Short-term Sequence Memory: Compressive effects of Recurrent Network Dynamics

Adam S. Charles (adamsc@princeton.edu)
Princeton University
Princeton, NJ, 08544, USA

Dong Yin dongyin@berkeley.edu
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Han Lun Yap (yaphanlun@gmail.com)
DSO National Laboratories
Singapore

Christopher J. Rozell (crozell@gatech.edu)
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA

abstract

Neural networks have become a ubiquitous as cognitive models in neuroscience and as machine learning systems. Deep neural networks in particular are achieving near-human performance in many applications. More recently, recurrent neural networks (RNNs) are being increasingly utilized, both as standalone structures and as layers of deep networks. RNNs are especially interesting as cortical networks are recurrent, indicating that recurrent connections are important in human-level processing. Despite their growing use, theory on the computational properties of RNNs is sparse. As many applications hinge on RNNs accumulating information dynamically, the ability of RNNs to iteratively compress information into the network is particularly critical. We thus present here non-asymptotic bounds on the network’s short-term memory (STM; the number of inputs that can be compressed into and recovered from a network state). Previous bounds on a random RNN’s STM limit the number of recoverable inputs by the number of network nodes. We show that when the input sequences are sparse in a basis or the matrix inputs is low-rank, the number of network nodes needed to achieve an STM grows as the overall information rate. Thus RNNs can efficiently store information embedded in longer input streams, shedding light on their computational capabilities.

Keywords: Short-term memory, RNNs, sparsity, low-rank

Introduction

Artificial neural networks have long been used as simplifying models of biological neural networks with the goal of better understanding fundamental cortical processes. With advances in machine learning and neuroscience, deep and recurrent neural networks are quickly achieving human-like performance in many machine learning tasks, further encouraging the study of cortical systems via their artificial counterparts (Pitts, 1943; Hopfield, 1982; Cadieu et al., 2014; Yamins et al., 2014; Majaj, Hong, Solomon, & DiCarlo, 2015; D.L.K.Yamins & DiCarlo, 2016). Recurrent networks are and have been particularly relevant to the modeling of cortical systems (Pitts, 1943; Hopfield, 1982) as biological neural networks are not feed-forward, and contain many recurrent connections. Interestingly, recurrent neural networks (RNNs) have seen a recent resurgence in cortical modeling, in part due to advances in training recurrent networks on experimental tasks (Sussillo & Abbott, 2009; DePasquale, Cueva, Rajan, Abbott, et al., 2018), and are also being increasingly applied in many machine learning applications (Jaeger, 2001; Luviso\v{c}i\u{c}us, 2012; Hinaut, Petit, Pointeau, & Dominey, 2014). In both cases, these networks essentially compress temporally evolving input stimuli over time into a single network state, which can relay that information upstream for use in data processing tasks, such as classification or prediction of future events. In RNNs, temporally evolving signals $s_t \in \mathbb{R}^L$ are iteratively input into a network evolving as

$$x_{t+1} = f(Wx_t + Zs_t + \varepsilon_t),$$

where $x_t \in \mathbb{R}^M$ is the state of the network at time $t$, $Z$ is the feed-forward input matrix, $W$ is the matrix of recurrent connections, $\varepsilon_t$ is potential noise in the system, and $f(\cdot): \mathbb{R}^M \rightarrow \mathbb{R}^M$ is a potential non-linearity (Jaeger, 2001; Wilson & Cowan, 1972; Amari, 1972; Sompolinsky, Crisanti, & Sommers, 1988; Maass, Natschläger, & Markram, 2002). The dynamics of RNNs accumulate information from the inputs $s_t$ over time into the network state $x_t$. While in general $f(\cdot)$ is a sigmoidal or thresholding function, much of the analysis of RNNs has started from the linear RNN, where $f(x) = x$. Additionally, a large portion of the RNN and LSM literature has shown that random connectivity (i.e. $W$ and $Z$ are random) yield complex dynamics useful for accumulating information. We thus consider linear, random RNNs in this work.

The information available in the state of the RNN to pass onto the later processing stages depends heavily on how efficiently the RNN dynamics compresses the inputs into the state. One way to quantify this ability of RNNs is to discuss the short-term memory of the network, or the number of past inputs that can be recovered from a network at any given time (Jaeger & Haas, 2004;
In our first results, we show that for the case where the input sequences are sparse, the number of nodes $M$ required to recover $N$-dimensional input vectors grows only as $M = O(K \log^2(ML))$ - a rate that depends linearly only on the overall sparsity $K$ and poly-logarithmically in the number of inputs $NL$. For the low-rank case, we find a similar bound where the number of nodes must satisfy $M = O(R(N+L) \log^2(NL))$. This bound is linear in the total number of degrees of freedom in the signal $R(N+L)$, and is again only poly-logarithmic in the number of inputs. Interestingly, both bounds also indicate the limitations of RNNs. Specifically, in each case, a coherence parameter $\mu$ appears, showing that random RNNs are only excellent at compressing and storing sequences that differ from Fourier vectors (either in the sparsity basis or from the left singular vectors in the low-rank case).

**Simulations**

To test our theory, we simulated a number of RNNs and attempted to empirically extract the driving input stream from the network state using LASSO and nuclear norm optimization programs as in (Charles et al., 2017). In particular, we test the effect of the compression levels, defined as $\gamma = M/LN$, and the complexity of the signal, defined as $\rho = K/M$ for sparse inputs and $\rho = (R(N+L))/M$ for low-rank inputs, on the recoverability of the inputs. Additionally, we test the effect of the input structure, specifically the level to which the input representation is coherent with a sinusoidal basis (i.e. if $\mu_S^2$ and $\mu_T^2$ are low or high).
Figure 2 depicts the results of 20 monte-carlo simulations for each parameter set, fixing $L = 40$ inputs and $N = 100$ time-steps while varying $M$ and either the input sparsity $K$ or the input rank $R$, as appropriate. We can see that for large parameter values, the relative mean-squared error (rMSE) is very low (essentially at the noise floor), indicating that the input can be recovered even when $M < NL$, i.e. the entirety of all the input streams have been successfully compressed into the single RNN. The simulation results also validate our theory's prediction that a high coherence with a sinusoidal basis limits the compressibility into RNNs.

![Figure 2: Simulated recovery of sparse and low-rank inputs from echo-state networks. For each plot, the intensity plots the relative mean-squared error (rMSE) over the different parameter settings. Top left: at low coherence values, the inputs are recoverable up to the noise floor for many parameter settings and $M < NL$. Top right: at high coherence the inputs are not recoverable for any $M < NL$. Bottom left: For low-rank inputs there is similarly a large range for which the theory holds. Bottom right: Similarly for low-rank inputs a high coherence value severely limits the recoverability of the RNN inputs.](image)

Finally, our main theory focuses on finite-length input streams. For infinite length inputs, an interesting effect arises where the length of recovery is a parameter left up to the recall system. When the recall system attempts to recover too long or too short of an input stream, the estimated inputs suffer from recall or omission errors respectively (e.g. Fig. 3). In this case, our theory actually predicts an optimal recall length for a given system (Fig. 3 bottom), which can be calculated as a function of the network parameters (Charles et al., 2014).

**Conclusions**

We demonstrated here two theorems which bound the STM of linear RNNs when the inputs are structured. Our theory predicts both that there can be very high streaming compression of such inputs into randomly connected networks as well as the limitation that the inputs must be incoherent in their representation from a set of sinusoids. When the inputs are too oscillatory, the bounds we prove revert back to the general case $M > N$.

By demonstrating STM bounds, recovery algorithms and accuracy guarantees, the culmination of these bounds (both for the single input networks and multi-
ple input networks), provide some of the most precise STM characterizations of RNNs. While these bounds only address the linear networks, we believe that the intuition gained from proving these bounds begins to provide a method for discussing the computational capabilities of neural networks. More specifically, despite our bounds being derived for a more abstract model, the predictions in terms of the number of items retrievable from the network state and of the existence of an optimal recall time might generalize beyond our simplifying assumptions. We thus believe that further investigation is warranted — both theoretically and empirically — to assess the applicability of these predictions in cortical systems.

References


