically assumed that $L = 1$ and that no additional statistics were known about the inputs $\boldsymbol{s}_n$, resulting in the STM bound where the number of inputs that could be remembered N is bounded by the number of nodes M. More recently, STM analysis have been taking into account the statistics of the inputs $\boldsymbol{s}$. Specifically, when the vector of inputs ($L = 1$) is K-sparse, the number of nodes needed to recover N inputs scales with $K\log(N) < N$ nodes [16], [17].

These results, based on the ideas of compressive sensing, have only addressed the case where one input at each time is input into the network and that the inputs were K-sparse. In many machine-learning applications, however, inputs are multi-dimensional, and can admit other low-dimensional structures outside of sparsity. For example, when each of the N L-dimensional inputs are stacked as a vector, that vector can be considered as sparse (in a spatio-input dictionary), however if the vectors $\boldsymbol{s}_n$ are concatenated into a matrix $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]$, that matrix might be low rank (i.e the L input streams are linearly correlated). In this work, we extend the results in [17] to both these cases. We first show that RNNs can store a large number of inputs that are joint sparse ($LN > M$), and then show that similar bounds hold for low-rank inputs.

To show these STM bounds, we follow the strategy of [17], using the tools of [19], [20] to prove that the past N length- L input vectors are recoverable from the network state at time N. For joint-sparse inputs, we show that the network dynamics satisfy the restricted isometry property (RIP). This property ensures that the network state approximately preserves differences between K-sparse input streams. Specifically, if the number of nodes in the network exceeds

$$M \geq C\frac{K}{\delta^2}\mu_S^2(\boldsymbol{\Psi})\log^5(NL)\log(\eta^{-1}),$$

where δ is the RIP constant, C is a universal constant, $\boldsymbol{\Psi}$ is the sparsity basis for the inputs, and $\mu(\cdot)$ is the coherence parameter measuring how different the columns of $\boldsymbol{\Psi}$ are from Fourier vectors, then with probability exceeding $1-\eta$, the inputs are recoverable via a LASSO procedure up to an error given by

$$\|\boldsymbol{S} - \widehat{\boldsymbol{S}}\|_F \leq \alpha\|\boldsymbol{\epsilon}\|_2 + \beta\frac{\|(\boldsymbol{S} - \boldsymbol{S}_K)\|_1}{\sqrt{K}},$$

for constants α and β, and where $\boldsymbol{S}_K$ is the best K-term approximation of $\boldsymbol{S}$.

For low-rank inputs, i.e. $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{V}$, where $\boldsymbol{Q} \in \mathbb{R}^{L\times R}$ and $\boldsymbol{V} \in \mathbb{R}^{R\times N}$, we require a different technique, instead directly showing that a solution to the KKT equations holds for a nuclear-norm optimization program. Specifically, generalise the golfing scheme in [21] to prove the existence of a solution. We thus show a similar STM bound to the joint-sparse case, where for rank-R $\boldsymbol{S}$, if

$$M \geq CR\left(N + \mu_L^2 L\right)\log^3(NL),$$

for a universal constant C and coherence parameter μ_L measuring how different the columns of the right low-rank decomposition matrix, $\boldsymbol{V}$, are from Fourier vectors, then the inputs are recoverable with high probability up to an error

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_F \leq \left(4\sqrt{\min(N,L)\frac{2LN+M}{M}} + 2\right)\epsilon.$$

This work leverages the tools developed in the compressive sensing literature to develop a theoretical understanding of RNNs, an important tool in machine learning. Our results, taken together, demonstrate that RNNs are very efficient in compressing long input streams into the network state. In both cases, the compression rate (how many inputs can be recovered from M nodes) is proportional to the underlying dimension of the inputs (the sparsity or rank), and only poly-logarithmic with the total number of inputs (LN). In terms of machine-learning tasks, this means that RNNs operating on structured, dynamic signals have access to long extents of the data history to make classification or prediction decisions.

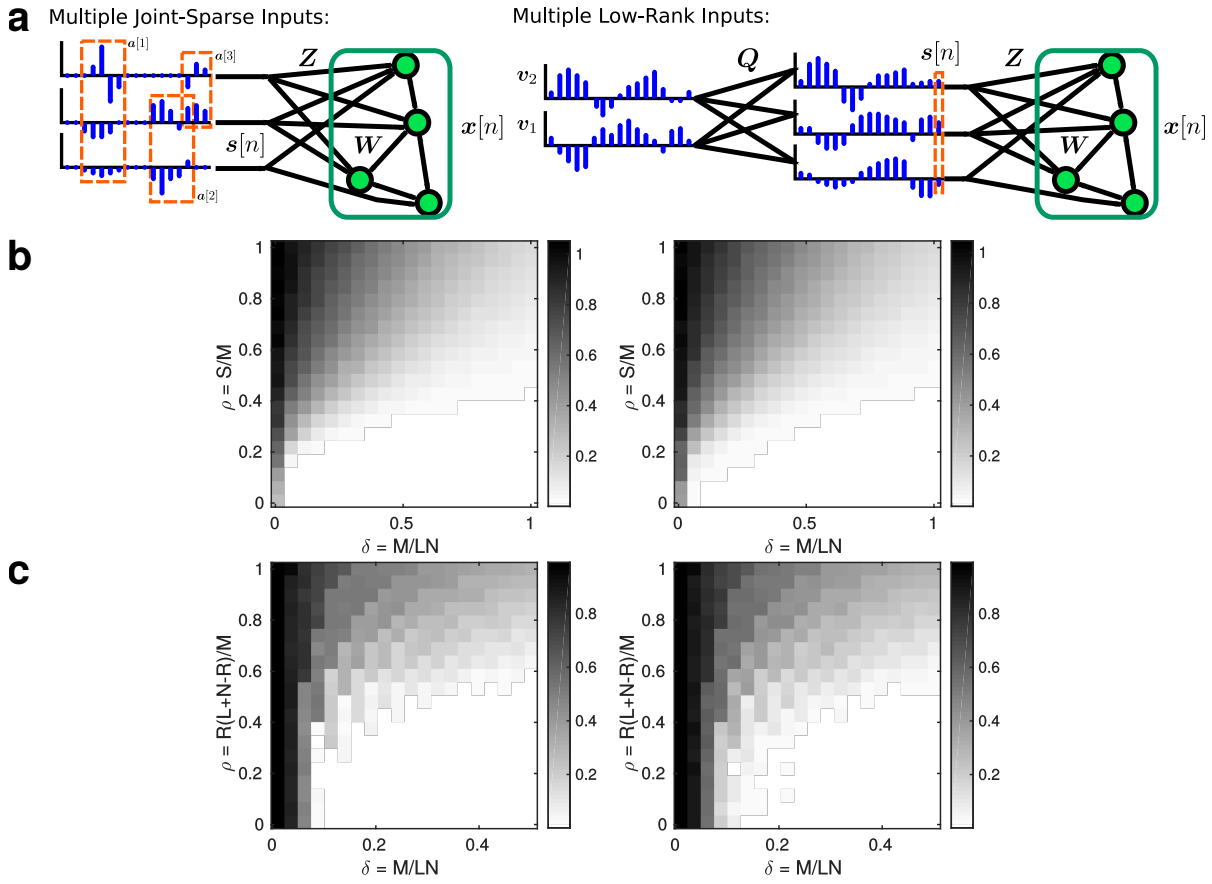

Fig. 1. (a) Network architecture for RNNs with sparse inputs and lowr-rank inputs. (b) Empirical relative mean-squared error (rMSE) calculated for a variety of parameter values (input sparsity and number of nodes) shows that the network state can recover inputs from many fewer nodes than predicted by theory that does not leverage the input statistics. Each pixel represents the average rMSE of 20 random trials with $L = 40$ and $N = 100$. (c) Similar results to (b) show that the recovery guarantees also hold for low-rank input statistics.

References

[1] H. Jaeger, "Short term memory in echo state networks," *GMD Report 152 German Nat. Research Center for Info. Tech.*, 2001.

[2] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neur. Net.: Tricks of the Trade*, 2012, pp. 659–686.

[3] X. Hinaut, M. Petit, G. Pointeau, and P. Dominey, "Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks," *Front. in neurorobotics*, vol. 8, 2014.

[4] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conf. on Comp. Vision and Pattern Rec.*, 2015, pp. 2625–2634.

[5] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc.*, 2013, pp. 6645–6649.

[6] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from eeg with deep recurrent-convolutional neural networks," in *Int. Conf. on Learning Rep.*, 2016.

[7] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *Adv. in Neur. Info. Proc. Sys.*, pp. 2431–2439.

[8] M. Massar and S. Massar, "Mean-field theory of echo state networks," *Physical Review E*, vol. 87, no. 4, p. 042809, 2013.

[9] O. Faugeras, J. Touboul, and B. Cessac, "A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs," *Front. in comp. neuro.*, vol. 3, 2009.

[10] K. Rajan, L. Abbott, and H. Sompolinsky, "Stimulus-dependent suppression of chaos in recurrent neural networks," *Physical Review E*, vol. 82, no. 1, p. 011903, 2010.

[11] M. Galtier and G. Wainrib, "A local echo state property through the largest lyapunov exponent," *arXiv preprint arXiv:1402.1619*, 2014.

[12] G. Wainrib, "Context dependent representation in recurrent neural networks," 2015, http://arxiv.org/pdf/1506.06602.pdf.

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

[14] O. White, D. Lee, and H. Sompolinsky, "Short-term memory in orthogonal neural networks," *Phys. rev. lett.*, vol. 92, no. 14, p. 148102, 2004.

[15] S. Ganguli, D. Huh, and H. Sompolinsky, "Memory traces in dynamical systems," *Proc. of the Nat. Acad. of Sci.*, vol. 105, no. 48, p. 18970, 2008.

[16] S. Ganguli and H. Sompolinsky, "Short-term memory in neuronal networks through dynamical compressed sensing," *Conf. on Neur. Info. Proc. Sys.*, 2010.

[17] A. S. Charles, H. L. Yap, and C. J. Rozell, "Short-term memory capacity in networks via the restricted isometry property," *Neur. comp.*, vol. 26, no. 6, pp. 1198–1235, 2014.

[18] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neur. comp.*, vol. 14, no. 11, pp. 2531–2560, 2002.

[19] H. Rauhut, "Compressive sensing and structured random matrices," *Theor. Found. and Num. Meth. for Sparse Rec.*, vol. 9, pp. 1–92, 2010.

[20] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Comms. Pure and Applied Math.*, vol. 61, no. 8, pp. 1025–1045, 2008.

[21] A. Ahmed and J. Romberg, "Compressive multiplexing of correlated signals," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 479–498, 2015.